The Alliance of Digital Humanities Organizations
The Association for Literary and Linguistic Computing
The Association for Computers and the Humanities
The Australasian Association for Digital Humanities
centerNet
The Society for Digital Humanities – Société pour l'étude des médias interactifs

Digital Humanities 2012

Conference Abstracts

University of Hamburg, Germany
July 16–22, 2012

Hamburg University Press
Publishing House of the Hamburg State and University Library
Carl von Ossietzky

Printed with the Support of the German Research Foundation

Editor
Jan Christoph Meister

Editorial Assistant
Katrin Schönert

Editorial Staff
Bastian Lomsché
Wilhelm Schernus
Lena Schüch
Meike Stegkemper

Technical Support
Benjamin W. Bohl
Daniel Röwenstrunk

Bibliographic information published by the *Deutsche Nationalbibliothek* (German National Library).
The *Deutsche Nationalbibliothek* lists this publication in the *Deutsche Nationalbibliografie*; detailed bibliographic data are available on the Internet at https://portal.dnb.de/
The *Deutsche Nationalbibliothek* stores this online publication on its Archive Server. The Archive Server is part of the deposit system for long-term availability of digital publications.
Also available open access on the Internet at:
Hamburg University Press – http://hup.sub.uni-hamburg.de
PURL: http://hup.sub.uni-hamburg.de/HamburgUP/DH2012_Book_of_Abstracts
For an elaborated version of this Book of Abstracts with color photos and zoom function, please visit: www.dh2012.uni-hamburg.de
ISBN 978-3-937816-99-9
© 2012 Hamburg University Press, publishing house of the Hamburg State and University Library Carl von Ossietzky, Germany
Cover design: Turan Usuk
Cover illustration: Dagmar Schwelle/laif
Printing house: Elbepartner, BuK! Breitschuh & Kock GmbH, Hamburg, Germany

The 24th Joint International Conference of
the Association for Literary and Linguistic Computing and
the Association for Computers and the Humanities

and

The 5th Joint International Conference of
the Association for Literary and Linguistic Computing,
the Association for Computers and the Humanities,
the Society for Digital Humanities – Société pour l'étude des médias interactifs,
for the first time organized together with
the Australasian Association for Digital Humanities and
centerNet

International Programme Committee

- Susan Brown (SDH/SEMI – Vice Chair)
- Arianna Ciula (ALLC)
- Tanya Clement (ACH)
- Michael Eberle-Sinatra (SDH/SEMI)
- Dot Porter (ACH)
- Jan Rybicki (ALLC)
- Jon Saklofske (SDH/SEMI)
- Paul Spence (ALLC – Chair)
- Tomoji Tabata (ALLC)
- Katherine Walter (ACH)

Local Organizing Committee

- Peer Bolten
- Imke Borchers
- Evelyn Gius
- Mareike Höckendorff
- Bastian Lomsché
- Jan Christoph Meister
- Marco Petris
- Wilhelm Schernus
- Lena Schüch
- Katrin Schönert
- Meike Stegkemper

Welcome DH2012 from the Vice President Research, University of Hamburg

Hans Siegfried Stiehl
University of Hamburg, Germany

With about 40,000 students, 680 professors and 4200 research staff, the University of Hamburg is one of Germany's largest universities which comprises six schools: Law; Business, Economics and Social Sciences; Medicine; Education, Psychology and Human Movement; Humanities; Mathematics, Informatics and Natural Sciences. As of this spring, these schools are home to about 300 collaborative research projects funded by the German Research Foundation (DFG), the Federal Ministry of Education and Research and the Seventh Framework Programme of the European Commission. This includes 28 DFG-Collaborative Research Centres, DFG-Research Training Groups, and DFG-Research Groups, as well as the Excellence Cluster 'Integrated Climate System Analysis and Prediction (CLISAP).'

From the mid-1990s on, researchers in the School of Humanities and the Department of Informatics at University of Hamburg began to explore the potential for cooperation in an emerging field then still referred to as 'Computational Philology.' Eventually in the School of Humanities the 'Arbeitsstelle Computerphilologie', one of the first institutions of its kind in Germany, was established.

Today the use of eScience and eHumanities approaches and technology has become part of the daily routine of an ever rising number of scholars and students. 'Digital Diversity: Cultures, Languages and Methods', the motto for this year's Digital Humanities conference, relates methodical and technical innovation to the traditional research agenda of the Humanities – a relation fostering the novel research paradigm DH that is of particular interest to University of Hamburg. Indeed, Digital Humanities methods play a vital role in some of our most advanced and prestigious research initiatives, such as the DFG-funded Collaborative Research Centre 'Manuscript Cultures in Asia, Africa and Europe.' In the context of this interdisciplinary research project the traditional focus on cultural diversity, which has been characteristic for our university from its very beginning in 1919, goes hand in hand with methodical and technical innovation. Projects like this demonstrate the relevance of spurring further exchange among the research paradigms of the humanities, informatics and of computational science. I am certain that the current conference is bound to make a significant contribution to further building bridges.

The University of Hamburg is therefore delighted to host the Digital Humanities 2012 conference, and it gives me great pleasure to welcome you to our university as well as to the Free and Hanseatic City of Hamburg. For us, the DH 2012 is one of the most important academic events in this year's calendar, and I wish the conference every success!

Chair of International Programme Committee

Paul Spence
Department of Digital Humanities, King's College London, UK

A recurring theme at Digital Humanities conferences in recent years has been the high number of submissions, and this year has continued the upward trend, with close to 400 submissions across the different categories: pre-conference workshops/tutorials, long papers, short papers, posters and multi-paper sessions (including panels). I take this as a sign that the field continues to grow and develop, and the quality of the submissions this year certainly made the job of the International Programme Committee challenging (in a good way), although thanks to the excellent facilities provided by our Hamburg hosts we have been able to expand the conference to five strands this year, meaning that this year's conference has more contributions, and by more participants, than most DH conferences in the past.

The theme for this year's conference was 'Digital Diversity: Cultures, languages and methods' and the conference schedule includes contributions on a wide range of topics, reflecting the increasing breadth in the field on all levels. The conference offers opportunities to explore new themes, acknowledges the increasing linguistic diversity of the field and reflects the growth of digital humanities centres and networks in new regions of the world. Both of our keynote speakers reflect on this diversity: Claudine Moulin will explore the challenges in developing interdisciplinary and transnational research structures, with particular consideration for the role of digital humanities; Masahiro Shimoda contemplates the relationship of the field to the wider humanities from a historical and cultural perspective.

I would like to thank all those who submitted proposals this year and all those who agreed to act as reviewers – your contributions on both fronts ensured that the conference continues to reflect the very best of digital scholarship in the humanities at this moment in time. We enlarged our group of reviewers this year, both in anticipation of increased submissions and in a concerted effort to build on the good work of previous PC chairs in broadening the geographic coverage of our reviewer pool.

I would like to give my thanks to the members of the International Programme Committee, who this year included: Arianna Ciula (ALLC), Tanya Clement (ACH), Michael Eberle-Sinatra (SDH-SEMI), Dot Porter (ACH), Jan Rybicki (ALLC), Jon Saklofske (SDH-SEMI), Tomoji Tabata (ALLC) and Katherine Walter (ACH). I would particularly like to thank the Vice Chair Susan Brown (SDH-SEMI) whose advice and good judgement were a great help throughout. Finally, I wish to thank the local organizers, in particular Jan Christoph Meister and Katrin Schönert, for their hard work and support in finding rooms, making practical suggestions and showing the energy and creativity which promise to make this an outstanding conference.

Welcome ashore!

Jan Christoph Meister and Katrin Schönert
University of Hamburg, Germany

The use of maritime metaphor is tempting in Hamburg. Our harbor is 824 years old this year, counts no. 3 in Europe and ranks among the top 15 in the world. From the days of the Hanseatic League to present, it has been the main driver of local economy and become Germany's 'Gateway to the World'. However, this year the *Freie und Hansestadt Hamburg*, the *Free and Hanseatic City of Hamburg*, is also the port of call for digital humanists. A hearty 'Welcome!' from your local organizers – drop anchor and come ashore: This is where DH 2012 happens!

Present day activity in Hamburg's port is all about cargo, but until the mid-20th Century its piers were also lined by ocean steamers that carried immigrants to the New World. Does the exchange of goods come before or after the exchange of people and ideas? In our globalized world where cultures meet and mingle across all domains – commerce, education, politics, knowledge – the philosophical question of primacy seems a bit old fashioned. As 21st century humanists we will of course not deny the importance of the material realm, and as digital humanists the relevance of technological advancement is part of our credo anyhow. On the other hand, we are by definition traditionalists. For our ultimate interest rests with people and cultures, their past, their present and their future, and with the various symbolic languages – natural, textual, visual, musical, material, abstract, concrete or performative – through which we communicate and reflect our human concerns. These are still the essential questions that motivate our endeavor, whether digital or traditional.

At the DH 2011 we gathered under the colorful California 'Big Tent' erected by our Stanford hosts – it was a marvelous experience to see how many we have become, how the field has grown into one of the most vibrant scientific communities, and how we collaborate in building bridges between the digital and the humanities. We speak one language!

And at the same time, we speak many. We're part of a unique intellectual culture that is only possible because of the multitude of human cultures that we come from and from which we contribute to our common cause. We carry intellectual cargo to a communal *agora* of ideas – ideas and concepts shipped from the many facetted cultural, philosophical, epistemological and methodological contexts and domains of the humanities. The annual DH conference is the biggest intellectual market place of our scientific community, and it is hard to imagine a venue where more could be on offer.

Our conference motto 'Digital Diversity – Cultures, languages and methods' underlines that which motivates and, at the same time, makes this intellectual exchange worthwhile: Diversity. It took us some time to discover its worth: For about two decades, when DH was still called Humanities Computing, we discussed whether what we practiced would not perhaps justify the formation of a discipline. For some of us this seemed a desirable status to attain; for others not. However, in today's perspective one thing is obvious: the reality of Digital Humanities cuts across traditional conceptions of carefully delineated disciplines and their traditional definition in terms of object domain, methodology, institution, degree course, journal, etc. Conceptually as well as institutionally, DH thrives on diversity, and what we do cannot be reduced to a single purpose, object domain or method. Conformity is absolutely not what DH is about, and it is puzzling why scholars outside our community still feel tempted to reduce DH to what the field itself has long transcended in theory as well as in practice: the mere application of computer technology.

One of the aims of DH conferences is of course to show case the many facets of contemporary digital humanities. Still, not every traditional humanist will be easily convinced. On that score, the history of Hamburg and its university may perhaps offer a good example for how single-mindedness, attributed or professed, can in the end be nevertheless subverted and brought to fruit. Hamburg politics at the beginning of the 20th century was dominated by the interests of merchants, bankers, and owners of shipping companies (mind you, it still is). Intellectual capital, it was held, was cheaper bought in than locally produced; so why invest in a university? It was only in 1919 when Werner von Melle, First Mayor

and an ardent educationalist, convinced his colleagues in the Senate that *their* business – the exchange of goods and money between nations – would thrive if the *Humanities'* business was looked after: the exchange of knowledge and ideas between cultures. The intellectual encounter with other cultures and languages through academic education, von Melle successfully argued, was necessary to sustain and further develop commerce with other nations. This eventually lead to the founding of the very university which, almost a century later, is today proud to host this DH conference. Our wish as local organizers is that the DH 2012 may present an equally persuasive example to those scholars who question the need to explore what at first glance might appear to be a foreign methodological paradigm.

Today Hamburg University is Germany's fifth largest in terms of student intake, and at the same time one which, true to the meaning of *universitas*, engages in teaching and research across the entire spectrum of the human, the social, the natural and the medical sciences. In the Faculty of the Humanities over one hundred degree courses are on offer, many of which focus on foreign languages and cultures. This diversity of cultures, languages and methods makes for an intellectual environment attracting over ten thousand students to the Humanities, that is close to one third of the university's total student population.

What if *Digital* Humanities became a topic of interest to each and every one of these students, if only in passing? And what if one would achieve this elsewhere too, at universities across the world? Ten years ago this would have sounded like a lunatic vision, but today DH certainly enjoys a strong increase in attention that has put its methodology on many people's agenda. This development has been strongly supported by research funding agencies and institution building, such as the formation of ADHO and of new DH associations across the world. Incidentally, by the time you read this the inaugural meeting of the DHD as a German chapter should be under way: it is scheduled to take place immediately prior to the DH 2012, and also at Hamburg University.

Perhaps even more impressive than the rise in external support is the tremendous internal dynamics of the DH community. The volume which you currently hold in hand (or read on screen) contains the abstracts of some 200 papers, panel sessions, posters and workshops. The Call for Papers attracted a record number of proposals, and even though the conference program was eventually extended to five parallel tracks, only half of the close to four hundred submissions could in the end be accommodated. The International Program Committee chaired by Paul Spence worked extremely hard in order to announce the results earlier than usual, but in the end our ambitious deadline could not be met: not only because of numbers involved, but more importantly because of the high quality of submissions which confronted reviewers with many hard choices. It is sad that so many excellent proposals could not be accepted, but this painstaking process also testifies to the very high standard that DH research has nowadays attained. As local organizers our deepest gratitude goes to everyone who submitted a proposal, whether successful or not, to the army of reviewers who dedicated their time to careful reviewing, and to Paul and the program committee for drawing up what is certainly one of the richest conference agendas in our field.

Thanks are also due to ADHO's conference coordinating committee and its chair, John Unsworth. The guidance that John and his team provide comes in two forms: That of carefully thought out and well documented protocols, and that of John's concise and immediate response to a frantic organizers' plea for help, mercy and redemption. The former is available to everyone via the ADHO website. The latter is reserved for those who have qualified through an act of folly committed about two years prior, an act usually referred to as a 'conference bid'. However, once you find yourself on the spot you may rest assured that advice and support will be granted magnanimously. The conference as well as the program committee and their chairs will be with you all the way, as they were with us. Both Paul and John have served as local organizers themselves, and through ADHO invaluable 'institutional knowledge' such as theirs can be passed on to others – not only the type of knowledge that is formalized in protocols, but also the personal experiences that one needs to draw on in moments of crisis. This is where organization building provides tangible benefits to a scientific community as ours.

Can one also say 'Thank you!' to a piece of software? It feels a bit like talking to your hammer, but then who of us knows the inventor of the hammer him- or herself? We do, however, know the person who invented ConfTool, the conference management software provided by ADHO which many previous DH organizers have used and which we have found to be absolutely indispensible. Incidentally, Harald Weinreich, its creator, lives and works in – Hamburg. (Honestly, we didn't know that when we bid

for DH 2012!) While ConfTool came to our aid as an administrative backend, the design of the public image, of the DH2012 logo and website, were the work of Leon and his developers. Indeed, making a DH conference happen is like producing a movie, and at the end you wish you could run a long list of credits. Our list would include Basti, Benjamin, Daniel, Lena, Meike and Willi who took on the technical production of the Book of Abstracts; Marco who contributed sysad wizzadry and firefighting, and Evelyn and Imke who specialized in what is perhaps best and with intentional opacity referred to as 'administrative diplomacy'. Many others deserve mentioning, but cannot be named here – so please just imagine those movie credits, accompanied by a suitably dramatic sound track ('Don't dream it's over' by *Crowded House* of 1986 will do just fine). To all of you whose names appear on that long, long list of names and who helped us make DH 2012 happen we extend a big, a VERY BIG thank you!

An important goal of this year's DH was to make sure that the conference motto would become meaningful in as many facets of conference reality as possible – through the composition of the academic and social program, through the choice of keynote speakers and their topics, and through regional representation in our team of international student conference assistants. Applications for the bursaries that were on offer were scrutinized by Elisabeth Burr, Chair of ADHO's Multi Cultural and Multi Lingual committee, and her team of reviewers, and awards were then made to twelve young DH scholars from all over the world. Each of them partners with one Hamburg student and we trust that this experience will inspire them to engage in joint, international DH student projects in the future. To the award recipients as well as to their local partners we say: welcome! It's great to have you on our team!

This ambitious project in particular could not have been realized without the very generous financial support of the DH 2012 granted by the Deutsche Forschungsgemeinschaft (DFG), and the equally generous support granted by the University of Hamburg whose President, Prof. Dieter Lenzen and Vice-President Research, Prof. Hans Siegfried Stiehl, were both extremely supportive. A welcome note is not the place to talk money; suffice it to say that without this support, the conference fee would probably have trebled. Our gratitude to both institutions is to that order, and more. We also thank Google Inc. who sponsor part of the social program.

In the light of the brief episode about our university's history mentioned above we particularly thank the Senator of Science and Research, Dr. Dorothee Stapelfeldt, for inviting the DH 2012 conference attendees to a welcome reception at Hamburg's City Hall. This is a great and exceptional honor not often bestowed on academic conferences in Hamburg, and it is a sign of acknowledgment of the role and importance of the humanities at large that is very much appreciated.

And with that there's only one more person we haven't said 'Thank you!' to. That person is – you. It's great you came and moored your vessel. Whether by boat or ship, the cargo you have brought to DH 2012 is what this convention is all about: The exchange of ideas. So discharge your bounty, come ashore, and help us hoist the banner of Digital Diversity at the DH 2012!

Your local organizers

Chris Meister, Katrin Schönert & team

Obituary for Gerhard Brey (1954-2012)

Harold Short
Dept of Digital Humanities King's College London, UK; Visiting Professor University of Western Sydney, Australia

We mourn the loss of our friend and colleague Gerhard Brey, who died in February this year after a short illness.

Gerhard was a remarkable man, with interests and expertise that bridged the humanities and the sciences, including computation. His humanistic background was extremely rich and varied. It included his fluency in several western European languages, including classical Greek and Latin, and also Sanskrit and Arabic. Part of the bridge was his interest and wide reading in the history of science. In terms of computation he developed considerable expertise across a range of programming languages and tools, from digital editing and typesetting to databases to text mining.

Gerhard was born in a small German town near the Austrian border, close to Salzburg. His early academic and professional career was in Germany, mainly in Munich, apart from a year spent studying in France. This was personally as well as professionally important, because it was at the university in Clermont-Ferrand that he met Gill, his wife and life-long companion. In 1996 Gerhard and Gill moved to England, and in 2001 he began working with Dominik Wujastyk at the Wellcome Institute. (For a longer and more detailed obituary, see the one by Wujastyk on the ALLC website at allc.org.)

Gerhard came to work at what is now the Department of Digital Humanities at King's College London in 2004, first as a part-time consultant, then as Research Fellow and Senior Research Fellow. His wide-ranging expertise in both humanities disciplines and technical matters meant that he was not only directly involved in numerous research projects, but was also consulted by colleagues in many more. While it is right to acknowledge Gerhard's formidable humanistic and technical range, it is the human being who is most sorely missed. He was deeply interested in people, and had a real gift for communication and engagement. We remember him especially for his calmness, patience, ready good humour and his self-deprecating sense of fun. As his head of department, I valued particularly his willingness to lend a hand in a crisis.

It was also Gerhard who took the lead in nurturing links with Japanese colleagues that led to a joint research project with the University of Osaka, and in the last years of his life he developed a passionate interest in Japanese culture, drawn in part by the work of our colleagues in Buddhist Studies, which spoke to his love of Sanskrit and Arabic literatures. As a colleague remarked, his gentle, quiet manner concealed a fine mind and wider and deeper learning than even his friends tended to expect. He will be more sorely missed than any of us can say.

Bursary Winners

DH 2012 Student Conference Assistant Bursaries

Kollontai Cossich Diniz (University of Sao Paulo, Sao Paulo, Brazil)

Peter Daengeli (National University of Ireland Maynooth, Dublin, Ireland)

Elena Dergacheva (University of Alberta, Edmonton, Canada)

Ali Kira Grotkowski (University of Alberta, Edmonton, Canada)

Dagmara Hadyna (Jagiellonian University, Kielce, Poland)

Biljana Djordje Lazic (University of Belgrade, Belgrade, Serbia)

Yoshimi Iwata (Doshisha University, Kyoto, Japan)

Jie Li Kwok (Dharma Drum Buddhist College, New Taipei City, Taiwan)

Jose Manuel Lopez Villanueva (Universidad Nacional Autonoma de Mexico, Mexico City, Mexico)

Alex Edward Marse (2CINEM, Atlanta, USA)

Gurpreet Singh (Morph Academy, Fatehgarh Sahib, India)

Rosa Rosa Souza Rosa Gomes (University of Sao Paulo, Sao Paulo, Brazil)

Table of Contents

List of Reviewers 1

Plenary Sessions

Dynamics and Diversity: Exploring European and Transnational Perspectives on Digital Humanities Research Infrastructures
Moulin, Claudine 9

Embracing a Distant View of the Digital Humanities
Shimoda, Masahiro 11

Pre-conference Workshops

Digital Methods in Manuscript Studies
Brockmann, Christian; Wangchuk, Dorji 15

Introduction to Stylomatic Analysis using R
Eder, Maciej; Rybicki, Jan 16

NeDiMAH workshop on ontology based annotation
Eide, Øyvind; Ore, Christian-Emil; Rahtz, Sebastian 18

Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts
Hinrichs, Erhard; Neuroth, Heike; Wittenburg, Peter 20

Here and There, Then and Now – Modelling Space and Time in the Humanities
Isaksen, Leif; Day, Shawn; Andresen, Jens; Hyvönen, Eero; Mäkelä, Eetu 22

Crowdsourcing meaning: a hands-on introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment
Petris, Marco; Gius, Evelyn; Schüch, Lena; Meister, Jan Christoph 24

Learning to play like a programmer: web mash-ups and scripting for beginners
Ridge, Mia 25

Introduction to Distant Reading Techniques with Voyant Tools, Multilingual Edition
Sinclair, Stéfan; Rockwell, Geoffrey 26

Towards a reference curriculum for the Digital Humanities
Thaller, Manfred 27

Free your metadata: a practical approach towards metadata cleaning and vocabulary reconciliation
van Hooland, Seth; Verborgh, Ruben; De Wilde, Max 28

Panels

Text Analysis Meets Text Encoding
Bauman, Syd; Hoover, David; van Dalen-Oskam, Karina; Piez, Wendell 33

Designing Interactive Reading Environments for the Online Scholarly Edition
Blandford, Ann; Brown, Susan; Dobson, Teresa; Faisal, Sarah; Fiorentino, Carlos; Frizzera, Luciano; Giacometti, Alejandro; Heller, Brooke; Ilovan, Mihaela; Michura, Piotr; Nelson, Brent; Radzikowska, Milena; Rockwell, Geoffrey; Ruecker, Stan; Sinclair, Stéfan; Sondheim, Daniel; Warwick, Claire; Windsor, Jennifer 35

Developing the spatial humanities: Geo-spatial technologies as a platform for cross-disciplinary scholarship
Bodenhamer, David; Gregory, Ian; Ell, Paul; Hallam, Julia; Harris, Trevor; Schwartz, Robert 41

Prosopographical Databases, Text-Mining, GIS and System Interoperability for Chinese History and Literature
Bol, Peter Kees; Hsiang, Jieh; Fong, Grace 43

Future Developments for TEI ODD
Cummings, James; Rahtz, Sebastian; Burnard, Lou; Bauman, Syd; Gaiffe, Bertrand; Romary, Laurent; Bański, Piotr 52

Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of Publication of Digitized Historical Prints
Geyken, Alexander; Gloning, Thomas; Stäcker, Thomas 54

Computational models of narrative structure
Löwe, Benedikt; Físseni, Bernhard; León, Carlos; Bod, Rens 57

Approaches to the Treatment of Primary Materials in Digital Lexicons: Examples of the New Generation of Digital Lexicons for Buddhist Studies
Nagasaki, Kiyonori; Tomabechi, Toru; Wangchuk, Dorji; Takahashi, Koichi; Wallman, Jeff; Muller, A. Charles 61

Topic Modeling the Past
Nelson, Robert K.; Mimno, David; Brown, Travis 64

Facilitating Research through Social-Document Networks
Pitti, Daniel; Simon, Agnès; Vitali, Stefano; Arnold, Kerstin 70

Digital Humanities as a university degree: The status quo and beyond
Thaller, Manfred; Sahle, Patrick; Clavaud, Florence; Clement, Tanya; Fiormonte, Domenico; Pierazzo, Elena; Rehbein, Malte; Rockwell, Geoffrey; Schreibman, Susan; Sinclair, Stéfan 72

Papers

Exploring Originality in User-Generated Content with Network and Image Analysis Tools
Akdag Salah, Alkim Almila; Salah, Albert Ali; Douglass, Jeremy; Manovich, Lev 79

Patchworks and Field-Boundaries: Visualizing the History of English
Alexander, Marc 82

Developing Transcultural Competence in the Study of World Literatures: Golden Age Literature Glossary Online (GALGO)
Alonso Garcia, Nuria; Caplan, Alison 84

Trees of Texts – Models and methods for an updated theory of medieval text stemmatology
Andrews, Tara Lee; Macé, Caroline 85

Mapping the Information Science Domain
Arazy, Ofer; Ruecker, Stan; Rodriguez, Omar; Giacometti, Alejandro; Zhang, Lu; Chun, Su 88

Words made Image. Towards a Language-Based Segmentation of Digitized Art Collections
Armaselu, Florentina 91

HisDoc: Historical Document Analysis, Recognition, and Retrieval
Baechler, Micheal; Fischer, Andreas; Naji, Nada; Ingold, Rolf; Bunke, Horst; Savoy, Jacques 94

Research infrastructures for Digital Humanities: The local perspective
Bärenfänger, Maja; Binder, Frank 97

Pelagios: An Information Superhighway for the Ancient World
Barker, Elton; Simon, Rainer; Isaksen, Leif 99

Putting TEI Tite to use – generating a database resource from a printed dictionary or reference type publication
Barner-Rasmussen, Michael 102

Digital Humanities in the Classroom: Introducing a New Editing Platform for Source Documents in Classics
Beaulieu, Marie-Claire; Almas, Bridget 105

DiaView: Visualise Cultural Change in Diachronic Corpora
Beavan, David 107

Catch + Release: Research and Creation of a Digital New Media Exhibition in the Context of a Cultural and Heritage Museum
Beer, Ruth 109

Opportunity and accountability in the 'eResearch push'
Bellamy, Craig 111

Connecting European Women Writers. The Selma Lagerlöf Archive and Women Writers Database
Bergenmar, Jenny; Olsson, Leif-Jöran 113

Stylometric Analysis of Chinese Buddhist texts: Do different Chinese translations of the 'Gandhavyūha' reflect stylistic features that are typical for their age?
Bingenheimer, Marcus; Hung, Jen-Jou; Hsieh, Cheng-en 115

Information Extraction on Noisy Texts for Historical Research
Blanke, Tobias; Bryant, Michael; Speck, Reto; Kristel, Conny 117

Modeling Gender: The 'Rise and Rise' of the Australian Woman Novelist
Bode, Katherine 119

Contextual factors in literary quality judgments: A quantitative analysis of an online writing community
Boot, Peter 121

Violence and the Digital Humanities Text as Pharmakon
Bradley, Adam James 123

Towards a bibliographic model of illustrations in the early modern illustrated book
Bradley, John; Pigney, Stephen 124

Automatic Mining of Valence Compounds for German: A Corpus-Based Approach
Brock, Anne; Henrich, Verena; Hinrichs, Erhard; Versley, Yannick 126

Networks of networks: a critical review of formal network methods in archaeology through citation network analysis and close reading
Brughmans, Tom 129

On the dual nature of written texts and its implications for the encoding of genetic manuscripts
Brüning, Gerrit; Henzel, Katrin; Pravida, Dietmar 131

Automatic recognition of speech, thought and writing representation in German narrative texts
Brunner, Annelen 135

Bringing Modern Spell Checking Approaches to Ancient Texts – Automated Suggestions for Incomplete Words
Büchler, Marco; Kruse, Sebastian; Eckart, Thomas 137

Designing a national 'Virtual Laboratory' for the humanities: the Australian HuNI project
Burrows, Toby Nicolas 139

Beyond Embedded Markup
Buzzetti, Dino; Thaller, Manfred 142

Myopia: A Visualization Tool in Support of Close Reading
Chaturvedi, Manish; Gannod, Gerald; Mandell, Laura; Armstrong, Helen; Hodgson, Eric 148

Translation Arrays: Exploring Cultural Heritage Texts Across Languages
Cheesman, Tom; Thiel, Stephan; Flanagan, Kevin; Zhao, Geng; Ehrmann, Alison; Laramee, Robert S.; Hope, Jonathan; Berry, David M. 151

Constructing a Chinese as Second Language Learner Corpus for Language Learning and Research
Chen, Howard 154

Social Curation of large multimedia collections on the cloud
Chong, Dazhi; Coppage, Samuel; Gu, Xiangyi; Maly, Kurt; Wu, Harris; Zubair, Mohammad 155

Sounding for Meaning: Analyzing Aural Patterns Across Large Digital Collections
Clement, Tanya; Auvil, Loretta; Tcheng, David; Capitanu, Boris; Monroe, Megan; Goel, Ankita 158

The Programming Historian 2: A Participatory Textbook
Crymble, Adam H.; MacEachern, Alan; Turkel, William J. 162

Multilingual and Semantic Extension of Folk Tale Catalogues
Declerck, Thierry; Lendvai, Piroska; Darányi, Sándor 163

Digital Language Archives and Less-Networked Speaker Communities
Dobrin, Lise M. 167

Language Documentation and Digital Humanities: The (DoBeS) Language Archive
Drude, Sebastian; Trilsbeek, Paul; Broeder, Daan 169

The potential of using crowd-sourced data to re-explore the demography of Victorian Britain
Duke-Williams, Oliver William 173

Sharing Ancient Wisdoms: developing structures for tracking cultural dynamics by linking moral and philosophical anthologies with their source and recipient texts
Dunn, Stuart; Hedges, Mark; Jordanous, Anna; Lawrence, Faith; Roueche, Charlotte; Tupman, Charlotte; Wakelnig, Elvira 176

Recovering the Recovered Text: Diversity, Canon Building, and Digital Studies
Earhart, Amy 179

Mind your corpus: systematic errors in authorship attribution
Eder, Maciej 181

Underspecified, Ambiguous or Formal. Problems in Creating Maps Based on Texts
Eide, Øyvind 185

A Frequency Dictionary of Modern Written and Oral Media Arabic
Elmaz, Orhan 188

Texts in Motion – Rethinking Reader Annotations in Online Literary Texts
Fendt, Kurt E.; Kelley, Wyn; Zhang, Jia; Della Costa, Dave 190

May Humanists Learn from Artists a New Way to Interact with Digital Technology?
Franchi, Stefano 192

A flexible model for the collaborative annotation of digitized literary works
Gayoso-Cabada, Joaquin; Ruiz, Cesar; Pablo-Nuñez, Luis; Sarasa-Cabezuelo, Antonio; Goicoechea-de-Jorge, Maria; Sanz-Cabrerizo, Amelia; Sierra-Rodriguez, Jose-Luis 195

HyperMachiavel: a translation comparison tool
Gedzelman, Séverine; Zancarini, Jean-Claude 198

Discrimination sémantique par la traduction automatique, expériences sur le dictionnaire français de Littré
Glorieux, Frédéric; Jolivet, Vincent 202

The Myth of the New: Mass Digitization, Distant Reading and the Future of the Book
Gooding, Paul Matthew; Warwick, Claire; Terras, Melissa 204

Designing Navigation Tools for an Environmental Humanities Portal: Considerations and Critical Assessments
Graf von Hardenberg, Wilko; Coulter, Kimberly 206

Processing Email Archives in Special Collections
Hangal, Sudheendra; Chan, Peter; Lam, Monica S.; Heer, Jeffrey 208

The Stylometry of Collaborative Translation
Heydel, Magda; Rybicki, Jan 212

Focus on Users in the Open Development of the National Digital Library of Finland
Hirvonen, Ville; Kautonen, Heli Johanna 215

The Rarer They Are, the More There Are, the Less They Matter
Hoover, David 218

Experiments in Digital Philosophy – Putting new paradigms to the test in the Agora project
Hrachovec, Herbert; Carusi, Annamaria; Huentelmann, Raphael; Pichler, Alois; Antonio, Lamarra; Cristina, Marras; Alessio, Piccioli; Lou, Burnard 221

Information Discovery in the Chinese Recorder Index
Hsiang, Jieh; Kong, Jung-Wei; Sung, Allan 224

Complex Network Perspective on Graphic Form System of Hanzi
Hu, Jiajia; Wang, Ning 228

A Computer-Based Approach for Predicting the Translation Time Period of Early Chinese Buddhism Translation
Hung, Jen-Jou; Bingenheimer, Marcus; Kwok, Jieli 230

Bridging Multicultural Communities: Developing a Framework for a European Network of Museum, Libraries and Public Cultural Institutions
Innocenti, Perla; Richards, John; Wieber, Sabine 232

Ptolemy's Geography and the Birth of GIS
Isaksen, Leif 236

Tracing the history of Noh texts by mathematical methods. Validitating the application of phylogenetic methods to Noh texts
Iwata, Yoshimi 239

Computing and Visualizing the 19th-Century Literary Genome
Jockers, Matthew 242

Using the Google Ngram Corpus to Measure Cultural Complexity
Juola, Patrick 245

'All Rights Worth Recombination': Post-Hacker Culture and ASCII Literature (1983-1993)
Katelnikoff, Joel 247

Evaluating Unmasking for Cross-Genre Authorship Verification
Kestemont, Mike; Luyckx, Kim; Daelemans, Walter; Crombez, Thomas 249

Literary Wikis: Crowd-sourcing the Analysis and Annotation of Pynchon, Eco and Others
Ketzan, Erik 252

Social Network Analysis and Visualization in 'The Papers of Thomas Jefferson'
Klein, Lauren Frederica 254

VariaLog: how to locate words in a French Renaissance Virtual Library
Lay, Marie Hélène 256

DeRiK: A German Reference Corpus of Computer-Mediated Communication
Lemnitzer, Lothar; Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Storrer, Angelika 259

Estimating the Distinctiveness of Graphemes and Allographs in Palaeographic Classification
Levy, Noga; Wolf, Lior; Dershowitz, Nachum; Stokes, Peter 264

Academic Research in the Blogosphere: Adapting to New Opportunities and Risks on the Internet
Littauer, Richard; Winters, James; Roberts, Sean; Little, Hannah; Pleyer, Michael; Benzon, Bill 268

Feeling the View: Reading Affective Orientation of Tagged Images
Liu, Jyi-Shane; Peng, Sheng-Yang 270

Characterizing Authorship Style Using Linguistic Features
Lucic, Ana; Blake, Catherine 273

Investigating the genealogical relatedness of the endagered Dagon languages
Moran, Steven; Prokic, Jelena 276

Landscapes, languages and data structures: Issues in building the Placenames Database of Ireland
Měchura, Michal Boleslav 278

Interoperability of Language Documentation Tools and Materials for Local Communities
Nakhimovsky, Alexander; Good, Jeff; Myers, Tom 280

Content Creation by Domain Experts in a Semantic GIS System
Nakhimovsky, Alexander; Myers, Tom 283

From Preserving Language Resources to Serving Language Speakers: New Prospects for Endangered Languages Archives
Nathan, David John 286

Retrieving Writing Patterns From Historical Manuscripts Using Local Descriptors
Neumann, Bernd; Herzog, Rainer; Solth, Arved; Bestmann, Oliver; Scheel, Julian 288

Distractorless Authorship Verification
Noecker Jr., John; Ryan, Michael 292

Cataloguing linguistic diversity: Glottolog/Langdoc
Nordhoff, Sebastian; Hammarström, Harald 296

Geo-Temporal Interpretation of Archival Collections Using Neatline
Nowviskie, Bethany; Graham, Wayne; McClure, David; Boggs, Jeremy; Rochester, Eric 299

Enriching Digital Libraries Contents with SemLib Semantic Annotation System
Nucci, Michele; Grassi, Marco; Morbidoni, Christian; Piazza, Francesco 303

The VL3: A Project at the Crossroads between Linguistics and Computer Science
Nuñez, Camelia Gianina; Mavillard, Antonio Jiménez 306

'Eric, you do not humble well': The Image of the Modern Vampire in Text and on Screen
Opas-Hänninen, Lisa Lena; Hettel, Jacqueline; Toljamo, Tuomo; Seppänen, Tapio 308

Electronic Deconstruction of an argument using corpus linguistic analysis of its on-line discussion forum supplement
O'Halloran, Kieran Anthony 310

Citygram One: Visualizing Urban Acoustic Ecology
Park, Tae Hong; Miller, Ben; Shrestha, Ayush; Lee, Sangmi; Turner, Jonathan; Marse, Alex 313

Towards Wittgenstein on the Semantic Web
Pichler, Alois; Zöllner-Weber, Amélie 318

Uncovering lost histories through GeoStoryteller: A digital GeoHumanities project
Rabina, Debbie L.; Cocciolo, Anthony 322

Workflows as Structured Surfaces
Radzikowska, Milena; Ruecker, Stan; Rockwell, Geoffrey; Brown, Susan; Frizzera, Luciano; INKE Research Group 324

Code-Generation Techniques for XML Collections Interoperability
Ramsay, Stephen; Pytlik-Zillig, Brian 327

Uncertain Date, Uncertain Place: Interpreting the History of Jewish Communities in the Byzantine Empire using GIS
Rees, Gethin Powell 329

Code sprints and Infrastructure
Reside, Doug; Fraistat, Neil; Vershbow, Ben; van Zundert, Joris Job 331

Digital Genetic Criticism of RENT
Reside, Doug 333

On the Internet, nobody knows you're a historian: exploring resistance to crowdsourced resources among historians
Ridge, Mia 335

Formal Semantic Modeling for Human and Machine-based Decoding of Medieval Manuscripts
Ritsema van Eck, Marianne Petra; Schomaker, Lambert 336

The Swallow Flies Swiftly Through: An Analysis of Humanist
Rockwell, Geoffrey; Sinclair, Stéfan 339

The Digital Mellini Project: Exploring New Tools & Methods for Art-historical Research & Publication
Rodríguez, Nuria; Baca, Murtha; Albrezzi, Francesca; Longaker, Rachel 342

Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research
Roe, Glenn H.; The ARTFL Project 345

Engaging the Museum Space: Mobilising Visitor Engagement with Digital Content Creation
Ross, Claire Stephanie; Gray, Steven; Warwick, Claire; Hudson Smith, Andrew; Terras, Melissa 348

Aiding the Interpretation of Ancient Documents
Roued-Cunliffe, Henriette 351

The Twelve Disputed 'Federalist' Papers: A Case for Collaboration
Rudman, Joseph 353

Writing with Sound: Composing Multimodal, Long-Form Scholarship
Sayers, Jentery 357

Intra-linking the Research Corpus: Using Semantic MediaWiki as a lightweight Virtual Research Environment
Schindler, Christoph; Ell, Basil; Rittberger, Marc 359

Corpus Coranicum: A digital landscape for the study of the Qu'ran
Schnöpf, Markus 362

The MayaArch3D Project: A 3D GIS Web System for Querying Ancient Architecture and Landscapes
Schwerin, Jennifer von; Richards-Rissetto, Heather; Agugiaro, Giorgio; Remondino, Fabio; Girardi, Gabrio 365

Multi-dimensional audio-visual technology: Evidence from the endangered language documentation
Sharma, Narayan P. 368

Contours of the Past: Computationally Exploring Civil Rights Histories
Shaw, Ryan Benjamin 370

Notes from the Collaboratory: An Informal Study of an Academic DH Lab in Transition
Siemens, Lynne; Siemens, Raymond 373

XML-Print: an Ergonomic Typesetting System for Complex Text Structures
Sievers, Martin; Burch, Thomas; Küster, Marc W.; Moulin, Claudine; Rapp, Andrea; Schwarz, Roland; Gan, Yu 375

Federated Digital Archives and Disaster Recovery: The Role of the Digital Humanities in Post-earthquake Christchurch
Smithies, James Dakin 380

Modeling Medieval Handwriting: A New Approach to Digital Palaeography
Stokes, Peter 382

A Digital Geography of Hispanic Baroque Art
Suárez, Juan-Luis; Sancho-Caparrini, Fernando 385

Approaching Dickens' Style through Random Forests
Tabata, Tomoji 388

Interfacing Diachrony: Visualizing Linguistic Change on the Basis of Digital Editions of Serbian 18th-Century Texts
Tasovac, Toma; Ermolaev, Natalia 392

Promise and Practice of Enhanced Publications to Complement Conventionally-Published Scholarly Monographs
Tatum, Clifford; Jankowski, Nicholas; Scharnhorst, Andrea 394

Culpeper's legacy: How title pages sold books in the 17th century
Tyrkkö, Jukka Jyrki Juhani; Suhr, Carla Maria; Marttila, Ville 396

The Differentiation of Genres in Eighteenth- and Nineteenth-Century English Literature
Underwood, Ted; Sellers, Jordan; Auvil, Loretta; Capitanu, Boris 397

Digital editions with eLaborate: from practice to theory
van Dalen-Oskam, Karina; van Zundert, Joris Job 400

Delta in 3D: Copyists Distinction by Scaling Burrows's Delta
van Zundert, Joris Job; van Dalen-Oskam, Karina 402

Wiki Technologies for Semantic Publication of Old Russian Charters
Varfolomeyev, Aleksey; Ivanovs, Aleksandrs 405

L'histoire de l'art à l'ère numérique – Pour une historiographie médiologique
Welger-Barboza, Corinne 407

Benefits of tools and applications for a digitized analysis of Chinese Buddhist inscriptions
Wenzel, Claudia 411

The ARTeFACT Movement Thesaurus: toward an open-source tool to mine movement-derived data
Wiesner, Susan L.; Bennett, Bradford; Stalnaker, Rommie L. 413

The electronic 'Oxford English Dictionary', poetry, and intertextuality
Williams, David-Antoine 415

Reasoning about Genesis or The Mechanical Philologist
Wissenbach, Moritz; Pravida, Dietmar; Middell, Gregor 418

The Digital Daozang Jiyao – How to get the edition into the Scholar's labs
Wittern, Christian 422

Posters

A Digital Approach to Sound Symbolism in English: Evidence from the Historical Thesaurus
Alexander, Marc; Kay, Christian 427

Collaborative Video and Image Annotation
Arnold, Matthias; Knab, Cornelia; Decker, Eric 429

Le Système modulaire de gestion de l'information historique (SyMoGIH): une plateforme collaborative et cumulative de stockage et d'exploitation de l'information géo-historique
Beretta, Francesco; Vernus, Pierre; Hours, Bernard 431

Realigning Digital Humanities Training: The Praxis Program at the Scholars' Lab
Boggs, Jeremy; Nowviskie, Bethany; Gil, Alexander; Johnson, Eric; Lestock, Brooke; Storti, Sarah; Swafford, Joanna; Praxis Program Collaborators 433

Supporting the emerging community of MEI: the current landscape of tools for note entry and digital editing
Bohl, Benjamin W.; Röwenstrunk, Daniel; Viglianti, Raffaele 435

'The Past Is Never Dead. It's Not Even Past': The Challenge of Data Provenance in the e-Humanities
Clark, Ashley M.; Holloway, Steven W. 438

The Social Edition: Scholarly Editing Across Communities
Crompton, Constance; Siemens, Raymond; The Devonshire MS Editorial Group 441

Courting 'The World's Wife': Original Digital Humanities Research in the Undergraduate Classroom
Croxall, Brian 443

The Academy's Digital Store of Knowledge
Czmiel, Alexander; Jürgens, Marco 445

Building a TEI Archiving, Publishing, and Access Service: The TAPAS Project
Flanders, Julia; Hamlin, Scott; Alvarado, Rafael; Mylonas, Elli 448

Author Consolidation across European National Bibliographies
Freire, Nuno 450

Historical Events Versus Information Contents – A Preliminary Analysis of the National Geographic Magazine
Fujimoto, Yu 453

'Tejiendo la Red HD' – A case study of building a DH network in Mexico
Galina, Isabel; Priani, Ernesto; López, José; Rivera, Eduardo; Cruz, Alejandro 456

Adaptive Automatic Gesture Stroke Detection
Gebre, Binyam Gebrekidan; Wittenburg, Peter 458

Towards a Transnational Multilingual Caribbean Digital Humanities Lab
Gil, Alexander 462

NUScholar: Digital Methods for Educating New Humanities Scholars
Graff, Ann-Barbara; Lucas, Kristin; Blustein, James; Gibson, Robin; Woods, Sharon 463

Latent Semantic Analysis Tools Available for All Digital Humanities Projects in Project Bamboo
Hooper, Wallace Edd; Cowan, Will; Jiao, David; Walsh, John A. 465

Machine Learning for Automatic Annotation of References in DH scholarly papers
Kim, Young-Min; Bellot, Patrice; Faath, Elodie; Dacos, Marin 467

An Ontology-Based Iterative Text Processing Strategy for Detecting and Recognizing Characters in Folktales
Koleva, Nikolina; Declerck, Thierry; Krieger, Hans-Ulrich 470

Integrated multilingual access to diverse Japanese humanities digital archives by dynamically linking data
Kuyama, Takeo; Batjargal, Biligsaikhan; Kimura, Fuminori; Maeda, Akira 473

Linguistic concepts described with Media Query Language for automated annotation
Lenkiewicz, Anna; Lis, Magdalena; Lenkiewicz, Przemyslaw 477

Virtual Reproduction of Gion Festival Yamahoko Parade
Li, Liang; Choi, Woong; Nishiura, Takanobu; Yano, Keiji; Hachimura, Kozaburo 480

Complex entity management through EATS: the case of the Gascon Rolls Project
Litta Modignani Picozzi, Eleonora; Norrish, Jamie; Monteiro Vieira, Jose Miguel 483

TextGrid Repository – Supporting the Data Curation Needs of Humanities Researchers
Lohmeier, Felix; Veentjer, Ubbo; Smith, Kathleen M.; Söring, Sibylle 486

RIgeo.net – A Lab for Spatial Exploration of Historical Data
Loos, Lukas; Zipf, Alexander 488

Automatic Topic Hierarchy Generation Using Wordnet
Monteiro Vieira, Jose Miguel; Brey, Gerhard † 491

Hypotheses.org, une infrastructure pour les Digital Humanities
Muscinesi, Frédérique 494

TXSTEP – an integrated XML-based scripting language for scholarly text data processing
Ott, Wilhelm; Ott, Tobias; Gasperlin, Oliver 497

Exploring Prosopographical Resources Through Novel Tools and Visualizations: a Preliminary Investigation
Pasin, Michele 499

Heterogeneity and Multilingualism vs. Usability – Challenges of the Database User Interface 'Archiv-Editor'
Plutte, Christoph 502

Medievalists' Use of Digital Resources, 2002 and 2012
Porter, Dot 505

Cross-cultural Approaches to Digital Humanities – Funding and Implementation
Rhody, Jason; Kümmel, Christoph; Effinger, Maria; Freedman, Richard; Magier, David; Förtsch, Reinhard 506

CWRC-Writer: An In-Browser XML Editor
Rockwell, Geoffrey; Brown, Susan; Chartrand, James; Hesemeier, Susan 508

The Musici Database
Roeder, Torsten; Plutte, Christoph 511

The TEICHI Framework: Bringing TEI Lite to Drupal
Schöch, Christof; Achler, Stefan 514

What Has Digital Curation Got to Do With Digital Humanities?
Schreibman, Susan; McCadden, Katiet Theresa; Coyle, Barry 516

Orbis Latinus Online (OLO)
Schultes, Kilian Peter; Geissler, Stefan 518

Semantically connecting text fragments – Text-Text-Link-Editor
Selig, Thomas; Küster, Marc W.; Conner, Eric Sean 520

The Melesina Trench Project: Markup Vocabularies, Poetics, and Undergraduate Pedagogy
Singer, Kate 522

Digital Edition of Carl Maria von Weber's Collected Works
Stadler, Peter 525

Data sharing, virtual collaboration, and textual analysis: Working on 'Women Writers In History'
van Dijk, Suzan; Hoogenboom, Hilde; Sanz, Amelia; Bergenmar, Jenny; Olsson, Leif-Jöran 527

Storage Infrastructure of the Virtual Scriptorium St. Matthias
Vanscheidt, Philipp; Rapp, Andrea; Tonne, Danah 529

Digital Emblematics – Enabling Humanities Research of a Popular Early Modern Genre
Wade, Mara R.; Stäcker, Thomas; Stein, Regine; Brandhorst, Hans; Graham, David 532

DTAQ – Quality Assurance in a Large Corpus of Historical Texts
Wiegand, Frank 535

The Digital Averroes Research Environment – Semantic Relations in the Editorial Sciences
Willems, Florian; Gärtner, Mattias 537

AV Processing in eHumanities – a paradigm shift
Wittenburg, Peter; Lenkiewicz, Przemyslaw; Auer, Erik; Lenkiewicz, Anna; Gebre, Binyam Gebrekidan; Drude, Sebastian 538

List of Reviewers

- Hiroyuki Akama
- Marc Alexander
- Peter Roger Alsop
- Deborah Anderson
- Vadim Sergeevich Andreev
- Tara Lee Andrews
- Simon James Appleford
- Stewart Arneil
- Rolf Harald Baayen
- Drew Baker
- David Bamman
- Piotr Bański
- Brett Barney
- Sabine Bartsch
- Patsy Baudoin
- Syd Bauman
- Ryan Frederick Baumann
- David Beavan
- Craig Bellamy
- Hans Bennis
- Anna Bentkowska-Kafel
- Alejandro Bia
- Hanno Biber
- Marcus Bingenheimer
- Tobias Blanke
- Gabriel Bodard
- Jeremy Boggs
- Peter Kees Bol
- Geert E. Booij
- Peter Boot
- Lars Broin
- Federico Boschetti
- Arno Bosse
- Matthew Bouchard
- William Bowen
- John Bradley
- David Prager Branner
- Gerhard Brey
- Anne-Laure Brisac
- Susan Brown
- Marco Büchler
- Marjorie Burghart
- Lou Burnard
- Elisabeth Burr
- Toby Nicolas Burrows
- Dino Buzzetti
- Olivier Canteaut
- Paul Caton
- Hugh Cayless
- Tom Cheesman
- Shih-Pei Chen
- Paula Horwarth Chesley
- Tatjana Chorney
- Neil Chue Hong
- Arianna Ciula
- Florence Clavaud
- Frédéric Clavert
- Tanya Clement
- Claire Clivaz
- Louisa Connors
- Paul Conway
- Charles M. Cooney
- David Christopher Cooper
- Hugh Craig
- Tim Crawford
- James C. Cummings
- Richard Cunningham
- Alexander Czmiel
- Marin Dacos
- Stefano David
- Rebecca Frost Davis
- John Dawson
- Marilyn Deegan
- Janet Delve
- Kate Devlin
- Joseph DiNunzio
- Quinn Anya Dombrowski
- Jeremy Douglass
- J. Stephen Downie
- David S. Dubin

- Stuart Dunn
- Alastair Dunning
- Amy Earhart
- Michael Eberle-Sinatra
- Thomas Eckart
- Maciej Eder
- Jennifer C. Edmond
- Gabriel Egan
- Øyvind Eide
- Paul S. Ell
- Deena Engel
- Maria Esteva
- Martin Everaert
- Kurt E. Fendt
- Franz Fischer
- Kathleen Fitzpatrick
- Julia Flanders
- Dominic Forest
- Fenella Grace France
- Amanda French
- Christiane Fritze
- Chris Funkhouser
- Jonathan Furner
- Richard Furuta
- Isabel Galina Russell
- Liliane Gallet-Blanchard
- David Gants
- Susan Garfinkel
- Kurt Gärtner
- Richard Gartner
- Alexander Gil
- Joseph Gilbert
- Sharon K. Goetz
- Mattew K. Gold
- Joel Goldfield
- Sean Gouglas
- Ann Gow
- Stefan Gradmann
- Wayne Graham
- Harriett Elisabeth Green
- Jan Gregory
- Nathalie Groß
- Gretchen Mary Gueguen
- Carolyn Guertin
- Ann Hanlon
- Eric Harbeson
- Katherine D. Harris
- Kevin Scott Hawkins
- Sebastian Heath
- Mark Hedges
- Serge Heiden
- Gerhard Heyer
- Timothy Hill
- Brett Hirsch
- Martin Holmes
- David L. Hoover
- Xiao Hu
- Lorna Hughes
- Barbara Hui
- Claus Huitfeldt
- László Hunyadi
- Leif Isaksen
- Aleksandrs Ivanovs
- Fotis Jannidis
- Matthew Jockers
- Lars Johnsen
- Ian R. Johnson
- Patrick Juola
- Samuli Kaislaniemi
- Sarah Whitcher Kansa
- John Gerard Keating
- Margaret Kelleher
- Kimon Keramidas
- Katia Lida Kermanidis
- Erik Ketzan
- Foaad Khosmood
- Douglas Kibbee
- Gareth Knight
- Fabian Körner
- Kimberly Kowal
- Steven Krauwer
- William Kretzschmar

- Michael Adam Krot
- Christoph Kuemmel
- Maurizio Lana
- Lewis Rosser Lancaster
- Anouk Lang
- John Lavagnino
- Alexei Lavrentiev
- Séamus Lawless
- Katharine Faith Lawrence
- Domingo Ledezma
- Caroline Leitch
- Piroska Lendvai
- Richard J. Lewis
- Thomas Lippincott
- Eleonora Litta Modignani Picozzi
- Clare Llewellyn
- Dianella Lombardini
- Elizabeth Losh
- Ana Lucic
- Harald Lüngen
- Kim Luychx
- Akira Maeda
- Simon Mahony
- Martti Makinen
- Kurt Maly
- Worthy N. Martin
- Javier Martín Arista
- Jarom Lyle McDonald
- Stephanie Meece
- Federico Meschni
- Adrian Miles
- Maki Miyake
- Jose Miguel Monteiro Viera
- Ruth Mostern
- Stuart Moulthrop
- Martin Mueller
- A. Charles Muller
- Trevor Muñoz
- Orla Murphy
- Frédérique Muscinesi
- Elli Mylonas
- Kiyonori Nagasaki
- Brent Nelson
- John Nerbonne
- Greg T. Newton
- Angel David Nieves
- Bethany Nowviskie
- Julianne Nyhan
- Daniel Paul O'Donnell
- Kazushi Ohya
- Mark Olsen
- Lisa Lena Opas-Hänninen
- Christian-Emil Ore
- Espen S. Ore
- John Paolillo
- Brad Pasanek
- Michele Pasin
- Susan Holbrook Perdue
- Santiago Perez Isasi
- Elena Pierazzo
- Wendell Piez
- Daniel Pitti
- Dorothy Carr Porter
- Andrew John Prescott
- Ernesto Priani
- Michael John Priddy
- Brian L. Pytlik Zillig
- Sebastian Rahtz
- Michael John Rains
- Stephen Ramsay
- Andrea Rapp
- Gabriela Gray Redwine
- Dean Rehberger
- Georg Rehm
- Allen H. Renear
- Doug Reside
- Jason Rhody
- Allen Beye Riddell
- Jim Ridolfo
- David Robey
- Peter Robinson
- Geoffrey Rockwell

- Nuria Rodríguez
- Glenn H. Roe
- Torsten Roeder
- Augusta Rohrbach
- Matteo Romanello
- Laurent Romary
- Lisa Rosner
- Charlotte Roueché
- Henriette Roued-Cunliffe
- Joseph Rudman
- Stan Ruecker
- Angelina Russo
- Jan Rybicki
- Patrick Sahle
- Patrick Saint-Dizier
- Jon Saklofske
- Gabriele Salciute-Civiliene
- Manuel Sánchez-Quero
- Concha Sanz
- Jentery Sayers
- Torsten Schaßan
- Stephanie Schlitz
- Desmond Schmidt
- Harry Schmidt
- Sara A. Schmidt
- Christof Schöch
- Susan Schreibman
- Charlotte Schubert
- Paul Anthony Scifleet
- Tapio Seppänen
- Ryan Benjamin Shaw
- William Stewart Shaw
- Lynne Siemens
- Raymond George Siemens
- Gary F. Simons
- Stéfan Sinclair
- Kate Natalie Singer
- Natasha Smith
- James Dakin Smithies
- Lisa M. Snyder
- Małgorzata Sokoł
- Paul Joseph Spence
- Michael Sperberg-McQueen
- Lisa Spiro
- Peter Anthony Stokes
- Suzana Sukovic
- Chris Alen Sula
- Takafumi Suzuki
- Elizabeth Anne Swanstrom
- Tomoji Tabata
- Toma Tasovac
- Aja Teehan
- Elke Teich
- Melissa Terras
- Manfred Thaller
- Ruck John Thawonmas
- Christopher Theibault
- Amalia Todirascu
- Kathryn Tomasek
- Marijana Tomić
- Charles Bartlett Travis
- Thorsten Trippel
- Charlotte Tupman
- Kirsten Carol Uszkalo
- Ana Valverde Mateos
- Karina van Dalen-Oskam
- Ron van den Branden
- H. J. van den Herik
- Bert van Elsacker
- Marieke van Erp
- Seth van Hooland
- Joris Job van Zundert
- Tomás Várdi
- John T. Venecek
- Silvia Verdu Ruiz
- Christina Vertan
- Paul Vetch
- Raffaele Viglianti
- John Walsh
- Katherine L. Walter
- Claire Warwick
- Robert William Weidman

- Corinne Welger-Barboza
- Willeke Wendrich
- Susan L. Wiesner
- Matthew Wilkens
- Perry Willett
- William Winder
- Andreas Witt
- Christian Wittern
- Mark Wolff
- Glen Worthey
- Clifford Edward Wulfman
- Vika Zafrin
- Douwe Zeldenrust
- Matthew Zimmerman
- Amélie Zöllner-Weber

Plenary Sessions

Dynamics and Diversity: Exploring European and Transnational Perspectives on Digital Humanities Research Infrastructures

Moulin, Claudine
moulin@uni-trier.de
Trier Centre for Digital Humanities, University of Trier, Germany

In my talk I would like to reflect on Digital Humanities and Research infrastructures from an international and interdisciplinary perspective. Preserving and documenting cultural and linguistic variety is not only one of the most important challenges linked to the development and future impact of Digital Humanities on scholarly work and methodologies, it is also one of the key elements to be considered in the field of policy making and research funding at the European and international level. I will explore this by reflecting on Digital Humanities and its outcomes from the perspectives of cultural history and linguistic and interdisciplinary diversity; I will also tackle key questions related to building multi-level inter- and transdisciplinary projects and transnational research infrastructures. In addition to considering how Digital Humanities can extend and transform existing scholarly practice I will also consider how it is fostering the emergence of new cultural practices that look beyond established academic circles, for example, interactions between Digital Humanities and works of art.

Biographical Note

Claudine Moulin studied German and English philology in Brussels and Bamberg, receiving post-doctoral research grants for manuscript studies in Oxford. She was a Heisenberg fellow of the Deutsche Forschungsgemeinschaft (DFG); since 2003 she holds the chair for German Philology/ Historical Linguistics at the University of Trier/Germany and is the Scientific Director of the Trier Centre for Digital Humanities. She has published monographs and articles in the domain of text editing, digital lexicography, glossography, historical linguistics, and manuscript and annotation studies. She is a founding member of the Historisch-Kulturwissenschaftliches Forschungszentrum (HKFZ Trier). She is a member of the Standing Committee for the Humanities of the European Science Foundation (ESF) and speaker of the ESF-Expert Group on Research Infrastructures in the Humanities. She was recipient of the Academy Award of Rhineland-Palatinate in 2010, Academy of Sciences and Literature, Mainz.

C. Moulin is co-editor of the linguistic journal *Sprachwissenschaft* and of the series 'Germanistische Bibliothek'; together with Julianne Nyhan and Arianna Ciula (et al.) she has published in 2011 the ESF-Science Policy Briefing on Research Infrastructures in the Humanities (`http://www.esf.org/research-areas/humanities/strategic-activities/research-infrastructures-in-the-humanities.html`).

Embracing a Distant View of the Digital Humanities

Shimoda, Masahiro
shimoda@l.u-tokyo.ac.jp
Department of Indian Philosophy and Buddhist Studies/ Center for Evolving Humanities, Graduate School of Humanities and Sociology, the University of Tokyo, Japan

How should cultures transmit what they believe to be of vital importance from their own culture in its period of decline to another culture on the rise? This question, taken as one of the most challenging by contemporary historians, might well be posed themselves by DH scholars for the purpose of recognizing the magnitude of the problem they have been confronting and the significance of the endeavor they have been undertaking in the domain of the humanities. The variety of efforts that cannot but be included with the aims of each individual project, when combined together in a single arena such as DH 2012, will in the end be found to have been dedicated to the larger project of the constructing of 'another culture on the rise' on a global scale in this age of drastic transformation of the medium of knowledge. To keep this sort of 'distant view' of DH in mind, while it seems to have no immediate, sensible influence on our day-to-day 'close views,' though, would be inevitable. Inevitable not only in the making of untiring efforts to improve research environments amidst bewildering changes of technologies, but also in positively inviting new, unknown, unprecedented enterprises into the domain of DH. DH in its diverse forms should therefore assimilate into its identity the ostensively incommensurate aspects of steadfastness and flexibility. Among the numerous approaches that might be taken for understanding this peculiar identity of DH, I would like to demonstrate the significance of Digital Humanities research of Buddhist scriptures appropriately placed in the longer view of the history of the humanities.

Biographical Note

Masahiro Shimoda is a Professor in Indian Philosophy and Buddhist Studies with a cross appointment in the Digital Humanities Section of the Center for Evolving Humanities at the University of Tokyo. He has been Visiting Professor at the School of Oriental and African Studies, University College London (2006), Visiting Professor at Stanford University (2010), and is presently Visiting Research Fellow at University of Virginia (2012). He is the president of Japanese Association for Digital Humanities established last September (2011), and the chair of the trans-school program of Digital Humanities at the University of Tokyo, which has started on the 1st April 2012 in the collaborative program among the Graduate School of Interdisciplinary Information Studies, the Graduate School of Humanities and Sociology, and the Center for Structuring Knowledge. As his main project, Shimoda has launched since 2010 with government granted budget 'the construction of academic Buddhist knowledge base in international alliance.' This multi-nodal project, comprising seven major projects of self-financed agencies (Indo-Tibetan Lexical Resources at University of Hamburg, Hobogirin project at École Français d'Étrême Orient, Pali Text Compilation Project at Dhammacai Institute in Thai, Digital Dictionary of Buddhism in Tokyo etc.) with SAT (Chinese Buddhist Text Corpus Project at the University of Tokyo) placed as their hub, aims at providing a variety of research resources for Buddhist studies such as the primary sources, secondary resources, catalogues, dictionaries, lexicons and translations, all databases interlinked to each other at a deep structural level.

Pre-conference Workshops

Digital Methods in Manuscript Studies

Brockmann, Christian
christian.brockmann@uni-hamburg.de
University og Hamburg, Germany

Wangchuk, Dorji
dorji.wangchuk@uni-hamburg.de
University of Hamburg, Germany

Manuscript Studies constitute one of the main research areas in the Humanities at the University of Hamburg. This has been underlined recently when the large multidisciplinary Centre for the Study of Manuscript Cultures (Sonderforschungsbereich 950, http://www.manuscript-cultures.uni-hamburg.de/) won a substantial grant from the Deutsche Forschungsgemeinschaft in May 2011. The centre can draw on experience aggregated by several other projects in Hamburg such as 'Forschergruppe Manuskriptkulturen in Asien und Afrika', 'Teuchos. Zentrum für Handschriften- und Textforschung' and 'Nepalese-German Manuscript Cataloguing Project (NGMCP)'. Manuscripts as the central and most important media for the dissemination of writing have influenced and even shaped cultures worldwide for millennia, and only the relatively recent advent of the printed book has challenged their predominance.

In the past few years, Manuscript Studies have profited greatly from the use of digital methods and technologies, ranging e.g. from the creation of better and more accessible images to electronic editions and from a broad range of special databases supporting this research to analytical tools. Whereas such areas as digital cataloguing and editing have received more extensive coverage, methods more specific to the study of manuscripts in particular deserve broader attention.

This workshop focuses on Manuscript Studies as a distinctive field, i.e. the study of manuscripts as a characteristic feature and expression of those cultures that are built on their use. It will examine recent developments in digital methods that can be applied across various manuscript cultures worldwide, and aim to make awareness and discussion of these accessible to a broader group of scholars. It focuses exclusively on new developments in its subject fields that rely on the digital medium or on recent advances in technology as applied to the study of manuscripts, with a penchant towards aspects beyond the scope of the individuals fields concerned with just one particular manuscript culture.

The workshop will consist of brief introductory presentations on current developments in these areas by international experts, short hands-on and demonstration units on multispectral imaging and computer-assisted script and feature analysis as well as discussions on expected future developments, application perspectives, challenges and possible fields of cooperation.

Joint speakers/demonstrators:

Jost Gippert (University of Frankfurt)

Lior Wolf (Tel-Aviv University)

Lorenzo Perilli (University of Rome, Tor Vergata)

Domenico Fiormonte (Università di Roma Tre)

Agnieszka Helman-Wazny (University of Hamburg) / Jeff Wallman (Tibetan Buddhist Resource Center, New York) / Orna Almogi (University of Hamburg, Centre for the Study of Manuscript Cultures)

Boryana Pouvkova / Claire MacDonald (University of Hamburg, Centre for the Study of Manuscript Cultures)

Daniel Deckers (University of Hamburg)

Arved Solth / Bernd Neumann (University of Hamburg, Centre for the Study of Manuscript Cultures)

Ira Rabin, Oliver Hahn, Emanuel Kindzorra (Centre for the Study of Manuscript Cultures, Federal Institute for Materials Research and Testing)

Introduction to Stylomatic Analysis using R

Eder, Maciej
maciejeder@gmail.com
Pedagogical University, Kraków, Poland

Rybicki, Jan
jkrybicki@gmail.com
Jagiellonian University, Kraków, Poland

1. Brief Description

Stylometry, or the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.), has been around at least since the middle of the 19th century, and has found numerous practical applications in authorship attribution research. These applications are usually based on the belief that there exist such conscious or unconscious elements of personal style that can help detect the true author of an anonymous text; that there exist stylistic fingerprints that can betray the plagiarist; that the oldest authorship disputes (St. Paul's epistles or Shakespeare's plays) can be settled with more or less sophisticated statistical methods.

While specific issues remain largely unresolved (or, if closed once, they are sooner or later reopened), a variety of statistical approaches has been developed that allow, often with spectacular precision, to identify texts written by several authors based on a single example of each author's writing. But even more interesting research questions arise beyond bare authorship attribution: patterns of stylometric similarity and difference also provide new insights into relationships between different books by the same author; between books by different authors; between authors differing in terms of chronology or gender; between translations of the same author or group of authors; helping, in turn, to find new ways of looking at works that seem to have been studied from all possible perspectives. Nowadays, in the era of ever-growing computing power and of ever-more literary texts available in electronic form, we are able to perform stylometric experiments that our predecessors could only dream of.

This half-day workshop is a hands-on introduction to stylometric analysis in the programming language R, using an emerging tool, a collection of Maciej Eder's and Jan Rybicki's scripts, which perform multivariate analyses of the frequencies of the most frequent words, the most frequent word n-grams, and the most frequent letter n-grams. One of the scripts produces Cluster Analysis, Multidimensional Scaling, Principal Component Analysis and Bootstrap Consensus Tree graphs based on Burrows's Delta and other distance measures; it applies additional (and optional) procedures, such as Hoover's 'culling' and pronoun deletion. As by-products, it can be used to generate various frequency lists; a stand-alone word-frequency-maker is also available. Another script provides insight into state-of-the-art supervised techniques of classification, such as Support Vector Machines, k-Nearest Neighbor classification, or, more classically, Delta as developed by Burrows. Our scripts have already been used by other scholars to study Wittgenstein's dictated writings or, believe it or not, DNA sequences!

The workshop will be an opportunity to see this in practice in a variety of text collections, investigated for authorial attribution, translatorial attribution, genre, gender, chronology. Text collections in a variety of languages will be provided; workshop attendees are welcome to bring even more texts (in either plain text format or tei-xml). No previous knowledge of R is necessary: our script is very user-friendly (and very fast)!

2. Tutorial Outline

During a brief introduction, (1) R will be installed on the users' laptops from the Internet (if it has not been already installed); (2) participants will receive CDs/ pendrives with the script(s), a short quickstart guide and several text collections prepared for analysis; (3) some theory behind this particular stylometric approach will be discussed, and the possible uses of the tools presented will be summarized. After that and (4) a short instruction, participants will move on to (5) hands-on analysis to produce as many different results as possible to better assess the various aspects of stylometric study; (6) additional texts might be downloaded from the Internet or added by the participants themselves. The results, both numeric and visualizations, will be analyzed. For those more advanced in R (or S, or Matlab), details of the script (R methods, functions, and packages) will be discussed.

3. Special Requirements

Participants should come with their own laptops. We have versions of scripts for Windows, MacOS and Linux. The workshop also requires a projector and Internet connection in the workshop room.

References

Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge UP.

Burrows, J. (1987). Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Clarendon Press.

Burrows, J. F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing 17(3): 267-287.

Craig, H. (1999). Authorial attribution and computational stylistics: if you tell authors apart, have you learned anything about them? Literary and Linguistic Computing 14(1): 103-113.

Craig, H., and A. F. Kinney, eds. (2009). Shakespeare, Computers, and the Mystery of Authorship. Cambridge: Cambridge UP.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Digital Humanities 2010: Conference Abstracts*. King's College London, pp. 132-135.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics* 6: 101-116.

Eder, M., and J. Rybicki (2011). Stylometry with R. *Digital Humanities 2011: Cconference Abstracts*. Stanford University, Stanford, pp. 308-311.

Eder, M., and J. Rybicki (2012). Do birds of a feather really flock together, or how to choose test samples for authorship attribution. *Literary and Linguistic Computing* 27 (in press).

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing* 19(4): 453-475.

Jockers, M. L., and D. M. Witten (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing* 25(2): 215-223.

Koppel, M., J. Schler, and S. Argamon (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1): 9-26.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator: stylometry in translation. In M. Oakley and M. Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins.

Oakes, M., and A. Pichler (2012). Computational Stylometry of Wittgenstein's *Diktät für Schlick*. *Bergen Language and Linguistic (Bells) Series*, (in press).

Rybicki, J., and M. Eder (2011). Deeper Delta across genres and languages: do we really need the most frequent words?. *Literary and Linguistic Computing* 26(3): 315-321.

NeDiMAH workshop on ontology based annotation

Eide, Øyvind
oyvind.eide@kcl.ac.uk
King's College, London, UK

Ore, Christian-Emil
c.e.s.ore@iln.uio.no
University of Oslo, Norway

Rahtz, Sebastian
sebastian.rahtz@oucs.ox.ac.uk
University of Oxford, UK

The aim of this workshop is to present and discuss current ontology based annotation in text studies and to give participants an introduction and updated insight to the field. One of the expected outcomes from the workshop is to throw light on the consequences and experiences of a renewed database approach to computer assisted textual work, based on the developments over the last decade in text encoding as well as in ontological systems.

1. The NeDiMAH Network

The Network for Digital Methods in the Arts and Humanities (NeDiMAH) is a research network running from 2011 to 2015, funded by the European Science Foundation, ESF. The network will examine the practice of, and evidence for, advanced ICT methods in the arts and humanities across Europe, and articulate these findings in a series of outputs and publications. To accomplish this, NeDiMAH provides a locus of networking and interdisciplinary exchange of expertise among the trans-European community of digital arts and humanities researchers, as well as those engaged with creating and curating scholarly and cultural heritage digital collections. NeDiMAH will work closely with the EC funded DARIAH and CLARIN e-research infrastructure projects, as well as other national and international initiatives. NeDiMaH includes the following Working Groups:

1. Spatial and temporal modelling,
2. Information visualisation,
3. Linked data and ontological methods,
4. Developing digital data
5. Using large scale text collections for research
6. Scholarly digital editions

The WGs will examine the use of formal computationally-based methods for the capture, investigation, analysis, study, modelling, presentation, dissemination, publication and evaluation of arts and humanities materials for research. To achieve these goals the WGs will organise annual workshops and whenever possible, the NeDiMAH workshops will be organised in connection with other activities and initiatives in the field.

The NeDiMAH *WG3, Linked data and ontological methods*, proposes to organise a preconference workshop 'Ontology based annotation' in connection with the Digital Humanities 2011 in Hamburg.

2. Motivation and background

The use of computers as tools in the study of textual material in the humanities and cultural heritage goes back to the late 1940s, with links back to similar methods used without computer assistance, such as word counting in the late nineteenth century and concordances from the fourteenth century onwards. In the sixty years of computer assisted text research, two traditions can be seen. One is that which includes corpus linguistics and the creation of digital scholarly editions, while the other strain is related to museum and archival texts. In the former tradition, texts are commonly seen as first class feasible objects of study, which can be examined by the reader using aesthetic, linguistic or similar methods. In the latter tradition, texts are seen mainly as a source for information; readings concentrate on the content of the texts, not the form of their writing. Typical examples are museum catalogues and historical source documents. These two traditions will be called form and content oriented, respectively. It must be stressed that these categories are not rigorous; they are points in a continuum.

Tools commonly connected to museum and archive work, such as computer based ontologies, can be used to investigate texts of any genre, be it literary texts or historical sources. Any analysis of a text is based on a close reading of it. The same tools can also be used to study texts which are read according to both the form oriented and the content oriented way (Eide 2008; Zöllner-Weber & Pichler 2007).

The novelty of the approach lies in its focus on toolsets for modelling such readings in formal systems. Not to make a clear, coherent representation of a text, but rather to highlight inconsistencies as well as consistencies, tensions as well as harmonies, in our readings of the texts. The tools used for such modelling can be created to store and show contradictions and inconsistencies, as well as providing the user with means to detect and examine such contradictions. Such tools are typically used in an iterative way in which results from one experiment may lead to adjustments in the model or in the way it is interpreted, similar to modelling as it is described

by McCarty (2005). The source materials for this type of research are to be found in the results of decades of digital scholarly editing. Not only in the fact that a wide variety of texts exist in digital form, but also that many of these texts have been encoded in ways which can be used as starting points for the model building. Any part of the encoding can be contested, in the modelling work as well as in the experiments performed on the model. The methods developed in this area, which the TEI guidelines are an example of, provide a theoretical basis for this approach.

In the end of the 1980ies Manfred Thaller developed Kleio, a simple ontological annotation system for historical texts. Later in the 1990s hypertext, not databases, became the tool of choice for textual editions (Vanhoutte 2010: 131). The annotation system Pliny by John Bradley (2008) was design both as a practical tool for scholars abut also because Bradley was interested in how scholars work when studying a text. One of the expected outcomes from this workshop is to throw light on the consequences and experiences of a renewed database approach in computer assisted textual work, based on the development in text encoding over the last decade as well as in ontological systems.

A basic assumption is that reading a text includes a process of creating a model in the mind of the reader. This modelling process of the mind works in similar ways for all texts, being it fiction or non-fictions (see Ryan 1980). Reading a novel and reading a historical source document both result in models. These models will be different, but they can all be translated into ontologies expressed in computer formats. The external model stored in the computer system will be a different model from the one stored in the mind, but it will still be a model of the text reading. By manipulating the computer based model new things can be learned about the text in question.

This method represents an answer to Shillingsburg's call for editions which are open not only for reading by the reader, but also for manipulation (Shillingsburg 2010: 181), and to Pichler's understanding of digital tools as means to document and explicate our different understandings and interpretations of a text (Zöllner-Weber & Pichler 2007).

A digital edition can be part of the text model stored in the computer system. As tools and representation shape thinking not only through the conclusions they enable but also through the metaphors they deploy (Galey 2010: 100), this model will inevitably lead to other types of question asked to the text. A hypothesis is that these new questions will lead to answers giving new insight into the texts of study. Some of these insights would not have been found using other methods.

There is a movement in the humanities from seeking local knowledge about specific cases (McCarty, Willard. *Humanities Computin*g. Basingstoke: Palgrave Macmillan, 2005) which in this respect are traditional humanities investigations into specific collections of one or a limited number of texts. The general patterns sought may rather be found on a meta-research level where one investigate into new ways in which research that has a traditional scope can be performed.

3. A description of target audience

Scholars interested in online and shared annotation of texts and media based on ontologies. Practice in the field is not a requirement. Knowlegde of the concept 'ontology' or 'conceptual model' can be an advantage.

The aim of this workshop is to present and discuss current ontology based annotation in text studies and to give the participant an introduction and updated insight in the field and also bringing together researchers. One of the expected outcomes from this workshop is to throw light on the consequences and experiences of a renewed database approach in computer assisted textual work, based on the developments over the last decade in text encoding as well as in ontological systems.

References

Bradley, J. (2008). Pliny: A model for digital support of scholarship. *Journal of Digital Information* 9(1). `http://journals.tdl.org/jodi/article/view/209/198` . Last checked 2011-11-01

Crane, G. (2006). What Do You Do with a Million Books? *D-Lib Magazine* 12(3). URL: `http://www.dlib.org/dlib/march06/crane/03crane.html` . (checked 2011-11-01).

Eide, Ø. (2008). The Exhibition Problem. A Real-life Example with a Suggested Solution. *Lit Linguist Computing* 23(1): 27-37.

Galey, A. (2010). The Human Presence in Digital Artefacts'. In W. McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: Open Book Publishers, pp. 93-117.

McCarty, W. (2005). *Humanities Computing*. Basingstoke: Palgrave Macmillan.

Moretti, F. (2005). *Graphs, maps, trees: abstract models for a literary history*. London: Verso.

Ryan, M.-L. (1980). Fiction, non-factuals, and the principle of minimal departure. *Poetics* 9: 403-22.

Shillingsburg, P. (2010). How Literary Works Exist: Implied, Represented, and Interpreted. In W. McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: Open Book Publishers, pp. 165-82.

Kleio-system, http://www.hki.uni-koeln.de/kleio/old.website/welcome.html , checked 2011-11-01.

Zöllner-Weber, A., and A. Pichler (2007). Utilizing OWL for Wittgenstein's Tractatus. In H. Hrachovec, A. Pichler and J. Wang (eds.), *Philosophie der Informationsgesellschaft / Philosophy of the Information Society. Contributions of the Austrian Ludwig Wittgenstein Society*. Kirchberg am Wechsel: ALWS, pp. 248-250.

Vanhoutte, E. (2010) Defining Electronic Editions: A Historical and Functional Perspective. In W. McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: Open Book Publishers, pp. 119-44.

Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts

Hinrichs, Erhard
erhard.hinrichs@uni-tuebingen.de
Eberhard Karls University Tübingen, Germany

Neuroth, Heike
neuroth@sub.uni-goettingen.de
Unversity of Göttingen, Germany

Wittenburg, Peter
Peter.Wittenburg@mpi.nl
Max-Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

---

Large research infrastructure projects in the Humanities and Social Sciences such as Bamboo (http://www.projectbamboo.org/), CLARIN (http://www.clarin.eu), DARIAH (http://www.dariah.eu/), eAqua (http://www.eaqua.net/index.php), Metanet (http://www.meta-net.eu), and Panacea (http://www.panacea-lr.eu) increasingly offer their resources and tools as web applications or web services via the internet. Examples of this kind include:

- Bamboo Technology Project (http://www.projectbamboo.org/infrastructure/)
- eAqua Portal (http://www.eaqua.net/portal/)
- Language Technology World Portal of MetaNet (http://www.lt-world.org)
- PANACEA platform (http://www.panacea-lr.eu/en/project/the-platform)
- TextGrid – eScience methods in Arts and Humanities (http://www.textgrid.de/en.html)
- VLO – Virtual Language Observatory (http://www.clarin.eu/vlw/observatory.php)
- WebLicht – Web Based Linguistic Chaining Tool (http://https://weblicht.sfs.uni-tuebingen.de/)

Such web-based access has a number of crucial advantages over traditional means of service provision via downloadable resources or desktop applications. Since web applications can be invoked from any browser, downloading, installation, and configuration of individual tools on the user's local computer is avoided. Moreover, users of web applications will be ensured to always use the latest

version of the software, since it will be updated on the host computer. It is exactly this ease of use that is of crucial advantage for eHumanities researchers, since configuration and updates of software often require computational skills that can ordinarily not be expected from humanities researchers.

The paradigm of service-oriented architectures (SOA) is often used as a possible architecture for bundling web applications and web services. While the use of web services and SOAs is quickly gaining in popularity, there are still a number of open technology and research questions which await more principal answers:

- Currently, web services and SOAs in the Digital Humanities often concentrate on written material. Will the current technology scale up to accommodate multimodal data like speech or video data as well?
- Currently, web services and SOAs typically process data in a synchronous fashion. How can very large data sets such as multimodal resources be processed in an asynchronous fashion?
- Currently, web services and SOAs tend to deliver analysis or search results in a non-interactive fashion, allowing user input only to initiate processing and to react to the processing result. How can the current applications be extended so as to allow dynamic user interaction **during** processing? Such considerations are of crucial importance for the eHumanities in order to support, inter alia, interactive annotation of text corpora, a desideratum for all text-oriented disciplines such as literary studies, history, and linguistics.
- Will web-based access over time completely replace stand-alone (downloadable) desktop or CLI applications, or will there always be a need for both: local **and** web-based applications?
- What is the impact of emerging technologies such as web sockets or cloud computing on existing web service environments?
- Currently, SOAs tend to be application or domain specific, catering to the data formats and services most relevant to particular user communities. What are the possibilities for generalizing such current practice and developing generic execution models and standards?
- How to generate knowledge from data, e.g. developing new digital methods and concepts such as new and adapted data structures, hierarchical data storage, data modeling, sorting and search algorithms, selection of data via metadata, and visualization tools?

1. Invited Speaker

- Eric Nyburg (Carnegie Mellon University, Pittsburgh): A Service-Oriented Architecture for Rapid Development of Language Applications

2. Accepted Papers

- Tara L. Andrews, Moritz Wissenbach, Joris J. Van Zundert and Gregor Middell – *Embracing research: consolidating innovation through sustainable development of infrastructure*
- Dorothee Beermann, Pavel Mihaylov and Han Sloetjes – *Linking annotations Steps towards tool-chaining in Language Documentation*
- Andre Blessing, Jens Stegmann and Jonas Kuhn – *SOA meets Relation Extraction: Less may be more in Interaction*
- Michael Scott Cuthbert, Beth Hadley, Lars Johnson and Christopher Reyes – *Interoperable Digital Musicology Research via music21 Web Applications*
- Emanuel Dima, Erhard Hinrichs, Marie Hinrichs, Alexander Kislev, Thorsten Trippel and Thomas Zastrow – *Integration of WebLicht into the CLARIN Infrastructure*
- Rüdiger Gleim, Alexander Mehler and Alexandra Ernst – *SOA implementation of the eHumanities Desktop*
- Thomas Kisler, Florian Schiel and Han Sloetjes – *Signal processing via web services: the use case WebMAUS*
- Chiara Latronico, Nuno Freire, Shirley Agudo and Andreas Juffinger – *The European Library: A Data Service Endpoint for the Bibliographic Universe of Europe*
- Przemyslaw Lenkiewicz, Dieter van Uytvanck, Sebastian Drude and Peter Wittenburg – *Advanced Web-services for Automated Annotation of Audio and Video Recordings*
- Scott Martens – *TüNDRA: TIGERSearch-style treebank querying as an XQuery-based web service*
- Christoph Plutte – *How to Turn a Desktop Application into a Web-Interface? – Archiv-Editor as an Example of Eclipse RCP and RAP Single Sourcing*
- Thomas Zastrow and Emanuel Dima – *Workflow Engines in Digital Humanities*

Here and There, Then and Now – Modelling Space and Time in the Humanities

Isaksen, Leif
leifuss@googlemail.com
University of Southampton, UK

Day, Shawn
day.shawn@gmail.com
Digital Humanities Observatory, Ireland

Andresen, Jens
jens.andresen@hum.au.dk
University of Aarhus, Denmark

Hyvönen, Eero
eero.hyvonen@tkk.fi
Aalto University, Finland

Mäkelä, Eetu
eetu.makela@aalto.fii
Aalto University, Finland

Spatio-temporal concepts are so ubiquitous that it is easy for us to forget that they are essential to everything we do. All expressions of Human culture are related to the dimensions of space and time in the manner of their production and consumption, the nature of their medium and the way in which they express these concepts themselves. This workshop seeks to identify innovative practices among the Digital Humanities community that explore, critique and re-present the spatial and temporal aspects of culture.

Although space and time are closely related, there are significant differences between them which may be exploited when theorizing and researching the Humanities. Among these are the different natures of their dimensionality (three dimensions vs. one), the seemingly static nature of space but enforced 'flow' of time, and the different methods we use to make the communicative leap across spatial and temporal distance. Every medium, whether textual, tactile, illustrative or audible (or some combination of them), exploits space and time differently in order to convey its message. The changes required to express the same concepts in different media (between written and performed music, for example), are often driven by different spatio-temporal requirements. Last of all, the impossibility (and perhaps undesirability) of fully representing a four-dimensional reality (whether real or fictional) mean that authors and artists must decide how to collapse this reality into the spatio-temporal limitations of a chosen medium. The nature of those choices can be as interesting as the expression itself.

This workshop allows those working with digital tools and techniques that manage, analyse and exploit spatial and temporal concepts in the Humanities to present a position paper for the purposes of wider discussion and debate. The position papers will discuss generalized themes related to use of spatio-temporal methods in the Digital Humanities with specific reference to one or more concrete applications or examples. Accepted papers have been divided into three themed sessions: Tools, Methods and Theory. This workshop is part of the ESF-funded NEDIMAH Network and organised by its Working Group on Space and Time. The group will also present its findings from the First NeDiMAH Workshop on Space and Time.

1. About NeDiMAH

NeDiMAH is examining the practice of, and evidence for, advanced ICT methods in the Arts and Humanities across Europe, and will articulate these findings in a series of outputs and publications. To accomplish this, NeDiMAH assists in networking initiatives and the interdisciplinary exchange of expertise among the trans-European community of Digital Arts and Humanities researchers, as well as those engaged with creating and curating scholarly and cultural heritage digital collections. NeDiMAH maximises the value of national and international e-research infrastructure initiatives by helping Arts and Humanities researchers to develop, refine and share research methods that allow them to create and make best use of digital methods and collections. Better contextualization of ICT Methods also builds human capacity, and is of particular benefit for early stage researchers. For further information see http://www.nedimah.eu.

The workshop will also be aligned and coordinated with ongoing work at the DARIAH Project (cf. `http://www.dariah.eu`). DARIAH is a large-scale FP7-project that aims to prepare the building of digital research infrastructure for European Arts and Humanities researchers and content/data curators.

2. Papers

2.1. Tools

Shoichiro Hara & Tatsuki Skino – *Spatiotemporal Tools for Humanities*

David McClure – *The Canonical vs. The Contextual: Neatline's Approach to Connecting Archives with Spatio-Temporal Interfaces*

Roxana Kath – *eAQUA/Mental Maps: Exploring Concept Change in Time and Space*

Kate Byrne – *The Geographic Annotation Platform: A New Tool for Linking and Visualizing Places References in the Humanities*

2.2. Methods

William A. Kretzschmar, Jr. & C. Thomas Bailey – *Computer Simulation of Speech in Cultural Interaction as a Complex System*

Karl Grossner –*Event Objects for Placial History*

Charles Travis – *From the Ruins of Time and Space: The Psychogeographical GIS of Postcolonial Dublin in Flann O'Brien's At Swim Two Birds (1939)*

2.3. Theory

Maria Bostenaru Dan – *3D conceptual representation of the (mythical) space and time of the past in artistic scenographical and garden installations*

Eduard Arriaga-Arango – *Multiple temporalities at crossroads: Artistic Representations of Afro in Latin America and the Hispanic World in the current stage of Globalization (Mapping Cultural emergences through Networks)*

Kyriaki Papageorgiou – *Time, Space, Cyberspace and Beyond, On Research Methods, Delicate Empiricism, Labyrinths and Egypt*

Patricia Murrieta-Flores – *Finding the way without maps: Using spatial technologies to explore theories of terrestrial navigation and cognitive mapping in prehistoric societies*

2.4. Discussion Objectives

- Bring together the experiences of researchers developing or using spatial or temporal methods in the Digital Humanities.
- Evaluate the impact of such methods in terms of addressing traditional Humanities questions and posing new ones.
- Explore non-investigative benefits, such as the use of spatial and temporal tools and visualization as means for contextualization.
- Identify where tools developed for spatial analysis may be applicable to temporal analysis (and vice versa).

2.5. Program Committee

- Jens Andresen, University of Aarhus
- Shawn Day, Digital Humanities Observatory
- Eero Hyvönen, Aalto University
- Leif Isaksen, University of Southampton
- Eetu Mäkelä, Aalto University

Crowdsourcing meaning: a hands-on introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment

Petris, Marco
marco.petris@uni-hamburg.de
University of Hamburg, Germany

Gius, Evelyn
evelyn.gius@uni-hamburg.de
University of Hamburg, Germany

Schüch, Lena
lena.schuech@googlemail.com
University of Hamburg, Germany

Meister, Jan Christoph
jan-c-meister@uni-hamburg.de
University of Hamburg, Germany

1. Context and description

Humanities researchers in the field of literary studies access and read literary texts in digital format via the web in increasing numbers – but, apart from *search* and *find*, the cognitive processing of a text still takes place outside the digital realm. The interest essentially motivating human encounters with literature hardly seems to benefit from the new paradigm: hermeneutic, i.e. 'meaning' oriented high-order interpretation that transcends a mere decoding of information. The main reason for this might be that hermeneutic activity is not deterministic, but explorative: in the scholarly interpretation of literature we are not looking for the right answer, but for new, plausible and relevant answers. Thus high-order hermeneutic interpretation requires more than the automated string- or word-level pattern analysis of the source object provided by most digital text analysis applications so far, namely the ability to add semantic markup and to analyse both the object-data and the metadata in combination. This requires markup that goes beyond the distinction between procedural vs. descriptive of Coombs et al. (1987) and even beyond the subdivision of descriptive markup into genuinely descriptive vs. performative introduced by Renear (2004). By semantic markup we rather mean a true hermeneutic markup as defined by Pietz (2010: paragraph 1):

By 'hermeneutic' markup I mean markup that is deliberately interpretive. It is not limited to describing aspects or features of a text that can be formally defined and objectively verified. Instead, it is devoted to recording a scholar's or analyst's observations and conjectures in an open-ended way. As markup, it is capable of automated and semi-automated processing, so that it can be processed at scale and transformed into different representations. By means of a markup regimen perhaps peculiar to itself, a text will be exposed to further processing such as text analysis, visualization or rendition. Texts subjected to consistent interpretive methodologies, or different interpretive methodologies applied to the same text, can be compared. Rather than being devoted primarily to supporting data interchange and reuse – although these benefits would not be excluded – hermeneutic markup is focused on the presentation and explication of the interpretation it expresses.

CLÉA (Collaborative Literature Éxploration and Annotation) was developed to support McGann's (2004) open-ended, discontinuous, and non-hierarchical model of text-processing and allows the user to express many different readings directly in markup. The web based system not only enables collaborative research but it is based on an approach to markup that transcends the limitations of low-level text description, too.[1] CLÉA supports high-level semantic annotation through TEI compliant, non-deterministic stand off markup and acknowledges the standard practice in literary studies, i.e. a constant revision of interpretation (including one's own) that does not necessarily amount to falsification. CLÉA builds on our open source desktop application CATMA[2].

In our workshop, we will address some key challenges of developing and applying CLÉA:

- We will discuss both the prerequisites mentioned above and their role in the development of CLÉA,
- present interdisciplinary use cases where a complex tagset that operationalizes literary theory (namely narratology) is applied,
- give a practical introduction in the use of CLÉA, and
- provide a hands-on session where participants can annotate their own texts.

Finally, we would like to engage participants in a design critique of CLÉA and a general discussion about requirements for text analysis tools in their fields of interest.

References

Coombs, J. H., A. H. Renear, and St. J. DeRose (1987). Markup Systems and the Future of Scholarly

Text Processing. *Communications of the ACM* (ACM) 30(11): 933–947. Available online at http://xml.coverpages.org/coombs.html (last seen 2011-10-31).

McGann, J. (2004). Marking Texts of Many Dimensions. In S. Schreibman, R. Siemans, and J. Unsworth (eds.), *A Companion to Digital Humanities*, 2004. Oxford: Blackwell, pp. 218-239. Online at http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-4&toc.depth=1&toc.id=ss1-3-4&brand=9781405103213_brand=default (last seen 2011-10-31).

Piez, W. (2010). Towards Hermeneutic Markup: An architectural outline. *King's College, DH 2010*, London. Available from: http://piez.org/wendell/papers/dh2010/index.html (last seen 2011-10-31).

Renear, A. H. (2004). Text Encoding. In S. Schreibman, R. Siemans, and J. Unsworth (eds.), *A Companion to Digital Humanities*, 2004. Oxford: Blackwell, pp. 218–239. Online at http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-5&toc.depth=1&toc.id=ss1-3-5&brand=default (last seen 2011-10-31).

Notes

1. We define this distinction as follows: description cannot tolerate ambiguity, whereas an interpretation is an interpretation if and only if at least one alternative to it exists. Note that alternative interpretations are not subject to formal restrictions of binary logic: they can affirm, complement or contradict one another. In short, interpretations are of a probabilistic nature and highly context dependent.
2. CLÉA is funded by the European Digital Humanities Award 2010, see http://www.catma.de

Learning to play like a programmer: web mash-ups and scripting for beginners

Ridge, Mia
m.ridge@open.ac.uk
Open University, UK

Have you ever wanted to be able to express your ideas for digital humanities data-based projects more clearly, or wanted to know more about hack days and coding but been too afraid to ask?

In this hands-on tutorial led by an experienced web programmer, attendees will learn how to use online tools to create visualisations to explore humanities data sets while learning how computer scripts interact with data in digital applications.

Attendees will learn the basic principles of programming by playing with small snippets of code in a fun and supportive environment. The instructor will use accessible analogies to help participants understand and remember technical concepts. Working in pairs, participants will undertake short exercises and put into practice the scripting concepts they are learning about. The tutorial structure encourages attendees to reflect on their experiences and consolidate what they have learned from the exercises with the goal of providing deeper insight into computational thinking.

The tutorial aims to help humanists without a technical background understand more about the creation and delivery of digital humanities data resources. In doing so, this tutorial is designed to support greater diversity in the 'digital' part of the digital humanities community.

Target audience: This tutorial is aimed at people who want to learn enough to get started playing with simple code to manipulate data, or gain an insight into how programming works. No technical knowledge is assumed. Attendees are asked to bring their own laptops or net books.

1. Tutorial structure

The tutorial will include:

- what a humanities data set is and how to access one
- how web scripting languages work (using JavaScript as an example)
- how to sketch out your ideas in pseudo-code

- the value of visualisation tools in understanding the shape of a data set
- prepared exercises: 'hello world', using script libraries for mash-ups, creating your first mash-up using a live cultural dataset (e.g. a timeline or map),
- how to find further resources and keep learning

Introduction to Distant Reading Techniques with Voyant Tools, Multilingual Edition

Sinclair, Stéfan
sgsinclair@gmail.com
McGill University, Canada

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

You have a collection of digital texts, now what? This workshop provides a gentle introduction to text analysis in the digital humanities using Voyant Tools, a collection of free web-based tools that can handle larger collections of texts, be they digitized novels, online news articles, twitter feeds, or other textual content. This workshop will be a hands-on, practical guide with lots of time to ask questions, so participants are encouraged to bring their own texts. In the workshop we will cover the following:

1. A brief introduction to text analysis in the humanities;
2. Preliminary exploration techniques using Voyant;
3. Basic issues in choosing, compiling, and preparing a text corpus;
4. Text mining to identify themes in large corpora;
5. Ludic tools and speculative representations of texts; and
6. Integrating tool results into digital scholarship.

This year's workshop will pay special attention to certain multilingual issues in text analysis, such as character encoding, word segmentation, and available linguistic functionality for different languages. The instructors will present in English, but can also present or answer questions in French and Italian.

This is intended as an introduction to text analysis and visualization. We hope for an audience with a range of interests and relevant competencies. Participants are expected to bring their own laptop and are encouraged to bring their own texts.

Towards a reference curriculum for the Digital Humanities

Thaller, Manfred
manfred.thaller@uni-koeln.de
Historisch-Kulturwissenschaftliche Informationsverarbeitung, Universität zu Köln, Germany

In late 2009 the *Cologne Centre for eHumanities* started an initiative to improve the cooperation between (mainly) German universities actively offering degree courses in Digital Humanities. Within three meetings the concepts of the participating universities have been compared and a discussion on possibilities for closer cooperation has been started. As a result:

A 'catalogue' (`http://www.cceh.uni-koeln.de/Dokumente/BroschuereWeb.pdf` in the context of `http://www.cceh.uni-koeln.de/dh-degrees-2011` – German only, so far) to document the degree programs that have actively contributed to the common work has been prepared. It includes ten BA programs, twelve MA / MSc programs, two certificates of DH based training as professional add-on qualification on top of regular degrees plus one embedded degree within the regular teaching of a Humanities study program. The universities of *Bamberg, Bielefeld, Darmstadt, Erlangen, Gießen, Göttingen, Graz, Groningen, Hamburg, Köln, Lüneburg, Saarbrücken* and *Würzburg* have contributed to this catalogue. What started as an initiative of Cologne has in the meantime become an integral part of DARIAH-DE, a general framework of projects for the establishment of an infrastructure for the Digital Humanities within Germany.

Parallel to that initiative, a discussion has been started which shall lead towards the identification of common elements of Digital Humanities curricula, making it easier for students to move between individual degrees and providing the ground work for the recognition of Digital Humanities as a general concept by the agencies responsible for the accreditation of university degrees within Germany. The German situation is probably different from that in many other countries, as two BA and one MSc program from the list above are offered not by Arts faculties, but by Computer Science faculties or institutes.

Both activities are results of an underlying process of 'comparing the notes' between people responsible directly for conceptualization and implementation of the degree courses. This process has so far been implemented mainly in Germany for pragmatic reasons: To make students aware of the existence of the field of Digital Humanities as a regular field of academic studies on the level of practical PR activities, you have to address a community which finds all of the participating universities as similarly logical choices for a place to study. It is also much easier to *start* the discussion of a core curriculum if during the first rounds all participants of the discussion operate under the same administrative rules for the establishment of degree courses.

We will organize a workshop attached to the Digital Humanities 2012 at Hamburg in order to extend this discussion to representatives of other university landscapes.

On the most fundamental level we would like to:

- Present the results of the German initiative.
- Invite presentations of the Digital Humanities degree courses existing within other countries.

On the level of practical co-operation we intend to discuss:

- The creation and maintenance of a database / catalogue of European degree courses in Digital Humanities.
- The possibility for improved exchange activities within existing funding schemes, within Europe e.g. ERASMUS, between different degree courses. This will require, among other things, the identification of elements in different curricula which could substitute courses required at a home institution.
- In a very exploratory way, the possibilities for the facilitation of genuine 'European Master' degrees in Digital Humanities, in the sense used by the European Commission.

On the conceptual level we hope:

- To initialize a discussion about possible terms of reference for Digital Humanities curricula, which transcend individual academic systems.
- To arrive at a working definition for the field covered by such curricula. We are covering, e.g., degree courses which try to combine archaeology with computer science elements as well as degree courses, which are closely related to computational linguistics. As these communities are differently related within different university landscapes, a common conceptual reference framework should be used.

As this is an initiative which emerges from ongoing work and is directed mainly at institutions and persons which have experience with the

implementation of Digital Humanities degrees, we do not intend to rely primarily on a call for papers.

We will not rely on a call for papers primarily. During April 2012 a set of documents will be sent to all institutions in Europe, and many beyond, which are known to organize a degree course in the Digital Humanities or a closely connected field, with an invitation to join the workshop. We hope for results leading to direct and *practical* cooperation within existing frameworks. So the *primary* target group of this workshop are the European academic institutions offering or planning degree courses in the Digital Humanities. This said, of course we also invite the wider community to join the conceptual discussions.

Participation of institutions we are not aware of, particularly from those which are currently only in the planning stages of Digital Humanities' degree courses, is very much hoped for. Please direct enquiries to manfred.thaller@uni-koeln.de to receive additional material from the preparatory round of discussions and supporting material before the start of the workshop.

The workshop will run for a full day. The following program is tentative, to be adapted to accommodate for proposals and explicit wishes from participants during the preparatory stage.

09:00 – 10:30 Setting the agenda – reports on existing degree courses.

11:00 – 12:30 What do we have in common? I: Parallels and differences in the scope of individual degree courses.

14:00 – 15:30 What do we have in common? II: Parallels and differences in the concept of 'Digital Humanities' underlying the individual courses.

16:00 – 17:30 Are there synergies? Creating a work program to facilitate discussions of exchange facilities and curricular coordination across national boundaries.

Free your metadata: a practical approach towards metadata cleaning and vocabulary reconciliation

van Hooland, Seth
svhoolan@ulb.ac.be
Université Libre de Bruxelles, Belgium

Verborgh, Ruben
ruben.verborgh@ugent.be
Ghent University, Belgium

De Wilde, Max
madewild@ulb.ac.be
Université Libre de Bruxelles, Belgium

1. Tutorial content and its relevance to the DH community

The early-to-mid 2000s economic downturn in the US and Europe forced Digital Humanities projects to adopt a more pragmatic stance towards metadata creation and to deliver short-term results towards grant providers. It is precisely in this context that the concept of Linked and Open Data (LOD) has gained momentum. In this tutorial, we want to focus on metadata cleaning and reconciliation, two elementary steps to bring cultural heritage collections into the Linked Data cloud. After an initial cleaning process, involving for example the detection of duplicates and the unifying of encoding formats, metadata are reconciled by mapping a domain specific and/or local vocabulary to another (more commonly used) vocabulary that is already a part of the Semantic Web. We believe that the integration of heterogeneous collections can be managed by using subject vocabularies for cross linking between collections, since major classifications and thesauri (e.g. LCSH, DDC, RAMEAU, etc.) have been made available following Linked Data Principles.

Re-using these established terms for indexing cultural heritage resources represents a big potential of Linked Data for Digital Humanities projects, but there is a common belief that the application of LOD publishing still requires expert knowledge of Semantic Web technologies. This tutorial will therefore demonstrate how Semantic Web novices can start experimenting on their own with non-expert software such as Google Refine. Participants of the tutorial can bring an export (or a subset) of metadata from their own projects or organizations. All necessary operations to reconcile metadata with

controlled vocabularies which are already a part of the Linked Data cloud will be presented in detail, after which participants will be given time to perform these actions on their own metadata, under assistance of the tutorial organizers. Previous tutorials have mainly relied on the use of the Library of Congres Subject Headings (LCSH), but for the DH2012 conference we will test out beforehand SPARQL endpoints of controlled vocabularies in German (available for example on `http://wiss-ki.eu/authorities/gnd/)` http://wiss-ki.eu/authorities/gnd/), allowing local participants to experiment with metadata in German.

This tutorial proposal is a part of the Free your Metadata research project.[1] The website offers a variety of video's, screencasts and documentation on how to use Google Refine to clean and reconcile metadata with controlled vocabularies already connected to the Linked Data cloud. The website also offers an overview of previous presentations. Google Refine currently offers one of the best possible solutions on the market to clean and reconcile metadata. The open-source character of the software makes it also an excellent choice for training and educational purposes. Both researchers and practitioners from the Digital Humanities are within cultural heritage projects inevitably confronted with issues of bad quality metadata and the interconnecting with external metadata and controlled vocabularies. This tutorial will therefore provide both practical hands-on information and an opportunity to reflect on the complex theme of metadata quality.

2. Outline of the tutorial

During this half day tutorial, the organizers will present each essential step of the metadata cleaning and reconciliation process, before focusing on a hands-on session during which each participant will be asked to work on his or her own metadata set (but default metadata sets will also be provided). The overview of the different features will approximately take 60 minutes:

- Introduction: Outline regarding the importance of metadata quality and the concrete possibilities offered by Linked Data for cultural heritage collections
- Metadata cleaning: Insight into the features of Google Refine and how to apply filters and facets to tackle metadata quality issues.
- Metadata reconciliation: Use of the RDF extension which can be in- stalled to extend Google Refine's reconciliation capabilities. Overview of SPARQL endpoints with interesting vocabularies available for Digital Humanists, in different languages.

After a break, the participants will have the opportunity to work individually or in group on their own metadata and to experiment with the different operations showcased during the first half of the tutorial. The tutorial organizers will guide and assist the different groups during this process. Participants will be given 60 minutes for their own experimenting and during a 45 minutes wrap-up, participants will be asked to share their the outcomes of the experimentation process. This tutorial will also explicitly try to bring together Digital Humanists will similar interests in Linked Data and in this way stimulate future collaborations between institutions and projects.

3. Target audience

The target audience consists both of practitioners and researchers from the Digital Humanities field who focus on the management of cultural heritage resources.

4. Special requests/equipment needs

Participants should preferably bring their own laptop and, if possible, have installed Google Refine. Intermediate knowledge of metadata creation and management is required.

Notes

1. See the projects website on `http://freeyourmetadata.org`.

Panels

Text Analysis Meets Text Encoding

Bauman, Syd
Syd_Bauman@Brown.edu
Brown University, USA

Hoover, David
david.hoover@nyu.edu
New York University, USA

van Dalen-Oskam, Karina
karina.van.dalen@huygens.knaw.nl
Huygens Institute, The Netherlands

Piez, Wendell
wapiez@mulberrytech.com
Mulberry Technologies, Inc., USA

1. Aim and Organization

The main aim of this panel discussion is to bring together text encoding specialists and text analysis researchers. Recent DH conferences have comprised, in addition to other activities, two distinct sub-conferences – one focusing on text encoding in general and TEI in particular, and the other on text analysis, authorship attribution, and stylistics. The separation between the two is so extreme that their participants often meet only at breaks and social events. This is reflected in the way text encoding specialists and text analysis scholars do their work as well: they hardly ever work together on the same projects. Because of this lack of connection, some of the long-promised benefits of markup for analysis remain unrealized. This panel takes a step toward bridging the gap between markup and analysis.

We focus on both the causes for the gap and possible solutions. What could and should markup do that it doesn't currently do? Why do analysts rarely work with the huge number of texts already encoded? How can text encoders and those who process encoded texts make their work more useful to text analysts, and how can text analysis specialists help encoders make their texts more useful? What opportunities currently exist for collaboration and cooperation between encoders and analysts, and how can more productive opportunities be created?

2. Panel Topic

The reasons for the present gap between markup and analysis are partly technical and partly non-technical, and arise from the disparate aims and methods of the two camps. While markup systems have generally been designed to meet the needs of (scholarly) publishing, markup adherents have often claimed that their markup is also useful for analytic purposes. However, the very concept of 'markup' itself is different for the two constituencies. XML, in particular, isn't 'markup' in the full sense, as used by specialists in text processing. Rather, it is a data structuring methodology that imposes a single unitary hierarchy upon the text. Consequently, it is a poor instrument for the complete interpretive loop or spiral, where we start with text (with or without markup), perform analysis, use markup to record or 'inscribe' our findings into the text, and then return to analysis at a higher level. This is largely because the inscription step is usually inhibited by any prior (XML) markup. Consider two flowcharts of document processing workflows at `http://piez.org/wendell/papers/dh2010/743_Fig2a.jpg` (what XML provides) and `http://piez.org/wendell/papers/dh2010/743_Fig2b.jpg` (what we need). (As noted on the page at `http://piez.org/wendell/papers/dh2010`, these images were presented as part of a paper delivered at Digital Humanities 2010 in London Piez 2010].) The difference between these is essentially that in the current (XML-based) architecture, extending and amending our document schemas and processing require re-engineering the system itself; the stable system (which is designed to support publishing not research) does not naturally sustain that activity. A system that supported markup in the sense that text analysis requires – which would support, among other possibilities, multiple concurrent overlapping hierarchies (including rhetorical, prosodic, narrative and other organizations of texts) and arbitrary overlap (including overlap between similar types of elements or ranges) – would also support incremental development of processing to take advantage of any and all markup that researchers see fit to introduce.

Part of the solution to this problem is in the emergence of standard methodologies for encoding annotations above or alongside one or more 'base' layers of markup, perhaps using standoff markup or alternatives to XML. The details are less important than the capabilities inherent in a data model not limited to unitary trees (see Piez 2010 for discussion; for a more wide-ranging critique of markup, see Schmidt 2010). Over the long term, given suitable utilities and interfaces, textual analysts may be able to use such systems productively; in the medium term this is more doubtful.

This leads to the non-technical part of the problem: largely because XML is not very well suited to their requirements, text analysis tools typically cannot handle arbitrary encoding in XML along with the texts themselves, while at the same time there is

not yet a viable alternative encoding technology, specified as a standard and suitable for supporting interchange. And so an analyst must begin by processing the encoded text into a usable form. While the markup specialist may be able (perhaps easily) to perform such a transformation on his or her XML texts using XSLT, this is usually more difficult for the text analyst, for several reasons. Moreover, we submit that many or most of these difficulties are not simply due to known limitations of current text-encoding technologies as described above, but will also persist in new, more capable environments.

Markup processing technologies such as XSLT are rarely among the text analyst's armamentarium, and the benefits of XSLT (and we imagine this would be the case with any successor as well) are often not clear enough to the text analyst to justify its significant learning curve. XSLT need not be difficult, but it can be challenging to assimilate – especially because it seems superficially similar to text analysis. Those who do learn XSLT will find some of the tasks most useful to them relatively easy (e.g., filtering markup and splitting/aggregating texts). XSLT 2.0 also includes regular expressions, grouping, and stylesheet functions, and handles plain text input gracefully, making it a much more hospitable environment than XSLT 1.0 for text analysis. Yet more often than not, these features serve only to make XSLT tantalizing as well as frustrating to those for whom markup processing cannot be a core competency.

Moreover, intimate familiarity with the encoding system is typically necessary to process it correctly, yet the text analyst is frequently working with texts that he or she was not a party to creating. Many texts are encoded according to the TEI *Guidelines*, but analysts are often not experts in TEI in general, let alone expert at a particular application of TEI by a particular project. But such expertise is often required, as even 'simple' word extraction from an arbitrary TEI text can be problematic. Certainly, plenty of TEI encodings make the task non-trivial. Consider in particular the task of ingesting an arbitrary unextended TEI P5 text and emitting a word list, assuming that we can ignore one important detail: word order. What assumptions are necessary? What information will the data provider need to provide to the transformation process? Below are some preliminary considerations toward a tool that would present the user with a form to fill out and return an XSLT stylesheet for extracting words (note that many of these considerations would benefit from collaboration between the text encoder and the text analyst):

- What metadata can be ignored? E.g., if a default rendition of display:none applies to an element, are its contents words?
- Which element boundaries always/never/sometimes imply word breaks? When 'sometimes', how can we tell?
- Which hierarchical content structures (if any) should be ignored? (E.g,. colophons, forwards, prefaces) Which trappings? (E.g., <interGrp>, <figDesc>)
- Which elements pointed to by an <alt> element (or child of a <choice>) get ignored, and which get included?

The complexity of this analysis compounds the difficulty already described. And it must be performed anew for every encoded text or set of texts the analyst wishes to consider.

Furthermore, the publishing focus noted above means that text-encoding projects rarely encode the elements most useful for analysis, and you can only get out of markup what the encoder puts in. An analyst interested in how names are used in literary texts, for example, will find that even projects that encode proper names often encode only the names of authors and publishers (not very interesting for text analysis), or only personal names (the analyst may also be interested in place names and names of other things). Likewise, adding the desired markup to encoded texts requires that the analyst conform to (or perhaps modify) the existing encoding scheme, imposing another learning curve above that for the XSLT needed to extract the names once they have been encoded. Consequently, scholars interested in text analysis typically find it is more efficient to use plain texts without any markup, and then apply an *ad hoc* system of manual tagging, or named entity recognition (NER) tools in combination with manual correction of tagging. The analyst who wants to extract the speeches of characters in already-encoded plays will sometimes discover that the speeches are encoded in such a way that automatic extraction is quite problematic (e.g., at Documenting the American South). It is a rare literary encoding project that provides texts with the dialog encoded for speaker so that the speech of each character can be extracted (even apart, again, from the overlap problem) – a kind of analysis made popular by Burrows's pioneering work (Burrows 1987), but very labor-intensive.

Finally, even if texts with the desired encoding are available and the analyst is willing to learn the basics of XSLT, typically the XSLT has to be rewritten or tweaked for each new collection of texts that is examined because of differences in encoding schemes. And it is just those ambitious encoding projects that are likely to encode more elements of

interest to the text analyst that are also more likely to have complex, individualized, and difficult-to-process encoding (e.g., the Brown University Women Writers Project). From the analyst's point of view, the process of using existing encoding may be more time-consuming and frustrating than doing the work manually. Surely this state of affairs is undesirable.

Yet there is another way forward besides text analysts learning new skills or text encoders offering their own text-extraction tools. While the problems described here add up to a formidable challenge, the very fact that we can enumerate them suggests that we are not entirely helpless. There is much work that can be done both to refine our understanding, and to develop tools and methodologies that will help to bridge this divide. We suggest beginning with closer collaboration between the two camps. If each supports, works with, and learns from the other, both sides will benefit, and we will have a better foundation of understanding on which to build the next generation of technologies – technologies that will be valuable for both camps.

References

Willa Cather Archive http://cather.unl.edu/

The Brown University Women Writers Project. http://www.wwp.brown.edu/

Documenting the American South. http://docsouth.unc.edu/

Piez, W. (2010). Towards hermeneutic markup: an architectural outline. DH2010, King's College London, July 9.

Burrows, J. (1987). *Computation into Criticism*. Oxford: Clarendon P.

Schmidt, D. (2010). The inadequacy of embedded markup for cultural heritage texts. *LLC* 25: 337-356.

Burnard, L., and S. Bauman (eds.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 1.7.0. 2010-07-06. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Designing Interactive Reading Environments for the Online Scholarly Edition

Blandford, Ann
a.blandford@ucl.ac.uk
University College London, UK

Brown, Susan
sbrown@uoguelph.ca
University of Guelph, Canada

Dobson, Teresa
teresa.dobson@ubc.ca
University of British Columbia, Canada

Faisal, Sarah
s.faisal@cs.ucl.ac.uk
University College London, UK

Fiorentino, Carlos
carlosf@ualberta.ca
University of Alberta, Canada

Frizzera, Luciano
dosreisf@ualberta.ca
University of Alberta, Canada

Giacometti, Alejandro
alejandro.giacometti.09@ucl.ac.uk
University College London, UK

Heller, Brooke
brooke.heller@gmail.com
University of British Columbia, Canada

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta, Canada

Michura, Piotr
zemichur@cyf-kr.edu.pl
Academy of Fine Arts in Krakow, Poland

Nelson, Brent
brent.nelson@usask.ca
University of Saskatchewan, Canada

Radzikowska, Milena
mradzikowska@gmail.com
Mount Royal University, Canada

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design, USA

Sinclair, Stéfan
stefan.sinclair@mcgill.ca
McMaster University, Canada

Sondheim, Daniel
sondheim@ualberta.ca
University of Alberta, Canada

Warwick, Claire
c.warwick@ucl.ac.uk
University College London, UK

Windsor, Jennifer
jwindsor@ualberta.ca
University of Alberta, Canada

PAPER 1

Introduction to Designing Interactive Reading Environments for the Online Scholarly Edition

In this panel, members of the Interface Design team of the Implementing New Knowledge Environments (INKE) project will present a set of ideas, designs, and prototypes related to the next generation of the online scholarly edition, by which we mean a primary text and its scholarly apparatus, intended for use by people studying a text.

We begin with 'Digital Scholarly Editions: An Evolutionary Perspective', which proposes a taxonomy of design features and functions available in a wide range of existing projects involving scholarly editions. 'Implementing Text Analysis E-reader Tools to Create Ad-hoc Scholarly Editions' draws on this taxonomy and examines how the strategies of ubiquitous text analysis can be applied to digital texts as a means of providing new affordances for scholarship. 'Visualizing Citation Patterns in Humanist Scholarship' looks at the use of citations in scholarly writing and proposes visual models of possible benefit for both writers and readers. 'The Dynamic Table of Contexts: User Experience and Future Directions' reports our pilot study of a tool that combines the conventional table of contents with semantic XML encoding to produce a rich-prospect browser for books, while 'The Usability of Bubblelines: A Comparative Evaluation of Two Prototypes' provides our results in looking at a case where two distinct prototypes of a visualization tool for comparative search results were created from a single design concept.

PAPER 2

Digital Scholarly Editions: A Functional Perspective

Sondheim, Daniel
sondheim@ualberta.ca
University of Alberta, Canada

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design, USA

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta, Canada

Frizzera, Luciano
dosreisf@ualberta.ca
University of Alberta, Canada

Windsor, Jennifer
jwindsor@ualberta.ca
University of Alberta, Canada

Definitions of the scholarly edition at times seem as diverse as their content; definitions may highlight usefulness (e.g., Hjørland, n.d.), reliability (e.g. Lyman, 2009), or editorial expertise (e.g. Price 2008). Nevertheless, it is clear that scholarly editions 'comprise the most fundamental tools in literary studies' (McGann 2001: 55). The move from print to screen has offered scholars an opportunity to remediate scholarly editions and to improve upon them (Werstine 2008; Shillingsburg 2006), perhaps in the hope of producing their 'fullest realization' (Lyman 2009: iii).

Online, distinctions between traditional types of scholarly editions are blurred. Variorum editions are increasingly becoming the norm (Galey 2007), since the economy of space is no longer as important a variable. The notion of a 'best version' of a text is also becoming less of an issue, as what constitutes 'the best' is open to interpretation and is often irrelevant with regard to questions that scholars would like to ask (Price 2008).

Rather than categorizing electronic scholarly editions, we propose to evaluate a selection of influential and/or representative examples on the basis of a series of functional areas that have been noted in relevant literature as having undergone substantial changes in their move to the digital

environment. These areas include (1) navigation, including browsing and searching; (2) knowledge-sharing, including public annotation; (3) textual analysis, including graphs and visualizations; (4) customizability of both interface and content; (5) side-by-side comparisons of multiple versions; and (6) private note-taking and markup.

This study reveals that although all of the functionalities available in the digital medium are not implemented in every digital scholarly edition, it is nevertheless the case that even the simplest amongst them offer affordances that are different than their printed counterparts. The fact remains, however, that due to variety of functionalities implemented in digital scholarly editions, we are still negotiating what Lyman's 'fullest realization' could be.

PAPER 3

Implementing Text Analysis E-reader Tools to Create Ad-hoc Scholarly Editions

Windsor, Jennifer
jwindsor@ualberta.ca
University of Alberta, Canada

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta, Canada

Sondheim, Daniel
sondheim@ualberta.ca
University of Alberta, Canada

Frizzera, Luciano
dosreisf@ualberta.ca
University of Alberta, Canada

Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design, USA

Sinclair, Stéfan
stefan.sinclair@mcgill.ca
McMaster University, Canada

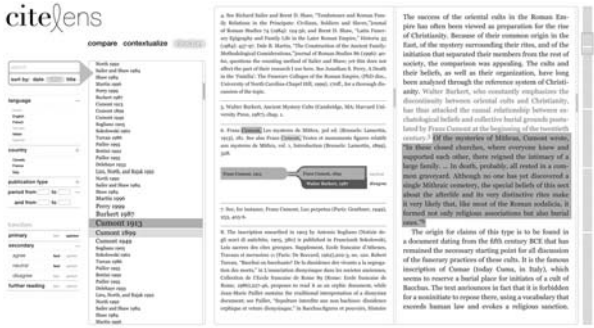
Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

With the proliferation of e-readers, the digitization efforts of organizations such as Google and Project Gutenberg, and the explosion of e-book sales, digital reading has become a commonplace activity. Although resources for casual reading are plentiful, they are insufficient to support in-depth scholarly research.

To alleviate this difficulty, we propose integrating Voyant tools, a user-friendly, flexible and powerful web-based text analysis environment developed by Stéfan Sinclair and Geoffrey Rockwell, with current e-reader technology, such as that used in the Internet Archive reader.

In this design, Voyant functions as a sidebar to e-readers, allowing the text to remain visible during analysis. Choosing from a list of tools allows scholars to create a custom text analysis tool palette and having more than one tool open allows cross-referencing between results. Tools can be dragged into a custom order and a scroll bar allows navigation between several tools at once. A Voyant tutorial is available and each individual tool also has its own instructions for use, as well as a help feature and an option to export results.

We anticipate this tool being of use to scholars in various fields who wish to use quantitative methods to analyze text. By implementing tools for textual analysis in an online reading environment, we are in effect offering scholars the ability to choose the kinds of analysis and depth of study that they wish; they will in effect be able to produce ad-hoc customized editions of digital text.



PAPER 4

Visualizing Citation Patterns in Humanist Monographs

Ilovan, Mihaela
ilovan@ualberta.ca
University of Alberta, Canada

Frizzera, Luciano
dosreisf@ualberta.ca
University of Alberta, Canada

Michura, Piotr
zemichur@cyf-kr.edu.pl
Academy of Fine Arts in Krakow, Poland

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design, USA

Sondheim, Daniel
sondheim@ualberta.ca
University of Alberta, Canada

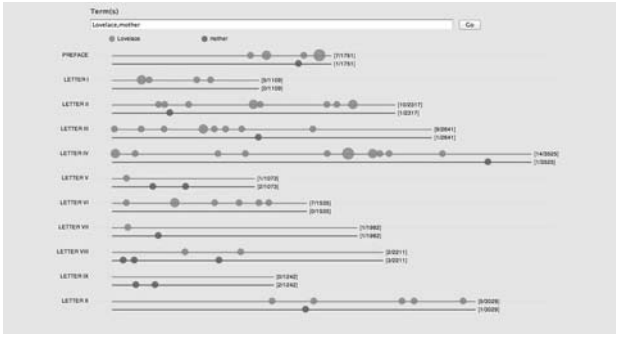
Windsor, Jennifer
jwindsor@ualberta.ca
University of Alberta, Canada

This paper documents the design and programming of a visualization tool for the analysis of citation patterns in extended humanist works such as monographs and scholarly editions. Traditional citation analysis is widely acknowledged as being ineffective for examining citations in the humanities, since the accretive nature of research (Garfield 1980), the existence of multiple paradigms, the preference for monographs (Thompson 2002), and the richness of non-parenthetical citations all present problems for interpretation and generalization.

To address this situation, we employ non-traditional methods of content and context analysis of citations, like functional classification (Frost 1979) and the exploration of the way in which sources are introduced in the flow of the argument (Hyland 1999). Acknowledging the richness of citations in humanist research, we employ graphic visualization for display and data inquiry, which allows us to carry out visual analytical tasks of the referencing patterns of full monographs.

We opted to provide three distinct views of the analyzed monograph. The first one – a faceted browser of the references, affords the comparison of different aspects of the references included and introduces the user to the structure and content of the monograph's critical apparatus. The second view contextualizes individual citations and highlights the fragments of text they support (see Figure 2); by representing both supported and non-supported portions of the monograph in their natural order, the view facilitates the linear reading of the way in which argument is built in the citing work. Finally, the third view of the visualization tool represents the internal structure of complex footnotes (see Figure 3) and visually highlights the relationship between different citations included in the same note, as well as their function in relation to the argument of the citing text.

In this presentation we will introduce the visualization tool, demonstrate its functionalities and provide the results of initial testing performed. We will also discuss the lessons learned from this early stage of the project.



PAPER 5

The Dynamic Table of Contexts: User Experience and Future Directions

Dobson, Teresa
teresa.dobson@ubc.ca
University of British Columbia, Canada

Heller, Brooke
brooke.heller@gmail.com
University of British Columbia, Canada

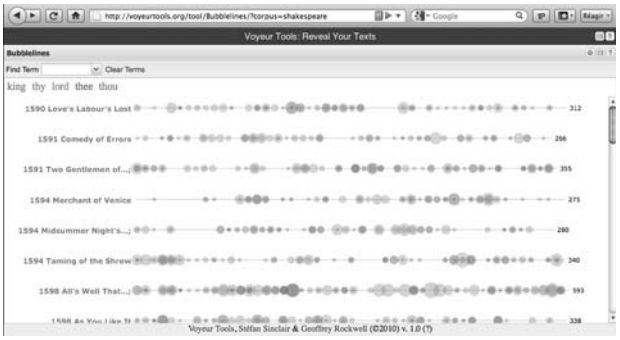
Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design, USA

Radzikowska, Milena
mradzikowska@gmail.com
Mount Royal University, Canada

Brown, Susan
sbrown@uoguelph.ca
University of Guelph, Canada

The Dynamic Table of Contexts (DToC) is a text analysis tool that combines the traditional concepts of the table of contents and index to create a new way to manipulate digital texts. More specifically, DToC makes use of pre-existing XML tags in a document to allow users to dynamically incorporate links to additional categories of items into the table of contents. Users may load documents they have tagged themselves, further increasing the interactivity and usefulness of the tool (Ruecker et al. 2009).

DToC includes four interactive panes, and one larger pane to view the text (see Figure 4). The first two panes are the table of contexts and the XML tag list: the former changes when a particular tag is selected in the latter, allowing the user to see where tokens of the tag fall in the sequence of the text. The third and fourth panes operate independently of each other but with the text itself: the 'Excerpts' pane highlights a selected token in the text and returns surrounding words, while the 'Bubbleline' maps multiple instances of a token across a pre-designated section (such as a chapter, or scene).



With twelve pairs of participants (n = 24), we completed a user experience study in which the participants were invited: 1) to complete in pairs a number of tasks in DToC, 2) to respond individually to a computer-based survey about their experience, and 3) to provide more extensive feedback during an exit interview. Their actions and discussion while working with the prototype were recorded. Data analysis is presently underway: transcribed interview data has been encoded for features of experience in XML; screen captures are supplementing our understanding of the participants' experience. In this paper we will report results and discuss future directions for development of the prototype.

PAPER 6

The Usability of Bubblelines: A Comparative Evaluation of Two Prototypes

Blandford, Ann
a.blandford@ucl.ac.uk
University College London, UK

Faisal, Sarah
s.faisal@cs.ucl.ac.uk
University College London, UK

Fiorentino, Carlos
carlosf@ualberta.ca
University of Alberta, Canada

Giacometti, Alejandro
alejandro.giacometti.09@ucl.ac.uk
University College London, UK

Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design, USA

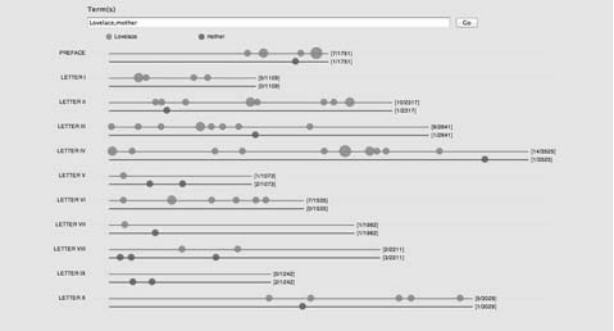
Sinclair, Stéfan
stefan.sinclair@mcgill.ca
McMaster University, Canada

Warwick, Claire
c.warwick@ucl.ac.uk
University College London, UK

In this paper, we explore the idea that programming is itself an act of interpretation, where the user experience can be profoundly influenced by seemingly minor details in the translation of a visual design into a prototype. We argue that the approach of doing two parallel, independent implementations can bring the trade-offs and design decisions into sharper relief to guide future development work. As a case study, we examine the results of the comparative user evaluation of two INKE prototypes for the bubblelines visualization tool. Bubblelines allows users to visually compare search results across multiple text files, such as novels, or pieces of text files, such as chapters. In this case, we had a somewhat rare opportunity in that both implementations of bubblelines began from a single design sketch, and the two development teams worked independently to produce online prototypes. The user experience study involved 12 participants, consisting of sophisticated computer users either working on or already having completed computer

science degrees with a focus on human-computer interaction. They were given a brief introduction, exploratory period, assigned tasks, and an exit interview. The study focused on three design aspects: the visual representation, functionality and interaction.

All users liked the idea of bubblelines as a tool for exploring text search results. Some users wanted to use the tools to search and make sense of their own work, e.g. research papers and computer programming codes. The study has shown that there was no general preference for one prototype over the other. There was however, general agreements of preferred visual cues and functionalities that users found useful from both prototypes. Similarly, there was an overall consensus in relation to visual representations and functionalities that users found difficult to use and understand in both tools. Having the ability to compare two similar yet somehow different prototypes has assisted us in fishing out user requirements summarized in the form of visual cues, functionalities and interactive experiences from a dimension that that would have been difficult to reach if we were testing a single prototype.



In summary, most participants preferred the visual appearance and simplicity of t1, but the greater functionality offered by t2. Apparently incidental design decisions such as whether or not search was case sensitive, and whether strings or words were the objects of search (e.g. whether a search for 'love' would highlight 'lover' or 'Love' as well as 'love') often caused frustration or confusion. We argue that the approach of doing two parallel, independent implementations can bring the trade-offs and design decisions into sharper relief to guide future development work.

Our plan is to take these requirements into account in order to generate a third prototype which we envision to evaluate with expert users where our focus would be on the ability of the tool in assisting experts in making sense of the data.

References

Garfield, E. (1980). Is Information Retrieval in the Arts and Humanities Inherently Different from That in Science? The Effect That ISI®'S Citation Index for the Arts and Humanities Is Expected to Have on Future Scholarship. *The Library Quarterly* 50(1): 40–57.

Galey, A. (2007) How to Do Things with Variants: Text Visualization in the Electronic New Variorum Shakespeare. *Paper presented at the Modern Language Association Annual Convention*, Chicago.

Frost, C.O. (1979). The Use of Citations in Literary Research: A Preliminary Classification of Citation Functions. *The Library Quarterly* 49(4): 399–414.

Hyland, K. (1999). Academic attribution: citation and the construction of disciplinary knowledge. *Applied Linguistics* 20(3): 341–367.

Lyman, E. (2009). *Assistive potencies: Reconfiguring the scholarly edition in an electronic setting*. United States-Virginia: University of Virginia.

Hjørland, B. (n.d.). Scholarly edition. Available at: `http://www.iva.dk/bh/core%20concepts%20in%20lis/articles%20a-z/scholarly_edition.htm` (accessed 20 October, 2011).

Lyman, E. (2009). *Assistive potencies: Reconfiguring the scholarly edition in an electronic setting*. United States-Virginia: University of Virginia. `http://www.proquest.com.login.ezproxy.library.ualberta.ca` (accessed Sept, 2010).

McGann, J. (2001). *Radiant textuality: literature after the World Wide Web*. New York: Palgrave.

Price, K. (2008). Electronic Scholarly Editions. In S. Schreibman and R. Siemens (eds.), *A Companion to Digital Literary Studies*. Oxford: Blackwell `http://digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-5` (accessed 25 October, 2011)

Radzikowska, M., S. Ruecker, S. Brown, P. Organisciak, and the INKE Research Group (2011). Structured Surfaces for JiTR. *Paper presented at the Digital Humanities 2011 Conference,* Stanford.

Rockwell, G., S. Sinclair, S. Ruecker, and P. Organisciak (2010). Ubiquitous Text Analysis. *Poetess Archive Journal* 2(1): 1-18.

Ruecker, S., S. Brown, M. Radzikowska, S. Sinclair, T. Nelson, P. Clements, I. Grundy, S. Balasz, and J. Antoniuk (2009). The Table of Contexts: A Dynamic Browsing Tool for Digitally

Encoded Texts. In L. Dolezalova (ed.), *The Charm of a List: From the Sumerians to Computerised Data Processing*. Cambridge: Cambridge Scholars Publishing, pp. 177-187.

Shillingsburg, P. L. (2006). *From Gutenberg to Google: electronic representations of literary texts*. Cambridge, UK; New York: Cambridge UP.

Thompson, J. W. (2002). The Death of the Scholarly Monograph in the Humanities? Citation Patterns in Literary Scholarship. *Libri* 52(3): 121-136.

Werstine, P. (2008). Past is prologue: Electronic New Variorum Shakespeares. *Shakespeare* 4: 208-220.

Developing the spatial humanities: Geo-spatial technologies as a platform for cross-disciplinary scholarship

Bodenhamer, David
intu100@iupui.edu
The Polis Center at IUPUI, USA

Gregory, Ian
i.gregory@lancaster.ac.uk
Lancaster University, UK

Ell, Paul
Paul.Ell@qub.ac.uk
Centre for Data Digitisation and Analysis at Queens University of Belfast, Ireland

Hallam, Julia
J.Hallam@liverpool.ac.uk
University of Liverpool, UK

Harris, Trevor
tharris2@wvu.edu
West Virginia University, USA

Schwartz, Robert
rschwart@mtholyoke.edu
Mount Holyoke College, USA

1. Significance

Developments in Geographic Information Systems (GIS) over the past few decades have been nothing short of remarkable. So revolutionary have these advances been that the impact of GIS on many facets of government administration, industrial infrastructure, commerce, and academia has been likened to the discoveries brought about by the microscope, the telescope, and the printing press. While concepts of spatial science and spatial thinking provide the bedrock on which a broad range of geospatial technologies and methods have been founded, the dialog between the humanities and geographic information science (GISci) have thus far been limited and have largely revolved around the use of 'off-the-shelf' GIS in historical mapping projects. This limited engagement is in stark contrast to the substantive inroads that GIS has made in the sciences and social sciences, as captured by the growing and valuable field of a social-theoretic informed Critical GIS. Not surprisingly, the humanities present additional significant challenges

to GISci because of the complexities involved in meshing a positivist science with humanist traditions and predominantly literary and aspatial methods. And yet it is the potential dialogue and engagement between the humanities and GISci that promises reciprocal advances in both fields as spatial science shapes humanist thought and is in turn reshaped by the multifaceted needs and approaches represented by humanist traditions. We use the term spatial humanities to capture this potentially rich interplay between Critical GIS, spatial science, spatial systems, and the panoply of highly nuanced humanist traditions.

The use of GIS in the humanities is not new. The National Endowment for the Humanities has funded a number of projects to explore how geo-spatial technologies might enhance research in a number of humanities disciplines, including but not limited to history, literary studies, and cultural studies. The National Science Foundation and National Institutes of Health also have supported projects related to spatial history, such as the Holocaust Historical GIS (NSF) and Population and Environment in the U.S. Great Plains (National Institute of Child Health and Human Development). Although successful by their own terms, these projects also have revealed the limits of the technology for a wider range of humanities scholarship, which an increasing body of literature discusses in detail. Chief among the issues are a mismatch between the positivist epistemology of GIS, with its demand for precise, measurable data, and the reflexive and recursive approaches favored by humanists and some social scientists (e.g. practitioners of reflexive sociology) who wrestle continually with ambiguous, uncertain, and imprecise evidence and who seek multivalent answers to their questions. The problem, it seems, is both foundational and technological: we do not yet have a well-articulated theory for the spatial humanities, nor do we have tools sufficient for the needs of humanists. Addressing these deficits is at the heart of much current work in GIScience and in the spatial humanities.

The panel, composed of scholars who have successfully blended geo-spatial technologies with cross-disciplinary methods in a variety of projects, will discuss how these digital tools can be bent toward the needs of humanities scholars and serve as a platform for innovative work in humanities disciplines. Emphasis will be on three important themes from the spatial humanities that also address the broader interests of the digital humanities:

1. Exploring the epistemological frameworks of the humanities and GISc to locate common ground on which the two can cooperate. This step often has been overlooked in the rush to develop new technology but it is the essential point of departure for any effort to bridge them. This venture is not to be confused with a more sweeping foundational analysis of ingrained methodological conceits within the sciences and the humanities, and certainly should not be misunderstood as a query about the qualitative approach versus the quantitative approach. Rather, what is desired here is for the technology itself to be interrogated as to its adaptability, in full understanding that the technology has, in its genesis, been epistemologically branded and yet still offers potential for the humanities. What is required is an appropriate intellectual grounding in the humanities and draws the technology further out of its positivistic homeland. The payoff for collaboration will be a humanities scholarship that integrates insights gleaned from spatial information science, spatial theory, and the Geospatial Web into scaled narratives about human lives and culture.
2. Designing and framing narratives about individual and collective human experience that are spatially contextualized. At one level, the task is defined as the development of reciprocal transformations from text to map and map to text. More importantly, the humanities and social sciences must position themselves to exploit the Geospatial Semantic Web, which in its extraordinarily complexity and massive volume, offers a rich data bed and functional platform to researchers who are able to effectively mine it, organize the harvested data, and contextualize it within the spaces of culture. Finding ways to make the interaction among words, location, and quantitative data more dynamic and intuitive will yield rich insights into complex socio-cultural, political, and economic problems, with enormous potential for areas far outside the traditional orbits of humanities research. In short, we should vigorously explore the means by which to advance translation from textual to visual communication, making the most of visual media and learning to create 'fits' between the messages of text (and numbers) and the capabilities of visual forms to express spatial relationships.
3. Building increasingly more complex maps (using the term broadly) of the visible and invisible aspects of a place. The spatial considerations remain the same, which is to say that geographic location, boundary, and landscape remain crucial, whether we are investigating a continental landmass or a lecture hall. What is added by these 'deep maps' is a reflexivity that acknowledges how engaged human agents build spatially framed identities and aspirations out of imagination and memory and how the multiple perspectives constitute a spatial narrative that complements

the verbal narrative traditionally employed by humanists.

After an introductory framing statement by the moderator, panelists will each take no more than 10 minutes to offer reflections and experiences drawn from their own projects that will address one or more of these themes. Several questions will guide these presentations:

1. What advantages did geo-spatial technologies bring to your research? What limitations? How did you overcome or work around the limits?
2. How did your project address the mismatch between the positivist assumptions of GIS and the multivalent and reflexive nature of humanities scholarship?
3. What lessons have you learned that would be helpful for other scholars who use these technologies?

Following the presentations, the moderator will guide a discussion among the panelists and with audience members to explore these themes further in an effort to distill an agenda or, more modestly, a direction for future work.

References

Bodenhamer, D., J. Corrigan, and T. Harris, eds. (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington: Indiana UP.

Daniels, S., D. DeLyser, J. Entrikin, and D. Richardson, eds. (2011). *Envisioning Landscapes, Making Worlds: Geography and the Humanities*. New York: Routledge.

Dear, M., J. Ketchum, S. Luria, and D. Richardson (2011). *GeoHumanities: Art, History, Text at the Edge of Place*. New York: Routledge.

Gregory, I., and P. Ell (2008). *Historical GIS: Technologies, Methodologies, and Scholarship*. Cambridge: Cambridge UP.

Knowles, A., ed. (2008). *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands, CA: ESRI Press.

Prosopographical Databases, Text-Mining, GIS and System Interoperability for Chinese History and Literature

Bol, Peter Kees
pkbol@fas.harvard.edu
Harvard University, USA

Hsiang, Jieh
jieh.hsiang@gmail.com
National Taiwan University, Taiwan

Fong, Grace
grace.fong@mcgill.ca
McGill University, Canada

PAPER 1

Introduction

1. Overview

Digital content and tools for Chinese studies have been developing very quickly. The largest digital text corpus currently has 400 million characters. There is now an historical GIS for administrative units and towns in China from 221 BCE to 1911. And there are general and specialized biographical and literary databases. This wealth of resources is constrained, however, by the histories of their development: all have been developed independently of each other and there has been no systematic effort to create interoperability between them. This panel brings together three innovators in Chinese digital humanities and shows how, by implementing system interoperability and content sharing, their separate approaches to content and tool development can be mutually supporting and result in more powerful applications for the study of China's history and literature. The goal of this session is both to introduce the projects the presenters lead and to demonstrate the advantages of sharing and interoperability across systems. Moreover, because their outputs are multi-lingual (Chinese, English translation, and pinyin) they are making the data from China's past accessible to a non-Chinese-reading public.

The China Biographical Database (CBDB) has been accumulating biographical data on historical figures, mainly from the 7th through early 20th century. It populates the database by applying text-mining procedures to large bodies of digital texts. Users

can query the system in terms of place, time, office, social associations and kinship, and export the results for further analysis with GIS, social networks, and statistical software. The Research Center for the Digital Humanities at National Taiwan University developed the Taiwan History Digital Library, a system for the spatial and temporal visualization of the distribution of Taiwan historical land deeds, creating a temporally-enabled map interface that allows users to discover relevant documents, examine them and track their frequency. Using CBDB code tables of person names, place names, and official titles for text mark-up, the Center is now applying this system to a compendium of 80,000 documents from the 10th through early 13th century that cover all aspects of government, including such diverse topics as religious activity, tax revenues and bureaucratic appointments. Users will be able to track the careers of individuals and call up their CBDB biographies through time and space as well as seeing when, where, and what the government was doing. This will be a model for incorporating the still larger compendia of government documents from later periods. Data will be discovered (e.g. the location and date of official appointments) and deposited into the CBDB. The Ming Qing Women's Writings (MQWW) project is a multilingual online database of 5000 women writers from Chinese history which includes scans of hitherto unavailable works and analyses of their content. Users can query the database by both person and literary content. Using APIs to create system interoperability between MQWW and CBDB, MQWW is building into its interface the ability to call up CBDB data on kinship, associations, and careers of the persons in MQWW. For their part CBDB users will be able to discover the writings of women through the CBDB interface.

PAPER 2

Chinese Biographical Data: Text-mining, Databases and System Interoperability

Bol, Peter Kees
pkbol@fas.harvard.edu
Harvard University, USA

Biography has been a major component of Chinese historiography since the first century BCE and takes up over half the contents in the twenty-five dynastic histories; biographical and prosopographical data also dominate the 8500 extant local histories. The http://isites.harvard.edu/icb/icb.do?keyword=k16229 is a collaborative project of Harvard University, Peking University, and Academia Sinica to collect biographical data on men and women in China's history (Bol and Ge 2002-10). It currently holds data on about 130,000 people, mainly from the seventh through early twentieth century, and is freely available as a http://isites.harvard.edu/icb/icb.do?keyword=k16229&pageid=icb.page76670 , as an http://cbdb.fas.harvard.edu/cbdbc/cbdbedit , and as an online query system with both http://59.124.34.70/cbdbc/ttsweb?@0:0:1:cbdbkmeng@@0.10566209097417267 and http://59.124.34.70/cbdbc/ttsweb?@0:0:1:cbdbkm@@0.10566209097417267 interfaces. Code and data tables cover the multiple kinds of names associated with individuals, the places with which they were associated at birth and death and during their careers, the offices they held, the educational and religious institutions with which they were associated, the ways in which they entered their careers (with special attention to the civil service examination), the people they were associated with through kinship and through non-kin social associations (e.g. teacher-student), social distinction, ethnicity, and writings (Fuller 2011).

The purpose of the database is to enable users, working in Chinese or English, to use large quantities of prosopographical data to explore historical questions (Stone 1972). These may be straightforward (e.g. What was the spatial distribution of books by bibliographic class during a certain period? How did age at death vary by time, space, and gender?) or complex (e.g. What percentage of civil officials from a certain place during a certain period were related to each other through blood or marriage? How did intellectual movements spread over time?). The standalone database and the online query system also include routines for generating genealogies of any extent and social networks of up to five nodes, finding incumbents in a particular office, etc. The standalone database (and next year the online system) can also output queries in formats that can be analyzed with geographic information system and social network analysis software (Bol 2012). Historical social network analysis is challenging but rewarding (Padgett & Ansell 1993; Wetherhall 1998).

We began to populate the database largely through text-mining and text-modeling procedures. We began with Named Entity Recognition procedures (e.g. finding text-strings that matched reign periods titles, numbers, years, months, and days to capture dates, with 99% recall and precision) written in Python. We then proceeded to write more complex Regular Expressions to identify, for example, the office a person assumed at a certain date, where the office title is unknown. Currently we are

implementing probabilistic procedures to identify the social relationship between the biographical subject and the names of people co-occurring in the biographical text (we have reached a 60% match with the training data).

An important outcome for the humanities generally is Elif Yamangil's (Harvard University: Computer Science) development of the 'Regular Expression Machine.' This is a graphical user interface (GUI) built within Java Swing library that allows a user to graphically design patterns for biographical texts, match these against the data and see results immediately via a user-friendly color-coding scheme. It consists of a workspace of three main components: (1) a view that displays the textual data currently used, (2) a list of active regular expressions that are matched against the data via color-coding, and (3) a list of shortcut regular expression parts that can be used as building-blocks in active regular expressions from (2). Additional facilities we have built into our product are (1) an XML export ability: Users can create XML files that flatten the current workspace (data and regular expression matched against) at the click of a button. This facilitates interfacing to other programs such as Microsoft Excel and Access for database input. (2) A save/load ability: Users can easily save/load the workspace state which includes the list of regular expressions and shortcuts and their color settings. (3) A handy list of pre-made regular expression examples: Numerous date patterns can be added instantly to any regular expression using the GUI menus. The point of building this application is to allow users with no prior experience with programming or computer science concepts such as regular expression scripting to experiment with data-mining Chinese biographical texts at an intuitive template-matching understanding level only, yet still effectively.

The CBDB project also accepts volunteered data. Our goal in this is social rather than technical. Researchers in Chinese studies have long paid attention to biography and in the course of their research have created tables, and occasionally databases, with biographical data. By offering standard formats and look-ups and queries for coding, we provide researchers with a permanent home for their data. Currently there are twelve collaborating research projects, among which are an extensive database of Buddhist monks, a collection of 5000 grave biographies, 30,000 individuals active in the ninth through tenth centuries. The more biographical data the project accumulates the greater the service to research and learning that explore the lives of individuals.

Humanists are exploring ways in which they can use data in quantity and looking for ways of making it accessible to others on the web. In Chinese studies – speaking now only of those online systems that include biographical data – these include the Tang Knowledge Database at the Center for Humanistic Research at Kyoto University, the Ming-Qing Archives at Academia Sinica and the National Palace Museum Taiwan, the databases for Buddhist Studies at Dharma Drum College, Ming Qing Women's Writings Database at McGill University, and the University of Leuven's database of writings by Christian missionaries and converts in China. The role of the China Biographical Database project in this environment is to provide an online resource for biographical data that others can draw on to populate fields in their own online systems. To this end we are currently developing (and will demonstrate this summer) a set of open Application Programming Interfaces that will allow other systems to incorporate data from the China Biographical Database on the fly. This will allow other projects to focus their resources on the issues that are of most interest to them while at the same time being able to incorporate any or all the data in CBDB within their own systems.

Figure 1: Online search results for Zhu Xi

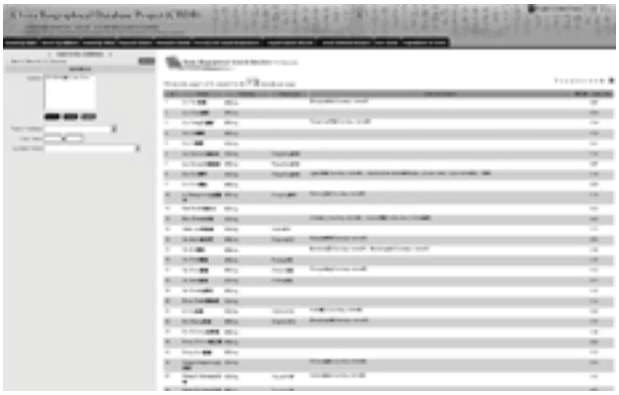


Figure 2: Persons in CBDB from Jianzhou in the Song period

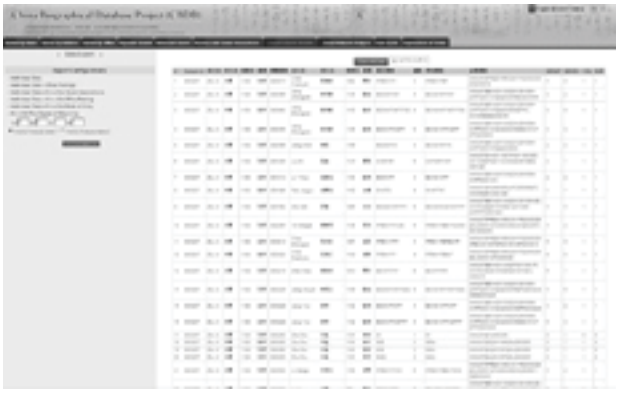


Figure 3: Kin to Zhu Xi (5 generations up and down, collateral and marriage distance of 1)

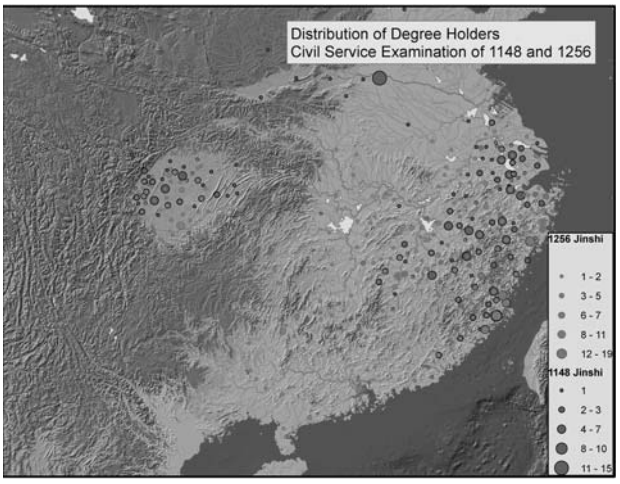


Figure 4: Civil Service Examination degree holders of 1148 and 1256

References

Bol, P. K. (2007). Creating a GIS for the History of China. In A. Kelly Knowles and A. Hillier (eds.), *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship*. Redlands, CA: ESRI Press, pp. 25-57.

Bol, P. K. (2012). GIS, Prosopography, and History. *Annals of GIS* 18(1): 3-15.

Bol, P. K., and Ge Jianxiong 葛剑雄 (2002-10). China historical GIS [electronic resource] = Zhongguo li shi di li xin xi xi = 中国历史地理信息系统. Version 1.0-5.0 Harvard University and Fudan University.

Fuller, M. A. (2011). *CBDB User's Guide*. 2nd edition. Harvard University: China Biographical Database.

Padgett, J. F., and Chr. K. Ansell (1993). Robust Action and the Rise of the Medici, 1400-1434. *American Journal of Sociology* 98: 1259-319.

Stone, L. (1972). Prosopography. In F. Gilbert, E. J. Hobsbawm and St. Richards Graubard (eds.), *Historical studies today*., xxi, 469. New York: Norton.

Wetherhall, Ch. (1998). Historical Social Network Analysis. In L. J. Griffin and M. van der Linden (eds.), *New methods for social history*. Cambridge, New York: Cambridge UP, p. 165.

PAPER 3

Context discovery in historical documents – a case study with Taiwan History Digital Library (THDL)

Hsiang, Jieh
jieh.hsiang@gmail.com
National Taiwan University, Taiwan

Research on pre-1900 Taiwanese history has suffered from the lack of primary historical documents. Most of the local government records, except for the Danxin Archives (an archive of about 20,000 official documents from the two northern Taiwan prefectures during the 19th century) are lost. Although there are quite a few records about Taiwan in the Imperial Court archives, they are kept in several different institutions, scattered among the volumes, and are hard to access. The situation with local documents such as land deeds is even worse. Some are published in local gazetteers or books, and some are family records kept by individual researchers for their own research use. There was an urgent need to collect primary documents and put them under one roof so that they can be better utilized by scholars.

In 2002 the Council of Cultural Affairs of Taiwan commissioned the National Taiwan University Library to select imperial court documents relevant to Taiwan from the Ming and Qing court archives, and the National Taichung Library to collect local deeds (especially land deeds). More than 30,000 court documents and 19,000 local deeds were collected and transcribed into full text. Most of the court documents and a portion of the land deeds were then published, from 2003 to 2007, into a book series of 147 volumes.

At the same time, CCA authorized the Research Center for Digital Humanities (RCDH) of NTU to

create a digital library so that the digital files of the found materials could be used on line. In addition to creating THDL, the Taiwan History Digital Library, RCDH also added a significant number of documents that were collected later. THDL now contains three corpuses, all in searchable full-text. The Ming and Qing Taiwan-related Court Document Collection, numbered at 45,722, are collected from about 300 different sources and includes memorials, imperial edicts, excerpts from personal journals of local officials, and local gazetteers. The Old Land Deeds Collection contains 36,535 pre-1910 deeds from over 100 sources, with new material added every month. The Danxin Archives includes about 1,151 government court cases with 19,557 documents. Together they form the largest collection of its kind, with over 100,000 documents, all with metadata and searchable full text. The three corpuses reflect 18th and 19th century Taiwan from the views of the central government, the local government, and grassroots respectively. THDL has been available for free over the Internet since 2009, and has already made an impact on research of Taiwanese history.

When designing THDL we encountered a challenge. Unlike conventional archives whose documents have a clear organization, the contents in THDL, being from many different sources, do not have predefined contexts. Although one can easily create a search-engine-like system that returns documents when given a query, such a system is cumbersome since the user has to peruse the returned documents to find the relevant ones. In THDL we use a different approach. We assume that documents may be related, and treat a query return (or any subset of documents) as a sub-collection – a group of related documents (Hsiang et al. 2009). Thus in addition to returning a list of documents as the query result, the system also provides the user with as many contexts about the returned sub-collection as possible. This is done in a number of ways. The first one is to classify the query return according to attributes such as year, source, type, location, etc. Visualization tools are provided to give the user a bird's eyes view of the distributions. Analyses of co-occurrences of names, locations and objects provide more ways to observe and explore the relationships among the actors and the collective meanings of the documents. For example, co-occurrence analysis of terms reveals the names of the associates mentioned most often in the memorials from a certain official, or the most frequently mentioned names during a historical event. Figure 1 is a snapshot of THDL after issuing the query '找洗字'. The map on the left shows the geographic locations of the documents in the query return. The top is a chronological distribution chart, and the left column is the post-classification of the query result according to year. We have also developed GIS techniques both for issuing a query and for analyzing/visualizing query results. For the land deeds, for instance, one can draw a polygon to query about all the land deeds appeared in that region (Ou 2011). (Such a query is quite impossible using keywords.) Figure 2 contains such an example. In order to go beyond syntactic queries, we further developed text mining techniques to explore the citation relations among the memorials and imperial edicts of the Ming Qing collection (Hsiang et al. 2012), and land transaction relations among the deeds in the land deed collection (Chen et al. 2011). Both projects have produced transitivity graphs that capture historical phenomena that would be difficult to discover manually.

In order to accomplish the above we developed a number of text-mining techniques. Term extraction methods were developed to extract more than 90,000 name entities from the corpuses (Hsieh 2011), and information such as the locations, dates, the four reaches of the land deeds, prices, and patterns. Matching algorithms were designed to find the transaction, citation and other relations mentioned above (Huang 2009; Chen 2011). GIS-related spatial-temporal techniques have also been developed (Ou 2011).

The contents of the China Biographical Database (CBDB) and Ming Qing Women's Writers (MQWW) described in other parts of this panel share a number of commonalities with the contents of THDL. In additional to being written in classical Chinese, the documents in each of the corpuses were collected from many different sources and spanned over hundreds of years. Furthermore, there are no intrinsic relations among the documents other than the obvious. To use such loosely knitted collections effectively in research would require a system that can help the user explore contexts that may be hidden among the documents. This is exactly what THDL provides. The designing philosophy of treating a query return as a sub-collection and the mining and presentation technologies developed for THDL can be adapted in the other two collections as well. Indeed, THDL's design, interface, and mining technologies are flexible enough to easily incorporate other Chinese language corpuses. The only assumption is that the documents in a collection should have well-structured metadata, which is important for the purpose of post-classification of a sub-collection. If the full-text of the content is also available, then more sophisticated analytical methods such as co-occurrence analysis can be deployed.

We have worked with the CBDB team of Harvard and built, within a month, a fully functional system from the THDL shell for Song huiyao (宋會要), a compendium of government records between 10th and 13th century China. (The content was jointly

developed at Harvard and the Academia Sinica of Taiwan.) The system has also incorporated 9,470 person names, 2,420 locations, and 3,366 official titles from CBDB and used them in co-occurrence analysis, classifications and other features. It also extracted, automatically, 11,901 additional terms from the corpus. To reduce the number of errors unavoidable from this automated process, we have designed a crowd-sourcing mechanism for users to make corrections and to add new terms they discovered when using the system. The new names obtained through this process will in turn be fed back to CBDB.

We have also built a prototype for MQWW with post-classification features. Significant enhancement to this system is being planned once we receive enough users' feedback. Incorporating the names and locations of CBDB into the system is also being studied.

While the text mining techniques that we have described here are designed for the Chinese language, the retrieval methodology of treating a query return as a sub-collection and the mechanisms for discovering and representing its collective meanings is universal. It provides a way to present and analyze the query return that seems better suitable for scholars to explore digital archives than the more conventional search engine that treats a query return as a ranked list.

Figure 1: Snapshot of THDL with query term "找洗字"

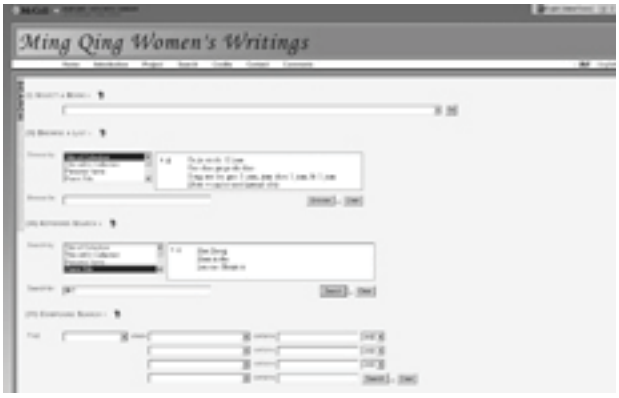


Figure 2: Query resulted from issuing a polygon (region) as a query

References

Chen S. P. (2011). *Information technology for historical document analysis*, Ph.D. thesis, National Taiwan University, Taipei, Taiwan.

Chen, S. P., Y. M. Huang, H. L. Ho, P. Y. Chen, and J. Hsiang (2011). Discovering land transaction relations from land deeds of Taiwan. *Digital Humanities 2011 Conference*, June 19-22, 2011. Stanford, CA, pp. 106-110.

Hsiang, J., S. P. Chen, and H. C. Tu (2009). On building a full-text digital library of land deeds of Taiwan. *Digital Humanities 2009 Conference*. Maryland, June 22-25, 2009, pp. 85-90.

Hsiang, J., S. P. Chen, H. L. Ho, and H. C. Tu (2012). Discovering relations from imperial court documents of Qing China. *International Journal of Humanities and Arts Computing* 6: 22-41.

Hsieh, Y. P. (2012), Appositional Term Clip: a subject-oriented appositional term extraction algorithm. In J. Hsiang (ed.), *New Eyes for Discovery: Foundations and Imaginations of Digital Humanities*. Taipei: National Taiwan UP, pp. 133-164.

Huang, Y. M. (2009). *On reconstructing relationships among Taiwanese land deeds*. MS Thesis, National Taiwan University, Taipei, Taiwan.

Ou, C.-H. (2011). *Creating a GIS for the history of Taiwan – a case study of Taiwan during the Japanese occupation era*. MS Thesis, National Taiwan University, Taipei, Taiwan.

PAPER 4

System Interoperability and Modeling Women Writers' Life Histories of Late Imperial China

Fong, Grace
grace.fong@mcgill.ca
McGill University, Canada

Recent scholarship has shown that Chinese women's literary culture began to flourish on an unprecedented level alongside the boom in printing industry in the late sixteenth century, and continued as a cultural phenomenon in the late imperial period until the end of the Qing dynasty (1644-1911) (Ko 1994; Mann 1997; Fong 2008; Fong & Widmer 2010). Over 4,000 collections of poetry and other writings by individual women were recorded for this period (Hu and Zhang 2008). These collections with their rich autobiographical and social contents open up gendered perspectives on and complicated many aspects of Chinese culture and society – unsuspected kinship and family dynamics, startling subject positions, new topoi and genres, all delivered from the experience of literate women. Yet, within the Confucian gender regime, women were a subordinated group and ideally to be located within the domestic sphere, and in many instances their writings had not been deemed worthy of systematic preservation. Perhaps less than a quarter of recorded writings in individual collections have survived the ravages of history; many more fragments and selections are preserved in family collections, local gazetteers, and various kinds of anthologies. Works by individual women have been difficult to access for research, as they have mostly ended up in rare book archives in libraries in China.

Aimed at addressing the problem of accessibility, the Ming Qing Women's Writings project (MQWW; http://digital.library.mcgill.ca/mingqing) is a unique digital archive and web-based database of historical Chinese women's writing enhanced by a number of research tools. Launched in 2005, this collaborative project between McGill and Harvard University makes freely available the digitized images of Ming-Qing women's texts held in the Harvard-Yenching Library, accompanied by the analyzed contents of each collection, which can be viewed online, and a wealth of literary, historical, and geographical data that can be searched (Fig. 1-4). However, MQWW is not a full-text database – we do not have the funding resources for that, but it is searchable based on an extensive set of metadata. In addition to identifying thematic categories, researchers can link between women based on exchange of their work and correspondence, obtain contextual information on family and friends, and note the ethnicity and marital status of the women writers, among other data fields. Each text within a collection is analyzed and identified according to author, title, genre and subgenre, whether it is a poem, a prose piece, or chapter of a verse novel. The MQWW digital archive contains more than 10,000 poems (majority by women) and also more than 10,000 prose pieces of varying lengths and genres (some by men), ranging from prefaces to epitaphs, and over 20,000 images of original texts. It clearly enables literary research; conferences, research papers, and doctoral dissertations have drawn on the resources provided, as well, the database has been serving as an important teaching resource in North America, China, Hong Kong, and Taiwan.

While the MQWW database contains basic information on 5,394 women poets and other writers, it was not originally designed with biographical research in mind. Yet, continuing research has pointed to how family and kinship, geographical location and regional culture, and social networks

and literary communities are significant factors affecting the education, marriage, and general life course of women as writers (Mann 2007; Widmer 2006). The current phase of the collaborative project with the China Biographical Database (CBDB), Harvard University, will develop the biographical data of MQWW for large-scale prosopographical study (Keats-Rohan 2007). It is guided by the potential for taking digital humanities research in the China field to a new stage of development by focusing on system interoperability. By this we mean that online systems are built so as to enable interaction with other online systems through an Application Programming Interface (API), which must be developed in both directions. Our methodological strength is based on the two robust databases with demonstrated long-term sustainability and scholarly significance. An API is being created for CBDB to enable it to interact with MQWW and vice versa. MQWW will retain its separate identity as a knowledge site with its unique digital archive of texts and multifunction searchable database. System interoperability will support cross-database queries on an ad hoc basis by individual users. We are developing an integrated search interface to query both databases.

My paper will be an experiment based on a project in progress. With the massive biographical data available in CBDB, I will formulate and test queries for kinship and social network analysis for women writers in MQWW for a life history model, or a prosopographical study, of writing women in late imperial China. Some of the questions I will address are:

- Can we map historical changes statistically to test the 'high tides' of women's writings (mid seventeenth century and late eighteenth century), which were arrived at through non-statistical observations?
- What graphs do the data yield for women's social associations when mapped according to geographical regions, historical periods, and male kin success in the civil examination system, and other defined parameters?
- How do the official assignments of male kin affect groups of women in their life cycle roles: as daughter, wife, mother, mother-in-law, grandmother?
- What social patterns can emerge, such as local and translocal marriage and female friendship, through social network analysis with data in CBDB?

The rich data for male persons in CBDB can offer possibilities for mapping the life history of women writers: the circulation of their books, the physical 'routes' of movement in their lives, their temporal and geographical experiences, and their subtle role in the social and political domain, and disclose unsuspected lines of family and social networks, political alliances, literati associations, and women's literary communities for further study and analysis. The experiments can show the probabilities and possibilities of the new system interoperability for identifying not only normative patterns, but also 'outlier' cases in marginal economic, geographical, and cultural regions, for which social, political, and historical explanations can be sought through more conventional humanities research.

Figure 1: Search poems of mourning by the keyword = dao wang (悼亡) using the English interface

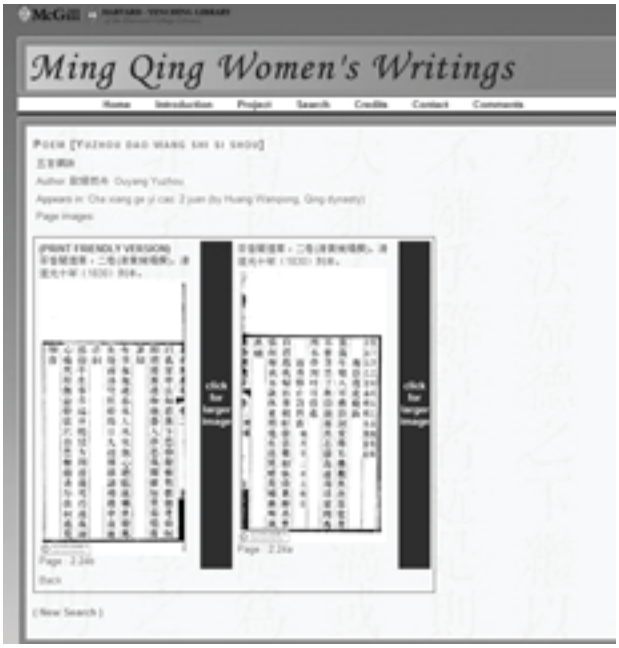


Figure 2: Search results displaying poem titles with the keyword = dao wang (悼亡) with links to digitized text images

Figure 3: Digitized texts of the first set of dao wang (悼亡) poems in the result list

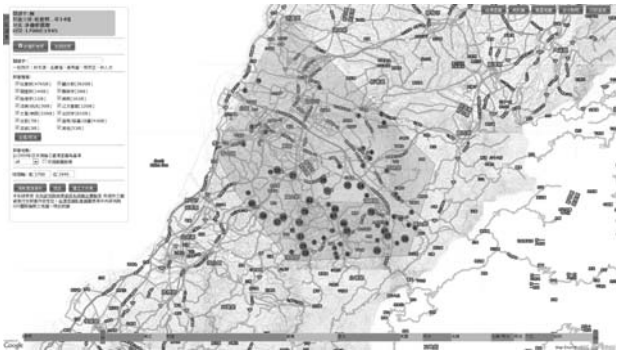


Figure 4: Search results for the woman poet Xi Peilan (1760-after 1829)

References

Fong, G. S. (2008). *Herself an Author: Gender, Agency, and Writing in Late Imperial China*. Honolulu: U of Hawaii P.

Fong, G. S., and E. Widmer, eds. (2010). *The Inner Quarters and Beyond: Women Writers from Ming through Qing*. Leiden: Brill.

Hu, W., and H. Zhang (2008). *Lidai funü zhuzuokao (zengdingben)* (Catalogue of women's writings through the ages [with supplements]). Shanghai: Shanghai guji chubanshe.

Keats-Rohan, K. S. B., ed. (2007). *Prosopography Approaches and Applications: A Handbook*. Oxford: Unit for Prosopographical Research, Linacre College, University of Oxford.

Ko, D. (1994). *Teachers of the Inner Chambers: Women and Culture in Seventeenth-Century China*. Stanford: Stanford UP.

Mann, S. (1997). *Precious Records: Women in China's Long Eighteenth-Century*. Berkeley, CA: U of California P.

Mann, S. (2007). *The Talented Women of the Zhang Family*. Berkeley, CA: U of California P.

Widmer, E. (2006). *The Beauty and the Book: Women and Fiction in Nineteenth-Century China*. Cambridge, MA: Harvard University Asia Center.

Future Developments for TEI ODD

Cummings, James
James.Cummings@oucs.ox.ac.uk
University of Oxford, UK

Rahtz, Sebastian
Sebastian.Rahtz@oucs.ox.ac.uk
University of Oxford, UK

Burnard, Lou
lou.burnard@tge-adonis.fr
TGE-Adonis, France

Bauman, Syd
Syd_Bauman@Brown.edu
Brown University, USA

Gaiffe, Bertrand
bertrand.gaiffe@atilf.fr
ATILF, France

Romary, Laurent
laurent.romary@inria.fr
INRIA, France

Bański, Piotr
bansp@02.pl
HUB & IDS, Poland

The purpose of this panel is to look at the application and future development of the literate programming system known as ODD which was developed for the Text Encoding Initiative (TEI) and underlies every single use of the TEI.

Though strongly influenced by data modelling techniques characteristic of markup languages such as SGML and XML, the conceptual model of the Text Encoding Initiative is defined independently of any particular representation or implementation. The objects in this model, their properties, and their relationships are all defined using a special TEI vocabulary called ODD (for One Document Does-it-all); in this way, the TEI model is used to define itself and a TEI specification using that model is, formally, just like any other kind of resource defined using the TEI. An application selects the parts of the TEI model it wishes to use, and any modifications it wishes to make of them, by writing a TEI specification (a TEI ODD document), which can then be processed by appropriate software to generate instance documents relevant to the given application. Typically, these instance documents will consist of both user documentation, such as project manuals for human use, and system documentation, such as XML schemas or DTDs, sets of Schematron constraints etc. for machine use. In this respect ODD is a sophisticated re-implementation of the 'literate programming' paradigm developed by Don Knuth in the 1970s reimagined as 'literate encoding'.

One of the requirements for TEI Conformance is that the TEI file 'is documented by means of a TEI Conformant ODD file which refers to the TEI Guidelines'. In many cases users employ pre-generated schemas from exemplar customizations, but they are usually better served if they use the TEI ODD markup language, possibly through the Roma web application, to constrain what is available to them thus customizing the TEI model to reflect more precisely their encoding needs.

Some of the mechanisms supporting this extensibility are relatively new in the TEI Guidelines, and users are only now beginning to recognize their potential. We believe that there is considerable potential for take-up of the TEI system beyond its original core constituencies in language engineering, traditional philology, digital libraries, and digital humanities in general. Recent additions to the TEI provide encoding schemes for database-like information about persons and places to complement their existing detailed recommendations for names and linguistic structures; they have always provided recommendations for software-independent means of authoring scientific documentation, but the ODD framework makes it easier for TEI documents to coexist with other specialist XML vocabularies as well as expanding it to encompass the needs of new specialised kinds of text. It has been successfully used for describing other XML schemas, notably the W3C ITS (Internationalisation Tagset) and ISO TC37 SC4 standards documents; more recently its facilities have greatly simplified the task of extending the TEI model to address the needs of other research communities, such as musicologists and genetic editors.

We believe that the current ODD system could be further enhanced to provide robust support for new tools and services; an important step is to compare and contrast its features with those of other 'meta-encoding' schemes and consider its relationship to ontological description languages such as OWL. The potential role of ODD in the development of the semantic web is an intriguing topic for investigation.

This panel brings together some of the world's most knowledgeable users and architects of the TEI ODD language, including several who have been responsible for its design and evolution over the years. We will debate its strengths, limitations, and future development. Each speaker will focus on one aspect, problem, or possible development relating to TEI ODD before responding to each

others suggestions and answering questions from the audience.

Lou Burnard will introduce the history and practical use of ODD in TEI, and describe its relevance as a means of introducing new users to the complexity of the TEI. Sebastian Rahtz will talk about the processing model for ODD, and the changes required to the language to model genuinely symmetric, and chainable, specifications. Bertrand Gaiffe will look at some of the core mechanisms within ODD, and suggest that model classes that gather elements (or other model classes) for their use into content models could be better as underspecified bags instead of sets. Syd Bauman will discuss co-occurrence constraints, pointing out that ODD's lack of support for this feature is a significant limitation, but also that it can often be worked around by adding Schematron to an ODD. Laurent Romary and Piotr Banski will describe issues in drafting ODD documents from scratch, in particular in the context of ISO standardisation work, introducing proposals to make ODD evolve towards a generic specification environment.

We believe that the DH2012 conference offers a useful outreach opportunity for the TEI to engage more closely with the wider DH community, and to promote a greater awareness of the power and potential of the TEI ODD language. We also see this as an invaluable opportunity to obtain feedback about the best ways of developing ODD in the future, thereby contributing to the TEI Technical Council's ongoing responsibility to maintain and enhance the language.

Organization

The 5 speakers will each give a 15-minute introduction to a problem or possible development that relates to TEI ODD. After this the organizer will moderate discussion between members of the panel on a number of questions before opening the discussion to questions from the audience.

Speakers

- Lou Burnard, lou.burnard@tge-adonis.fr, TGE-Adonis
- Syd Bauman, syd_bauman@brown.edu, Brown University Center for Digital Scholarship
- Bertrand Gaiffe, bertrand.gaiffe@atilf.fr, ATILF
- Sebastian Rahtz, sebastian.rahtz@oucs.ox.ac.uk, University of Oxford
- Laurent Romary & Piotr Bański (Piotr speaking), laurent.romary@inria.fr, bansp@02.pl, Inria, HUB & IDS

Organizer James Cummings, james.cummings@oucs.ox.ac.uk, University of Oxford

References

Burnard, L., and R. Rahtz (2004). RelaxNG with Son of ODD. *Extreme Markup Languages 2004*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.7139&rep=rep1&type=pdf

Romary, L. (2009). ODD as a generic specification platform, *Text encoding in the era of mass digitization – Conference and Members' Meeting of the TEI Consortium* http://hal.inria.fr/inria-00433433

TEI By Example (2010). Customising TEI, ODD, Roma. *TEI By Example*, http://tbe.kantl.be/TBE/modules/TBED08v00.htm

TEI Consortium, eds. (2007). TEI P5 Guidelines Chapter on "Documentation Elements". *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TD.html

TEI Consortium, eds. (2007). TEI P5 Guidelines Section on "Implementation of an ODD System". *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html#IM

TEI Consortium, eds. (2007). Getting Started with TEI P5 ODDs. *TEI-C Website*. http://www.tei-c.org/Guidelines/Customization/odds.xml

Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of Publication of Digitized Historical Prints

Geyken, Alexander
geyken@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Gloning, Thomas
Thomas.Gloning@germanistik.uni-giessen.de
CLARIN-D, Germany

Stäcker, Thomas
staecker@hab.de
Herzog August Bibliothek Wolfenbüttel, Germany

PAPER 1

Introduction

1. Problem Statement

It has been and still is one of the core aims in German Digital Humanities to establish large reference corpora for the historical periods of German, i.e. Old and Middle High German, Early New High German (ENHG), New High German and Contemporary German (NHG). This panel focusses on NHG and ENHG. A reference corpus of the 17th to 19th century German is currently being compiled by the Deutsches Textarchiv project at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). The Herzog August Bibliothek Wolfenbüttel (HAB) constructs text corpora encompassing the period between the 15th and the 18th century, with a focus on ENHG. Apart from core activities like these, that usually are carried out by institutions for long time research such as academies, research centers and research libraries, many digital resources are created by individual scholars or small project teams. Often, however, such resources never show up in the pool of publicly available research corpora.

Building an integrated corpus infrastructure for large corpora of ENHG and NHG faces three main problems:

1. Despite the growing acceptance of annotation standards such as the TEI, there are **different 'encoding dialects'** and baseline formats used in different contexts (e.g. TextGrid Baseline Encoding, the TEI-Encoding recommended by HAB, the ›base format‹ of the DTA, and others). Neither one has gained wider acceptance nor have they been checked against each other. As a result, repositories of digital documents do not apply consistent, interoperable encoding. In addition, users cannot draw on these resources in an integrated way.
2. There is **no approved system of evaluation** for both the quality of corpus texts and the quality of the encoding. In addition, there is **no reputation system** for crediting individual researchers' accomplishments with respect to the production and annotation of corpus texts. As a consequence, there is **no 'culture of sharing corpus resources' in a collaborative way** in the research community.
3. The vision of an **integrated** historical corpus of German and an integrated research platform is in **conflict** with the **principle of local ascription** of scholarly work. While the user would like to have *one* place to find *all* the available corpus texts and *one* platform to run the available corpus tools, each academy, each institute, each library, each research project and even individual researchers need to produce work and resources ascribable in a local or even in a personal way.

2. Panel Topics

Well-established **infrastructure projects** such as TextGrid, CLARIN or DARIAH can contribute enormously to the process of integration by establishing methods of interoperation, a system of quality assurance and credit, and a set of technical practices that allow to integrate resources of different origin, to credit the producers of resources in an appropriate way, to ensure public access in a persistent way and thereby to involve a greater scholarly community.

The **proposed panel**, organized by the Deutsches Textarchiv at the BBAW, the Special Interest Group 'Deutsche Philologie' ('German Philology') in CLARIN-D and the HAB Wolfenbüttel, will demonstrate how efforts for community-building, text aggregation, the technical realization of interoperability and the collaboration of two large infrastructure centers (BBAW and HAB) is set to work in the digital publication platforms of the Deutsches Textarchiv (BBAW) and AEDit (HAB).

Technical requirements include tools for long-term archiving as well as the implementation and documentation of reference formats. Interfaces

need to be standardized, easy to handle and well documented to minimize the user's effort to upload their resources. Transparent criteria have to be developed regarding the expected quality, the encoding level and requirements for interoperability of texts in the repository. The DTA provides a large corpus of historical texts of various genres, a heterogeneous text base encoded in one consistent format (i.e. DTA 'base format'). Facilities for quality assurance (DTAQ) and for the integration of external text resources (DTAE) have been developed.

PAPER 2

The DTA 'base format'

Haaf, Susanne
haaf@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutsches Textarchiv, Germany

Geyken, Alexander
geyken@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutsches Textarchiv, Germany

---

The DTA 'base format' consists of about 80 TEI P5 <text>-elements which are needed for the basic formal and semantic structuring of the DTA reference corpus. The purpose of developing the ›base format‹ was to gain coherence at the annotation level, given the heterogeneity of the DTA text material over time (1650-1900) and text types (fiction, functional and scientific texts). The restrictive selection of 'base format' elements with their corresponding attributes and values is supposed to cover all annotation requirements for a similar level of structuring of historical texts. We will illustrate this by means of characteristic annotation examples taken from the DTA reference corpus.

We will compare the DTA 'base format' to other existing base formats considering their different purposes and discuss the usability of the DTA 'base format' in a broader context (DTAE, CLARIN-D). A special adaptation of the oXygen TEI-framework which supports the work with the DTA 'base format' will be presented.

PAPER 3

DTAE

Thomas, Christian
thomas@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutsches Textarchiv, Germany

---

DTAE is a software module provided by the DTA for external projects interested in integrating their historical text collections into the DTA reference corpus. DTAE provides routines for uploading metadata, text and images, as well as semiautomatic conversion tools from different source formats into the DTA 'base format'. The text- and metadata are indexed for lemma-based full text search and processed with tools for presentation in order to offer parallel views of the source image, the XML/TEI encoded text as well as a rendered HTML presentation layer. In addition, external contributors can integrate the processed text into their own web site via <iframe> and use the DTA-query-API for full text queries. DTAE demonstrates how interchange and interoperability among projects can work on a large scale. The paper illustrates these issues by example of five selected cooperation projects. Possibilities (and limits) of the exchange of TEI documents will be discussed as well as the difficult, but worthwhile task of converting other/older text formats into DTA 'base format'.

PAPER 4

DTAQ

Wiegand, Frank
wiegand@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutsches Textarchiv, Germany

---

DTAQ is a browser-based tool to find, categorize, and correct different kinds of errors or inconsistencies occuring during the XML/TEI-transcription resp. the optical character recognition of texts. Using a simple authentication system combined with a fine-grained access control system, new users can easily be added to this Quality Assurance system. The GUI of the tool is highly customizable and provides various views of source images, XML-transcriptions, and HTML-presentations. The backend of DTAQ is built upon many open source packages. Using *Perl* as a glue language, the system runs on *Catalyst*, connects to

a *PostgreSQL* database via the *DBIx::Class* ORM and builds its web pages with *Template Toolkit*. The frontend makes heavy use of *jQuery* and *Highcharts JS* to create a very interactive and responsive user interface.

PAPER 5

The project AEDit

Stäcker, Thomas
staecker@hab.de
Herzog August Bibliothek Wolfenbüttel, Germany

In the field of digital humanities there is a great demand for trustworthy platforms or repositories that aggregate, archive, make available and disseminate scholarly texts or databases. AEDit (Archiving, Editing, Disseminating) aims at establishing such a platform at the HAB Wolfenbüttel for documents with a focus on ENHG. Partners and contributors of the initial phase are four outstanding editorial projects from the German academy program. Central issues will be addressed, e.g. creating persistent identifiers at word level by IDs or Xpointer, determining the relation of community specific markup to basic text encoding by means of stand-off markup, defining a basic format for texts and databases, integrating research data originating from biographic, bibliographic or subject-related and lexical databases, examining ways of interconnection with already existing editorial tools such as TextGrid and last but not least setting up an institutional framework for supporting scholars who are willing to participate or publish online.

PAPER 6

CLARIN-D – Historical Corpora, Collaboration, Community building

Gloning, Thomas
Thomas.Gloning@germanistik.uni-giessen.de
CLARIN-D, Germany

The task of the special interest group ›Deutsche Philologie‹ in Clarin-D is to support infrastructure centers (IDS, BBAW, MPI, HAB) in building up, enriching and integrating German language resources (e.g. corpora, tools, dictionaries, best practices in research methodology). One goal in the field of historical corpora is to create an infrastructure that allows users (i) to integrate their own documents into historical reference corpora; (ii) to gain proper credit and reputation in doing so; (iii) to reuse the documents in local corpora together with expert corpus technology. – In this talk I shall demonstrate usage scenarios that show how the DTA extension infrastructure (DTAE) can be used to create specialized corpora for specific research topics. These usage scenarios shall serve as prototypes for collaboration and community building in ENHG and NHG corpus research.

References

CLARIN-D http://clarin-d.net/

DTA http://www.deutschestextarchiv.de/

DTA 'base format', Description [in German]:

http://www.deutschestextarchiv.de/doku/basisformat

DTAE http://www.deutschestextarchiv.de/dtae

DTAQ http://www.deutschestextarchiv.de/dtaq

Herzog August Bibiliothek Wolfenbüttel http://www.hab.de/

Wolfenbütteler Digitale Bibliothek (WDB) http://www.hab.de/bibliothek/wdb/

Geyken, A., et al. (2011). Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In S. Schomburg, C. Leggewie, H. Lobin, and C. Puschmann (eds.), *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, 20./21. September 2010. Beiträge der Tagung. 2., ergänzte Fassung. Köln: HBZ, pp. 157-161.

Jurish, B. (2010). More than Words: Using Token Context to Improve Canonicalization of Historical German. *Journal for Language Technology and Computational Linguistics (JLCL)* 25(1).

Jurish, B. (2012) *Finite-state Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam 2012 (urn:nbn:de:kobv:517-opus-55789).

Unsworth, J. (2011) Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI. *Journal of the Text Encoding Initiative* 1 (http://jtei.revues.org/215 , 29. 8. 2011).

Bauman, S. (2011). Interchange vs. Interoperability. Presented at Balisage: The Markup

Conference 2011, Montréal, Canada, August 2 - 5, 2011. In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, vol. 7. doi:10.4242/BalisageVol7.Bauman01.

Computational models of narrative structure

Löwe, Benedikt
b.loewe@uva.nl
University of Amsterdam, The Netherlands

Físseni, Bernhard
bernhard.fisseni@uni-due.de
University of Duisburg-Essen, Germany

León, Carlos
carlos.leon@uni-hamburg.de
University of Hamburg, Germany

Bod, Rens
L.W.M.Bod@uva.nl
University of Amsterdam, The Netherlands

PAPER 1

Introduction

A question of particular interest to the Computational Narrative community is the question of the notion of *structural equivalence* of stories (Löwe 2010, 2011). On the one hand, it is closely related to research in other areas such as the study of analogical reasoning in cognitive science and psychology; on the other hand, a solution to the question of when stories can count as structurally similar underlies a number of potential computational applications for narrative databases.

In our panel, the four speakers will discuss various foundational issues that have to be dealt with before a structural theory of narrative similarity can be developed. The majority of these issues have to do with the empirical validation of proposed formal representations; the aim is to develop (1) a methodology that allows to determine and investigate those aspects of narratives that are computationally and cognitively relevant for the comparison of stories and (2) a formal framework that allows to represent narratives with regard to these aspects and also allows to encode the necessary algorithm (formalization guidelines). The presentations will report on joint projects of the panelists in this field, and part of the purpose of the panel is to present the results of these projects to the *Digital Humanities* community.

These tasks are approached using the following empirical and computational methods: First, (quasi-)*experimental studies* are used to determine the relevant dimensions and trainability of analysis systems (Fisseni, Bod below). Secondly,

computational representation and simulation is used to evaluate representational formalisms, and will be experimentally evaluated in a final step (León, below).

1. Theoretical Background

The field of *computational models of narrative* goes back to the 1970s and has produced numerous computational representations of narrative structure (e.g. Lehnert 1981; Turner 1994; León 2010). Its roots lie in the structuralist school of narratology (Barthes, Genette, Greimas, Todorov, among others) that started with Vladimir Propp's study of Russian folk tales (Propp 1928), and it was greatly successful with the methods of modern computational linguistics:

There is now a considerable body of work in artificial intelligence and multi-agent systems addressing the many research challenges raised by such applications, including modeling engaging virtual characters [...] that have personality [...], that act emotionally [...], and that can interact with users using spoken natural language (Si, Marsella & Pynadath 2005: 21).

Recently, there has been an increased interest in developing theoretical foundations of what is called *shallow story understanding* in this community: high-level structural analysis of the narrative as opposed to understanding 'deeply', i.e., with background knowledge. The intersection of narratives and computation is also being considered in the field of Digital Humanities or the application of computer software to narrative analysis. In this context, we assume that theory of narrative structures is a prerogative to computational treatment of narratives. All work presented here is concerned with validating and extending existing theories empirically. Even though non-structural factors may influence judgment of stories, they should evidently be excluded in our formalization of structural similarity. Potentially, one will have to reconsider the notion of 'structural core' and its differentiation from 'mere' accidental features such as motifs or style (the latter is discussed by Crandell et al. 2009, presented at DH 2009).

Two **central themes of the entire panel** are the questions (1) *Is there a structural core of narratives and can we formally approximate it?* and (2) *Are structural similarity judgments a 'natural kind' or rather a trained skill?* The basis for discussing these issues will be prepared in this presentation and further developed in three following presentations.

PAPER 2

Narrative Similarity and Structural Similarity

Löwe, Benedikt
b.loewe@uva.nl
University of Amsterdam, The Netherlands

This first presentation will introduce the notions and concepts that we shall deal with: the distinction between the narrative and its *formalization* (or *annotation*), various levels of granularity, and various dimensions of similarity. We shall discuss the human ability to identify a *structural core* of a narrative and discuss intersubjectively in what respects two narratives are structurally the same.

We discuss the question whether this *structural core* exists and how to approach it. In particular, we shall discuss a number of methodological issues that create obstacles when trying to determine this *structural core* (Löwe 2011; Fisseni & Löwe 2012).

PAPER 3

Empirically Determining 'Optimal' Dimensions and Granularity

Fisseni, Bernhard
bernhard.fisseni@uni-due.de
University of Duisburg-Essen, Germany

Dimensions that can be easily brought into focus by an adequate instruction are highly relevant for our implementations and can presumably also be annotated with high reliability and inter-annotator agreement by test subjects (see Bod, below). These dimensions may also arguably be considered important for the reception of narratives. As different dimensions can be relevant for different tasks, the setting presented to test subjects must be varied to trigger different granularities and (presumably) focus different dimensions. For example, taking the role of a magazine editor should focus different notions than considering movies in an informal setting.

Preliminary experiments (Block et al. submitted; Bod et al. 2012; Fisseni & Löwe 2012) show that naive test subjects do not have a clear preformed concept

of story similarity that privileges the structural core of stories. Therefore, work will have to be done to determine how to focus structural aspect and control other, non-structural aspects.

PAPER 4

A Computational Framework for Narrative Formalizations

León, Carlos
carlos.leon@uni-hamburg.de
University of Hamburg, Germany

Even with the most recent advances of Artificial Intelligence, completely automatic formalizations of narrative texts are still impossible, but it is well possible to develop and process formal representations of stories computationally. In this presentation, we shall focus on implementing a computational instantiation of the set of different formalizations. This instantiation will be used to formalize stories and check their structural similarity under human supervision. In order to do this, a mixed methodology will be applied: computational versions of the defined formal systems will be implemented in the form of several structured descriptions of the stories, along with information about their respective granularities. The dimensions that are modeled should be those that can be easily accessed (see Fisseni, above) and reliably annotated (see Bod, below).

A mixed human-computer process for acquisition of one of the candidate formalizations has been successfully tested by the author (León 2010; León & Gervás 2010); hence, a computational tool will assist human users during the formalization process, iteratively creating partial structures according to the defined granularity. It may also be interesting to use techniques from knowledge representation and natural language processing to formalize at least some guidelines and thus test their consistency and usability. While these guidelines may not unambiguously define how to formalize each story, they will be used to maximize the consensus among the formalizers (see Bod, below).

PAPER 5

Inter-Annotator Agreement for Narrative Annotations

Bod, Rens
L.W.M.Bod@uva.nl
University of Amsterdam, The Netherlands

A way to measure the quality of guidelines and formal representation derived by applying them is inter-annotator agreement, which is used to assess the quality of linguistic structural annotations such as treebanks (see e.g. Carletta et al. 1997; Marcu et al. 1999). We intend to apply inter-annotator agreement to the formal study of narratives (Bod et al., 2011, 2012).

As Propp's formal analysis of Russian folktales (Propp 1928) has profoundly influenced Computational Narratology, we ran a pilot experiment in which external users are annotating several Russian folktales with a subset of Propp's definitions, to establish the viability of the methodology (Bod et al. 2012). After a training process, test subjects were expected to have a basic knowledge about Propp's formal system. In the main phase of the experiment, they were to apply their understanding of the formal system to other stories. The results indicate that Propp's formal system is not easily taught (or learnt), and that this may have to do with the structural constraints of the system: Its functions and roles are so highly mutually dependent that variation is great.

Hence, similar experiments with more 'modern' and formal representations (such as those by León, above) are planned. These experiments will also profit from the preliminary studies (see Fisseni, above) which try to determine which dimensions can be triggered in test subjects and how to achieve this. Then it will be possible to measure agreement between test subjects (using standard statistics), which should provide an insight in the reliability of the guidelines and the viability of the formal representation.

References

Afanas'ev, A. N. (1973). *Russian Fairy Tales*. Translation by N. Guterman from the collections of A. Afanas'ev. Folkloristic commentary by R. Jakobson. New York: Pantheon.

Afanas'ev, A. N. (1985). Shabarsha.Translated by Kathleen Cook. In *The Three Kingdoms*.

Russian Fairy Tales From A. Afanasiev's Collection, illustrated by A. Kurkin. Moscow: Raduga Publisher.

Barzilay, R., and M. Elhadad (1997). Using Lexical Chains for Text Summarization. In I. Mani and M. Maybury (eds.), *Intelligent Scalable Text Summarization. Proceedings of a Workshop sponsored by the ACL*. Somerset, NJ: ACL, pp. 10-17.

Block, A., B. Fisseni, C. León, B. Löwe, and D. Sarikaya (submitted). *Narrative summarization and its correspondence to Proppian functions*.

Bod, R., B. Löwe, and S. Saraf (2011). How much do formal narrative annotations differ? A Proppian case study. In C. Ess and R. Hagengruber (eds.), *The computational turn: Past, presents, futures?, Aarhus University, July 4–6, 2011*. Münster: MV-Wissenschaft, pp. 242-245.

Bod, R., B. Fisseni, A. H. Kurji, and B. Löwe (2012). Objectivity and reproducibility of Proppian narrative annotations. To appear in the proceedings of the workshop on Computational Models of Narratives, Istanbul, 26-27 May 2012.

Brants, T. (2000). Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings Second International Conference on Language Resources and Evaluation LREC-2000*.

Carletta, J. C., A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics* 23(1): 13-31.

Conroy, J. M., and D. P. O'Leary (2001). Text summarization via hidden Markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01. New York, NY: ACM, pp. 406-407.

Crandell, C., E. Gong, R. Kraus, T. Lieu, and J. Mason-Marshall (2009). Modulating Style (and Expectations): An Experiment with Narrative Voice in Faulkner's The Sound and the Fury. Talk at *Digital Humanities 2009*, Maryland.

Dyer, M. G. (1983). *In-depth understanding: A computer model of integrated processing for narrative comprehension*. Artificial Intelligence Series. Cambridge MA: MIT Press.

Fisseni, B., and B. Löwe (2012). Which dimensions of narratives are relevant for human judgments of story equivalence? To appear in the proceedings of the workshop on Computational Models of Narratives, Istanbul, 26-27 May 2012.

Kupiec, J., J. Pedersen, and F. Chen (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM Press, pp. 68-73.

Lehnert, W. G. (1981). Plot units and narrative summarization. *Cognitive Science* 4: 293-331.

León, C. (2010). *A computational model for automated extraction of structural schemes from simple narrative plots*. Ph.D. thesis, Universidad Complutense de Madrid.

León, C., and P. Gervás (2010). Towards a Black Box Approximation to Human Processing of Narratives based on Heuristics over Surface Form. Paper at the *AAAI 2010 Fall Symposium on Computational Models of Narrative November 11-13*, 2010. Arlington, VA.

Lin, C.-Y., and E. Hovy (1997). Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing*, ANLC '97. Stroudsburg, PA: ACL, pp. 283-290.

Löwe, B. (2010). Comparing Formal Frameworks of Narrative Structures. In M. Finlayson (ed.), *Computational models of narrative. Papers from the 2010 AAAI Fall Symposium*, vol. FS-10-04 of *AAAI Technical Reports*, pp. 45–46.

Löwe, B. (2011). Methodological Remarks about Comparing Formal Frameworks for Narratives. In P. Allo and G. Primiero (eds.), *Third Workshop in the Philosophy of Information, Contactforum van de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten*. Brussel: KVAB, pp. 10-28.

Marcu, D. (1998). Improving Summarization through Rhetorical Parsing Tuning. In E. Charniak (ed.), *Proceedings of the Sixth Workshop on Very Large Corpora*. Montréal: Université de Montréal.

Marcu, D., M. Romera, M., and E. Amorrortu (1999). Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. In M. A. Walker (ed.), *Towards Standards and Tools for Discourse Tagging. Proceedings of the Workshop*, pp. 71-78.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19: 302-330.

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM*, 38: 39-41.

Passonneau, R., N. Habash, and O. Rambow (2006). Inter-annotator Agreement on a Multilingual Semantic Annotation Task. In *Proceedings LREC-2006*, pp. 1951-1956.

Propp, V. Ya. (1928). *Morfologiya skazki*. Leningrad: Akademiya.

Rumelhart, D. E. (1980). On evaluating story grammars. *Cognitive Science* 4: 313-316.

Schank, R. C. (1982). *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge: Cambridge UP.

Si, M., S. C. Marsella, and D. V. Pynadath (2005). Thespian: using multi-agent fitting to craft interactive drama. In M. Pechoucek, D. Steiner, and S. Thompson (eds.), *AAMAS '05: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems* (2005), pp. 21-28.

Turner, S. (1994). *The creative process. A computer model of storytelling*. Hillsdale, NJ: Lawrence Erlbaum.

Approaches to the Treatment of Primary Materials in Digital Lexicons: Examples of the New Generation of Digital Lexicons for Buddhist Studies

Nagasaki, Kiyonori
nagasaki@dhii.jp
International Institute for Digital Humanities, Japan

Tomabechi, Toru
tomabechi@dhii.jp
International Institute for Digital Humanities, Japan

Wangchuk, Dorji
dorji.wangchuk@uni-hamburg.de
University of Hamburg, Germany

Takahashi, Koichi
snb44191@nifty.ne.jp
University of Tokyo, Japan

Wallman, Jeff
jeffwallman@tbrc.org
Tibetan Buddhist Resource Center, New York, USA

Muller, A. Charles
acmuller@jj.em-net.ne.jp
University of Tokyo, Japan

PAPER 1

Introduction

Recently, several projects have emerged with the shared aim of creating online digital lexicons for Buddhist studies. There are many possibilities in the creation of digital lexicons for Buddhist studies, since the Buddhist religion itself is so extensively multilingual and multicultural, and also has an unusually broad variety and number of readers. As seen in Father Roberto Busa's initiation of his Thomas Aquinas lexicon project in 1949, the digital lexicon has been a basic resource in the digital humanities. The field of Buddhist studies was one of the first to have its own newly-created comprehensive digital lexicons, entitled 'Digital Dictionary of Buddhism (`http://www.buddhism-dict.net/ddb`).' Under continuous development on the web for more than 15 years, it has been

one of the successful examples of the creation of new online reference resources from various technical and scholarly perspectives. However, recent developments in the area of ICT have given those who aim to make other type of digital lexicons new opportunities to attempt to realize their ideals. Most important in this regard is the spread of collaborative frameworks on the Web. Moreover, TEI P5 has been aiding in the development of such frameworks. Thus, the movement toward the creation of online lexicons has become steadily more visible.

One basic difficulty seen in the construction of digital lexicons for Buddhist studies is that of making the decisions of selection between such a diversity of resources, methodologies, and potential users, along with the possibilities of improvement of primary texts as the sources of individual entries due to the ongoing discovery of ancient manuscripts. Textual sources are written in a range of languages including Sanskrit, Pali, Tibetan, Chinese, and so on; other types of materials such as pictures, statues, maps, etc., also need to be treated. Methodologies need to be diverse because the field of Buddhist studies includes the approaches of philosophy, literature, history, culture, psychology, and a number of other disciplinary approaches. Going far beyond specialists in Buddhist studies, users of such lexicons also include scholars from many other fields, as well as followers of Buddhist religion, and general users. On the other hand, many primary sources have gradually come to be distributed on the Web so that online lexicons can easily refer to them.

In this presentation, we will focus on the way of treating primary sources in each project. This is important, because there is still considerable debate among these projects regarding optimal approach. This includes a transitional result of the ITLR project and Bauddha Kośa project, the results of the work of the Tibetan Buddhist Research Center. We expect comments from researchers and practitioner of other fields.

PAPER 2

Indo-Tibetan Lexical Resource (ITLR): The Underlying Principle, Policy, and Practice of Employing Primary and Secondary Sources

Wangchuk, Dorji
dorji.wangchuk@uni-hamburg.de
University of Hamburg, Germany

The collaborative project 'Indo-Tibetan Lexical Resource' (henceforth ITLR) has been initiated with the sole aim of creating a digital lexical resource that will benefit both academics and non-academics who are engaged in the study of Buddhist (i.e. mainly but not exclusively Indic and Tibetan) textual and intellectual cultures. The ITLR database will include Indic words (or phrases), terms, names with their corresponding Tibetan – and occasionally also Chinese and Khotanese – translations, their etymologies and explanations found primarily in Indic and Tibetan sources, metonyms or synonyms, enumerative categories and classifications, modern renderings, related discussions found in modern academic works – all substantiated with primary and secondary sources. Recognizing the advantages of a digital lexical resource over a printed one, the aim of the ITLR from the very outset has been to create a research tool that is continually improvable, extendable, easily accessible, and reliable.

Reliability is a key issue in dealing with or employing sources and resources. One of the causes of frustration in the field of Buddhist Studies seems to be not the lack of relevant lexicographical resources per se but the lack of a comprehensive and reliable up-to-date lexical resource. Of course, while we cannot speak of reliability in absolute terms, maximizing the degree of reliability has been one of the envisioned goals of the ITLR project. Our success will depend not only on the availability of financial, technical, and human resources but also on our competence, cautiousness, and perseverance, not to speak of on how we use primary and secondary sources.

In this presentation, we will discuss the underlying practice, policy, and principle of employing primary and secondary sources for the ITLR project. It will be argued that while the absence or presence of source-references in a lexical work in itself might not

necessarily indicate its reliability or unreliability, the lack of verifiable evidences, as in the case of most existing digital Tibetan dictionaries, would often undermine its credibility.

PAPER 3

The *Bauddha Kośa* Project: developing a new digital glossary of Buddhist technical terminology

Koichi Takahashi
snb44191@nifty.ne.jp
University of Tokyo, Japan

Buddhist technical terms are occasionally so profound that it is difficult to translate them into modern languages. Thus, many of scholars have been making efforts to compose specialized dictionaries in the Buddhist terminology. Added to that, today there is movement to develop digital dictionaries which can be browsed on the internet. In this situation, our project attempts to make a new type of digital glossary of the Buddhist technical terms, which is called the *Bauddha Kośa*, using XML as the data framework.

The basic methodology of our project is, rather than applying definitions to terms by ourselves, to extract statements to explain the meanings of words from the classical texts written in various languages. Then we add the historical rendering for them and the annotations written in Sanskrit, some of which are available only in the Chinese or Tibetan translations today, on the statements quoted in our glossary. At the same time, we translate these sentences into modern languages in order to propose the more appropriate and intelligible translation equivalents. In other words, the new glossary consists of citations from the classical works and their translation equivalents.

As for the policy to digitize this glossary, our project follows TEI P5 (`http://www.tei-c.org/Guidelines/P5/` , [2011/09/13]). Although TEI P5 provides adequate elements to encode a general dictionary or glossary in modern Western languages, our project occasionally faced the difficulty to digitize our glossary by using the Guidelines of TEI P5 because of its peculiarity that it mainly consists of citations demanding to show the information about sources. (This issue was reported at the poster session of OSDH 2011 on September 13, 2011.)

In this way, some issues to solve remain before accomplishing our purpose, but we are preparing the model of the new digital glossary depending on the *Abhidharmakośabhāsya* (ed. by P. Prahan 1967), one of the most important glossaries of the Buddhist technical terms composed by Vasubandhu in 5c. This text has two Chinese translations and one Tibetan, and some historical commentaries. In this presentation, I will argue on a few issues to develop the scheme for our glossary *Bauddha Kośa* by using TEI P5 from the philological viewpoint.

PAPER 4

A Dynamic Buddhist Lexical Resource based on Full Text Querying and Tibetan Subject Taxonomies

Wallman, Jeff
jeffwallman@tbrc.org
Tibetan Buddhist Resource Center, New York, USA

To accurately define terms in a lexical resource one must be able to identify the context of those terms in literature. As inquiry into the Tibetan Buddhist lexicon progresses, the Tibetan Buddhist Resource Center (TBRC) offers a framework to evaluate lexical terms (technical terms, concepts, subjects, keywords) in a wide variety of contexts across a massive corpus of source Tibetan texts. Through its preservation and cataloging process, TBRC has developed a method to classify individual works within larger collections according to indigenous Tibetan subject classifications. These indigenous subjects are then organized according to a broader, more generalized framework of knowledge-based taxonomies. Each taxonomy is structured around a series of 'heap spaces' – groupings of similar topics, rather than hierarchical structures.

The source literature corpus includes a burgeoning TEI-compliant eText repository as well as scanned source xylographs, manuscripts and modern reprints, spanning the range of the Tibetan literary heritage. The corpus is the largest online repository of Tibetan materials in the world. The metadata framework and text corpus is the basis of an integrated library resource being developed at TBRC. Typing in a technical term in the library, a researcher can see the entry in the database of topics, the location of each topic in a taxonomy, and the relevant

associated works. Extending from this controlled entry point, the researcher can then discover and markup terms in pages, by issuing full-text queries across the eText repository.

References

Muller, A. Ch. (2011). The Digital Dictionary of Buddhism: A Collaborative XMLBased Reference Work that has become a Field Standard: Technology and Sustainable Management Strategies. *Digital Humanities 2011 Conference Abstracts*. June 2011, pp. 189-190.

Nagasaki, K., et al. (2011). Collaboration in the Humanities – Through the Case of Development of the ITLR Project –. *IPSJ Symposium Series Vol. 2011, No. 8: The Computers and the Humanities Symposium*. Dec 2011, pp. 155-160.

Topic Modeling the Past

Nelson, Robert K.
rnelson2@richmond.edu
University of Richmond, USA

Mimno, David
david.mimno@gmail.com
Princeton University, USA

Brown, Travis
travisrobertbrown@gmail.com
University of Maryland, College Park, USA

PAPER 1

Introduction

The enormous digitized archives of books, journals, and newspapers produced during the past two decades present scholars with new opportunities - and new challenges as well. The possibility of analyzing increasingly large portions of the historical, literary, and cultural record is incredibly exciting, but it cannot be done with conventional methods that involve close reading or even not-so-close skimming. These huge new text archives challenge us to apply new methods. This panel will explore one such method: topic modeling.

Topic modeling is a probabilistic, statistical technique that uncovers themes and topics and can reveal patterns in otherwise unwieldy amounts of text. In topic modeling, a 'topic' is a probability distribution over words or, put more simply, a group of words that often co-occur with each other in the same documents. Generally these groups of words are semantically related and interpretable; in other words, a theme, issue, or genre can often be identified simply by examining the most common words in a topic. Beyond identifying these words, a topic model provides proportions of what topics appear in each document, providing quantitative data that can be used to locate documents on a particular topic or theme (or that combine multiple topics) and to produce a variety of revealing visualizations about the corpus as a whole.

This panel will, first and foremost, illustrate the interpretative potential of topic modeling for research in the humanities. Robert K. Nelson will analyze the similarities and differences between Confederate and Union nationalism and patriotism during the American Civil War using topic models of two historic newspapers. Travis Brown will explore techniques to tailor topic model generation using historical data external to a corpus to

produce more nuanced topics directly relevant to particular research questions. David Mimno (chief maintainer of the most widely used topic modeling software, MALLET) will describe his work using topic modeling to generate – while respecting copyright – a new scholarly resource in the field of Classics that derives from and organizes a substantial amount of the twentieth-century scholarly literature.

The panel will also address methodological issues and demonstrate new applications of topic modeling, including the challenge of topic modeling across multi-lingual corpora, the integration of spatial analysis with topic modeling (revealing the constructedness of space, on the one hand, and the spatiality of culture, on the other), and the generation of visualizations using topic modeling useful for 'distant reading.' The panel thus addresses issues of multilingualism, spatial history, data mining, and humanistic research through computation.

PAPER 2

Modeling Nationalism and Patriotism in Civil War America

Nelson, Robert K.
rnelson2@richmond.edu
University of Richmond, USA

Scholars of the American Civil War have productively attended to particular keywords in their analyses of the conflict's causes and its participants' motivations. Arguing that some words carried extraordinary political and cultural weight at that moment, they have sought to unpack the deep connotations of terms that are especially revealing and meaningful. To take a couple of recent examples, Elizabeth R. Varon frames *Disunion!: The Coming of the American Civil War, 1789-1859* (unsurprisingly) around the term 'disunion': 'This book argues that 'disunion' was once the most provocative and potent word in the political vocabulary of Americans' (Varon 2008: 1). Similarly, in *The Union War* Gary W. Gallagher emphasize the importance of 'Union,' arguing that 'No single word in our contemporary political vocabulary shoulders so much historical, political, and ideological meaning; none can stir deep emotional currents so easily' (Gallagher 2011: 46). Others studies have used terms like 'duty,' 'honor,' 'manliness,' 'freedom,' 'liberty,' 'nation,' 'republic,' 'civilization,' 'country,' and 'patriotism' to analyze the ideological perspectives and cultural pressures that shaped the actions and perspectives of soldiers and civilians during the Civil War (Linderman 1987; Prior 2012; Gallagher 1997: 73).

Together, the production of enormous digital archives of Civil War-era documents in the past decade and the development of new sophisticated text-mining techniques present us with an opportunity to build upon the strengths of this approach while transcending some of its limitations. While unquestionably insightful, arguments that have relied heavily upon keyword analyses are open to a number of critiques. How do we know that the chosen keywords are the best window through which to examine the issues under investigation? How can we know – especially in studies which rely upon keyword searches in databases – that we have not missed significant evidence on the topic that does not happen to use the exact terms we look for and analyze? Does the selection of those words skew our evidence and predetermine what we discover? Topic modeling addresses these critiques of the keyword approach while offering us potentially even greater insights into the politics and culture of the era. First, as a 'distant reading' approach it is comprehensive, allowing us to analyze models that are drawn not from a selection but from the entirety of massive corpora. Second, as it identifies word distributions (i.e. 'topics'), topic modeling encourages – even forces – us to examine larger groups of related words, and it surfaces resonant terms that we might not have expected. Finally and perhaps most importantly, the topics identified by this technique are all the more revealing because they are based on statistical relationships rather than *a priori* assumptions and preoccupations of a researcher.

This presentation will showcase research into Union and Confederate nationalism and patriotism that uses topic modeling to analyze the full runs of the Richmond *Daily Dispatch* and the *New York Times* during the war – taken together a corpus consisting of approximately 90 million words. It will make three interrelated arguments drawn from a combination of distant and close readings of topic models of the *Dispatch* and the *Times*.

First, I will argue that Confederates and Yankees used the same patriotic language to move men to be risk their lives by fighting for their countries. Distinct topic models for the *Dispatch* and the *Times* each contain topics with substantially overlapping terms (see table below) – terms saturated with patriotism. Typically celebratory of the sacrifices men made in battle, the patriotic pieces (often poems) in these similar topics from each paper aimed to accomplish the same thing: to evoke a love of country, God, home, and family necessary to move men to risk their

lives and believe that it was glorious to die for their country.

Table 1: The top 24 predictive words for two topics from the Dispatch and the Times, with the words they shared in bold

Second, I will suggest that southerners (or at least the editor of the *Dispatch*) developed a particularly vitriolic version of Confederate nationalism to convince southern men to kill northerners. The *Dispatch* was full of articles that insisted that northerners were a foreign, unchristian, and uncivilized people wholly unlike southerners; in vicious editorials the *Dispatch's* editor maintained that northerners were infidels and beasts who it was not only okay but righteous to kill. The remarkably similar signatures evident in a graph of two topics (Figure 1) – one consisting of patriotic poetry aimed at moving men to die, the other of vicious nationalistic editorials aimed at moving them to kill – from a model of the *Dispatch* suggests, first, the close relationship between these two topics and, second, the particular moments when these appeals needed to be made: during the secession crisis and the early months of the war in the spring and summer of 1861 when the army was being built, immediately following the implementation of the draft in April 1862, and at the end of the war in early 1865 as Confederates struggled to rally the cause as they faced imminent defeat.

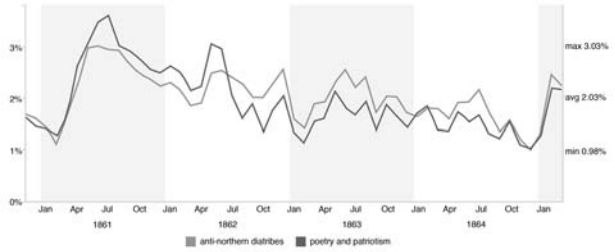


Figure 1

Finally, I will argue the kind of nationalism evident in the *Dispatch* was not and could not be used by northerners for the same purpose. While northerners and southerners used the same language of patriotism, there is no analog to the vicious nationalistic topic from the *Dispatch* in the topic model for the *New York Times*. Unionists insisted that the South was part of the United States and southerners had been and continued to be Americans – though traitorous Americans, to be sure. As a title of one article in the *Times* proclaimed, this was 'Not a War against the South.' It was a war against traitors, the *Times* insisted, and the 'swords [of Union soldiers] would as readily seek a Northern heart that was false to the country as a Southern bosom' ('Not a War against the South,' 1861). Northern nationalism is evident in the model for the *Times* in a more politically inflected topic on Unionism and a second topic consisting of patriotic articles and poems. The graphs of these two topics (Figure 2) with spikes during elections seasons suggest the instrumental purpose of nationalistic and patriotic rhetoric in the *Times*: not to draw men into the army but rather to drive them to the polls. The editor of the *Times* (correctly, I think) perceived not military defeat but flagging popular will as the greatest threat the Union war effort, and victory by copperhead Democrats who supported peace negotiations would have been the most potent expression of such a lack of will.

In briefly making these historical and historiographic arguments about nationalism and patriotism and about dying and killing in the American Civil War, this presentation aims to demonstrate the interpretative potential of topic modeling as a research methodology.

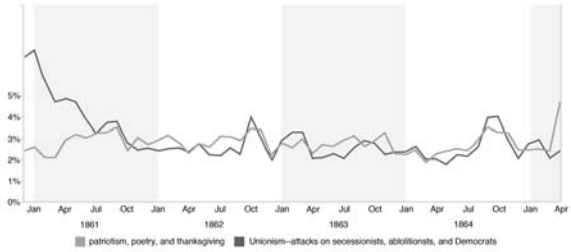


Figure 2

References

Gallagher, G. W. (1997). *The Confederate War*. Cambridge: Harvard UP.

Gallagher, G. W. (2011). *The Union War*. Cambridge: Harvard UP.

Linderman, G. F. (1997). *Embattled Courage: The Experience of Combat in the American Civil War*. New York: Free Press.

Not a War Against the South. (1861). *New York Times*. 10 May. Available at: `http://www.nytimes.com/1861/05/10/news/not-a-war-against-the-south.html` [Accessed on 13 March 2012].

Prior, D. (2010). Civilization, Republic, Nation: Contested Keywords, Northern Republicans, and the Forgotten Reconstruction of Mormon Utah. *Civil War History* 56(3): 283-310.

Varon, E. R (2008). *Disunion!: The Coming of the American Civil War, 1789-1859*. Chapel Hill: U of North Carolina P.

PAPER 3

Telling New Stories about our Texts: Next Steps for Topic Modeling in the Humanities

Brown, Travis
travisrobertbrown@gmail.com
University of Maryland, College Park, USA

Latent Dirichlet Allocation (LDA) topic modeling has quickly become one of the most prominent methods for text analysis in the humanities, with projects such as the work by Yang et al. (2011) on Texas newspapers and Robert Nelson's *Mining the Dispatch* (Nelson 2010) demonstrating its value for characterizing large text collections. As an unsupervised machine learning technique, LDA topic modeling does not require manually annotated training corpora, which are often unavailable (and prohibitively expensive to produce) for specific literary or historical domains, and it has the additional benefit of handling transcription errors more robustly than many other natural language processing methods. The fact that it accepts unannotated (and possibly uncorrected) text as input makes it an ideal tool for exploring the massive text collections being digitized and made available by projects such as Google Books and the HathiTrust Digital Library.

LDA is an example of a generative model, and as such it has at its heart a 'generative story,' which is a hypothetical narrative about how observable data are generated given some non-observable parameters. In the LDA story, we begin with a set of topics, which are simply probability distributions over the vocabulary. The story then describes the process by which new documents are created using these topics. This process (which has been described many times in the topic modeling literature; see for example the original presentation by Blei et al. (2003)) is clearly not a realistic model of the way that humans compose documents, but when we apply LDA topic modeling to a set of documents we assume that it is a useful simplification. After making this assumption, we can essentially 'play the story in reverse,' using an inference technique such as Gibbs sampling to learn a set of topic distributions from our observed documents. Despite the simplicity of the generative story, the method can produce coherent, provocative, and sometimes uncannily 'insightful' characterizations of collections of documents.

While LDA topic modeling has a clear value for many applications, some researchers in the fields of information retrieval and natural language processing have described it as 'something of a fad' (Boyd-Graber 2011), and suggest that more attention should be paid to the broader context of generative and latent variable modeling. Despite the relatively widespread use of LDA as a technique for textual analysis in the humanities, there has been little work on extending the model in projects with a literary or historical focus. In this paper I argue that extending LDA – to incorporate non-textual sources of information, for example – can result in models that better support specific humanities research questions, and I discuss as examples two projects (both of which are joint work by the author and others) that add elements to the generative story specified by LDA in order to perform more directed analysis of nineteenth-century corpora.

The first of these projects extends LDA to incorporate geographical information in the form of a gazetteer that maps place names to geographical coordinates.[1] We propose a *region topic model* that identifies topics with regions on the surface of the Earth, and constrains the generative story by requiring each toponym to be generated by a topic whose corresponding region contains a place with that name, according to the gazetteer. This approach provides a distribution over the vocabulary for each geographical region, and a distribution over the surface of the Earth for each word in the vocabulary. These distributions can support a wide range of text analysis tasks related to geography; we have used this system to perform toponym disambiguation on Walt Whitman's *Memoranda During the War* and a collection of nineteenth-century American and British travel guides and narratives, for example.

The second project applies a supervised extension of LDA (Boyd-Graber & Resnik 2010) to a collection of Civil War-era newspapers.[2] In this extension the model predicts an observed response variable associated with a document – in our case contemporaneous historical data such as casualty rates or consumer price index – on the basis of that document's topics. We show that this approach can produce more coherent topics than standard LDA, and it also captures correlations between the topics discovered in the corpus and the historical data external to the corpus.

Both of these projects preserve the key advantages that the unsupervised nature of LDA topic modeling entails – the ability to operate on large, unstructured, and imperfectly transcribed text collections, for example – while adding elements of supervision that improve the generated topics and support additional kinds of analysis. While we believe that our results in these experiments are interesting in their own right,

they are presented here primarily as examples of the value of tailoring topic modeling approaches to the available contextual data for a domain and to specific threads of scholarly investigation.

This work was supported in part by grants from the New York Community Trust and the Institute of Museum and Library Services.

References

Blei, D. M., A. Ng, and M. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Boyd-Graber, J. (2011). Frequently Asked Questions. http://www.umiacs.umd.edu/~jbg/static/faq.html (accessed 23 March 2012).

Boyd-Graber, J., and P. Resnik (2010). Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation. In *Proceedings of Empirical Methods in Natural Language Processing*. Cambridge, MA, October 2010.

Nelson, R. K. (2010). *Mining the Dispatch*. http://dsl.richmond.edu/dispatch/ (accessed 23 March 2012).

Speriosu, M., T. Brown, T. Moon, J. Baldridge, and K. Erk (2010). Connecting Language and Geography with Region-Topic Models. In *Proceedings of the 1st Workshop on Computational Models of Spatial Language Interpretation*. Portland, OR, August 2010.

Yang, T., A. Torget, and R. Mihalcea (2011). Topic Modeling on Historical Newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland, OR, June 2011. http://www.aclweb.org/anthology/W11-1513 (accessed 23 March 2012).

Notes

1. Joint work by the author with Jason Baldridge, Katrin Erk, Taesun Moon, and Michael Speriosu. Aspects of this work were presented by Speriosu et al. (2010) and will appear in an article in an upcoming special issue of *Texas Studies in Literature and Language*.
2. Joint work by the author with Jordan Boyd-Graber and Thomas Clay Templeton. We have also presented results from this work at the 2011 Chicago Colloquium on Digital Humanities and Computer Science on experiments using casualty rates as the response variable.

PAPER 4

The Open Encyclopedia of Classical Sites: Non-consumptive Analysis from 20th Century Books

Mimno, David
david.mimno@gmail.com
Princeton University, USA

Traditional scholarship is limited by the quantity of text that a researcher can read. Advances in large-scale digitization and data analysis have enabled new paradigms, such as 'distant reading' (Moretti 2000). These data-driven approaches, though not approaching the subtlety of human readers, offer the ability to make arguments about entire intellectual fields, from collections spanning hundreds of years and thousands of volumes. Unfortunately, although such corpora exist, the current legal environment effectively prohibits direct access to material published after 1922, even for the great majority of works that are not commercially available (Boyle 2008). This paper explores the feasibility of scholarly analysis on the limited, indirect view of texts that Google Books can legally provide.

The proposed Google Books settlement (Google, Inc. 2011) presents the concept of 'non-consumptive' use, in which a researcher does not read or display 'substantial portions of a Book to understand the intellectual content presented within the Book.' The most common mode of access supported by archives such as JStor and Google Books is keyword search. When a user provides a query, the search engine ranks all documents by their relevance to a specific user-generated query and then displays short text 'snippets' showing query words in context. This interface, though useful, is not adequate for scholarship. Even if researchers have a specific query in mind, there is no guarantee that they are not missing related words that are also relevant. Word count histograms (Michel et al. 2011) suffer similar problems, and are also vulnerable to ambiguous words as they do not account for context.

Another option is the application of statistical latent variable models. A common example of such a method for text analysis is a statistical topic model (Blei, Ng Jordan 2003). Topic models represent documents as combinations of a discrete set of topics, or themes. Documents may be combinations of multiple topics; each topic consists of a probability distribution over words.

Statistical topic models have several advantages over query-based information retrieval systems. They organize entire collections into interpretable, contextually related topics.

Semantically related words are grouped together, reducing the chance of missing relevant documents. Instances of ambiguous words can be assigned to different topics in different documents, depending on the context of the document. For example, if the word 'relief' occurs in a document with words such as 'sculpture' or 'frieze, ' it is likely to be an artwork and not an emotion.

Topic modeling has been used to analyze large-scale book collections published before 1922 and therefore available in full-text form (Mimno & McCallum 2007). In this work I present a case study on the use of topic modeling in digitized corpora protected by copyright that we cannot access in their entirety, in this case books on Greco-Roman and Near-Eastern archeology that have been digitized by Google. The resulting resource, the Open Encyclopedia of Classical Sites, is based on a collection of 240-character search result snippets provided by Google. These short segments of text represent a small fraction of the overall corpus, but can nevertheless be used to build a representation of the contents of the books.

The construction of the corpus involved first selecting a subset of the entire books collection that appeared relevant to Greco-Roman and Near-Eastern archeology. I then defined a set of query terms related to specific archeological sites. These terms were then used to construct the corpus of search result snippets. In this way I was able to use a non-consumptive interface (the search engine) to create a usable sub-corpus without recreating large sections of the original books.

Preprocessing was a substantial challenge. I faced problems such as identifying language in highly multilingual text, recognizing improperly split words, and detecting multi-word terms. This process required access to words in their original sequence, and therefore could not be accomplished on unigram word count data.

Finally I trained a topic model on the corpus and integrated the resulting model with the Pleiades geographic database. This alignment between concepts and geography reveals many patterns. Some themes are geographically or culturally based, such as Egypt, Homeric Greece, or Southern Italy. Other themes cut across many regions, such as descriptions of fortifications or research on trade patterns.

The resulting resource, the Open Encyclopedia of Classical Sites, links geography to research literature in a way that has not previously been available. Users can browse the collection along three major axes. First, they can scan through a list of topics, which succinctly represent the major themes of the book collection, and the specific sites and volumes that refer to each theme. Second, they can select a particular site or geographic region and list the themes associated with that site. For example, the city of Pylos in Greece is the site of a major Mycenaean palace that contained many Linear B tablets, is associated with a character in the Homeric epics, and was the site of a battle in the Peloponnesian war. The topics associated with the site distinguish words related to Linear B tablets, Mycenaean palaces, characters in Homer, and Athens and Sparta. Finally, users can select a specific book and find the themes and sites contained in that volume.

This project provides a model both for what is possible given large digital book collections and for what is feasible given current copyright law. Realistically, we cannot expect to analyze the full text of books published after 1922. But we should also not be satisfied with search and simple keyword histograms. Short snippets provide sufficient lexical context to fix many OCR-related problems and support semantically enriched searching, browsing, and analysis.

Funding

This work was supported by a Google Digital Humanities Research grant.

References

Blei, D. M., A. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993-1022.

Boyle, J. (2008). *The Public Domain*. New Haven: Yale UP.

Google, Inc. (2011). Amended Settlement Agreement. `http://www.googlebooksettlement.com` , accessed Mar 24, 2012.

Michel, J., Y. Shen, A. Aiden, A. Veres, M. Gray, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Aiden. (2011). *Quantitative analysis of culture using millions of digitized books*. Science 331(6014): 176-82.

Mimno, D., and A. McCallum (2007). *Organizing the OCA*. In *Proceedings of the Joint Conference on Digital Libraries*. Vancouver, BC, June 2007.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review* (Jan/Feb): 54-68.

Facilitating Research through Social-Document Networks

Pitti, Daniel
dpitti@Virginia.edu
Institute for Advanced Technology in the Humanities, University of Virginia, USA

Simon, Agnès
agnes.simon@bnf.fr
Bibliothèque nationale de France, France

Vitali, Stefano
vitali.stefano@gmail.com
La Soprintendenza archivistica per l'Emilia-Romagna, Italy

Arnold, Kerstin
k.arnold@barch.bund.de
Bundesarchiv, Germany

People live and work in socio-historical contexts: over the course of their lives they produce a range of records that document their lives and the contexts in which they lived: birth records, correspondence, books, articles, photographs, works of art, films, notebooks, collected artifacts, school records, manuscripts, employment records, etc. Social relations between people, relations among documents, and relations among people and the artifacts created by them constitute a vast social-document network[1] that can be drawn on by scholars needing to reconstruct and study the lives, works, and events surrounding historical persons.

In the past, cultural heritage professionals have largely viewed document networks and social networks in isolation, with a primary focus on documents. Increasingly, in parallel with (and perhaps inspired by) the emergence of social computing on the Web, cultural heritage professionals and scholars are expanding their focus to social-document networks. Scholarly projects that focus on social-document networks or social networks, such as Research-oriented Social Environment (RoSE)[2] and The Crowded Page, [3] have begun to emerge. The cultural heritage communities (library, archive, and museum) have begun to make explicit the implicit networks found in the descriptions of books, articles, manuscripts, correspondence, art objects, and other artifacts in their care, and the authority files used to describe people who created or are documented in the resources described. Particularly important examples are the Virtual International Authority File (VIAF, a collaboration between OCLC Research and major national and research libraries)[4] and WorldCat Identities[5].

This panel will focus on three national and one international cultural heritage projects facilitating humanities research by providing innovative access to social-document networks that provide access to both resources and the socio-historical contexts in which the resources were created.

1. Interconnecting French cultural heritage treasures on the Web: data.bnf.fr – France

Agnès Simon (Curator, Département Information bibliographique et numérique, Bibliothèque nationale de France (BnF)), will discuss data.bnf.fr, a service the BnF is developing to facilitate the discovery of its holdings, interrelated and interconnected to other resources on the Web.

The BnF is the most important heritage institution in France, with a history going back to the 16th century. It has fourteen divisions located over many sites. A large variety of materials are processed in different catalogues, reflecting not only the history of the materials themselves but also that of the methods and technologies used for their description. However, the descriptions are maintained in disparate BnF catalogues and databases, complicating discovery of these resources and their interrelation.

data.bnf.fr will leverage both new conceptual models, for organization of the bibliographic information (such as FRBR), and Linked Data technologies, to provide integrated access to both the holdings of the BnF and related resources available on the Web. A high-level ontology has been designed to make the bibliographic, archival, or other metadata models used in the BnF interoperable. RDF is used to express and expose data extracted from the various library descriptive systems and other complementary systems available on the Web. Incorporating other Semantic web-related ontologies, data.bnf.fr provides a scalable foundation for interconnecting diverse resources, particularly cultural heritage.

The project's effectiveness is made possible by innovative use of the semantics and quality of the structured data contained in the various BnF catalogues. Persons, corporate bodies, and works, accurately identified through authority files, constitute nodes in a network of descriptions of related resources, regardless of the type. Descriptions of persons and corporate bodies are enhanced with information about those entities

from Encoded Archival Description (EAD) finding aids. Other Internet resources are used in a similar, complementary way. The bibliographic data is remodeled and collocated according to FRBR categories, and displayed in a user-friendly way, with direct links to the digitized material from the Gallica digital library, whenever they exist.

data.bnf.fr began in July 2011 with a significant amount of information and is continuously broadening its scope and exploring new ways to collaborate with ongoing initiatives in other cultural heritage institutions in France, namely current work remodeling the French Archives' databases to better fit in the Web landscape of the French and European cultural heritage treasures.

2. Social Networks and Archival Context (SNAC) Project – United States

Daniel Pitti, Associate Director of the Institute for Advanced Technology in the Humanities, University of Virginia, will describe the SNAC project.

The initial phase of the two-year SNAC research and demonstration project began in May 2010 with funding from the National Endowment for the Humanities (U.S.). The project's objectives are to:

- extract data describing creators of and others documented in records from EAD-encoded descriptions of records (finding aids),
- migrate this data into Encoded Archival Context-Corporate Bodies, Persons, Families (EAC-CPF)-encoded authority descriptions,
- augment authority descriptions with additional data from matching library and museum authority records,
- and, finally, use the resulting extracted and enhanced archival authority descriptions to build a prototype system that provides integrated access to the finding aids (and thus the records) from which the descriptions of people were extracted and to the socio-historical contexts in which the records were created.

In the initial phase of SNAC, the primary source of data was 28,000 finding aids. In the second phase, the source data will vastly expand and include not only descriptions of archival records but also original archival authority descriptions. The number of finding aids will increase to more than 148,000, and up to two million OCLC WorldCat collection-level descriptions will be added. The National Archives and Records Administration (NARA), British Library, Smithsonian Institution, BnF, and Archives nationales will contribute nearly 500,000 archival authority descriptions.

While SNAC's immediate objectives are to significantly refine and improve the effectiveness of the methods used in building an innovative research tool, its long-term objective is to provide a solid foundation of both methods and data for establishing a sustainable national archival program cooperatively governed by and maintained by the professional archive and library community.

3. Catalogo delle Risorse archivistiche (CAT) – Italy

Stefano Vitali, Soprintendente archivistico per l'Emilia Romagna (Supervising Office for the Archives in Emilia Romagna Region), will describe the Catalogo delle Risorse archivistiche (CAT). CAT provides integrated access to archival resources held in national, regional, and local repositories. Access to these holdings will be provided via a central system based on archival descriptions of the custodians and creators of archival records.

CAT will sketch a general map of the national archival heritage, providing initial orientation to researchers and guiding them towards more informative resources available in the systems participating in the National Archival Portal. It will contain descriptive records of both the current custodians and original creators of archival records. Data *harvesting* techniques based on the OAI-PMH protocol will be used to aggregate descriptive data from systems distributed throughout Italy. In addition, CAT will explore direct submission of XML-based descriptions of archival repositories and creators, as well as direct data entry into the CAT maintenance interface.

CAT's goal is to provide a comprehensive list of all of the creators of archival records (persons, corporate bodies, and families) held in Italian repositories, and a comprehensive guide to the custodians (institutional and non-institutional) of the archival records. In addition to providing access to and context for archival records in Italy, CAT will be a bridge to other catalogs and descriptive systems in the cultural heritage domain, include the National Library System.

The Archives Portal Europe: Research and Publication Platform for European Archives– Europe

Kerstin Arnold, Scientific Manager, Bundesarchiv, and Leader of the Europeana interoperability work package of APEx, will describe the APEx project and its predecessor, APEnet.

The Archives Portal Europe has been developed within APEnet, a collaboration of nineteen European national archives and the Europeana Foundation, to build a common access point for searching and researching archival content[6]. Funded by the European Commission in the *e*Content*plus*programme, APEnet began in January 2009 and celebrated the release of Archives Portal Europe 1.0 in January 2012. At the moment the portal contains more than 14.5 million descriptive units linked to approximately 63 million digitized pages of archival material. By joining the materials of currently 62 institutions from 14 European countries, the portal has become a major actor on the European cultural heritage scene.

The tasks achieved in APEnet will be taken one step further with its successor, APEx, which recently has held its kick-off meeting at The Hague, The Netherlands. Funded by the European Commission in the ICT Policy Support Programme (ICT-PSP), APEx will inlcude 28 European national archives plus ICARUS (International Centre for Archival Research) as project partners.

While APEnet focused on integrating access to EAD-encoded archival finding aids contributed by the participating national archives and to EAG (Encoded Archival Guide)-encoded descriptions of the national archives themselves, APEx will increase the number of participating archives, enhance and improve the training of archival staff in participating institutions, and improve the quality of the integrated access system, in particular via Web 2.0 functionality and examining Linked Data approaches to be adapted within the Archives Portal Europe.

A primary objective will be incorporating EAC-CPF, the international standard for describing record creators and the people documented in them. The resulting archival authority descriptions will enhance access to archival records, and provide socio-historical context for understanding them.

Notes

1. In the describing the RoSE project to Daniel Pitti, Alan Liu described this network as a 'social-document graph.'
2. RoSE: `http://transliteracies.english.ucsb.edu/category/research-project/rose`
3. The Crowded Page: `http://www.crowdedpage.org/`
4. VIAF: `http://viaf.org/`
5. WorldCat Identities: `http://www.worldcat.org/identities/`
6. Archives Portal Europe: `http://www.archivesportaleurope.eu/` and APEnet: `http://www.apenet.eu/`

Digital Humanities as a university degree: The status quo and beyond

Thaller, Manfred
manfred.thaller@uni-koeln.de
Universität zu Köln, Cologne, Germany

Sahle, Patrick
sahle@uni-koeln.de
Universität zu Köln, Cologne, Germany

Clavaud, Florence
Florence.Clavaud@enc.sorbonne.fr
Ecole Nationale des Chartes, Paris, France

Clement, Tanya
tclement@ischool.utexas.edu
University of Texas, Austin, USA

Fiormonte, Domenico
fiormont@uniroma3.it
Università Roma Tre, Rome, Italy

Pierazzo, Elena
elena.pierazzo@kcl.ac.uk
King's College, London, UK

Rehbein, Malte
malte.rehbein@uni-wuerzburg.de
Universität Würzburg, Würzburg, Germany

Rockwell, Geoffrey
geoffrey.rockwell@ualberta.ca
University of Alberta, Edmonton, Canada

Schreibman, Susan
schreibs@tcd.ie
Trinity College, Dublin, Ireland

Sinclair, Stéfan
stefan.sinclair@mcgill.ca
McGill University, Montreal, Canada

As Desmond Schmidt's recent posting on Humanist revealed (Humanist 24 September 2011), there have been an increasing number of positions advertised in the Digital Humanities (DH). This uptick in hiring has been, while possibly not uniform across disciplines and countries, impressively diverse. It seems that the impact of Digital Humanities as a recognized field of study is increasing.

Over the past decade a plethora of training activity in the form of summer schools, workshops, and focused training events has taken place. There has also been work in building academic programs. While the field has seen a flowering of individual seminars

or modules as part of more traditional curricula over the past decade, the recent announcement of many new degree programs marks a new phase in the discipline. These degree programs, from the undergraduate to the PhD level, are extremely diverse, reflecting not only individual institutional needs and opportunities, but also national scholarly cultures and circumstances.

In Germany, for instance, the national branch of the European project DARIAH (Digital Research Infrastructures for the Arts and Humanities) is addressing current teaching issues through an analysis of DH related teaching in traditional degrees, having created a brochure on existing German DH programs (Sahle 2011), working toward a national reference curriculum, and bringing this discussion to the international level.

This session continues the dialogue begun at DH 2010 and DH 2011, as well as in recent publications (see references), around Digital Humanities education by bringing together Digital Humanities educators who have been pivotal in originating, designing, or teaching degree courses in the field. The panelists will bring a wide international perspective to the discussion in order to move beyond individual academic landscapes.

This panel will also take an historical approach. One should not forget that over a decade ago it seemed that digital humanities programs (or humanities computing as it was then known) would continue to flourish. In the status report on 'Computing in Humanities Education: A European Perspective', published in 1999, twenty-five degree programs of various brands of Digital Humanities were presented and discussed. Nine have survived. Of these, five represent various brands of Computational Linguistics. In two workshops aimed at an international curriculum for 'History and Computing' in the early nineties, fifteen study programs at European universities were discussed. Of these, one has survived as is, a second survived with a changed focus. Of at least six Italian degree courses created in the nineties, only one has survived. It is also unfortunate, that only three institutions in the UK are continuing their Digital Humanities degrees.

To ensure that new degree courses succeed, we should analyze why programs and courses did not survive. Some reasons for past failures are:

- Trying to start a degree course with insufficient resources;
- Starting courses that are dependent on one person;
- Focusing a degree only on a small specific branch, thus limiting the number of participants;
- Unclear profiles, making it difficult for students to see which opportunities a specific degree offers;
- Computer science orientation with a high level of mathematical background frightening away potential Humanities students;
- Very shallow requirements, giving courses a poor reputation;

With these potential pitfalls in mind, panelists will address the following broad questions, (particularly at the national level) about the current state DH education at university level:

1. Are there degree courses which are explicitly labeled 'Digital Humanities' (or with an equivalent label, like 'Humanities Computing' or 'eHumanities') in your country?

2. Are there degree courses which are not explicitly labeled Digital Humanities but are dedicated to some form of interdisciplinary study between a Humanities subject and the IT tools needed for this field (e.g. degree courses dedicated to IT methods in philology, linguistics, history, cultural studies, archeology, art history etc.)

3. What degree courses exist which are flavors of library and information science degrees, directed explicitly at Humanities' graduates or including a significant amount of Digital Humanities content?

4. Are there any degree courses offered at Computer Science faculties that are targeting humanities students with DH-flavored degrees?

5. Are there programs or courses at the intersection of DH and other emerging fields of study? Are 'digital preservation' or 'game studies' real DH programs or DH inflected?

6. On which levels do such course programs exist? (BA, MA, PhD, vocational add-on qualifications after a first degree, etc.)

7. Are such programs well established or recently created? How big is the student demand? How visible are these courses outside the DH community?

8. How are they organized? By a DH Department, a DH Center in collaboration with a more traditional department, or as part of the teaching offered by departments that support traditional academic fields?

9. Are there required / optional courses on Digital Humanities embedded in other Humanities degree programs (beyond "computer literacy" courses)?

10. What content is taught in these classes or degree programs? Is there some consensus between different institutions within a national context on the content of a 'DH degree'? Is there possibly a

consensus on such content within a specific subfield – i.e. 'Computing in Archeology' – even if there is no generalized consensus on DH-curricula as such

11. What can or should be done to arrive at a core DH curriculum with a clearly defined bundle of knowledge and skills?

12. Do differences in culture and language make a case for 'national' or 'local' DH, or is DH a 'universal' field in terms of methodologies? Can DH tools, research and curricula be culturally neutral, regardless of where they are designed and produced? How can we best address the problem of cultural diversity within DH?

13. Can a consistent understanding of Digital Humanities as a field be traced in the courses which claim to teach it?

Perspectives for the Digital Humanities look particularly bright at the moment. We should see this as a chance to lay the groundwork for more stable curricula that are comparable at least within, preferably across, different national landscapes in academia. As digital humanities pedagogy continues to flourish, there is a clear desire in the community to compile resources and compare experiences – this panel will be of broad interest to conference participants.

References

Clement, T., et al. (2010). Digital Literacy for the Dumbest Generation – Digital Humanities Programs 2010. *Paper presented at the DH 2010*. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-815.html . [references there have been skipped here] (all online resources in this bibliography accessed 23 March 2012).

de Smedt, K., et al. (1999). *Computing in Humanities Education: A European Perspective*. Bergen http://www.hd.uib.no/AcoHum/book/ .

Drucker, J., et al. (2002). *Final Report for Digital Humanities Curriculum Seminar*, University of Virginia http://jefferson.village.virginia.edu/hcs/dhcs/ .

Fiormonte, D. (2010). The International Scenario of Digital Humanities. In T. Numerico et al. (eds.), *L'umanista digitale*. Bologna: Il Mulino, pp. 201-210.

Gouglas, S., et al. (2006) Coding Theory: Balancing Technical and Theoretical Requirements in a Graduate-Level Humanities Computing Programme. *Mind Technologies: Humanities Computing and the Canadian Academic Community*, pp. 245-256.

Gouglas, S. et al. (2010) *Computer Games and Canada's Digital Economy: The Role of Universities in Promoting Innovation* http://ra.tapor.ualberta.ca/~circa/?page_id=307 .

Hanlon, C. (2005). History on the Cheap: Using the Online Archive to Make Historicists out of Undergrads. *Pedagogy* 5(1): 97-101 http://muse.jhu.edu/journals/pedagogy/v005/5.1hanlon.pdf .

Brett, D. H., ed. (2012, forthcoming). *Digital Humanities Pedagogy: Practices, Principles, Politics*. Ann Arbor: U of Michigan P.

Kirschenbaum, M. (2010). What is Digital Humanities, and What's it Doing in English Departments? *ADE Bulletin* 150 http://bit.ly/hQQ1vI

Norcia, M. (2008). Out of the Ivory Tower Endlessly Rocking: Collaborating across Disciplines and Professions to Promote Student Learning in the Digital Archive. *Pedagogy* 8(1): 91-114 http://muse.jhu.edu/journals/pedagogy/v008/8.1norcia.pdf .

Orlandi, T. (2007). *Ultimo bilancio dell'informatica umanistica* http://rmcisadu.let.uniroma1.it/~orlandi/pubbli/informatica/montevarchi.pdf .

Parodi, M. (2009). Oltre le due Culture. *Informatica Umanistica* http://www.ledonline.it/informatica-umanistica/ .

Ragone, G., et al. (2011). *Lo statuto dell'Informatica umanistica*. Session at the Conference "Dall'Informatica umanistica alle culture digitali", Rome 2011 http://digilab.uniroma1.it/news/news.aspx?IDnews=82 .

Rockwell, G. (1999). *Is humanities computing an academic discipline?* http://www.iath.virginia.edu/hcs/rockwell.html .

Rockwell, G., ed. (2009). *The Academic Capacity of the Digital Humanities in Canada* (http://tapor.ualberta.ca/taporwiki/index.php/The_Academic_Capacity_of_the_Digital_Humanities_in_Canada

Sahle, P. (2011). *Digitale Geisteswissenschaften*. Cologne. [Printed catalog on study programs in Germany] (online version: http://www.cceh.uni-koeln.de/Dokumente/BroschuereWeb.pdf .)

Sinclair, S., and S. W. Gouglas (2002). Theory into Practice: A Case Study of the Humanities Computing Master of Arts Programme at the University of Alberta. *Arts and Humanities in Higher Education* 1(2): 167-183 http://ahh.sagepub.com/content/1/2/167 .

Spiro, L. (2011). *Knowing and Doing: Understanding the Digital Humanities Curriculum*. Paper presented at the DH 2011. See blogpost for further information: http://digitalscholarship.wordpress.com/2011/06/20/making-sense-of-134-dh-syllabi-dh-2011-presentation/

Svensson, P. (2010). The Landscape of Digital Humanities. *DHQ: Digital Humanities Quarterly* 4(1) digitalhumanities.org/dhq/vol/4/1/000080/000080.html (digitalhumanities.org/dhq/vol/4/1/000080/000080.html).

Papers

Exploring Originality in User-Generated Content with Network and Image Analysis Tools

Akdag Salah, Alkim Almila
alelma@ucla.edu
University of Amsterdam, New Media Studies Department, The Netherlands

Salah, Albert Ali
asalah@boun.edu.tr
Bogazici University (BU), Computer Science Department, Turkey

Douglass, Jeremy
jeremydouglass@gmail.com
University of California, San Diego (UCSD), Visual Arts Department & Software Studies Lab, USA

Manovich, Lev
manovich@ucsd.edu
University of California, San Diego (UCSD), Visual Arts Department & Software Studies Lab, USA

In this paper, we visit the question of originality in the context of creative influence across social networks, using with the tools of network analysis and digital image analysis. Our data sets come from a specialized online social network site (oSNS) called deviantArt, which is dedicated to sharing user-generated artworks. Currently, deviantArt (dA) is the 13th largest social network in the U.S. with 3.8 million weekly visits. Operating since 2000, dA is the most well known site for user-generated art; has more than 15 million members, and hosts a vast image archive of more than 150 million images. The social network infrastructure allows us to gather all sorts of rich data from this platform: which artist is following which other artist, what comments are posted under each artwork and by whom, the temporal unfolding of the connection structure, demographic information about artists, category and user generated tag information for artworks, number of hits per artwork and per artist, etc.

Our main research question: how can we define 'originality' in computable terms when analyzing such a massive dataset of visual artworks? Apart from its obvious relevance, we choose dA over other online platforms of art, because 1) unlike (online) museums, dA contains many derivative works, and at the same time provides sufficient information to trace these to their origins; 2) it has a category structure that defines stylistic boundaries; 3) it encompasses many artistic styles, some of which coming with cultural and artistic peculiarities that create unique analysis opportunities (e.g. 'stock' images, see below).

For this specific case study, we perused the vast archive of dA and focus on a subset of content, created specifically for being highly volatile and usable. These are the so-called 'stock' images; images that are freely available for other members to use. We extracted a set of stock images, linked to a second set of art works created through the use of these stock images, as well as a temporal cross-section of the dA network that encapsulates the social interactions of the producers and consumers of these stock images. Using this dataset, it becomes possible to trace the transformation of 'visual' ideas, both in terms of similarities in the image space, and in terms of distances in the social interaction space.

1. Construction the Social Interaction Space

A social network is a graph representation of social relations. Graphs are the most popular and well-researched data structure for representing and processing relational data. In a graph, each node represents one entity (a person in a social network; a researcher or a work in a citation network, etc.) and the edges (or arcs, if they are directed) of the graph represent some relation. One can also indicate the strength of the relation by associating weights with the edges of the graph. Network graphs are very useful for studying social networks.[1]

The dA network is very rich, not only because it consists of 15 million nodes and billions of edges, but because it represents a multigraph of relations: Different social interaction patterns superpose different arcs on the node structure, and these are amenable to joint analysis, as well as individual inspection. Furthermore, each node (i.e. artist) is complemented with demographic information, and with site statistics showing the popularity of each member and each individual image. Obviously, the 're-creation' of specialized networks from this immense set is crucial in clarifying and then interpreting it via network analysis and visualization tools.

As a preliminary research for this study, we have crawled the dA site to extract a subset of 'professional' members. Professional members are the paying subscribers, and thus contain the highest concentration of users that make a living out of producing artworks, as well as being more regular in their usage of the site in general. In order to obtain a vibrant core of the dA network, we have used a number of assumptions about these members. This helped us reach a manageable and relevant set of

users. The first heuristic is the subscription status; the paying members of the site are more serious users and have access to more services. These can be automatically determined through scraping. Our first data reduction followed these members, and we thus obtained a network with 103.663 vertices and about 4.5 million arcs, the latter representing a user being 'watched' by another user (average degree is 43.25). (Buter et al. 2011) These data are further scrutinized with a k-Core algorithm that helped us to identify the most important members of the network. In this work we use this core-network as our starting point. Thus out of 103.663 vertices, i.e. members, we extracted a core of 3402 members. Among these, we focus on those who publish stock images, which constitute 642 members with 13762 images. To further analyze how these stock images are reused by other members, we employ digital image analysis.

2. Image Tool Analysis

Image analysis is a vibrant research area at the intersection of signal processing, computer vision, pattern recognition and machine learning. From medical image processing to image retrieval and biometrics, image analysis techniques are used in a variety of applications. Among these applications, the Digital Humanities community would find the scattered work done around automatic artwork analysis techniques interesting (Hurtut, 2010). These are used in a number of applications such as virtual restoration, image retrieval, studies on artistic praxis, and authentication. The scientific community also shows an interest in the ever expanding professional image archives from museums, galleries, professional artist networks, etc. (Chen 2001, 2005, 2007).

One of the authors, Lev Manovich has founded *Software Studies Initiative* at UCSD in 2007 to apply image analysis techniques for the analysis of digital content that were not necessarily based on institutionalized archives (Manovich 2008). For example, in *How to Analyze Million Manga Pages*, Manovich, Douglass and Huber showed how image analysis techniques could be applied to answer humanities research questions, especially the ones that have to deal with the fast growing user-generated image sets, and their cultural dimensions (Manovich et al. 2011).[2] In our dA project, we make use of these tools in order to first extract 'features' (quantified descriptions of visual properties) from the images that we have located as stock image collections of core deviantArt members. As a second set, we download and analyze the images of members who indicate that they have used these images via the commenting tool of deviantArt. In Fig. 1, one can see an example of how a stock-image and the images that make use of this stock image look. The research questions in this study are 1) to determine in what (measurable) terms to define 'originality,' (i.e., how we can rate and compare different ways in which dA members use a stock image); 2) to determine if we can scale this definition for larger image sets.

Figure 1: An example of a stock image (right, upper corner), and images that use this stock.

Today it is possible to submit image queries to commercial search engines, where the subsequent image retrieval is not only based on the metadata, but also on the content of the images (for examples, see Li et al. 2007a, 2007b; Wang et al. 2008). Here, the 'semantic' component is retrieved from the images themselves: first, some low quality features of images are extracted; these features are used for tagging the images according to a pre-defined image vocabulary. A different approach to image retrieval is suggested by (Hertzmann et al. 2001), where parts of images themselves are used to locate images similar to those parts. Hertzmann terms this practice as finding 'image analogies'. Our data set is based on such 'image analogies', but for us, it is more important to find the most 'original' works – i.e., images that uses the stock image creatively, changing to arrive at different content and visual style than the original version.

3. Conclusion

deviantArt is listed among the top ten most visited websites in the category of art. With 32 million unique visitors, dA has both the world's biggest artist community and the largest active art audience. To thoroughly understand dA's cultural and artistic value, one has to use computational methods because of its massive scale. In this study, we use subsets of dA images (stock images and images which use them) to explore how computational analysis can be used to define 'originality' in user-generated art

in computable terms. To select and then analyze relevant subsets of the data, we designed a work flow which uses network and image analysis methods together. To the best of authors' knowledge, this is the first project that combines network and image analysis from a humanities point of view. We submit our research as an example that might inspire and guide other researchers who have to deal with huge archives of user generated content.

References

Buter, B., N. Dijkshoorn, D. Modolo, Q. Nguyen, N. van Noort, B. van de Poel, A. A. Akdag Salah, and A. A. Salah (2011). Explorative visualization and analysis of a social network for arts: The case of deviantArt. *Journal of Convergence* 2(2): 87-94.

Chen, C., H. Wactlar, J. Wang, and K. Kiernan (2005). Digital imagery for significant cultural and historical materials. *Int. Journal on Digital Libraries* 5(4): 275-286.

Chen, H. (2001). An analysis of image retrieval tasks in the field of art history. *Information Processing and Management* 37(5): 701-720.

Chen, H. (2007). A socio-technical perspective of museum practitioners' image-using behaviors. *The Electronic Library* 25(1): 18-35.

Hertzmann, A., C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin (2001). Image analogies. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 327-340.

Hurtut, T. (2010). 2D Artistic images analysis, a content-based survey. `http://hal.archives-ouvertes.fr/docs/00/45/94/01/PDF/survey.pdf` (accessed 10 January 2011).

Manovich, L. (2008). Cultural Analytics: Analysis and Visualization of Large Cultural Data Sets. `http://www.manovich.net/cultural_analytics.pdf` (accessed 10 January 2011).

Manovich, L., J. Douglass, and W. Huber (2011). Understanding scanlation: how to read one million fan-translated manga pages. *Image & Narrative* 12(1): 206-228. `http://www.imageandnarrative.be/index.php/imagenarrative/article/viewFile/133/104` (accessed 10 January 2011).

Li, Q., S. Luo, and Z. Shi (2007). Semantics-based art image retrieval using linguistic variable. *Fuzzy Systems and Knowledge Discovery* 2: 406-410.

Li, Q., S. Luo, and Z. Shi (2007). Linguistic expression based image description framework and its application to image retrieval. *Soft Computing in Image Processing*, pp. 97-120.

Wang, W., and Q. He (2008). A survey on emotional semantic image retrieval. *ICIP 15th IEEE International Conference on Image Processing*, pp. 117-120.

Notes

1. In our work we use *Pajek* for network analysis (`http://vlado.fmf.uni-lj.si/pub/networks/pajek`). We also make use of *R* for building networks (`http://cran.r-project.org/`), *Sci2* for calculating basic network measurements (`http://https://sci2.cns.iu.edu/user/index.php`), as well as *Gephi* for visualizing the networks (`http://www.gephi.org`).
2. Software Studies Initiative created *Image-Plot*, the first set of image analysis and image visualization tools for humanities researches, `http://lab.softwarestudies.com/p/imageplot.html`

Patchworks and Field-Boundaries: Visualizing the History of English

Alexander, Marc
marc.alexander@glasgow.ac.uk
University of Glasgow, UK

1. Introduction

This paper uses the database of the *Historical Thesaurus of English* (Kay et al. 2009; hereafter abbreviated to *HT*) to visualize change in the history of English, and in particular in the English lexicon. The *HT*, published in 2009, is the world's largest thesaurus and the most complete thesaurus of English, arranging into hierarchical semantic categories all the recorded meanings expressed in the language from Anglo-Saxon times to the present. The underlying *HT* database (see Kay & Chase 1987; Wotherspoon 1992), held at the University of Glasgow, is a massive computational resource for analyzing the recorded words of English with regards to both their meaning and dates of use.

The present paper describes a methodology of visualizing English using data from the *HT*, developed from an earlier pilot project (Alexander 2010). By combining visualization techniques with the high-quality humanities data provided in the *HT*, it is possible to give scholars a long-range view of change in the history, culture and experiences of the English-speaking peoples as seen through their language.

2. The Data

The data stored within the *HT* is a fine-grained conceptual hierarchy containing almost all of the recorded words in English, arranged semantically. Each category of words is nested within other, wider categories, so that, for example, the verb category *Live dissolutely* is within *Licentiousness*, itself adjacent to *Guilt* and *Rascalry* and within the wider category *Morality*. This hierarchical structure differs from the organization of many other thesauri; *HT* categories relate to others not just linearly, but can operate either horizontally (on the same hierarchical level) or vertically (on a higher or lower level, either containing or being contained by another category). In addition, each concept is able to contain a series of subcategories within itself, separate from the main sequence. It is this complex hierarchical structure which helps make the *HT* database so useful for visualization: each individual point in the hierarchy can contain both word entries for the concept represented by that point, and also all the conceptual descendants which follow it, each surrounded by siblings of similar meaning.

The size of the *HT* also makes it amenable to computational analysis. The current version of the database (as of mid 2011) contains 793,747 entries, compared to *OED2*'s 616,500 (Algeo 1990: 137), all within 236,346 categories, each representing a distinct concept. Taking into account each field stored within it, the database itself contains approximately 22.7 million pieces of data.

3. Visualization

The normal metaphor for visualizing hierarchy is a tree-like system, like that often used in organisation charts. *HT* data is, however, far too large to be used in such a way – even a spider-like tree or hypertree could not represent the thesaurus, whose largest category alone (the adverb *Immediately*) contains over 250 synonyms.

Instead, the present paper will describe an alternative way of displaying the *HT* hierarchy, representing each category as a nested rectangle on a plane. This technique produces a 'treemap' (see Shneiderman 2009), wherein each entry in a hierarchy is represented by a rectangle which is large enough to contain smaller rectangles representing its descendants while simultaneously being itself small enough to nest within further rectangles representing its parent categories. In short, a treemap structure takes the organisational chart metaphor of *senior is up* and replaces it with *senior is big*.

Doing this for the full hierarchy of the *HT* lets us view the semantic structure of the English language as a whole. Figure 1 is therefore a treemap showing all of present-day English in the *HT*, with every word represented by a small dot of ink. Those black dots are present-day words which originated in Old English, and white dots represent those which entered the language much more recently. The map is arranged by semantic field, so words in close semantic proximity are also physically close to one another on the diagram. An interactive display of this visualization allows an analyst to see which areas correspond to which categories and lexemes.

4. Patchworks and Field-Boundaries

One effect which is visible from Figure 1 is that modern English has a 'patchwork' effect in the visualization. Areas such as *Physics* and *Chemistry* are quite light, as are parts of *Number*

(which includes *Mathematics*) and *Language*. This phenomenon makes it possible to visually view the areas where the English language has 'stretched' in order to accommodate new vocabulary. These light patches within the greater patchwork are areas of recent lexical innovation, made up of clusters of words first cited in the *OED* and recorded in the *HT* from recent years. Therefore, we would expect this patchwork effect in areas affected by rapid social, technological or academic growth, such as *Computing* (inside *Number*, adjacent to *Mathematics*), *Physics, Chemistry, Linguistics, Communication, Travel*, and so on. Conversely, darker and therefore older patches cover existence in *Time and Space, Creation, Causation, Faith, Emotion*, and the parts of *Number* which refer to *Arithmetic* or *Enumeration*.

This effect is pronounced in present-day English, but if other selections of the data are taken, it reduces somewhat. If the present-day data is thought of as a 'slice' of the *HT*, then other such slices can be taken between the Old English period and the present day. Other slices reveal that this patchwork effect, with its pronounced edges between new and old semantic fields, do not occur throughout the rest of the history of the language. Proceeding backwards from the present day through the Late Modern, Early Modern and Middle English periods, there is a visible reduction in the patchwork effect observed above, so that by Middle English almost no delineated rectangular patches of innovation can be found. The paper compares these slices and offers an interpretation of various areas of patchwork and patchwork-like effects across time, which, as a result of rapid growth, generally indicates fundamental and rapid progress in the context of the language.

5. Conclusion

In the ways outlined above, such visual displays of *HT* data can provide useful entry points to a large, complex lexicographical and lexicological dataset. Firstly, in a pedagogical sense, these displays can give students and the public a new way of looking at humanities data, and of exploring them interactively. Secondly, as computer displays and online dictionary interfaces become more polished, new ways of encouraging exploration of lexical data online are needed to replace the lost experience of browsing a printed dictionary, rather than only providing users with a blank search interface. And finally, such visualizations can point analysts towards areas of possible semantic, lexical or cultural interest, whether areas of trauma in the history of the language, areas of rapid growth, or areas of relative stability. The paper will conclude by demonstrating some of these, including 'slices' of the data giving visualisations of various points in the history of English (eg in the time of Chaucer or Shakespeare), and in illustrating lacuna in our knowledge of earlier periods of the language, and will outline areas of future development of this approach.

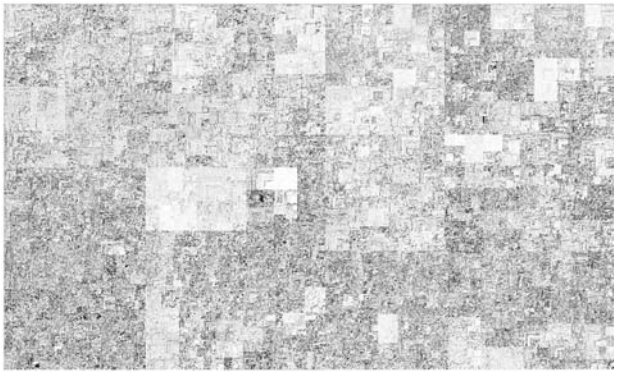


Figure 1: Present-day English from the HT, shaded by first cited date

References

Alexander, M. (2010). 'The Various Forms of Civilization Arranged in Chronological Strata': Manipulating the *Historical Thesaurus of the OED*. In M. Adams (ed.), *'Cunning passages, contrived corridors': Unexpected Essays in the History of Lexicography*. Monza: Polimetrica.

Algeo, J. (1990). The Emperor's New Clothes: The second edition of the society's dictionary. *Transactions of the Philological Society* 88(2): 131-150.

Kay, C., and T. J. P. Chase (1987). Constructing a Thesaurus Database. *Literary and Linguistic Computing* 2(3): 161–163.

Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon (2009). *Historical Thesaurus of the OED*. Oxford: Oxford UP.

Shneiderman, B. (2009). Treemaps for space-constrained visualization of hierarchies. `http://www.cs.umd.edu/hcil/treemap-history/index.shtml` (accessed 26 October 2011).

Simpson, J., and E. Weiner, eds. (1989). *The Oxford English Dictionary*, 2nd. ed. Oxford: Oxford UP.

Wotherspoon, I. (1992). Historical Thesaurus Database Using Ingres. *Literary and Linguistic Computing* 7(4): 218-225.

Developing Transcultural Competence in the Study of World Literatures: Golden Age Literature Glossary Online (GALGO)

Alonso Garcia, Nuria
nalonsog@providence.edu
Providence College, USA

Caplan, Alison
acaplan@providence.edu
Providence College, USA

The technological demands of today's global society call into question the use of traditional approaches to language and literary studies and challenge educators to think about ways to integrate multimedia creatively in their scholarship and teaching. The project presented in this paper and currently under construction offers an innovative approach to studying literature that employs tools from the field of the digital humanities.

The *Comprehensive Digital Glossary of Golden Age Spanish Literature* or *GALGO* (*GoldenAgeLiteratureGlossaryOnline*) is a searchable Spanish-English online glossary that consists of select words, from the most commonly studied literary texts of the 16th and 17th centuries, whose meanings and multiple connotations reflect important linguistic and cultural concepts. The selection of words is broadly consistent with the *keyword theory* developed by Raymond Williams in the mid-20th century. Keywords are lexical units that comprise a network of associations, triggering 'different memories and imaginings,' centered around dominant cultural notions and practices (Hart et al. 2005: 6). Keywords illustrate ways of perceiving the world and reflect the shared interest or ideology of a particular society and find expression in a common vocabulary (Williams 1983: 15). Analyzing keywords is an exercise in examining their usages in context and in considering the evolution of their meanings over time (Burgett & Hendler 2007: 1). Keywords research is rooted in historical semantics and focuses on how meaning is constructed and altered through conflict and negotiation among certain social groups and movements; it is about exploring the possibilities of language and unlocking meaning.

GALGO is a vocabulary of keywords that explores the canonical prose texts of Golden Age Spanish literature – *El abencerraje*, *Lazarillo de Tormes*, Miguel de Cervantes' twelve *Novelas ejemplares*, and principal *comedias* by Calderón de la Barca, Cervantes, Lope de Vega and Tirso de Molina – from the perspective of a history of ideas. The digital glossary tags socially charged words that reverberate through the texts and beyond. The glossary stimulates students to evaluate critically their own assumptions and reflect on philosophical and ethical questions that continue to be of current importance. The polysemic nature of the terms selected, words such as *esfuerzo*, *fuerza*, *gentileza*, and *honra*, offers valuable insight into the complex, idiosyncratic Golden Age Spanish lexicon. We believe that identifying the different uses of the same word in a literary work is both essential to reading comprehension and to an understanding of societal attitudes of the period.

GALGO is organized in such a way that the specific contextualized definition is given in each instance where the keyword is tagged. The literary works have been digitized and are available in the glossary in complete electronic versions. The box containing the definition can remain open and is moveable, and users are able to compare and contrast immediately the word's varied meanings in the text. Furthermore, with the support of a search algorithm, the glossary's multilayered design gives students the option to view the same keyword in all of the texts in the database. GALGO also contains a Javascript snippet that will highlight and group together each distinct definition of a search term across texts, thereby allowing students to chart the prevalence of a particular usage of the word.

The methodology of *GALGO* combines the rich word inventory of literary concordances with the ideological interpretations of medieval and early modern manuscript glosses. The entries reveal a word's different shades of meaning and, borrowing from the exegetical tradition popularized in the Middle Ages, function as brief explanatory notes that give broader significance to literary passages. For example, in the novella *El abencerraje*, the noun *gentileza* and its adjective form *gentil* are defined in the following ways, depending on the context:

gentileza:

Unrequired goodwill shown by the more powerful to the less powerful (magnanimity).

gentil:

1. Of cultivated manner and appearance.

2. The best of its kind (exemplary).

When necessary, an additional feature called *Expansion of meaning* appears alongside the definition in order to identify further semantic

connections. The entry for *honra* in the line '*...tuyo es mi corazón, tuya es mi vida, mi honra y mi hacienda*' from *El abencerraje* reads:

Definition: Public esteem received in acknowledgement of one's strength of character, specifically here, a woman's chastity and/or marital fidelity.

Expansion of meaning: In this period, honor (good reputation) was based on the following factors: social status, material wealth, strength of character, achievements, and in the case of a woman, her chastity and/or marital fidelity. Here, in pledging *vida*, *honra*, and *hacienda*, the noblewoman was committing everything she was and had to the man.

By integrating qualitative analyses of literary texts with quantitative computer derived methods, *GALGO* promotes a novel form of active reading and engagement in language and culture. In the last two decades, the pedagogical approach to foreign language study has moved beyond a primary emphasis on developing functional communication skills to an equally compelling interest in acquainting students with representative literary and cultural texts in the target language. *The Modern Language Association*'s 2007 report places the acquisition of transcultural and translingual competencies, namely the ability to reflect on how realities are shaped differently across languages, historical periods and cultures, at the core of the foreign languages discipline. *GALGO* engages learners in the analysis of classic literary texts in a meaningful way, allowing them to test their hypotheses as to possible lexical and cultural preconceptions through immediate feedback.

References

Burgett, B., and G. Hendler (2007). *Keywords for American Cultural Studies*. New York: New York UP.

Hart, D. P., S. E. Jarvis, W. P. Jennings, and D. Smith-Howell (2005). *Political Keywords. Using Language that Uses Us*. New York: Oxford UP.

MLA Ad Hoc Committee on Foreign Languages (2007). *Foreign Languages and Higher Education: New Structures for a Changed World*. New York.

Williams, R. (1983). *Keywords: A Vocabulary of Culture and Society*. 2nded. New York: Oxford UP.

Trees of Texts – Models and methods for an updated theory of medieval text stemmatology

Andrews, Tara Lee
tara.andrews@arts.kuleuven.be
Katholieke Universiteit Leuven, Belgium

Macé, Caroline
caroline.mace@arts.kuleuven.be
Katholieke Universiteit Leuven, Belgium

The construction of a stemma codicum from the manuscript witnesses to a given text remains a somewhat divisive issue in philology. To a classical philologist, it is a necessary first step toward creating a critical edition of the text;[1] to a scholar who adheres to the principles of new philology, the assumptions inherent in the construction of a stemma are so fundamentally unsound that the exercise is pointless.[2]

The advent of digital philology demands a fresh look at the practice of stemmatology; indeed there has been a great deal of work in the past two decades on the subject, and in particular on the applicability of similar work from the field of evolutionary biology on the reconstruction of phylogenetic relationships between species.Among the numerous works on this subject are: [3] More recent work has begun to overcome the limitations of these techniques when applied to manuscript texts, such as the biological assumption that no extant species is descended from another, or the assumption that any ancestor species has precisely two descendants, or the converse assumption that any descendant species has precisely one ancestor.[4]

Despite this considerable methodological progress, the fundamental assumptions at the core of stemmatology have yet to be formalised, or even seriously questioned. The purpose of this paper is to revisit these assumptions and simplifications; to propose a model for the representation of texts in all their variations that is tractable to computational analysis; to set forth a method for the evaluation of stemma hypotheses with reference to text variants, and vice versa; and to discuss the means by which we can provide the field of stemmatology with a formal underpinning appropriate for the abilities that modern computational methods now provide us to handle vast quantities of data.

1. Modelling a text tradition

The first step to computational analysis of any text tradition is to model it programmatically. This act of object modelling is distinct from, and orthogonal to, text encoding – the component texts may reach us in a variety of forms, including TEI XML and its critical apparatus module, CSV-format alignment tables, a base text and a list of variants exported from a word processor, or even individual transcriptions that remain to be collated. All of these must be parsed into a uniform computational object model for further programmatic analysis.

Our model is relatively simple, taking into account only those features of the text that are computationally useful for our purposes. We define a tradition as the union of all the versions of a text that exist, represented by its witnesses. A collation is the heart of the tradition, representing in a unified way the entire text with all its known variations. We have adapted the variant graph model proposed by Schmidt and Colomb[5] for this purpose. Whereas the Schmidt/Colomb model is a single graph, where readings are represented along edges, ours is a pair of graphs, in which readings are represented by vertices. The set of vertices is identical in both graphs; in the 'sequence' graph, the witnesses themselves are represented as a set of paths. Each witness path is directed and acyclic, taking a route through the reading vertices from beginning to end, according to the order of its text. The witness path may branch, if for example the text contains corrections by the original scribe(s) or by later hands; the branches are labeled appropriately. If transpositions are present, the graph as a whole is not directed and acyclic (DAG); we have therefore found it most useful to represent transposed readings as separate vertices, linked through a relationship as described below. A DAG may then be trivially transformed into an alignment table, with transpositions handled either as separate rows in the table or as variant characters within the same row.

Figure 1: A portion of a variant graph, with directed witness paths and undirected relationships

The second graph concerns relationships between readings, or types of variation that occur within the text. We may specify that two different readings are orthographic or syntactic variants of the same root word, or that they have some semantic relationship to each other. We may also specify transpositions in this way. The result is an unconnected and undirected graph that adds extra information to the sequence graph without compromising its properties. The relationship information allows association and regularization of reading data, producing thereby more exact versions of alignment tables for our statistical procedures.

2. Modelling a stemma hypothesis

The goal of our research is to arrive at an empirical model of medieval text transmission, and set forth a formalized means of deducing copying relationships between texts. As such, we have two tasks before us. We must be able to evaluate a stemma hypothesis according to the variation present in a text tradition, and we must be able to evaluate variants in a tradition according to a given stemma hypothesis.

The problem becomes computationally tractable once we realize that a stemma hypothesis is also a DAG. No matter the complexity of the stemma, no matter the degree of 'horizontal transmission', a stemma can be represented as a connected and rooted DAG, where each extant or hypothesized manuscript is a vertex, and each edge describes a source relationship from one manuscript to another. (In our model, any stemma has a single archetype; where multiple archetypes exist in a text tradition, we must either consider a common theoretical origin for these archetypes, or speak of multiple stemmata.)

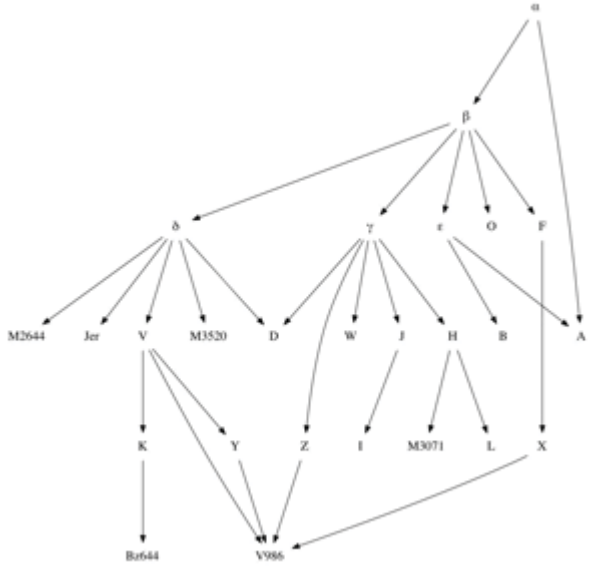


Figure 2: An arbitrarily complex stemma, represented as a DAG

In cooperation with the Knowledge Representation and Reasoning group within the KU Leuven Computer Science faculty,[6] we have developed a method for evaluating sets of variants against a stemma hypothesis of arbitrary complexity, testing

whether any particular set of variants aligns with the stemma hypothesis, and measuring the relative stability of particular readings (defined in terms of how often they are copied vs. how often they are altered). In conjunction with the variant graphs of our text model, we are also able programmatically to draw some empirical conclusions about the relationships between readings as they are altered within the text.

3. An empirical approach to text variation

The idea of distinguishing 'text-genealogical' variants from the coincidental sort is not new;[7] with the exception of U. Schmid (2004), however, none of the approaches taken have been empirical, in accordance with text relationships known from outside the variant tables. Given the means to quickly evaluate a stemma in detail according to the variants present in the text, distinguishing those variants that follow the stemma from those that do not, we should now easily be able to arrive at an empirical classification of genealogically 'significant' and 'insignificant' variants. This threatens to create a problem of circular logic: how to retrieve the genealogical variants in the real world? We build hypotheses on the basis of variation; we accept or reject variants as 'genealogical' based on the hypothesis. How do we break the cycle?[8]

We may adopt two approaches to anchor our methods in external results. First we test our methods on artificial traditions that have been created for the testing of stemmatological methods. These traditions are convenient in that the true stemmata are known, but they cannot reflect the true conditions under which medieval texts were copied. Thus we must also use 'real' text traditions, whose true stemmata are not fully known; in many medieval traditions, however, some paratextual information (e.g. colophon notations, physical damage or other physical characteristics of the texts) survives and can be used to establish or verify copying relationships. In our presentation we will discuss the results obtained on three artificial traditions, and three real medieval traditions with credible stemma hypotheses derived from external characteristics.

4. Conclusion

When examining a medieval tradition, we can rarely be certain what the scribe has copied in a text, and what has changed coincidentally or independently. To date, stemmatic reasoning has relied on the question of whether the scholar finds it likely that a given variant was preserved in copies. Given the models and methods for analysis proposed herein, we have an opportunity to remove this limitation of 'assumed likelihood' and take all the evidence of a text into account, to build a statistical model independent of the constraints of evolutionary biology, using statistical probability rather than scholarly instinct alone.

References

Blockeel, H., B. Bogaerts, M. Bruynooghe, B. de Cat, S. de Pooter, M. Denecker, A. Labarre, and J. Ramon (2012). Modeling Machine Learning Problems with FO(·). *Submitted to the 28th International Conference on Logic Programming*.

Cameron, H. D. (1987). The upside-down cladogram. Problems in manuscript affiliation. In H. M. Hoenigswald and L. F. Wiener (eds.), *Biological metaphor and cladistic classification : an interdisciplinary perspective*. Philadelphia: U of Pennsylvania P.

Cerquiglini, B. (1989). *Éloge de la variante : histoire critique de la philologie,* Paris: Éd. du Seuil.

Howe, C. J., A. C. Barbrook, M. Spencer, P. Robinson, B. Bordalejo, and L. Mooney (2001). Manuscript Evolution. *Trends in Genetics* 17**:** 147-52.

Maas, P. (1957). *Textkritik*. Leipzig: Teubner.

Macé, C., and P. Baret (2006). Why Phylogenetic Methods Work: The Theory of Evolution and Textual Criticism. In C. Macé, P. Baret, A. Bozzi, and L. Cignoni (eds.), *The Evolution of Texts: Confronting Stemmatological and Genetical Methods*. Pisa, Rome: Istituti Editoriali e Poligrafici Internazionali.

Pasquali, G. (1962). *Storia della tradizione e critica del testo*. Firenze: F. Le Monnier.

Reeve, M. D. (1986). Stemmatic method: 'Qualcosa che non funziona?' In P. F. Ganz (ed.), *The role of the book in medieval culture : proceedings of the Oxford International Symposium, 26 September-1 October 1982*. Turnhout: Brepols.

Roos, T., and Y. Zou (2011). Analysis of Textual Variation by Latent Tree Structures. *IEEE International Conference on Data Mining*, 2011 Vancouver.

Salemans, B. J. P. (2000). *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian, Way: The Case of Fourteen Text Versions of Lanseloet van Denemerken*. Ph.D., Katholieke Universiteit Nijmegen.

Schmid, U. (2004). Genealogy by chance! On the significance of accidental variation (parallelisms). In P. T. van Reenen, A. den Hollander, and M.

van Mulken (eds.), *Studies in Stemmatology II*. Amsterdam: Benjamins.

Schmidt, D., and R. Colomb (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies* 67: 497-514.

Spencer, M., and C. J. Howe (2001). Estimating distances between manuscripts based on copying errors. *Literary and Linguistic Computing* 16: 467-484.

P. T. van Reenen, A. den Hollander, and M. van Mulken, eds. (2004). *Studies in Stemmatology I.,* Amsterdam: Benjamins.

West, M. L. (1973). *Textual Criticism and Editorial Technique: Applicable to Greek and Latin Texts*. Stuttgart: Teubner.

Notes

1. Maas 1957; Pasquali 1962; West 1973.
2. See, among many references on this question Cerquiglini 1989; Reeve 1986.
3. Cameron 1987; Howe et al. 2001; Macé & Baret 2006; Spencer & Howe 2001.
4. Roos & Zou 2011.
5. Schmidt & Colomb 2009.
6. Blockeel et al. 2012.
7. Salemans 2000. See also papers by Wachtel, van Mulken, Schmid, Smelik, Schøsler, and Spencer et al in van Reenen et al. 2004.
8. See the critique of concerning the work of Salemans.

Mapping the Information Science Domain

Arazy, Ofer
ofer.arazy@ualberta.ca
University of Alberta, Canada

Ruecker, Stan
sruecker@ualberta.ca
Illinois Institute of Technology, USA

Rodriguez, Omar
omar.rodrigueza@gmail.com
University of Alberta, Canada

Giacometti, Alejandro
alejandro.giacometti@gmail.com
University College London, UK

Zhang, Lu
lu.eva.zhang@gmail.com
University of Alberta, Canada

Chun, Su
schun@ualberta.ca
University of Alberta, Canada

In the scholarly community, there has been for the last fifteen years an increasing interest in interdisciplinary projects, where researchers with different disciplinary backgrounds form teams to collaborate in addressing problems that are either not clearly within a single research domain, or that can best be approached by combining a variety of approaches. However, there also remain a number of research areas where interdisciplinary approaches seem to be warranted, but scholars in single disciplines have continued to work largely in isolation from their colleagues in other domains. We focus on one specific scientific field – information science – which is investigated in various disciplines: library and archival studies, computer science, management (and specifically, management information systems), and engineering, to name just a few.

Prior works in the area have discussed the disconnect between various areas within the broad field of information science, relying primarily on co-citation techniques and content analysis. These studies concluded that although all domains within this area are focused on the concepts of information, people, and (information and communication) technologies, the disciplines remain disjunct in terms of their disciplinary recognitions and topic of research. Despite some similarities in research methodology ('research approach' and 'level of analysis') (Glass et al. 2004), there are substantial differences in

terms of co-citation patterns (Ellis et al. 1999; Sawyer & Huang 2007), indicating the fields are distinct. Furthermore, although the research in these various areas may often appear to be similar on the surface, the contents of research are varied as well (Ellis et al. 1999). Glass et al. (2004) concluded that 'overall, we see that there was minimal topic overlap' among the various disciplines, and Sawyer and Huang (2007) summarize that these areas of scholarship 'have overlapping, but substantively different, research foci.'

Despite these interesting findings, extant research suffers from three primary limitations. First, previous studies are limited in scope, often focusing on just a subset of the academic units that publish articles in the domain. Second, prior research often relied on manual analysis, limiting the number of articles that are studied. Third, previous research in the area made little usage of information visualization techniques; recent advancements in the field of visualization have provided tools for exposing intricate relationships in data, and have the potential to reveal interesting patterns in interdisciplinary science.

Furthermore, although research on earlier data describes a disjoint domain, recent years have seen what appear to be the first signs of convergence in the various sub-fields within the broad information science discipline. We see more and more cross-departmental collaborations, journals with a much broader scope (e.g. *JASIST*), and there is a move across North America to amalgamate these various departments under one institutional roof (often referred to as the 'information school' or simply 'iSchool').

The objective of this paper is to address these previous limitations and map developments in the information science domain in the last fifteen years, using novel visualization techniques. We revisit the question of the divergence/convergence within the sub-disciplines, and propose a scalable method for analyzing (through both statistical analysis and advanced visualizations) the semantic and usage patterns in information science research articles. We focus our attention on three pre-defined topics that cross boundaries and are investigated in many sub-domains. For example the notion of 'Trust' was defined as 'designing tools that would help alleviate the problem of trust in e-commerce, e.g. rating and reputation mechanisms', and was represented using the following keywords: 'trust', 'online', 'web', 'e-commerce', 'electronic commerce', 'e-business', 'trustworthiness', 'rating', 'reputation', 'reputation mechanism', 'system', and 'information system'.

Our method involved the following steps: (a) defining each of the three topic areas – 'trust', 'text mining', and 'web services' - through a series of search keywords; (b) running these keywords on the Thompson Reuters Web of Knowledge (`http://wokinfo.com/` http://wokinfo.com/) corpus of research articles, repeating the query separately for the 15 years under investigation, and choosing the 500 highest ranked (in terms of matching the query) research articles at each time period; (c) analyzing the relationships between the 500 articles in each time period using both co-citation and content analysis; and (d) developing novel visualization to expose how the patterns of relationships develop over the fifteen years.

Traditionally, citation patterns have been visualized using a network diagram, where the articles are represented as nodes and the citations are shown as lines connecting the nodes (e.g. Shen et al. 2006). Others have developed variations on this form of visual representation (Nakazono et al. 2006; Brandes & Willhalm 2002; Ibekwe-SanJuan & SanJuan 2009).

Our approach builds on the Eigenfactor method (Bergstrom & West 2008), which applies an algorithm similar to Google's PageRank (Brin & Page 1998) to determine which nodes in a networks (in this case, journals) are most important, and how nodes connect with one another. The importance depends on where a journal resides in this mesh of citation links: the more citations a journal receives, especially from other well connected journals, the more important the journal is.

We model our visualization technique after the *well-formed eigenfactor* method (Stefaner 2009) that utilizes Eigenfactor, and was used in the past to track the evolution of the Neuroscience domain (`http://well-formed-data.net/archives/331/neuroscience-infoporn`). That analysis pre-determined the association of journals to disciplines, and then analyzed the relationships between disciplines as they evolve over time. We, however, sought a more organic approach, which recognizes that a journal could publish articles belonging to various areas (and even a specific article may belong to more than one area). Thus, we employed the classification of articles to categories provided by the Web of Knowledge.

Building on prior work in the field (Ellis et al. 1999; Glass et al. 2004), we analyze relationships both in terms of content and citation patterns. For each of the three topics, and for each of the time periods, we analyzed the 500 research articles to create similarity matrixes – in terms respectively of content and citation pattern – and then aggregated the data to calculate similarities between Web of Knowledge categories. These categories similarity matrixes were the input for our visualization method.

The results of our analysis are 3 pairs of visualizations, where each pair represents the evolution in the information science domain in terms of content and citation pattern for one topic. Figures 1a-c below depict the evolution in the 'trust' area, when using citation patterns as the method of analysis.

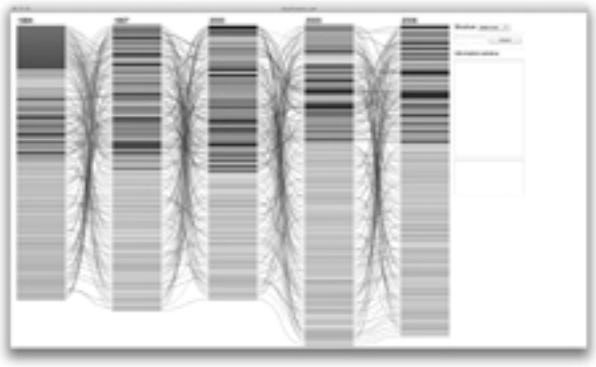


Figure 1a: the evolution in the 'Trust' topic; each line represents a Web of Science category; each cluster is made up of similar categories, corresponding to a specific scientific discipline

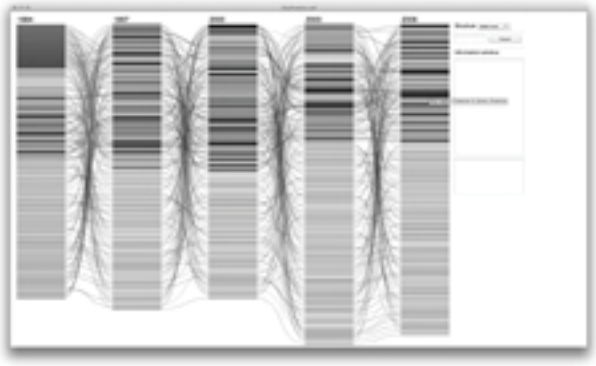


Figure 1b: the 'Information Science & Library Science' category, and how it evolved in terms of clustering with other categories

Figure 1c: the cluster that includes the 'Information Science & Library Science' category; on the right: the list of categories in this cluster; each colored line represents a category and its origins

Figure 2, below, represents the visualization for the same topic – 'Trust' – this time based on the content of the research articles.

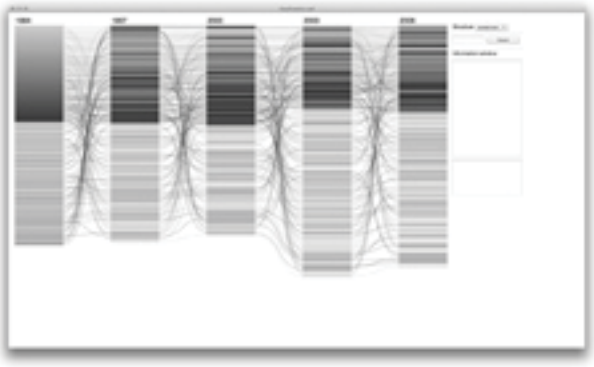


Figure 2: the evolution in the 'Trust' topic, when analysis is based on the contents of articles

There are two interesting findings from our analysis. First, we see substantial differences between co-citation patterns and the patterns based on content (i.e. comparing Figures 1a and 2a). While the topical analysis shows one primary cluster, the co-citation visualization depicts many small clusters. Thus, while there is a substantial convergence in terms of contents, the information science field remains disjoint in terms of citation patterns. Some of the primary clusters correspond to the domains of management information systems, computer engineering, medical informatics, health information systems, physics, law, and computer science. Second, there are differences between topics. While the 'Trust' topic seem to have emerged from one area, and then broken into smaller sub-areas over the years (see Figure 1a), the 'Text mining' topic remains relatively consistent in terms of its decomposition to clusters.

We conclude that our novel data analysis and visualization methods have revealed patterns that have never been described before, highlighting the differences between topics of investigation within the broad information science discipline. In our next iteration of the project will include (a) the analysis of a larger corpus of research articles, trying to capture this entire domain, (b) refined data analysis methods, and (c) the exploration of new visualization techniques, which would help us to discover additional hidden patterns.

Funding

This research was supported in part by SSHRC, the Social Sciences and Humanities Research Council of Canada.

References

Bergstrom, C. T., and J. D. West (2008). Assessing Citations with the Eigenfactor™ Metrics. *Neurology* 71: 1850-1851.

Brandes, U., and T. Willhalm (2002). Visualization of Bibliographic Networks with a Reshaped Landscape Metaphor. *Proceedings of the Symposium on Data Visualisation,* Barcelona, Spain, May 27-29, 2002.

Brin S., and L. Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30: 107-117.

Ellis, D., D. Allen, and T. Wilson (1999). Information Science and Information Systems: Conjunct Subjects Disjunct Disciplines. *Journal of the American Society for Information Science and Technology* 50(12): 1095-1107.

Glass, R. L., V. Ramesh, and I. Vessey (2004). An Analysis of Research in Computing Disciplines. *Communications of the ACM* 47(6): 89-94.

Ibekwe-SanJuan, F., and E. SanJuan (2009). The Landscape of Information Science: 1996-2008. *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries,* Austin, TX, USA, June 15-19, 2009).

Nakazono, N., K. Misue, and J. Tanaka (2006). NeL2: Network Drawing Tool for Handling Layered Structured Network Diagram. *Proceedings of the 2006 Asia-Pacific Symposium on information Visualisation, – Volume 60* (Tokyo, Japan). K. Misue, K. Sugiyama, and J. Tanaka (eds.). ACM International Conference Proceeding Series, vol. 164. Australian Computer Society, Darlinghurst, Australia, pp. 109-115.

Sawyer, S., and H. Huang (2007). Conceptualizing Information, Technology, and People: Comparing Information Science and Information Systems Literatures. *Journal of the American Society for Information Science and Technology* 58(10): 1436-1447.

Shen, Z., M. Ogawa, S. T. Teoh, and K. Ma (2006). BiblioViz: a System for Visualizing Bibliography Information. *Proceedings of the 2006 Asia-Pacific Symposium on information Visualisation - Volume 60* (Tokyo, Japan). K. Misue, K. Sugiyama, and J. Tanaka, Eds. ACM International Conference Proceeding Series, vol. 164. Australian Computer Society, Darlinghurst, Australia, 93-102.

Stefaner, M. (2009). Well-Formed Eignefactor, `http://well-formed-data.net/archives/192/well-formed-eigenfactor`

Words made Image. Towards a Language-Based Segmentation of Digitized Art Collections

Armaselu, Florentina
armaselu@ymail.com
ZoomImagine, Germany

1. Introduction

Within the framework of computer vision, the field of image segmentation is becoming increasingly important for a set of applications aiming to an understanding of the visual content, like object recognition and content-based image retrieval. The goal of the segmentation process is to identify contiguous regions of interest in a digitized image and to assign labels to the pixels corresponding to each identified area. While most of today's searching capabilities are based on metadata (keywords and caption) attached to the image as a whole, one can observe a tendency towards more meaning-oriented approaches intending to split the analyzed image into semantically annotated objects.

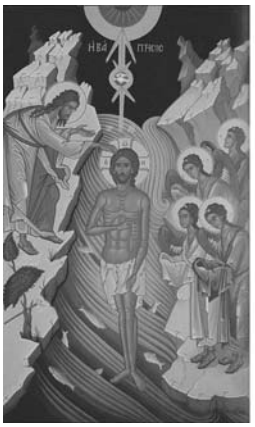
From this perspective, research efforts have been dedicated, on one hand, to the task of automated image annotation (Barnard et al. 2003; Johnson 2008; Leong & Mihalcea 2009; Leong et al. 2010), and on the other hand, to the construction of semi-automatically annotated datasets used for supervised image processing (MSRC-v2; Russel, Torralba et al. 2008; CBCL StreetScenes; PASCAL VOC 2007; Welinder et al. 2010). The visual data used for automated annotation may consist, for instance, in images extracted from the Web and the surrounding text in the html pages used as a source of candidate labels (Leong and Mihalcea; Leong et al.). The datasets are mainly derived from digital photographs with the goal to allow recognition of objects belonging to a number of classes like *person, animal, vehicle, indoor (Pascal VOC), pedestrians, cars, buildings, road* (CBCL StreetScenes), *birds species* (Caltech- UCSD Birds 200) or to categories interactively defined by users in a Web-based annotation environment (LabelMe).

Our proposal is related to the latter group of applications, i.e. the construction of an interactively annotated dataset of art images (to our knowledge a category less represented in the field), taking into account the textual descriptions of the images intended for analysis. The set of images for

segmentation will include digitized icons on themes described in the *Painter's Manual* by Dionysius of Fourna. The choice was not arbitrary, given the highly canonical nature of the Byzantine and Orthodox iconography, almost unchanged for centuries, and the close relationship between the text and the pictorial content of the associated icon. Another reason concerns the potential for subsequent applications of learning algorithms to the segmentation of new sets of icons depicting the same themes (and therefore relatively similar content), starting from the initial annotated training set. The choice of a largely circulated text like the *Painter's Manual*, translated in several languages, may also serve as a basis for multilingual annotation of the dataset.

2. The Experiment

The project consists in the use of two types of software, one intended to the segmentation of the visual content, the other to the processing of the textual fragments in order to provide the user with candidate labels for the annotation of the segments. The experiment described below makes use of two open source Java packages, *GemIdent* – a statistical image segmentation system, initially designed to identify cell phenotypes in microscopic images (Holmes et al. 2009), and the *Stanford parser*, a natural language statistical parser for English, which can be adapted to work with other languages such as German, French, Chinese and Arabic. Figure 1 presents the original image, *The Baptism of Christ* (Nes 2004: 66), and the corresponding text from the *Painter's Manual* book.



Christ standing naked in the midst of the Jordan, the Forerunner is on the bank of the river to the right of Christ, looking up, with his right hand resting on the head of Christ, and h is left hand upraised. Above is heaven, and out of it the Holy Spirit in a ray of light descends on to the head of Christ, and in the ray are these words: "This is my beloved Son in whom I am well pleased." To the left angels are standing devoutly with their wings outspread over their garments. Below the Forerunner in the Jordan is a naked old man lying bent up, who looks back at Christ in fear and holds an urn from which pours water; fish surround Christ.

Figure 1 a: 'The Baptism of Christ'. b: *The Baptism of Christ* (Dionysius of Fourna 1996: 33). Greek variant. Cretan school. Egg tempera on pine. 78.5 x 120 cm (1989) (Nes 2004: 66)

The image processing involved two phases of *colour* and *phenotype* (class or category of objects to be distinguished) training and a phase of segmentation (classification) (see Fig. 2). In the phase of colour training (a), the user can define a set of colours, according to their importance in the analyzed image, by selecting sample points on the image for each colour. We defined four colours: *Amber*, *Blue*, *Brown*, and *Yellow*. The phenotype training (b) implied the choice of representative points for every object in the picture that should be identified by the program in the segmentation phase. From the processed text with the Stanford parser, we used the following nouns as objects labels: *angel*, *bank*, *Christ*, *fish*, *Forerunner*, *Holy_Spirit*, *Jordan*, *ray*, and *word*. For each phenotype, a colour for the corresponding segment was also chosen (b, c). The program required as well the specification of a NON-phenotype, i.e. an area in the image which does not include any of the above defined objects, in our case the black background representing the sky in the picture.

Figure 2: *GemIdent*. a. Colour training; b. Phenotype training; c. Segmented image

One can observe that the result of the segmentation process (c) is not perfect. There are small regions belonging to a segment but included in another object: for instance, the light blue (*Forerunner*) and red (*Christ*) areas inside the angel objects. *GemIdent* allows error correction and re-segmentation in order to improve the segmentation process. Although more tests are needed, the results obtained so far seem to indicate a performance depending on the colour selection (tests with same theme pictures, using the trained colour set without colour training on the new pictures, showed less satisfactory results), and on the number and position of the sample points for the phenotype training (ranging from 20 points for small objects like *word* to 150 for larger objects like *bank*; selection of sample points from border areas adjacent to different objects seem also particularly important).

On the other hand, the experiment supposed the use of the *Type dependency Viewer* to visualize the grammatical relationships produced by the Stanford parser, according to the typed dependencies model (Marneffe & Manning 2011). Figure 3 presents the parsing results on the Dionysius of Fourna's text. The directed edges in the graphs show the

dependencies between a *governor* element (or *head*) and a *dependent*, both corresponding to words in the sentence. The edge labels express the grammatical relations between the two words. For example, *nsubj* refers to a *nominal subject* relation between a verb (*standing*, *surround*) and the syntactic subject of a clause (*Christ*, *fish*). Similarly, *pobj*, *dobj* connect a *preposition* (*in*, *of*, *to*) with its *object* - the head of a noun phrase following the preposition (*midst*, *Jordan*, *head*, *Christ*) – or, respectively, a *verb (surround)* and its *direct object (Christ)*. Other relations may refer to adverbial, prepositional or relative clause modifiers (*advmod, prep, rcmod*), determiners (*det*), coordination (*cc*), conjunct (*conj*), undetermined dependency (*dep*), etc.

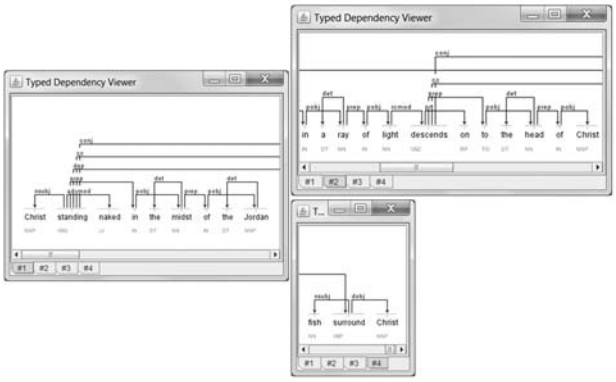


Figure 3: *Stanford parser* results displayed using the *Typed dependency viewer*. Excerpts. *The Baptism of Christ* (Dionysius of Fourna 1996: 33)

A parallel between the pictorial and textual representation of *The Baptism of Christ* may be further explored. The introductory sentence describing Christ, as a subject of the clause, standing in the midst of the Jordan, is actually pointing to the central figure of the painting. Prepositional or adverbial phrases (such as *to the right*, *to the left*, *above*, *below*) act as mediators positioning the other elements in the icon relatively to the central figure. Likewise, *Christ* occurrences in the text as direct or prepositional object (of verbs like *rest*, *descend*, *look back, surround*) are emphasising the 'focal point' projection towards which all the other agents' actions in the pictorial composition seem to converge.

3. Further development

Since the project is a work in progress, further development will be required:

- more tests, involving new sets of images and associated art-historical description or annotation in another language (if translated variants of the text and an adapted parser for that language are available); tests using similar tools, as the Fiji/ImageJ segmentation plugins, result analysis and comparison with other approaches like, for example, the computer-assisted detection of legal gestures in medieval manuscripts (Schlecht et al. 2011);
- extension of the segmentation process and encoding to include language-based information, i.e. not only objects labels (*nouns*) but also dependencies reflecting the objects relationships in the textual description (relative positioning, actions agents and objects, by means of *verbs*, adjectival, adverbial, prepositional *modifiers*), as they are captured by the pictorial content;
- implementation of a semi-automatic tool combining segmentation and parsing capabilities as in the presented experiment; the tool and images corpus will be available on line so that potential users (students, teachers, researchers in art history or image analysis) can reproduce the results or use them in their own applications.

The DH 2012 presentation will focus on the preliminary stage aspects of the project, e.g. general framework, goals and expectations, experiments and partial results, tools and data choices, model design, more than on an overall quantitative and qualitative results analysis.

References

Barnard, K., et al. (2003). Matching Words with Pictures. *Journal of Machine Learning Research* 3.

CBCL StreetScenes. `http://cbcl.mit.edu/software-datasets/streetscenes` (accessed 25 March 2012).

Typed Dependency Viewer. `http://tydevi.sourceforge.net/` (accessed 25 March 2012).

MSRC-v2 image database, 23 object classes. Microsoft Research Cambridge. `http://research.microsoft.com/en-us/projects/ObjectClassRecognition` (accessed 25 March 2012).

Dionysius of Fourna (1996). *The Painter's manual*. 1989, 1996, Torrance, CA : Oakwood Publications.

Fiji, `http://fiji.sc/wiki/index.php/Fiji` (accessed 25 March 2012).

Holmes, S., A. Kapelner, and P. P. Lee (2009). An Interactive Java Statistical Image Segmentation System: GemIdent. *Journal of Statistical Software* 30(10).

Johnson, M. A. (2008). *Semantic Segmentation and Image Search*. Ph.D. dissertation, University of Cambridge.

Leong, C. W., and R. Mihalcea (2009). Exploration in Automatic Image Annotation using Textual Features. *Proceedings of the Third*

Linguistic Annotation Workshop, ACL-IJCNLP 2009, Singapore, pp. 56-59.

Leong, C. W., R. Mihalcea, and S. Hassan (2010). The Mining for Automatic Image Tagging. *Coding 2010*, Poster Volume, Beijing, pp. 647-655.

De Marneffe, M.-C., and C. D. Manning (2011). *Stanford typed dependencies manual.* http://nlp.stanford.edu/software/dependencies_manual.pdf (accessed 25 March 2012).

Nes, S. (2009). *The Mystical Language of Icons*. Michigan/Cambridge, UK: Eerdmans Publishing.

The PASCAL VOC. (2007). http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007 (accessed 25 March 2012).

The Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml (accessed 25 March 2012).

Russell, B. C., A. Torralba, K. P. Murphy, and W. T. Freeman (2008). *LabelMe: a database and web-based tool for image annotation*. International Journal of Computer Vision 77(1-3): 157-173.

Schlecht J., B. Carqué, and B. Ommer (2011). Detecting Gestures in Medieval Images. *IEEE International Conference on Image Processing (ICIP 2011)*, September 11-14, Brussels, Belgium. http://hci.iwr.uni-heidelberg.de/COMPVIS/research/gestures/ICIP11.pdf (accessed 25 March 2012).

Welinder, P., et al. (2010). *Caltech-UCSD Birds 200,* California Institute of Technology, CNS-TR-2010-001. http://vision.caltech.edu/visipedia/CUB-200.html (accessed 25 March 2012).

HisDoc: Historical Document Analysis, Recognition, and Retrieval

Baechler, Micheal
micheal.baechler@unifr.ch
University of Fribourg, Switzerland

Fischer, Andreas
afischer@iam.unibe.ch
University of Bern, Switzerland

Naji, Nada
nada.naji@unine.ch
University of Neuchatel, Switzerland

Ingold, Rolf
rolf.ingold@unifr.ch
University of Fribourg, Switzerland

Bunke, Horst
bunke@iam.unibe.ch
University of Bern, Switzerland

Savoy, Jacques
jacques.savoy@unine.ch
University of Neuchatel, Switzerland

The HisDoc project aims at developing automatic generic tools to support cultural heritage. More precisely, we have designed and implemented a system that performs a fully automated conversion of manuscripts into the corresponding electronic format and provides effective search capabilities. To demonstrate the effectiveness of the proposed solution, we have conducted a large-scale experiment using medieval handwritten manuscripts written in Middle High German. The corpus used in our experiments is based on a well known medieval epic poem called Parzival and attributed to Wolfram von Eschenbach. The first version dates to the first quarter of the thirteenth century. Currently, we can find several versions (with variations) but the St. Gall collegiate library cod. 857 is the one used for experimental evaluation.

The complete system does not make strong assumptions about the current corpus, and it can be adapted with little effort to handle other types of documents and languages. The proposed HisDoc system is subdivided into three main components, namely: image analysis, text recognition and information retrieval.

The image analysis module basically has two essential goals, namely image enhancement, which aims at improving image quality in order to make

handling all the subsequent processing steps an easier task. The second goal is to perform layout analysis that provides a structural description of the scanned document pages. This meta-information may correspond to the number of columns, the number of lines per column, the location of the headline, etc. It is known that these old manuscripts suffer from ink bleeding, holes and stitches on parchments in addition to the artifacts surrounding the non-uniform handwritten text which represents a difficult challenge at this phase as well as in the phases that follow.

The resulting digital images and their meta-information are then passed to the text recognition phase, the second module of HisDoc, in order to produce the corresponding digital transcription. Here, flexible and robust recognition systems based on Hidden Markov Models (Marti & Bunke 2001) and Neural Networks (Graves et al. 2009) are used for the task of automating the transcription of historical texts (Fischer et al. 2009). In this context, flexibility means that the system can be adapted to new writing styles without great effort, and robustness means that the recognizer should attain a high rate of correctly recognized words. Besides the automatically generated version, an error-free transcription was created manually with the help of experts. This latter version forms our ground-truth text and is used to assess the performance levels in the different stages of our project.

For the analysis and recognition of historical documents, only little work has been reported in the literature so far, (Bunke & Varga 2007; Likforman-Sulem et al. 2007) present surveys of approaches to off-line handwriting recognition and on different state-of-the-art text line extraction methods respectively. Commercial systems with a high recognition accuracy are available only for restricted domains with modern handwritten scripts, e.g., for postal address (Srihari et al. 1996) and bank check reading (Gorski et al. 2001).

The HisDoc recognition module has evolved in terms of performance and achieved its current word-accuracy which is close to 94% thus a word-error rate of around 6% using Hidden Markov Models and character similarity features (Fischer et al. 2010). As can be seen, it is unfortunately impossible to obtain flawless recognition especially with the existence of all of the aging and decorative issues mentioned above. The level of the error rate depends on various factors such as the accuracy of the recognition system, the quality of the contrast between the background and the ink, and the regularity of the handwriting. Finally, the size of the training set also plays a role, but relatively good performance can be achieved with a relatively small training set (around 6,000 words).

In order to reduce the human effort needed to generate training samples, HisDoc's text recognition module also includes methods for efficient ground-truth creation and text-image alignment in case of an existing electronic text edition (Fischer et al. 2011).

Performing effective searches on the resulting transcriptions is the goal of the third and last module in HisDoc, the information retrieval (IR) module. The cascading effects of manuscript quality, graphical characteristics and recognition error rate will have an impact on the performance level. But we must also consider additional factors particularly related to medieval writing. Our corpus was handwritten during the thirteenth century, when spelling was not standardized which introduces yet an additional challenge. Moreover, the co-existence of different spellings referring to the same entity in the same manuscript would also reduce the retrieval effectiveness. Besides that, one should also keep in mind the fact that grammar used in medieval languages was clearly different once compared to that of our modern days, which allowed more flexibility for the writer, thus varying from one region to another, or even from one writer to another residing in the same region. All these grammatical and orthography-related matters (spelling variations, punctuation, initials' capitalization, etc.) in addition to the challenges faced in the first two modules of HisDoc would absolutely burden the retrieval process as they impose their negative impact on retrieval effectiveness causing the performance level to fall if the retrieval system is built the conventional way. To quantify this, the retrieval effectiveness is compared to that obtained using the error-free transcription.

In HisDoc's retrieval module, certain techniques were introduced in order to integrate the possible variants to accordingly enhance the retrieval effectiveness by allowing some sort of soft-matching between the query terms representing the user's information needs and four different text surrogates of the transcription, each of which incorporates a certain intensity level of variants' inclusion. As a concrete example, the term '*Parzival*' appeared in the original manuscript as '*Parcifal*', '*Parcival*' and '*Parzifal*'. All of these variants are possible and must be considered as correct spellings. At this lexical level, one might also consider the inflectional morphology where various suffixes were possibly added to nouns, adjectives and names to indicate their grammatical case. With this additional aspect, the proper name '*Parcival*' may also appear as '*Parcivale*', '*Parcivals*' or '*Parcivalen*', increasing the number of potential spelling variants that we need to take into account.

Regarding text representations, we have implemented whole words representation, short overlapping sequences of characters within each

word (*n*-gram) or the first *n* characters of each word (trunc-*n*). During this indexing stage, we have also evaluated the effect of different word normalization procedures such as stemming. This procedure tries to extract the stem from a surface form by automatically removing its inflectional suffixes (number, gender) (light stemmer) or both inflectional and derivational suffixes (aggressive stemmer). As for search strategies, we have evaluated the classical vector-space model *tf idf* and its variants as well as several probabilistic models.

Based on our experiments and using a well-known evaluation methodology (Voorhees & Harman, 2005), using either short (single term) or long queries (three terms), we found that probabilistic IR models tend to produce better retrieval effectiveness (Naji & Savoy 2011). Regarding the stemming procedure, aggressive stemming tends to produce slightly better retrieval performance when facing with longer queries. On the other hand, ignoring the stemming normalization with short queries usually offers the best performance.

The presence of compound words was also analyzed in order to improve the quality of retrieval. The German language is known for this compound construction which occurs more frequently than in English. The main concern here is the fact that the same concept can be expressed with or without a compound form (e.g., *Bankpräsident* or *Präsident der Bank*). Applying an automatic decompounding strategy may provide some successful improvement, particularly for longer queries.

We have assessed the various recognition corpora against the ground-truth version. Compared to this error-free version, the simplest transcription surrogate (the classical output of recognition system, i.e., including no variants) shows a mean degradation in retrieval performance around -10.24% for single-term queries. Considering systematically three or seven variants per input word usually tends to cause the retrieval effectiveness to decrease significantly (from -5.42% for 3-term queries to -64.34% with single-term queries). Including more possible terms per recognition is therefore not an effective solution. We have thus proposed a wiser strategy that includes word variants according to their likelihood probabilities obtained during the recognition stage. This approach produces clearly better retrieval performance. To somehow illustrate our achievement, we can compare the best results obtained using two versions of an English language corpus having character error rates of 5% and 20%. The retrieval performance degradation was around -17% and -46% respectively (Voorhees & Harman, 2005). While our best results for the Parzival corpus, which has a word-error rate of around 6%, the retrieval performance degradation was only limited to around -5%.

The best practices found and the conclusions drawn from our experiments can then be applied to further manuscripts, hence making them digitally accessible and effectively searchable with the least cost possible by partially or totally eliminating the need to having them manually transcribed which in turn will result in saving a lot of resources (time, human effort, finances, etc.). With these documents being completely searchable via digital means, more user's information needs can thus be satisfied via the facilities provided by web-service-based search engine.

Acknowledgement

The authors wish to thank Prof. Michael Stoltz, Dr. Gabriel Viehhauser (University of Bern, Switzerland), Prof. Anton Näf and Mrs. Eva Wiedenkeller (University of Neuchatel) for their valuable support. This research is supported by the Swiss NSF under Grant CRSI22_125220.

References

Bunke, H., and T. Varga (2007). Off-line Roman Cursive Handwriting Recognition. In B. Chaudhuri (ed.), *Digital Document Processing: Major Directions and Recent Advances* Berlin: Springer, pp. 165-173.

Fischer, A., M. Wüthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz (2009). Automatic Transcription of Handwritten Medieval Documents. *Proceedings of the 15th International Conference on Virtual Systems and Multimedia*, pp. 137-142.

Fischer, A., K. Riesen, and H. Bunke (2010). Graph Similarity Features for HMM-Based Handwriting Recognition in Historical Documents. *Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition*, pp. 253-258.

Fischer, A., E. Indermühle, V. Frinken, and H. Bunke (2011). HMM-Based Alignment of Inaccurate Transcriptions for Historical Documents. *Proceedings of the 11th International Conference on Document Analysis and Recognition*, pp. 53-57.

Gorski, N., V. Anisimov, E. Augustin, O. Baret, and S. Maximor (2001). Industrial Bank Check Processing: The A2iA Check Reader. *International Journal on Document Analysis and Recognition* 3(4): 196-206.

Graves, A., M. Liwicki, S. Ferdnandez, R. Bertolami, H. Bunke, and J. Schmidhuber (2009). A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence 31(5): 855-868.

Likforman-Sulem, L., A. Zahour, and B. Taconet (2007). Text Line Segmentation of Historical Documents: A Survey. *International Journal on Document Analysis and Recognition* 9(2): 123-138.

Marti, U.-V., and H. Bunke (2001). Using a Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System. *International Journal of Pattern Recognition and Artificial Intelligence* 15(1): 65-90.

Naji, N., and J. Savoy (2011). Information Retrieval Strategies for Digitized Handwritten Medieval Documents. *Proceedings of the Asian Information Retrieval Symposium*, Dubai, December 2011, LNCS #7097. Berlin: Springer, pp. 103-114.

Srihari, S. N., Y. Shin, and V. Ramanaprasad (1996). A System to Read Names and Addresses on Tax Forms. In *Proceedings of the IEEE* 84(7): 1038-1049.

Voorhees, E. M., and D. K. Harman(2005). *TREC. Experiments and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press.

Research infrastructures for Digital Humanities: The local perspective

Bärenfänger, Maja
Maja.Baerenfaenger@germanistik.uni-giessen.de
Justus-Liebig-University Gießen, Germany

Binder, Frank
frank.binder@germanistik.uni-giessen.de
Justus-Liebig-University Gießen, Germany

A number of research infrastructure projects are currently constructing the next generation research environments for digital humanities (Wynne 2010; ESFRI 2011). They are also conducting activities to interact with potential users, early adopters and key multipliers in the field. With this paper, we report on our experience of conducting such activities at the local institutional level in order to (1) identify potential users, to (2) survey their needs and current practice with regard to digital humanities, to (3) try and stimulate them towards adopting research infrastructures where appropriate and, (4), where the case, to identify remaining barriers. We report on our work in progress that is taking these communicative activities to the local level of several humanities faculties of a large German university.

Research infrastructures (Wittenburg et al. 2010; ESFRI, 2011) are complementing a portfolio of established models of supporting computationally intensive research processes in the humanities: (1) the professional intermediary and (2) competence centers (Edmonds 2005; McCarthy & Kirschenbaum, 2003). Some recent infrastructure initiatives have already engaged in bringing research infrastructures to the user, for instance TextGrid and D-SPIN (Binder et al. 2010), others have contacted prospective users to sharpen the research infrastructures' visions and priorities (e.g. Herold et al. 2010). We believe that these activities need to be complemented at the local levels of universities as well and thereby taking a local perspective on the users and the conditions under which they are carrying out their research activities.

A consideration of the local perspective has two initial foci: The first one lies on general local conditions. These are conditions concerning the local status quo and local infrastructures in digital humanities (henceforth: DH). The second focus has to lie on the local user itself, or better: potential or real users of DH infrastructures, their knowledge about DH infrastructures, their attitude towards DH infrastructures, their requirements

and wishes, especially concerning local support structures. Consecutively, a consideration of the local perspective has to lead to an appropriate local DH infrastructure strategy. In the following, we will outline the questions which have to be considered in an analysis of the local perspective, and the methods which can be used for this analysis, before we will conclude with the elements of a local DH user strategy.

An analysis of the local conditions should begin with a detailed analysis of local DH-related research activities: What is already happening (maybe without naming it DH)? Which current and past projects have an affinity with DH? Who are the experts? Which DH support structures already exist, and which current and future roles do local libraries and computing centers have? Methods to be used in this stage are web-based inquiries (on current and past research activities and on existing DH support structures) and expert interviews (on current and future DH strategies). In a second step, it is necessary to further shed light on the (potential or real) users of DH infrastructures, e.g. the local researchers in the humanities. Do they know about existing national and international DH infrastructures? Do they use national or international DH infrastructures? Do they cooperate with DH centers? Which attitude do they have towards DH infrastructures? Which apprehensions or hopes do they have? What requirements and wishes concerning DH infrastructures do they have? Expert users should also be asked whether they are willing or able to serve as professional intermediaries for their local colleagues. This survey on local users should differentiate between experts and non-experts. This could be realized by a qualitative analysis of interviews with two kinds of groups, an expert group and a non-expert group. Furthermore, an online-survey could be used to quantify the findings of the qualitative study.

The analysis of the local conditions and users may then be used to develop a local DH user-strategy which should include elements like local community building, local public relations activities and local support structures. Several such elements have already been described and implemented in DH infrastructure projects and initiatives like TextGrid, D-Spin, CLARIN, and DARIAH. Established community building events like summer schools (e.g. Binder et al. 2010), or the THATCamp series, can also be instantiated as light-weight events, such as a local 'Data Day' (Woutersen-Windhouwer 2011). Instruments such as newsletters or mailing lists can be used and adapted for local contexts. Apart from these existing event types and strategic elements, a local perspective on DH infrastructures should also be complemented by local support structures like a DH competence center and local professional intermediaries.

At the Justus-Liebig University Giessen, we conducted a survey of the local DH-related research activities and local DH support structures. We found that more than 25 ongoing or recently completed projects in two of five faculties for humanities could be regarded as DH related projects. This is a surprisingly large amount of projects, if you consider the size of the faculties observed. For the categorization procedure, we defined digital humanities with regard to methods and digital practices. Projects which are classified as DH related projects therefore have to use computer-based methods (e.g. 'Die Ordnung von Wissen in Texten'), use or produce digital linguistic corpora (e.g. 'English as a lingua franca in science and business communication'), digital archives (e.g. the 'Prometheus' image archive), digital editions (e.g. the digital edition 'Chronik des Gettos Lodz Litzmannstadt'), or other digital material (e.g. 'e-campus Altertum').

The crucial point is that most of these projects do not define themselves as DH projects, which is fatal if you want to build up a local digital humanities community. This finding, which complements previous observations and discussions (e.g. Joula 2008; Svensson 2010; Prescott 2012), is most important for a local DH strategy. What has to be done first, is to make researchers aware that they are already working in the field of digital humanities, and that they could gain a lot from identifying themselves as part of a local and global DH community. We think that this first step, becoming aware of the field of digital humanities and becoming aware that one is already part of the game, is one of the most underestimated and at the same time most important points, if we would like the digital humanities community to grow and if digital humanities research infrastructures should be used by more researchers than at present time.

References

Binder, F., S. Ahlborn, J. Gippert, and H. Lobin (2010). *Connecting with User Communities: D-SPIN Summer School 2010*. Poster at SDH +NEERI 2010. Vienna/Austria. `http://www.dspin-sommerschule.de/` (accessed 22nd March 2012).

Edmond, J. (2005). The Role of the Professional Intermediary in Expanding the Humanities Computing Base. *Literary and Linguistic Computing* 20(3): 367-80.

ESFRI (European Strategy Forum on Research Infrastructures) (2011). *Strategy Report on Research Infrastructures – Roadmap*

2010. Luxembourg: Publications Office of the European Union. http://dx.doi.org/10.2777/23127 (accessed 22nd March 2012).

Herold, A., T. Warken, F. Binder, and G. Gehrke (2011). *D-SPIN Report R3.4 Case Studies, Final Results*. http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R3.4.pdf (accessed 22nd March 2012).

Juola, P. (2008). Killer Applications in Digital Humanities. *Literary and Linguistic Computing* 23(1): 73-83.

McCarthy, W., and M. Kirschenbaum (2003). Institutional Models for Humanities Computing. *Literary and Linguistic Computing* 18(4): 465-89.

Prescott, A. (2012). Consumers, creators or commentators? Problems of audience and mission in the digital humanities. *Arts and Humanities in Higher Education* 11(1-2): 61-75.

Svensson, P. (2010). The Landscape of Digital Humanities. *Digital Humanities Quarterly* 4(1). http://www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html (accessed 22nd March 2012).

Wittenburg, P., N. Bel, L. Borin, G. Budin, N. Calzolari, E. Hajicova, K. Koskenniemi, L. Lemnitzer, B. Maegaard, M. Piasecki, J. Pierrel, S. Piperidis, I. Skadina, D. Tufis, R. van Veenendaal, T. Váradi, and M. Wynne (2010). Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. http://www.lrec-conf.org/proceedings/lrec2010/summaries/679.html (accessed 22nd March 2012).

Woutersen-Windhouwer, S. (2011). *Long-term experiences with repositories*. Invited talk at CLARIN-D tutorials. http://www.clarin.eu/system/files/woutersen-experiences.pdf (accessed 22nd March 2012).

Wynne, M. (2010). *The humanities research environment of the future*. Opening keynote at the D-SPIN Summer School 2010 http://www.clarin.eu/node/3286 (accessed 22nd March 2012).

Pelagios: An Information Superhighway for the Ancient World

Barker, Elton
e.t.e.barker@open.ac.uk
Open University, UK

Simon, Rainer
Rainer.Simon@ait.ac.at
Austrian Institute of Technology, Austria

Isaksen, Leif
leifuss@googlemail.com
University of Southampton, UK

On-line resources that reference ancient places are multiplying rapidly, bringing huge potential for the researcher provided that they can be found; but users currently have no way of easily navigating between them or comparing their contents. Pelagios, a growing international cooperative of online ancient world projects,[1] addresses the problems of discovery and reuse with the twin aims of helping digital humanists to make their data more discoverable, and of empowering real-world users (scholars, students and the general public) to find information about particular ancient places and visualize it in meaningful ways. While the project focuses on the ancient world, the methodology and tools developed will be of interest to anyone working with data containing spatial references. The Pelagios family purposefully includes partners maintaining a wide range of different document types including texts, maps and databases. In doing so we take some of the first steps towards building a Geospatial Semantic Web for the Humanities.[2]

In this paper we discuss two of the major workflows underpinning Pelagios. First, we address the method by which the partners prepare their data so that it can be linked together in an open and transparent manner: i.e. what are the processes that you should undertake if you want to make your data Pelagios compliant? Second, we consider the various ways in which the results can be visualized, paying particular attention to the tools and technologies used and the problems encountered. We end with a brief description of our visualization services – both a Graph Explorer and various embeddable web widgets – which, we believe, demonstrate the value of taking a lightweight Linked Open Data approach to addressing problems of discoverability, interconnectivity and reusability of online resources. At the same time, however, we use this paper to

discuss real-world practical concerns as well as engage in deeper speculation about the significance of this type of approach for escaping the 'siloing' mentality that inhibits many other data integration initiatives.

1. Ontology

The first part of the paper will sketch out and reflect upon the architectural aspects of Pelagios, in particular the requirements necessary to maximize the exposure and interconnectivity of the data themselves. The structure of the data is targeted at helping the user groups of each partner accomplish two kinds of task:

A) Discovering the references to places within a single document (text, map, database);

B) Discovering all the documents (texts, maps, databases) that reference a specified place.

In order to achieve this we (i) use a common RDF model to express place references (the OAC model); and (ii) align all local place references to the Pleiades Ancient World Gazetteer. The OAC annotation ontology provides an extremely lightweight framework for associating global concepts (such as places) with specific documents (and fragments of them).[3] An OAC annotation is a set of RDF triples which identify a target document (by means of its URI) and the body of the annotation itself, in our case a URI identifying a specific place in Pleaides. An example is given below:

ex:ann1 rdf:Type oac:Annotation

dcterms:title "Example annotation"

oac:hasTarget <some resource>

oac:hasBody <http://pleiades.stoa.org/places/[PLEIADES ID]>

The decision to use the OAC model typifies our pragmatic lightweight approach, which has not been to reinvent the wheel but to reduce, reuse and recycle wherever possible. Using a publicly available, lightweight core ontology permits modular extensions for different kinds of document so that details specific to each type do not add unnecessary complexity for users wishing to publish data in conformance with the core ontology. We have also found that Pelagios partners are afforded a great deal of flexibility in implementation. For example, the RDF may be expressed in a number of different formats (RDF/XML, Turtle, RDFa, SPARQL, etc.) and the simplicity of the ontology allows partners to focus on the considerably more challenging task of aligning local place referencing systems with the global Pleiades gazetteer.

Pelagios has been documenting the various processes by which each partner has identified place references and aligned those references to Pleiades: one significant outcome of the project will be a 'cookbook' guide for those looking to adopt a lightweight Linked Open Data approach to related domains, which we outline here. The Pelagios process, however, has greater extensibility than provision for digital classicists. Although we are using URIs for places in the ancient world, the OAC ontology is equally applicable to other gazetteers (including those based on modern placenames, such as GeoNames) or even non-spatial entities such as periods or people.

2. The Pelagios Explorer and Embeddable Widgets

The second part of the paper assesses the possibilities for data exploration once different projects have adopted the core model for representing place references. Here we discuss the technologies exploited in developing the Pelagios Explorer, a prototype Web application that makes discovery and visualization of the aggregated data simple, and various web widgets that can sit on partner websites to provide a window onto the Pelagios linked data world. By aggregating Pelagios partners' place metadata in a Graph Database, the Explorer supports a number of common types of query using visually-oriented interaction metaphors ('which places are referenced in these datasets?', 'what is the geographical footprint of these datasets?', 'which datasets reference a particular place') and displays results using a graph-based representation.[4] In addition, the Pelagios Explorer exposes all data available in the visualizations through an HTTP API to enable machine-access. For their part the widgets that we are currently developing provide more specific views on Pelagios data than the graph explorer allows (such as an overview of the data about a particular place). But equally we are exploring ways of making these widgets easy to configure to, customise for and embed in external Websites, thereby maximizing the potential reuse value of the data.

There are at least two different user groups whose interests and concerns we address. For the 'super users' interested in contributing data, adhering to the two principles of the OAC model and Pleiades URIs is the essential step in order to prepare data for use in these various Pelagios visualizations. With these basic issues established, we go on to discuss more detailed control over how data can be represented in a generic interface, for instance using names and titles to label data, or structuring a dataset hierarchically into subsets, such as individual books, volumes, chapters, and pages. Toolset and methods are still at

an early stage but some useful resources have already been made available for general use via our blog.[5]

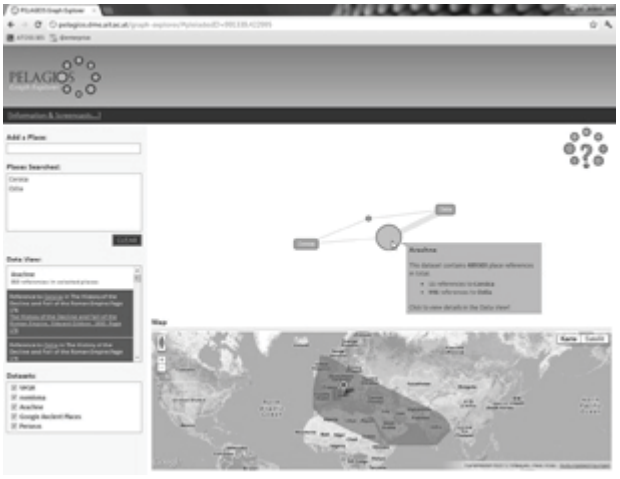
Our second user group represents scholars, students and members of the public who may wish to discover the resources that reference an ancient place of interest. We consider how the Pelagios framework is beginning to bring together an enormous diversity of online data – such as books that reference places, and archaeological finds discovered there – which a user can search through, combine and visualize in various ways depending on their needs.[6] In particular, we consider the challenges, both technological and intellectual, regarding the development, production and use of such tools or services that aim to provide a useful and intuitive resource for non technically-minded subject specialists.

We conclude with a brief reflection on the processes by which the Pelagios family has been developing. In particular we stress the digital services that have made the coordination of such an international initiative possible, and outline the challenges that remain to anyone wishing to add their data to the Pelagios multiverse.[7]

Funding

This work has been supported by JISC (the Joint Information Systems Committee) as part of the Geospatial and Community Outreach programme (15/10) and the Resource and Discovery programme (13/11).

Screenshots of the Pelagios Explorer:



References

Elliott, T., and S. Gillies (2009). Digital Geography and Classics. *DHQ: Digital Humanities Quarterly* 3(1).

Harris, T. M., L. J. Rouse, and S. Bergeron (2010). The Geospatial Semantic Web, Pareto GIS, and the Humanities. In D. J. Bodenhamer, J. Corrigan, and T. M. Harris (eds.), *The Spatial Humanities: GIS and the Future of Scholarship*. Bloomington: Indiana UP.

Sanderson, R., B. Albritton, R. Schwemmer, and H. van de Sompel (2011). SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination. *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries*. Ottawa, Canada.

Schich, M., C. Hidalgo, S. Lehmann, and J. Park (2009). The network of subject co-popularity in Classical Archaeology. In *Bolletino di Archaeologia On-line*.

Notes

1. Pelagios includes: Arachne, `http://www.arachne.uni-koeln.de` ; CLAROS, `http://explore.clarosnet.org` ; GAP, `http://googleancientplaces.wordpress.com` ; Nomisma, `http://nomisma.org` ; Open Context, `http://opencontext.org` ; Perseus, `http://www.perseus.tufts.edu` ; Pleiades, `http://pleiades.stoa.org` ; Ptolemy Machine, `http://ptolemymachine.appspot.com` ; SPQR, `http://spqr.cerch.kcl.ac.uk` ; Ure museum, `http://www.reading.ac.uk/Ure`.
2. See Harris et al. (2010).
3. OAC is a powerful information model in its own right and has recently been used as the basis for the SharedCanvas annotation system. See Sanderson et al. (2011)
4. Network visualizations are ideal for representing bipartite networks such as this. See, for example: Schich et al. (2009).
5. `http://pelagios-project.blogspot.com/p/useful-resources.html`

6. See screencasts at: http://pelagios.dme.ait.ac.at/screencasts/exploring-datasets.html , and http://pelagios.dme.ait.ac.at/screencasts/exploring-places.html .
7. A vision that coincides well with the discussion of the future of Classical scholarship by Elliott and Gillies (2009).

Putting TEI Tite to use – generating a database resource from a printed dictionary or reference type publication

Barner-Rasmussen, Michael
mbr@hum.ku.dk
Danish National Research Foundation Centre for Language Change In Real Time, Denmark

The motivation to write this paper stems from an experience we had a few years ago. We needed to digitize a large reference work called Danmarks Stednavne (DS 1922-2006, 'Place Names of Denmark') which contains most, if not all, of the place names in Denmark.

We wanted and needed the end result to be in a database-style format, that is, capturing or otherwise marking the place name of each record, and including all the sources, scholarly interpretations, historical names and dates, and so on. But we discovered that no one was willing to sign up for the task, even though we had a quite decent budget (capped at approximately $20 a page and 10,000 pages in total).

We were uniformly informed that the task was too complex.

This paper presents the solution we came up with: we bought a TEI Tite (Trolard 2009) encoded digitization procedure from an outside vendor, but did the post processing into a database format ourselves.

Cost and practicality have weighed heavily and are important aspects of the methodology presented. It might seem more interesting to try and create some genetic, self-learning algorithm to automatically eke out the semantics of a large collection of reference entries and mark them up appropriately. We are eagerly awaiting such a marvelous algorithm. However, in our particular circumstances, digital humanities is about developing methods and tools that leverage accessible and affordable competencies and resources in order to reach a practical goal cost-effectively. In this case, it will curate (and thus save) a vital piece of Danish cultural heritage, resulting in a new digital resource usable by anyone with an interest in the origin and history of place names in Denmark, and providing Denmark's place-name authorities with a modern repository for future place-name registration and research (the first, rough

presentation of which can be viewed at `http://danmarksstednavne.navneforskning.ku.dk/`).

Since the task had been deemed 'highly complex' by the vendors we had approached, we put some effort into the analysis. This paper presents the analysis, along with some 'mid-term' results since we have now parsed approximately 50% of the source material.

To date, approximately two thirds of Denmark is covered by the DS publication, which spans 25 (soon to be 26) volumes published in the period from 1922 to 2006. The individual place-names articles list a selection of the earliest known source types, historical forms of the name, and pronunciations, and provide scholarly interpretations and annotation of the name's origin and meaning. An example is given below.

Sivdel. IV. [*syʁdi̯ˀəl*]. MB Siuff deel; MK 1794 Syv Deel. — D e l , se Indl.; samme Sted som følgende.

Figure 1: An entry in DS

TEI Tite is a subset of the TEI created to standardize off-site text encoding somewhat, so that vendors and buyers can arrange for it by simply agreeing to adhere to TEI Tite at both ends. It captures all the printed material, all characters, and all structural divisions (like chapters, line breaks, etc.).

Given this, our initial data analysis showed us that it should be possible to parse individual 'records' (entries in DS) by 'switching' on typographical cues – for example, using the semi-colon character ';' to differentiate between different source references, long hyphens to extract the interpretation, and so on.

The lexicography of the publication has never been formally specified, but it was quickly obvious that at least each individual volume had a fairly consistent lexicography; thus, a volume-by-volume parsing might reasonably succeed. Another practical problem was that since the publication has been ongoing for more than 90 years, the volumes exhibited small differences that are significant for parsing and must be dealt with individually.

The prospect of having to code 25 parsers gave us pause, however, so we set ourselves the task of producing a single program that could be configured from the outside to allow for (more or less) minute differences in lexicography and even new lexicographical instances – the roman numeral after the place name in Figure 1 was such an undiscovered item.

Very briefly stated, what we came up with is a modular, outside-of-code configurable, easily extended program/platform for parsing TEI Tite-formatted XML representations of dictionary or reference data, or textual data that exhibit 'semantic typesetting'.

The general idea is that a module takes care of only one singular task in the whole parsing procedure, operates via a standard interface, and is callable outside the code via a configuration file.

The program is written in a modern high level programming language (C#) so that the skills needed to produce additional modules are readily available on most university campuses and in the population at large.

So: no super computers, no state-of-the-art statistical analysis, no genetic algorithms or natural language comprehension, no crowd-sourcing the production of new knowledge – at least not while digitizing existing sources of knowledge. Moreover, we claim no 'new forms of scholarly inquiry', not immediately at least, derived maybe, but not new and probably not that many new questions posed that could not have been answered albeit very laboriously before.

We do, however, believe that the methodology would likely add value to various TEI-encoded digital artifacts at not only a reasonable cost, but also by utilizing skills and competencies that are readily available. Billed hours for the parsing activity ended up in the low hundreds (~450 hours), so cost effectiveness would appear acceptable.

The paper presentation will present and discuss the following elements vital/interesting to the methodology.

1. 80-20 The roughly 50.000 place name entries in DS are of more or less the same lexicography but of vastly differing length and complexity, so the program is designed to quickly pick up all 'low hanging fruits', parsing the simpler entries in their entirety and handing them off to the database team.
2. Scoping of the modules Examples from the code will be presented to illustrate that there is almost no lower limit as to how little a single module should do. The salient point being the number of volumes that the module would be relevant for is more important than doing a lot of stuff all at once.
3. When in doubt ask Another key element of the design is that every choice the program has to make is estimated as to how long it would take to produce working code versus simply asking the user/parser. As it turned out this principle has been the biggest time and cost saver on the project since it eliminates much of the risk of making the code brittle and having the programmer(s) take off on a quest for 'a clever hack'™. Differentiating

between source references and name instances will be demonstrated and discussed.

4. Iterative, iterative, iterative Writing the parser modules proved to be at least as much data analysis as the, err, data analysis.
5. Creating the framework and remembering what was done Taking into account item 4. above the framework was designed so it 'remembers' the steps taken on a given volume or subset of the publication so that 'reparative' steps/modules can be added on as they become evidently needed.
6. Lessons learned

1. Tools, as they are often envisioned in DH publications are most often 'done deals', completed software artifacts that can accomplish this thing or that with humanistic digital artifacts. The 'tool' this paper presents is the TEI Tite, a programming language and a methodology for putting the three together to achieve a specific end. As programming languages become more and more 'user friendly' the barrier to entry into this type of interactive, iterative, collaborative tool building lowers to the point where a project like ours became possible and this seems to be the way of the future.
2. Also, last but verily not least: We believe that if it has lexicography, it can be parsed.
   However, they are not stringent (database) schema and cannot be fully automatically parsed into such.
   Enter our technology/methodology for iteratively parsing lexicographically organized digital data (in the form of a TEI Tite encoded digitization).
   The digital version of DS is paid for by the DigDag-project (`http://www.digdag.dk`)

From the DigDag website:

The DigDag project, short for Digital atlas of the Danish historical-administrative geography, is a research project funded by The National Programme for Research Infrastructure under the Danish Agency for Science, Technology and Innovation.

Running from 2009 to 2012, the aim of the project is the establishment of a database structure which can tie the Danish administrative districts together geographically and historically. This will provide a historical-administrative GIS platform for the digital humanities allowing to:

- establish a historical-administrative research infrastructure of Denmark c. 1600–
- form the backbone of future historical and administrative research
- create a powerful search engine for use in the service functions of archives, collections and libraries

As source data will be used Det digitale matrikelkort, the digital cadastre, and the basic unit will be the ejerlav, the townland.

Participants in the project are a range of major Danish research institutions with a focus on historical and administrative research together with the major Danish heritage institutions and archives.

References

Trolard, P. (2009). TEI Tite — A recommendation for off-site text encoding. `http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html`

DS 1922-2006. Danmarks Stednavne vol 1-25. Volumes 1-15 published by the Place Names Committee, Volumes 16-24 by the Department of Name Research, Volume 25 by the Department of Name Research at University of Copenhagen. See also `http://da.wikipedia.org/wiki/Danmarks_Stednavne` and/or `http://nfi.ku.dk/publikationer/trykte_serier/danmarks_stednavne/`

Bolter, J. D. (2001). *Writing Space: Computers, Hypertext, and the Remediation of Print*. Second Edition. Mahwah, NJ: Lawrence Erlbaum.

Levy, P. (1998). *Becoming Virtual: Reality in the Digital Age*. Da Capo Press.

Weinberger, D. (2002). *Small Pieces Loosely Joined: A Unified Theory of the Web*. New York: Perseus Publishing.

Lowy, J. (2005). *Programming .NET Components: Design and Build .NET Applications Using Component-Oriented Programming*. Sebastopol, CA: O'Reilly Media; 2 edition.

Digital Humanities in the Classroom: Introducing a New Editing Platform for Source Documents in Classics

Beaulieu, Marie-Claire
Marie-Claire.Beaulieu@Tufts.edu
Tufts University, USA

Almas, Bridget
bridget.almas@tufts.edu
Tufts University, USA

In this paper, I will introduce a new online teaching and research platform on which students can collaboratively transcribe, edit, and translate Latin manuscripts and Greek inscriptions, creating vetted open source digital editions. This platform was created by a group of faculty, software developers, and library professionals at Tufts University to enable students to work with original untranslated documents. By doing so, students not only develop their own language skills and research abilities, but they also contribute to the efforts of the scholarly community worldwide to meet the challenge of publishing large numbers of primary source documents online while preserving high editorial standards as described by Robinson (2010). The platform will be made available as open-source software and since it is language-independent, it can be used to edit and translate any source document in any language and any Humanities field.

The platform is an adaptation of Image J and Son of Suda Online (SoSOL). It allows students to transcribe ancient texts directly from a high-resolution digital image while mapping the text onto the lettering. The transcriptions thus produced are automatically encoded in accordance with the EpiDoc markup guidelines, a standard developed by an international consortium of scholars for the accurate description of inscriptions and manuscripts in TEI XML (see Sosin 2010). This ensures greater visibility for the collection and fosters scholarly dialogue since all documents published under the EpiDoc standard are compatible and can be exploited together in an open-access collaborative context (see Bodard 2008). The students' work therefore contributes to a worldwide effort to make ancient manuscripts and inscriptions available online. To link all these resources and present the web interface, we used the CITE services ('Collections, Indexes, and Texts, with Extensions') developed by the Homer Multitext Project (see Blackwell and Smith 2009). We intend to investigate ways to take advantage of the chain of authority supported by the SoSOL environment to facilitate the process of the long-term archiving of the resources in the Tufts institutional Fedora repository. We will make the data from this project available to the broader humanities community through the Perseus Project.

Initially, the students will be working with Latin manuscripts (ranging in date between the 12th and 19th centuries) held in the Tufts University Library which have been digitized and published online. The students will also edit and translate ancient Greek funerary inscriptions which relate to Marie-Claire Beaulieu's research on the poetics of Greek epitaphs. The collection created by the students will form a cohesive dataset of Latin and Greek texts which will be used at Tufts for ongoing research projects on ancient literature but will also be a valuable addition to corpora being developed by scholars worldwide such as the EpiDoc databases (see Cayless et al. 2009, esp. 34).

Students will begin their work with a high resolution image of a Latin or Greek document published in our online collection. After receiving training on the decipherment of ancient Greek and Latin scripts, they will transcribe the texts and map them onto the image of the document using the software described above. The transcription step is important since it allows for careful review: thanks to the direct visual link created by the software between the image of the document and the transcription, the instructor and the audience can instantly evaluate the accuracy of the transcription (see figures 1 and 2).

The second step is to produce a normalized edition of the texts, including the resolution of abbreviations and lacunae, wherever possible. At this stage, two important digital tool sets are integrated with each text. The Alpheios Greek and Latin tools provide interactive lexical and grammatical analysis of each word, and the Perseus tools compute dynamic word lists and help the students and their instructors keep track of the progress accomplished in language learning (see figure 3). The Perseus tools can also be used to design personalized reading lists and exams for each student. Although this project uses Latin and Greek language tools, the design can easily accommodate the integration of similar tools for any language.

The third and final step is translation and annotation, which is crucial for making the texts accessible to a wide-ranging audience of students and scholars in other disciplines as well as the general public. At this stage, students will translate the texts into English and gather as much information as possible on each text, including its date and historical and

literary significance. They will establish parallels with other similar texts, which can be implemented as hyperlinks. The entries are finally rounded out with a complete bibliography of the scholarship available on each text and credits naming all contributors to the entry.

After their publication online, the entries will be linked to the SAKAI/Trunk software, an open-access collaborative learning management resource that allows students to design personalized electronic portfolios. These portfolios can be used by instructors to evaluate a student's overall performance and progress and can also serve in applications for grants, graduate school admission, or in the context of a job search. An entry in this online collection should feature prominently in a student's portfolio, since it constitutes a valuable and quantifiable language learning experience as well as a piece of original scholarship on a previously unedited and/or untranslated text. Throughout this process, students participate in the efforts of the scholarly community worldwide while perfecting their own language and research skills, thus becoming true citizen scholars.

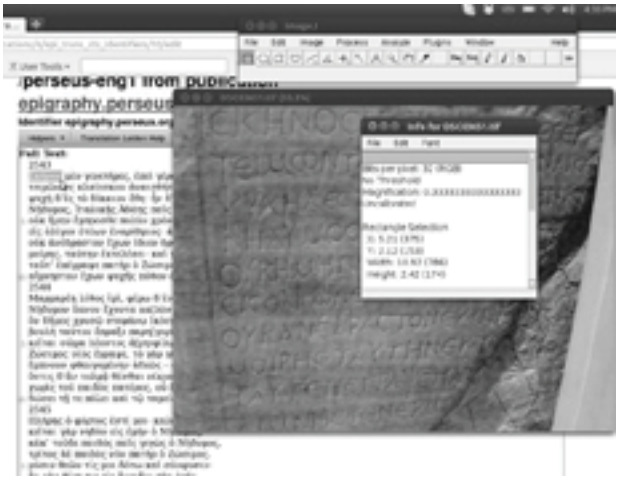


Figure 1: The transcription step

Figure 2: The submission and evaluation step

Figure 3: The Alpheios tools provide interactive lexical and grammatical analysis of each word

References

Blackwell, Chr., and D. N. Smith (2009). Homer Multitext – Nine Year Update. *Digital Humanities 2009 Conference Abstracts*, (June 2009): 6-8. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf (accessed March 9, 2012).

Bodard, G. (2008). The *Inscriptions of Aphrodisias* as Electronic Publication: A User's Perspective and a Proposed Paradigm. *Digital Medievalist* 4. http://www.digitalmedievalist.org/journal/4/bodard/ (accessed March 9, 2012).

Cayless, H., Ch. Roueché, T. Elliott, and G. Bodard. (2009). Epigraphy in 2017. *Digital Humanities Quarterly* 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000030.html (accessed March 9, 2012).

Robinson, P. (2010). Editing Without Walls. *Literature Compass*, 7: 57-61. http://dx.doi.org/10.1111/j.1741-4113.2009.00676.x (accessed March 9, 2012).

Sosin, J. (2010). 'Digital Papyrology.' Congress of the International Association of Papyrologists, Geneva, Switzerland. http://https://exchange.tufts.edu/owa/redir.aspx?C=582e4a9690a04632b2997bc4c6836f47&URL=http%3a%2f%2fwww.stoa.org%2farchives%2f1263 (accessed March 12, 2012).

DiaView: Visualise Cultural Change in Diachronic Corpora

Beavan, David
d.beavan@ucl.ac.uk
University College London, UK

1. Introduction

This paper will introduce and demonstrate *DiaView*,[1] a new tool to investigate and visualise word usage in diachronic corpora. DiaView highlights cultural change over time by exposing salient lexical items from each decade or year, and providing them to the user in an effortless visualisation. This is made possible by examining large quantities of diachronic textual data, in this case the Google Books corpus (Michel et al. 2010) of one million English books. This paper will introduce the methods and technologies at its core, perform a demonstration of the tool and discuss further possibilities.

2. Applicable Corpora

Key to the success of any large-scale cultural inspection is a large corpus of writing, both in terms of chronological span, and also depth of sampling from each year. Two corpora stand out as candidates for this approach: The Corpus of Historical American English (COHA) (Davies 2010) and the Google Books corpus. The COHA dataset, while being 400 million words and having been more rigorously compiled, is restricted in terms of its availability. Google Books, on the other hand, is 155 billion words, spanning from the sixteenth century to the present, although the precise corpus make-up is less lucid. Google have provided n-gram frequency lists for download, and consequently this visualisation has been based upon the Google Books English One Million data set. It should be stressed that the techniques DiaView are applicable to many other data sets, from very focused specialised corpora to large-scale newspaper archives. The choice of corpus used for this demonstration is mainly driven by two factors: public availability and re-use in addition to the need for general content leading to a wider public appreciation of the results this tool provides.

3. Established Query Methods

The standard tools available to current corpus users revolve around two methods: word searches or the comparison of frequencies (either over time or between two focussed points). These tools are in widespread use, and while well understood and used, they do not present the whole linguistic picture to their users.

The *Google Books n-gram Viewer*,[2] for instance is very capable at allowing users to chart the usage of lexical items over time. While illuminating, it implies that users already know what they are looking for (the lexical item) and also that frequency (the number of times that word is used) provides enough data to make assertions. Choosing the term of interest at the outset means this resource delivers a powerful search function, but it fails at browsing or exploration. In other terms you only find what you look for.

The alternative method, based upon comparing frequencies over time, is best exemplified (using the same corpus) by the *Top Words 1570-2008*.[3] This tool meticulously visualises the most frequent words in the Google Books corpus and charts their rise and fall across every decade the corpus covers. Word frequency alone does not tell us enough. Just because a lexical item is frequently occurring in a particular time span, it may not be of interest, particularly if you find it frequently elsewhere. That word soon downgrades from important to common and suddenly becomes much less interesting.

DiaView takes a different approach, one that promotes browsing (not searching) and transcends basic word frequency methods (it strives to deliver salience).

4. DiaView Concept

DiaView is designed to operate at a more opportunistic conceptual level than the examples above; in this case summarising and browsing data takes precedence over specific focussed searches. DiaView aims to complement other approaches, and as such will operate as a gateway to other methods, particularly lexical searches.

The tool divides its corpus into a number of chronological divisions, i.e. years or decades. In each, statistical measures are used to extract those lexical elements that are salient or focussed on that time-span. This does not necessarily imply frequently used words, but, rather, words which are found predominantly in that time, a possible key into the cultural issues experienced or explored at that moment in time.

5. DiaView Method

A relational database has been used to realise DiaView, however other technologies are equally applicable to producing similar results. Equally, any number of statistical measures or comparators can

be used to ascertain the salient words for each year, *mutual information* or *log-likelihood* are good candidates for use.

For this demonstration DiaView requires word frequency lists for each year. In this case the corpus was trimmed to 1900-1950. Fundamentally this is a tabulated list of lexical items, followed by the number of occurrences in each year.

DiaView then creates an aggregated view of these frequency lists, gathering each *type* (individual word) and the total number of occurrences across the entire corpus (summing the data above). A further optional stage is to cull this list; either by disposing of the least frequently used types (e.g. remove words not used more than ten times each year) or by concentrating on the most frequently used types (e.g. keep the 100,000 most often used words).

Extracting salient terms from each year is now performed. Taking each type (identified previously) in turn, its frequency in each year is inspected and compared to its global frequency. Here the statistical measure is used to gauge its connection to each particular year. This removes the dependence of raw frequencies: a word may occur few times or a great number, what is of prime importance is if its distribution is skewed or focussed on a particular chronological range. For each year a ranked list of salient types are created.

The visualisation is created by extracting the salient types from each year and displaying the top 25 in descending order. Each type in each year can be used to hyperlink other resources, such as the Google Book n-gram Viewer.

6. Future possibilities

While DiaView offers new ways to view large data sets, it is open to further enhancements. Access to corpora divided by genre would add valuable benefits to the visualisation, allowing users to narrow down the corpus to include material only of their choice. Linguistic stemming could also be used to gather words around their root form, e.g. to cluster run, running, runs, ran etc.

References

Davies, M. (2010). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at `http://corpus.byu.edu/coha/` .

Michel, J., Y. Shen, A. Aiden, A. Veres, M. Gray, W. Brockman, The Google Books Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Aiden (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* (Published online ahead of print: 12/16/2010)

Notes

1. `http://www.scottishcorpus.ac.uk/corpus/diaview/`
2. `http://books.google.com/ngrams`
3. `http://pages.cs.wisc.edu/~gleicher/Hacks/tops.html`

Catch + Release: Research and Creation of a Digital New Media Exhibition in the Context of a Cultural and Heritage Museum

Beer, Ruth
rbeer@ecuad.ca
Emily Carr University of Art and Design, Vancouver, Canada

We will present, *Catch + Release: Mapping Stories of Geographic and Cultural Transition* (2009-2012), a research and creation project funded by the Social Sciences and Humanities Council of Canada. It is a case study of a collaborative, interdisciplinary digital new media practice in community engagement and informal learning. It looks at local/global environmental, socio-cultural and historical interplays related to multi-cultural community perspectives, including First Nations, on changes due to the demise of the fishing industry. Historically, fishing was one of the primary reasons for the multi-cultural immigration that helped to foster the settlement, economic development, and social growth on the west coast of Canada.

We will discuss our interactive digital new media exhibition *Catch + Release* installed in The Gulf of Georgia Cannery Museum National Historic Site, a social history museum in Steveston, British Columbia, a small city on Canada's Pacific coast. The exhibition proposes to advance an encounter with history as critical inquiry into contemporary life of the region. The researchers' backgrounds in new media art, new media education, cultural theory, and interactive design, come together to explore how research and creation contributes to informed discourse that engages with the region's cultural history and social, cultural and ecological present. Our new media exhibition that uses sensor technologies and immersive interactivity to overlay stories from the community together with those of the museum, is intended to foster a sense of belonging and deeper understanding of historical and current circumstances. While our study is localized, these conditions and stories resonate in other coastal communities that once relied on fishing, thus, our project addresses cultural and marine environmental sustainability in many similar contexts.

This project draws on qualitative inquiry informed by practice-based research methodologies in education and visual art and is impacted by the fields of new media art, cultural studies, museum studies and the history of art as a discipline vital to the analysis of how history, including the past and just-past, is mediated through images, objects, and experiences (Benjamin 1968).

Our project at the intersection of socially-engaged interactive new media art and pedagogy, views the museum as an essentially social institution that requires interdisciplinary analysis as an informal site of learning and as space for promoting critical thinking. The Gulf of Georgia Cannery Museum National Historic Site, a museum since the closing of the cannery in 1970 and that still features the intact production line, is an important pedagogical space for school groups, tourists and local visitors, and subsequently a major source of regional identity.

For the *Catch + Release* exhibition, we designed situation-specific, new media interactive art experiences for contributing diverse perspectives; catching and releasing stories that challenge the seamlessness of those historical narratives typically represented by museum displays of artifacts and information panels. The exhibition is intended to generate polyphony of diverse voices to make viewers feel a part of their regional museum, and exposes them to multiple viewpoints about their (local) relationship to the (global) ecology-making visitors more aware of their relationship to and impact on the local ecology and history. In this way, the project did not seek to create an exhibition in the space, but rather used the space to exhibit the place in which it is situated.

Our presentation will describe the conceptual basis and rationale for the inclusion of the new media artworks that comprise the exhibition. They include: interactive projections of live-stream visualization of data from underwater sensors of Canada's ocean observatory; clusters of light and sound digital sensor activated interactive sculptures that reference the marine geoscape and communication devices used by fisherman; and a montage of video interviews with people whose stories expose sometimes oppositional views on the fishing industry. The polyphony of multiple and diverse voices argues for a more open-ended and complicated way of dealing with the representation of history. We will discuss how immersive technology can provide unique opportunities for teaching and learning experiences and embodied ways of knowing through a rich multi-sensory environment that engages the viewer with how culture and memory come together in new forms.

Our study is also intended to stimulate interest the historical role of the museum by inviting the community to participate in the creation and

complication of historical narratives. Cultural and heritage museums, specifically, have become multi-purpose, spaces where museum professionals strive to expand their programmes and define new communication approaches capable of engaging broader publics (Prior 2008). In this context, recent interactive technologies are seen to offer museums an ability to create a sense of spectacle and excite audiences (Hemsley, Cappellini & Stanke 2005). Increasingly, however, the reification of technologized display raises concerns about the type and quality of museum experience (Griffiths 1999). Despite the potential of multi-media displays to address the diverse needs of different audiences or audience members by combining information and physical spaces and presenting interactive storytelling, there is still a tendency to rely on basic interactive terminals with mouse, keyboard and monitor. While this may be an efficient way to deliver information, our project contributes aesthetic and expressive possibilities for actively engaging audiences. Recent discourse around new media art practices and new museums has encouraged an interest in technology beyond fixed modalities of story authoring, to sensory-driven, interactive processes of learning that allow visitors to engage with interactive art forms and digital technologies to construct their own learning experience in the museum (Henry 2010; Rogoff 2010). However more research is needed to assess new models for establishing and maintaining dialogues with visitors and new ideas about how to connect these dialogues with the broader audiences (Cutler 2010). Contemporary new media artwork that invites users to interpret or experience content and function through a range of physical and sensorial interactions are seen to be extremely promising (Hornecker & Buur 2006). Their ability to author embodied, responsive engagements are well suited to the museums' pedagogical programs.

As contemporary art is transformed through the incorporation of new media practice, works like *Catch + Release* offer a potentially innovative approach to learning. If as Ellsworth (2005) contends, new digital media contribute to pedagogic disruption and call for new 'routes' of relational thinking, they establish the ground of shared social interest drawing together artists, museum professionals and educators to discuss how the public becomes co-creators of meaning (or active learners) through experiential, participatory engagement (Bishop 2006).

Social history museums have begun to invite artistic inquiry, presenting temporary digital new media art exhibitions such as *Catch + Release* that advance the idea of a complex cultural and social history, challenge conventional displays and provide alternate modes of visitors' participation, meaning making and knowledge construction (Bishop 2006), in these new spaces of artistic exchange or encounter.

References

Benjamin, W. (1968). The Work of Art in the Age of Mechanical Reproduction. In H. Arendt (ed), *Illuminations*. Translated from German by H. Zohn. New York: Shocken Books, pp. 217-251.

Bishop, C., ed. (2006). *Participation*. Cambridge: MIT Press.

Cutler, A. (2010, Spring). What is to be done, Sandra? Learning in cultural institutions in the 21st century. *Tate Papers 13*. http://www.tate.org.uk/research/tateresearch/tatepapers/10spring/cutler.shtm (accessed 1 October 2011).

Ellsworth, E. (2005). *Places of Learning: Media, Architecture, Pedagogy*. New York: Routledge.

Griffiths, A. (1999). Media technology and museum display: A century of accommodation and conflict. *MIT communications forum*. http://web.mit.edu/comm-forum/papers/griffiths.html#1 (accessed 1 October 2011).

Hemsley, J., V. Cappellini, and G. Stanke (2005). *Digital Applications for Cultural and Heritage Institutions*. Burlington: Ashgate.

Henry, C. (2010). *The Museum Experience: The Discovery of Meaning*. Reston: National Art Education Association.

Hornecker, E., and J. Buur (2006). *Getting a Grip on Tangible Inter action: A Framework on Physical Space and Social Interaction, Proceedings of the SIGCHI 2006 Conference on Human Factors in Computing System*. New York: ACM Press, pp. 437-446.

Prior, N. (2008). Having One's Tate and Eating It: Transformations of the Museum in a Hypermodern Era. In McClellan, A. (ed.), *Art and its Publics: Museum Studies at the Millenium*. Hoboken, NJ: Wiley-Blackwell, pp. 51-76.

Rogoff, I. (2010, Summer). Practicing research: Singularising knowledge. *MaHKUzine,* 9: 37-42. http://www.mahku.nl/download/maHKUzine09_web.pdf (accessed 1 October 2011).

Opportunity and accountability in the 'eResearch push'

Bellamy, Craig
txt@craigbellamy.net
Victorian eResearch Strategic Initiative (VeRSI), Australia

In this paper I will present an examination of the institutional and epistemological tensions between particular national 'eResearch' infrastructure agendas (or 'eScience', 'cyber-infrastructure') and particular aspects of the digital humanities. Whilst much of the digital humanities positions itself within the research 'infrastructures' of the humanities (journals, academic departments, conferences, libraries, and sober ethics committees) – and is partly responsible for building the 'human capital' to work in the humanities – eResearch has largely emerged outside of the perspectives and training of the digital humanities, primarily driven by a 'big science' agenda (ie. an emphasis on mass data storage and infrastructures that largely support scientific methods and ways of collaborating). This has created numerous complexities for the digital humanities, particularly in the UK and Australia where it may, for better or worse, be emerging as a competing set of discourses and practices to the digital humanities (Examples will be given in the presentation).

Admittedly, the eResearch agenda *has* created many opportunities for research in the humanities, however, the way in which this agenda has been institutionalised in a number of countries means that it doesn't always serve the needs of the humanities. This is because eResearch largely exists *outside* of humanities research structures and is principally measured and driven by different accountability metrics. Its technical development also sits within rarefied institutional hierarchies between 'service staff and researcher', and biases a mistaken industrial utility; neither of which seem particularly useful for the contemporary challenges of digital humanities nor humanities research. As Geoffrey Rockwell states:

[there are] dangers in general and especially the issue of the turn from research to research infrastructure [...] we need to be careful about defining the difference and avoid moving into the realm of infrastructure [...] those things we are still studying (Rockwell 2010: 5).

Through presenting key examples, in this presentation I will broadly explore the institutionalisation of eResearch-infrastructures in the humanities over the past decade, and in particular, reference the countries in which I have the most experience, Australia and the UK. The landscape in Australia is particularly problematic as unlike other countries, the eResearch agenda is still at its height whilst the digital humanities is still developing its own institutional addenda.

The main tension appears to be a 'two cultures' one; it is the misunderstandings between applied computing, largely focused upon meeting the many practical needs of large-scale scientific endeavours, and the digital humanities that has its own particular understandings of the efficacy of computing within its heterogeneous research endeavours. Many infrastructure investors unavoidably claim a 'research enabling' or even a research pedigree for their work, but the exact nature of this research and how it helps us understand humans society and culture is, on occasions, yet to be determined (and this is far from an easy task and is largely an experimental practice; rarely a utilitarian one). Plus the institutional positioning of eResearch infrastructure in university service divisions, remote national services, and monolithic government and science-led programmes, means that the tradition of critique, and synthesis of eResearch infrastructure within contemporary digital humanities scholarship, is barely possible.

1. What can be done?

As a trained historian and long-time digital humanities advocate who has benefited from investments in eResearch – and indeed, I am employed by a particularly flexible strategic eResearch programme in the State of Victoria in Australia – I caution against retreating too eagerly from the 'infrastructure turn' as there are still healthy opportunities in many countries between the cracks of otherwise clumsy agendas. However these opportunities need to be positioned within a research-led digital humanities agenda and not a science led-agenda (and there is a perhaps a dark side to the **#alt-ac** movement in the United States if new digital outputs are not well supported within a humanities research setting) (Nowviskie 2011).

Perhaps a better approach for the humanities than cambering to a science led 'e-infrastructure' funding model, along with its often abrasive tectonic plates of incongruous collaborations, would be to lobby for the funding model to strengthen digital humanities research. The digital humanities has a sophisticated international network of centres, undergraduate and graduate degrees, associations, conferences, journals, and research accountability structures that are largely *internal* to the humanities and is thus

much better equipped to lead computing in the humanities than eResearch (and there are some positive institutional developments in this direction). And if led by the digital humanities, new research infrastructures such as data and text centres, virtual environments, and digital libraries would be more relevant to humanities research, thus insuring their long term sustainability. But this would require 'e-infrastructure' to be institutionalised in a much more responsive way; in a way that isn't unequally coupled with the needs of science. Again Geoffrey Rockwell states:

Perhaps things like the Text Encoding Initiative Guidelines are the real infrastructure of humanities computing, and the consortia like the TEI are the future of light and shared infrastructure maintenance (Rockwell 2010: 5).

I would like to think that this is because the TEI and derivatives such as EpiDoc exist within a deeply scholarly and vibrant international research culture that is both embedded within and accountable to humanities research; this is not always the case with eResearch infrastructure. However, for the digital humanities to take a greater lead in terms of guiding the implementation of eResearch infrastructure, in its various institutional settings, would require the digital humanities to be strengthened institutionally to rise to the challenge, especially in countries where 'eResearch' is much stronger than the digital humanities. All infrastructure, despite its veneer of utilitarian simplicity, is 'among the complex and expensive things that society creates' (Hauser 2011). 'e-Infrastructure' for the humanities may provide opportunities, but aspects of the present model in various countries lacks a complex humanities research environment and is wedded to an empirical, engineering, and industrial instrumentalism that is often at odds or even hostile to the humanities. It is not that eResearch does not do some things very well, it is the promise of research that it doesn't do particularly well. The goals of eResearch infrastructures are often so monumental; that they should perhaps be a set of research questions in themselves rather than practical goals. Based on my experience within eResearch and the digital humanities I will propose a number of alternative directions in this presentation and seek insight from others in the audience from various national and institutional contexts.

References

Barjak, F., J. Lane, M. Poschen, R. Proctor, S. Robinson, and G. Weigand (2010). e-Infrastructure adoption in the social sciences and humanities: cross-national evidence from the AVROSS survey. Inf*ormation, Communication and Society* 13(5): 635-651.

Capshew, J. H., and K. A. Rader (1992). Big science: price to the present. *Osiris* 7: 2-25, http://www.jstor.org/stable/301765

Edwards, P., S. Jackson, J. Bowker, and K. Knobel (January, 2007). Understanding infrastructure: dynamics, tension, design. *Report of a Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures, Rice University*, http://cohesion.rice.edu/Conferences/Hewlett/emplibrary/UI_Final_Report.pdf

Hauser, Th. (2011). Cyberinfrastructure and data management (presentation). Research Computing, University of Bolder, Colorado, http://www.stonesoup.org/meetings/1106/work3.pres/2b-CI-DM-TH.htm

Katz, R. N. (2008). The tower and the cloud: higher education in the age of cloud computing educause, http://net.educause.edu/ir/library/pdf/PUB7202.pdf

Nowviskie, B. (2011). #alt-ac Alternative academic careers for humanities scholars, http://nowviskie.org/2010/alt-ac/

Rockwell, G. (14 May 2010 **)**. As Transparent as Infrastructure: On the research of cyberinfrastructure in the humanities'. *Connexions*, http://cnx.org/content/m34315/1.2/

http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf

Svennsson, P. (2011). From optical fibre to conceptual cyberinfrastucture. *DHQ: Digital Humanities Quarterly* 5(1), http://digitalhumanities.org/dhq/vol/5/1/000090/000090.html

Turner, G. (September, 2008), Report from the HASS capability workshop, Old Canberra House, Australian National University, 15 August 2008 (unpublished report).

Turner, G. (2009). Towards and Australian Humanities Digital Archive. A report of a scoping study of the establishment of a national digital research resource for the humanities, Australian Academy of the Humanities, http://www.humanities.org.au/Portals/0/documents/Policy/Research/Towards_An_Australian_Digital_Humanities_Archive.pdf

Unsworth, J., Chair (2006) 'Our cultural commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, American council of learned societies.

Connecting European Women Writers. The Selma Lagerlöf Archive and Women Writers Database

Bergenmar, Jenny
jenny.bergenmar@lir.gu.se
University of Gothenburg, Sweden

Olsson, Leif-Jöran
leif-joran.olsson@svenska.gu.se
University of Gothenburg, Sweden

1. Introduction

Arguments for investing time and resources in digital scholarly editions are often based upon the cultural and literary dignity of the authorship (Shillingsburg 2004). In the case of Selma Lagerlöf, as in the case of for example Ibsen, one can argue that her authorship is one of the most important in Swedish literary history and one of the most well known internationally of all Swedish authors. Following this line of argument, literary scholarly editions tend to be focused on single, canonized authorships, and since letters to the author and translations of his/her work are not usually included, the authorship is represented in splendid isolation.

2. From Work to Reception

We will discuss why and how the Selma Lagerlöf Archive (SLA), which makes the Selma Lagerlöf collection at the National Library of Sweden accessible and provides digital scholarly editions of her work, aims to create a broader historical context, including the intensive contacts within networks of European women translators and writers which were important for her success in Europe. What the research on Selma Lagerlöf until now lacks, is a more complete mapping of these patterns of dissemination, and also, of course, evidence of reception of her work in different countries. This has consequences for the Selma Lagerlöf Archive too. Charting the translations is not a small task, and investigating legal rights to be able to publish these translations online is perhaps even more difficult. This also needs to be a collective effort, since the source material is spread over Europe and written in different languages.

3. Collaborative Methods

The Selma Lagerlöf Archive's participation in the COST Action ISO901 'Women Writers in History: Towards a New Understanding of European Literary Culture' (2009-2013), facilitates the discovery of unknown or forgotten connections between European women in the literary field. By standardizing 'author' and 'work', but not 'reception', the Selma Lagerlöf Archive and the *WomenWriters* database (`http://www.databasewomenwriters.nl/`) can exchange information on these levels. 'Reception' is expressed as a relation between 'author' and 'work'. Thus the 'author' can function in different roles – as 'reader', 'critic', 'admirer' etc. 'Work' can be specified by type – 'letter', obituary' etc. This is a flexible model, since it covers many different kinds of research material, and there is no need to agree on a common definition of 'reception', which would not be adequate for all research material and research questions of the projects using the database.

By way of microservices that are currently being developed, the metadata in the xml-databases of SLA, can be shared within the agreed frame of Women Writers-database, in addition to our other APIs. This collaboration benefits the SLA in two ways. Considering the international reach of Women Writers-database, it may function as an important output to the otherwise limited target group of SLA. SLA can also collect the relevant data on translations or receptions registered in the *WomenWriters* database and thereby making the time consuming task of finding and registering this information unnecessary.

The development of microservices follows the collective work carried out in COST Action 'Interedition' (van Zundert 2011), where the Selma Lagerlöf Archive also participated (although Sweden was not formally a member). The modular microservices developed will vary in implementation language and platform, but will be web-accessible to other services using REST-like interfaces. As for now, the Selma Lagerlöf Archive has not yet developed any standard format tool for visualizing the connections between European women writers, but evidently this will be an important task for the future and a powerful method for rewriting the European map of literary history.

4. Comparing Translations and Reception Documents

An inventory of translations is a first step towards discovering the mechanisms of dissemination of this – and other – authorships. But a qualitative analysis of the translations to different languages and of reception documents such as reviews, are also

necessary for understanding the different national and cultural contexts in which the authorship was used. A possible tool in the qualitative analysis of the European translations of Selma Lagerlöf's work is the development of Juxta and CollateX, which was also previously carried out in 'Interedition'. The modularisation initiated in the previous COST Action 'Open Scholarly Communities on the Web' known as the Gothenburg model, had the objective to transform the collation from a 'black box' to a clear process with discrete chainable steps.

The web service versions of the collation step of Juxta and CollateX are primarily used interactively for comparing variants of Selma Lagerlöf's work online within our own interface. But in addition to this, it allows for any electronic texts to be compared, thus creating the possibility of comparing different translations. Collation of translations is usually as laborious as collation of variants, but in this way we can more effectively trace how Selma Lagerlöf's work has been transformed in different translations. Since the German translations were often used as sources for other European translations (for example to Czech and Russian), it is important to examine to what extent the translations were rewritten to accommodate to the cultural climate. For example, some of the German translations were explicitly directed towards the *Heimatkunstbewegung,* which can be described as a precedent to the Blut und Boden-literature in the 1930s (Ljung Svensson 2011). It is a question still unexplored, how this ideological rhetoric was translated into other languages, nations and ideologies.

The possibility of comparing electronic texts through collation may also be applied for other purposes than showing the deviations between two editions or translations. It can be used to discover similarities between texts. One ambition of the COST Action 'Women Writers in History' is to bring data together, which has, up until now, been dispersed among different archives, physical and digital. This allows for a quantitative approach in estimating the reach of female writers activities in Europe. The collected data will also be used qualitatively, for example in exploring patterns in the reception of women writers by critics, and if these patterns correspond to specific themes in the literary texts. Provided that the reception documents (for example reviews) are in the same language and in an electronic format, reviews might also be compared through the microservices, in order to find out if certain ways of describing female authors, or focusing on certain themes in their texts, recur. The same method – although still at an experimental stage – may be used to compare texts by different female authors as a shortcut to discovering common topics or themes. The advantage of this method is that the researchers don't have to know beforehand what to search for, as when you search electronic texts for certain keywords. Similarities may appear that researchers would not have imagined.

5. Conclusion

In order to revitalize the field of scholarly editing, the author's work must be presented as nodes in cultural networks, involving different contexts and countries, where the texts were received, used and transformed, as well as different agents and institutions, participating in the dissemination of the authorship. This kind of scholarly editing is feasible if existing databases and tools are used to share and compare data. This method of working emphasizes both the text as a social object and digital scholarly edition as a collaborative activity (Robinson 2009).

References

Ljung Svensson, A. S. (2011). *Jordens dotter. Selma Lagerlöf och den tyska hembygdslitteraturen/Die Tochter der Erde. Selma Lagerlöf und die deutsche Heimatliteratur um 1900,* Göteborg och Stockholm: Makadam förlag.

Robinson, P. M. W. (2009). The Ends of Editing. *Digital Humanities Quarterly* 3(3). http://digitalhumanities.org/dhq/vol/3/3/000051/000051.html .

Shillingsburg, P. (2004). Hagiolatry, Cultural Engineering, Monument Building, and Other Functions of Scholarly Editing. In R. Mondiano, L. F. Searle, and P. Shillingsburg (eds.), *Voice, Text, Hypertext. Emerging Practices in Textual Studies*. Seattle, Washington: U of Washington P, pp. 412-423.

van Zundert, J., et al. (2011). Interedition: Principles, Practice and Products of an Open Collaborative Development Model for Digital Scholarly Editions. *Book of Abstracts DH 2011*, June 19-22, Stanford, http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-227.xml .

Stylometric Analysis of Chinese Buddhist texts: Do different Chinese translations of the 'Gandhavyūha' reflect stylistic features that are typical for their age?

Bingenheimer, Marcus
m.bingenheimer@gmail.com
Temple University, USA

Hung, Jen-Jou
Jenjou.Hung@gmail.com
Dharma Drum Buddhist College, Taiwan

Hsieh, Cheng-en
chengen.xie@ddbc.edu.tw
Dharma Drum Buddhist College, Taiwan

Buddhist Hybrid Chinese is a form of Classical Chinese that is used in the translation of Buddhist scriptures from Indian languages to Chinese between the 2nd and the 11th century CE. It differs from standard Classical Chinese of the period in vocabulary (esp. the use of compounds and transcriptions of Indian terms), register (esp. the inclusion of vernacular elements), genre (esp. the use of prosimetry), and rarely even syntax (at times imitating the syntax of the Indian original). Texts in Buddhist Hybrid Chinese are central to all traditions of East Asian Buddhism, which is practiced in China, Korea, Japan and Vietnam.

No comprehensive linguistic description of Buddhist Hybrid Chinese has been attempted so far and perhaps never will, due to the great diversity between translation idioms that at times use different Chinese terms for one single Indian term, and in other cases one single Chinese term for different Indian terms. In as far as Buddhist Hybrid Chinese has been described, the research generally concentrates on grammatical particles (e.g. Yu 1993), single texts (e.g. Karashima 1994), single terms (e.g. Pelliot 1933) or even single characters (e.g. Pulleyblank 1965). The stylometric study of Buddhist Hybrid Chinese – as that of Classical Chinese in general – has only just begun. Only since 2002, when the Chinese Buddhist Electronic Text Association (CBETA) distributed the texts in XML are the canonical texts available in a reliable digital edition.[1]

The Chinese Buddhist canon was printed first in the 10th century and regarding texts before that date its contents have been relatively consistent since then. The currently most widely referenced edition (the Taishō edition, published 1924-34) is based on a Korean edition from the 14th century. It contains ca. 2200 texts from India and China. Due to insufficient and unreliable bibliographic information for texts translated before the 7th century, the attributions to individual translators – where they exist at all – are often questionable. This again has an impact on the dating of the early texts, as they are usually dated via their translator(s). Since most stylometric methods, including those for authorship attribution, were developed for European languages, they often rely on easily parsable word-boundaries, which in the case of Buddhist Hybrid Chinese do not exist. Our wider aim is therefore to develop methods to identify stylistic clues for certain eras in Chinese translations from Indian texts. Can we, based on stylometric features, find a way to date Chinese Buddhist texts or at least to meaningfully corroborate or contradict traditional attributions?

In this study we have compared three translations of the same text, i.e. the *Gandhavyūha* section (ch. Ru fajie pin 入法界品) of the *Avatamsakamsūtra* (ch. Huayan jing 華嚴經). The *Gandhavyūha,* which contains a long narrative of the quest of the young man Sudhana to visit spiritual teachers, was translated into Chinese three times:

T. 278 by Buddhabhadra 佛陀拔陀羅et. al. (Chang'an 418-20 CE)

T. 279 by Śiksānanda實叉難陀 et. al. (Chang'an 695-699 CE)

T. 293 by Prajña 般若et. al. (Chang'an 796-8 CE)

Our task in this particular case was to develop an algorithm that can demonstrate that the T.278 was translated three to four hundred years earlier than T.279 and T.293, and show which of its features can identify a translation idiom that is earlier or at least different from that of T.279 and T.293. Can it be shown that the two Tang dynasty translations (T.279 and T.293) truly are more closely related to each other than to the translation from the Eastern Jin (T.278)?

Our approach here combines a general statistical weighing of n-grams with a focus on grammatical particles (*xuci* 虛詞). A ranking of their importance for our corpus must factor in occurrence as well as variance. The algorithm must also provide for the fact that characters that function as particles can also be used in nominal or verbal compounds. These instances must be filtered out by applying a list of compounds from a large dictionary of Buddhist terms

(Soothill & Hodous 1937). The algorithm for this is developed in the first section.

The following sections describe the sampling procedure and the preparation of the corpus. Although ostensibly all versions of the same Indian text, the three translations differ greatly in length, mainly because the volume of the Indian *Gandhavyūha* expanded between the 5th and the 8th centuries. To counter this problem and to produce enough samples for our analysis, each translation will be divided into sub-divisions of equal length. Then, the frequencies of grammatical particles in these divisions will be calculated and used for defining the stylometric profile of the three translations. We will therefore deal with text clusters on which we can use Principle Component Analysis (PCA), which we have used in a previous study (Hung, Bingenheimer, Wiles 2010). Using PCA on the extracted profiles and plotting the values of first and second components in 2-d charts we are able to discern clearly that T.279 and T.293 are closer to each other and more distant/different from T.278. The two Tang dynasty translations seem indeed to differ from the Jin dynasty translation in its use of particles, and the first and second component of the PCA analysis result shows, which particles create the distinction.

Thus stylometric analysis can give us a better understanding of the translation styles of Buddhabhadra, Śiksānanda and Prajña. All translators have several other translations attributed to them and comparing their *Gandhavyūha* translation to the rest of their corpus, and then again their corpora with each other, could in the future help us to improve our algorithms that ideally would be able to describe and demarcate the work of different translators. The general aim is to get a first handle on the quantitative analysis of the corpus written in Buddhist Hybrid Chinese and extract significant features, which can then be used for a more accurate linguistic description of the idiom.

What the analysis does not account for is changes in the Indian text. The Eastern Jin translation was translated from a somewhat different version of the Indian text than the two Tang translations 300-400 years later. This does, however, not impact our analysis. It is possible to distinguish how grammatical particles were used by different translators, because they reflect different styles of Buddhist Hybrid Chinese, which is what we are looking to describe. Even taking into account that the Sanskrit text of the *Gandhavyūha* has evolved between the 5th and the 8th century, its grammar could not have changed to the degree as there are changes in the translation idiom.

References

Hung, J.-J., M. Bingenheimer, and S. Wiles (2010). Quantitative Evidence for a Hypothesis regarding the Attribution of early Buddhist Translations. *Literary and Linguistic Computing* 25(1): 119-134.

Karashima, S. 幸嶋靜志 (1994). *Chōagonkyō no gengo no kenkyū – onshago bunseki o chushin toshite* 長阿含經の原語の研究 – 音写語分析を中心として[Study of the language of the Chinese Dīrghāgama]. Tokyo: Hirakawa平河出版社.

Pelliot, P. (1933). Pāpīyān > 波旬 Po-siun. *T'oung Pao* (Sec. Series), 30 (1-2): 85-99.

Pulleyblank, E. G.. (1965). The Transcription of Sanskrit K and Kh in Chinese *Asia Major* 11 (2): 199-210.

Soothill, W. E. and L. Hodous (1937). *A Dictionary of Chinese Buddhist Terms*. London: Kegan. [Reprint Delhi: Motilal, 1994]. Digital as XML/TEI file at `http://buddhistinformatics.ddbc.edu.tw/glossaries/`.

Yu, L. 俞理明 (1993). *Fojing wenxian yuyan* 佛経文献語言 [The Language of the Buddhist Scriptures]. Chengdu: Bashu shushe巴蜀書社.

Notes

1. The CBETA edition is an openly available digital edition of the Chinese Buddhist Canon (the texts can be downloaded in various formats at `http://www.cbeta.org/`.

Information Extraction on Noisy Texts for Historical Research

Blanke, Tobias
tobias.blanke@kcl.ac.uk
King's College London, UK

Bryant, Michael
michael.bryant@kcl.ac.uk
King's College London, UK

Speck, Reto
reto.speck@kcl.ac.uk
King's College London, UK

Kristel, Conny
c.kristel@niod.knaw.nl
NIOD, Amsterdam, The Netherlands

The European Holocaust Research Infrastructure (EHRI)[1] project aims to create a sustainable Holocaust Research Infrastructure that will bring together documentary evidence from dispersed archives for historical research. EHRI involves 20 partner organisations in 13 countries. It aims to provide open access to Holocaust material such as documents, objects, photos, film and art. One of the challenges of the project is that the dispersed archives of interest to EHRI often do not have the means to sufficiently digitise their resources. Even if the resources are digitally available, they remain inaccessible to searching and browsing by researchers.

For EHRI, we investigated how we can use open source OCR infrastructure developed at King's College London for the Ocropodium project (Bryant et al. 2010) so that its output can feed the semantic extraction of metadata useful for research discovery and analysis. Current commercial OCR technology does not serve well such specific research interests in historical document collections, as it cannot be easily customised. Most commercial OCR software products are proprietary 'black boxes' which provide digitisation staff with little scope for understanding their behaviour and customising parameters under which they run. At the source level, there is a marked reluctance of OCR software manufacturers to allow access to their code even in a collaborative environment.

In the context of Ocropodium, we developed a workflow tool Ocropodium Web Processing (OWP), with which various open source OCR tools can be combined to create the best possible OCR solution for specific document types. The OWP workflow environment is based on the principles of visual programming environments (Cox & Gauvin 2011). It allows the user to build custom workflows comprised of discrete processes. The workflows need not be purely linear; instead they take the form of a graph, specifically a directed acyclic graph (DAG). The DAG is comprised of connected nodes, which each perform a discrete function. Nodes can be thought of as functions which take one or more inputs and evaluate these to produce a result. Thus, OWP allows archive staff to embed further services beyond OCR. For this paper, we experimented with information extraction (IE) services to semantically enrich archival descriptions. We present our experiments to evaluate how common off-the-shelf IE services behave against potentially noisy OCR'd texts.

Semantically enriched library and archive federations have recently become an important part of research in digital libraries (Kruk & McDaniel 2009). Especially so, as research users often have more demands on semantics than is generally provided by archival metadata. For instance, in archival finding aids place names are often only mentioned in free-form narrative text and not especially indicated in controlled access points for places. Researchers would like to search for these locations. Place name extraction from the descriptions might support this. For the future EHRI infrastructure, we want to use IE services to enrich the researchers' experience. In our experiments, we concentrated on extracting names and places facets, both immensely important for Holocaust research.

Our experiments demonstrate the principal workflow using IE tools. In our proof-of-concepts, we did not address larger problems of IE from historical texts, which are manifold. As for OCRing, off-the-shelf commercial IE software has often problems with delivering acceptable results here. For the problems of extracting semantic information from low quality textual data, please compare (Packer et al. 2010), while (Warner & Clough, 2009) describe plans for a larger extraction project from the UK National Archives. Instead of concentrating on improving the IE itself, we were mainly interested in mapping and evaluating the current state-of-the-art. For the presentation at DH2012, we will deliver exact evaluations of using off-the-shelf IE tools against various levels of underlying 'dirty' OCR text, commonly encountered in OCRed historical resources.

Even the off-the-shelf IE tools have already delivered encouraging results. For instance, we OCR'd PDF files of survivor testimonies sent by the Wiener Library,[2] an EHRI partner. The documents were typical fairly low resolution (612x790) grey-scale scans of typed documents. Due to the low resolution

we needed to do some advanced preprocessing of the images and scaled the images by a factor of four using an anti-aliasing filter to approximate a typical resolution from 300 DPI scan and finally binarised them (converting a colour or grey-scale image to one containing only black and white). After binarisation, additional filters were applied to deskew the images and remove edge noise.

The resulting transcript produced by the open source Tesseract OCR engine (>Smith 2007) was fairly low quality, with around 90% character accuracy. We ignored further possible improvements by combining Tesseract with more advanced binarisation and pre-processing tools. As said, in this particular experiment, we were interested how standard IE services would react to low-quality textual input. We used the off-the-shelf OpenCalais service by ThompsonReuters to extract semantic information (Goddard & Byrne 2010). Even this standard setup has proven to produce useful results. OpenCalais proved successful at detecting the presence of a personal name in the transcript, even when the OCR was imperfect. For example, it detected that 'Dmulaltsr Tappe' (Dr. Walter Tappe) was a name. It also marked up several instances of places, such as Berlin and Wilmersdorf. Other incorrectly OCR'd locations such as 'Slchsischestraeae' (Schlesisches) were also marked up as places, due to the (correctly OCR'd) preceding phrase 'lived in'. Further semantic data marked up by OpenCalais included positions ('lawyer', 'auditor', 'actor') and industry terms ('food'). In several OCR transcripts it detected the topic as 'politics'. Social tags given included 'Berlin', 'Geography' and 'Geography of Europe'. We repeated our experiment with other IE tools such as the open source Annie tool from GATE (>Bontcheva et al., 2002) and could further improve our results, especially as we could rely on advanced gazetteers of place and person names of the Holocaust that are based on long-running community projects.[3]

These initial successful results will mean that we will develop our OCRing of finding aids into a full service in the context of the EHRI project. We think we can thereby significantly enhance the research experience of using historical archives.

References

Bontcheva, K., H. Cunningham, D. Maynard, V. Tablan, and H. Saggion (2002). Developing reusable and robust language processing components for information systems using GATE. In *Database and Expert Systems Applications. Proceedings 13th International Workshop*. Berlin: Springer, pp. 223-227.

Bryant, M., T. Blanke, M. Hedges, and R. Palmer (2010). Open Source Historical OCR: The OCRopodium Project Research and Advanced Technology for Digital Libraries. In M. Lalmas et al. (eds.), *Research and advanced technology for digital libraries*. Berlin: Springer.

Cox, P. T., and S. Gauvin (2011). Controlled dataflow visual programming languages. *Proceedings of the 2011 Visual Information Communication - International Symposium*. Hong Kong: ACM.

Goddard, L., and G. Byrne (2010). Linked Data tools: Semantic Web for the masses. *First Monday* 15, 1.

Kruk, S. R., and B. McDaniel (2009). *Semantic Digital Libraries*. Berlin: Springer.

Packer, T. L., J. F. Lutes, A. P. Stewart, D. W. Embley, E. K. Ringger, K. D. Seppi, and L. S. Jensen (2010). Extracting person names from diverse and noisy OCR text. *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. Toronto, ON, Canada: ACM.

Smith, R. (2007). An Overview of the Tesseract OCR Engine. Document Analysis and Recognition. Proceedings 9th International Conference. Berlin:Springer, pp. 629-633.

Warner, A., and P. Clough (2009). *A Proposal for Space Exploration at The National Archives. 2011*. Available: http://ir.shef.ac.uk/cloughie/papers/York2009.pdf

Notes

1. http://www.ehri-project.eu/
2. http://www.wienerlibrary.co.uk/
3. http://resources.ushmm.org/hsv/

Modeling Gender: The 'Rise and Rise' of the Australian Woman Novelist

Bode, Katherine
katherine.bode@anu.edu.au
Australian National University, Australia

1. Overview

Drawing on the *AustLit: the Australian Literature Resource* database,[1] this paper uses quantitative and computational methods to offer new insights into literary history: in this case, gender trends in the contemporary Australian novel field.

AustLit is currently the most comprehensive and extensive online bibliography of a national literature. Created in 2001, *AustLit* merged a number of existing specialist databases and bibliographies, and has subsequently involved well over a hundred individual researchers – from multiple Australian universities and the National Library of Australia – in an effort 'to correct unevenness and gaps in bibliographical coverage' and continually update the collection. The database has received significant government and institutional funding and support, and includes bibliographical details on over 700,000 works of Australian literature and secondary works on that literature. In addition to its high degree of comprehensiveness, *AustLit* is well suited to quantitative and computational analysis due to its construction according to established bibliographical standards and fields, which gives the data a high degree of consistency across the collection. This study focuses on Australian novels, the most comprehensively recorded aspect of *AustLit*. Even so, *AustLit* notes that its 'coverage of some popular fiction genres such as westerns and romances, and of self-published works, is representative rather than full'.[2]

3. Methodology

Although *AustLit* is generally used by researches for individual queries, it also has a guided search function that allows a large number of results to be extracted as tagged text. Using this function I constructed searches that returned information, including the author, date and place of publication and genre, for all Australian novels published between 1945 and 2009. I then used Apple OS X command lines to organize the data before exporting it to CSV format. Once in this format, I searched all the author names in *AustLit* to discover any pseudonyms and to determine if the authors were male or female. I explored the data in Excel, using the pivot table function to query the data and the chart function to visualize of trends over time.

My approach to this data is based on a combination of book history's awareness of the contingency and limitations of data, and the paradigm of computer modeling as outlined by Willard McCarty in *Humanities Computing*.[3] Although there has been virtually no conversation between the two fields, the methodological underpinnings of book history and the digital humanities together signal important directions for developing a critical and theoretically aware approach to working with data in the humanities. Quantitative book historians explicitly acknowledge the limited and mediated nature of data, and provide clear and detailed accounts of the origins, biases and limitations of the literary historical data they employ. In presenting their results, such scholars insist that quantitative results provide a particular perspective, suited to the investigation of some aspects of the literary field but not others, and offer indications rather than proof of literary trends. As Robert Darnton writes, 'In struggling with [literary data], the historian works like a diagnostician who searches for patterns in symptoms rather than a physicist who turns hard data into firm conclusions'.[4]

Although developed for use with language, McCarty notion of modeling – as an exploratory and experimental practice, aimed not at producing final and definitive answers but at enabling a process of investigation and speculation – can be adapted for analysis and visualization of literary historical data, and employed in ways that resonate productively with quantitative book historical methods. Specifically, I used a modeling approach to explore a series of hypotheses about trends in Australian literary history. For instance, the gender trends in authorship in a given period might lead me to suppose a particular hypothesis (such as the influence of second-wave feminism on constructions of authorship). In enabling modification of or experimentation with data – for instance, subtracting particular publishers or genres – modeling provides a way of testing such ideas, with the results of different models either fulfilling and strengthening, or challenging, the original hypothesis. This methodological framework shows how relatively simple computational processes in Excel can enable a process of thinking with the computer that reveals trends, conjunctions and connections that could not otherwise be perceived.

This combination of book historical and digital humanities methods avoids what I see as limitations in other 'distant readings' of the literary field,

including Franco Moretti's well-known 'experiment' in literary history: *Graphs, Maps, Trees*.[5] In particular, the skeptical approach to data encouraged by book history avoids the portrayal of quantitative results as objective and transparent. Meanwhile, McCarty's notion of modeling – as a means of thinking with the computer – avoids the assumption that the computer is a passive tool in the process of analysis and interpretation.

4. Results

Extracting data from this online archive, visualizing and modeling it, has allowed me to address questions of long-standing interest to literary scholars and to explore new and hitherto unrecognized trends in Australian literary history. This paper illustrates the potential of such research for literary studies with a case study that responds to the conference theme of 'digital diversity'. A major issue in literary historical discussions of 'diversity' is the presence – or absence – of women authors in the literary field. In contemporary literary studies (in Australia and other national contexts), such discussion has focused on the period from the late 1960s to the late 1980s, when second-wave feminism is said to have had its major impact. This period is cited as a time of substantial growth in women's presence in the literary field, leading to increased diversity and a deconstruction of the canon of male authors. Visualisations of the *AustLit* data allows me to explore whether this perception of gender trends is warranted, while the process of modeling enables investigation of the causes of the trend.

In relation to gender trends from the late 1960s to the late 1980s, the results of this study support the longstanding perception that women's involvement in the Australian novel field increased at this time (see Figure 1, showing the proportion of Australian novels by men and women from 1945 to 2009). However, the outcomes of the modeling process significantly complicate the standard interpretation of this trend, and the relationship between politics and the literary field it implies. Contextualising this growth in women's writing in relation to the genre of Australian novels shows that the largest area of growth was in romance fiction, a form of writing usually seen as inimical to second-wave feminism.

Another challenge to existing interpretations of the relationship between gender, politics and fiction emerges in the past two decades, when gender trends in authorship are explored in the context of gender trends in critical reception. In the 1970s and 1980s, even as romance fiction represented the largest proportion of women's publishing, critical attention to Australian women writers in newspapers and academic journals increased (see Figure 2, depicting the proportion of men and women among the top twenty most discussed Australian novelists in these different forums and overall). The proportion of Australian novels by women has continued to grow in the past two decades, such that they write well over half of all Australian novels published in the 2000s (see Figure 1), and the proportion of romance fiction in this field has declined. Yet in this same period, women writers have been less and less the subject of critical discussion in newspapers. These results resonate with those recently presented by Vida, an advocacy group for women writers, showing the same prevailing focus on male authors in the book reviews of a wide range of major American and European newspapers and magazines in 2010 and 2011, including *The Atlantic*, the *London Review of Books*, the *New York Review of Books* and the *Times Literary Supplement*.[6]

Although women continue to be well represented in academic discussion of Australian literature, the focus in that forum is on nineteenth-century authors, whereas both contemporary and historical male authors are discussed. Despite the impact of feminism on the Australian literary field, these results suggest that the longstanding perception of the great author as a man is still influential. Presence or representation is not, I argue, the only issue in determining the success of diversity in the literary field. Other constructions – relating to the genre in which fiction is written and gender trends in its reception – shape the field in ways that only come into view on the level of trends: when the literary field is approached as a system, using data-rich and computer-enabled methods.

Notes

1. AustLit. (2001-). AustLit. `http://www.austlit.edu.au/` (accessed 25 March 2012).
2. AustLit. (2001-). About Scope. `http://www.austlit.edu.au/about/scope` (accessed 25 March 2012).
3. McCarty, W. (2005). Humanities Computing. London: Palgrave Macmillan.
4. Darnton, R. (1996). The Forbidden Best-Sellers of Pre-Revolutionary France. New York: Norton, p. 169.
5. Moretti, F. (2005). Graphs, Maps, Trees: Abstract Models for a Literary History. London: Verso.
6. Vida: Women in Literary Arts (2010). The Count 2010. `http://www.vidaweb.org/the-2011-count` (accessed 19 June 2011); Vida: Women in Literary Arts (2011). The Count 2010. (accessed 9 March 2012).

Contextual factors in literary quality judgments: A quantitative analysis of an online writing community

Boot, Peter
peter.boot@huygens.knaw.nl
Huygens Institute of Netherlands History, The Netherlands

1. Introduction

Some literary works and authors acquire enduring reputation, whereas others do not. What causes the difference is a contested issue. Fishelov (2010) distinguishes between a 'beauty party' and a 'power party', where the beauty party argues for the existence of aesthetic qualities, inherent in the work, while the power party argues that social power structures determine the outcome of canonisation processes. As argued by Miall (2006), the question should be decided empirically. However, because there is no objective measure of literary quality (if there were one, the discussion would be over) empirical investigation of the issue has been fraught with difficulty. Both parties can lay some claim to empirical evidence. For the power party, Rosengren (1987: 295-325) and Van Rees and Vermunt (1996: 317-333), among others, have shown that reputations to some extent are made by newspaper reviewers. For the other side, some of the evidence that Miall adduces suggests that foregrounded linguistic features (deviant from normal, non-literary usage) may be responsible for some of the (emotional) response that literary reading evokes. It is long way, however, from that finding to showing that this sort of textual properties can actually explain a work's longer-term reputation.

As a way forward in this debate, I propose to look into online writing communities, i.e. websites where communities of amateur writers publish and discuss their poems and stories. The sites offer the potential for empirical and quantitative research into literary evaluation because they contain large quantities of works and responses to these works. Text analysis can be combined with statistical analysis of number of responses and ratings.

In this paper I will look at Verhalensite, a Dutch-language online writing community, unfortunately no longer active. At the basis of the paper is a site download that contains ca. 60,000 stories and poems, written by 2500 authors. The texts received 350,000 comments and the comments drew another 450,000 reactions. 150,000 comments were accompanied by a rating in terms of 1 to 5 stars. I reported about Verhalensite and its research potential in (Boot 2011a). In (Boot 2011b) I discuss available data and use them in an analysis of the possibility to predict long-term activity on the site based on activity and response in the first four weeks of site membership.

2. Context factors

I focus here on the role of context factors in determining commenters' response to the works published on the site. I look at pairs of one author and one commenter, where the commenter has responded to at least one of the author's works. I select only those pairs where both author and commenter have published at least ten works. This makes it possible to compute linguistic similarity between authors' and commenters' texts and the similarity of their genre preferences. There are 49,437 of author-commenter pairs fulfilling these requirements, and we can compute statistical relationships between the contextual factors and the response variables. As an example I give correlations between some of the context variables and the number of comments the commenter has given the author, presumably reflecting the commenter's opinion of the author's works:

Variable	Partial correlation	Variable grouping
Similarity creative texts	0.06	Author-commenter similarity
Similarity commentary texts	0.11	
Difference in numbers of poems	-0.11	
Same preferred genre	0.09	
Replies by author to comments (fraction)	0.07	Author networking activity
Comments by author on others	0.15	
Comments by author on commenter	0.57	
Ratings by author	0.13	
Prose writers	0.08	Text properties

Table 1: Partial correlations with number of comments from commenter to author, given author productivity, commenter's average number of comments per week and overlap in the commenter's and author's active periods on the site

The first four variables are measures of similarity between the author's and commenter's texts. The correlations show that the closer commenter and author are in terms of language, in the amount

of poetry they write, and in genre, the more likely the commenter is to respond to the author's works. I use two aspects of textual similarity. The similarity between authors' and commenters' creative texts (poems and stories) is computed using Linguistic Inquiry and Word Count (LIWC), argued by Pennebaker (e.g. Pennebaker & Ireland 2011: 34-48) to reflect important psychological properties. LIWC counts words in grammatical, psychological and semantic categories (e.g. pronouns, positive emotions, or work-related words). I use the cosine distance over all LIWC categories as a measure of textual similarity. For the commentary texts, I created a number of site-specific vocabulary lists corresponding to certain aspects of the response: compliments (good, nice, cool), greetings (hi, welcome, grtz), critical (not necessarily negative) discourse (dialogue, stanza, suspense), negative response (disappointing, boring, missing), and some others. I computed frequencies for each of these categories. Textual similarity between author and commenter was computed from the weighted differences between their frequencies in the site-specific categories.

The next four variables are measures of the author's networking activity, which was shown by Janssen (1998: 265-280) to be an important factor in determining the amount of critical attention that an author receives. They represent respectively the fraction of received comments that the author replies to (usually to say thank you, sometimes with a longer reaction), the number of times the author comments on others' works, the number of comments the author has given to the specific commenter (the quid-pro-quo effect), and the number of ratings that the author has given. All four have positive influence on the amount of comments. The next variable shows that if the shared genre is prose (and not poetry), the commenter is more likely to rate the author's works highly. At present I have no hypothesis as to why this should be the case.

In terms of the discussion about literary evaluation, the numbers are interesting because none of the context variables reflect properties of the texts (except for the prose/poetry distinction). They show that to some extent literary evaluation depends on linguistic agreement between author and evaluator and on an author's networking activities. While at a general level this is perhaps hardly surprising, it is interesting to obtain a measure of the (relative) strengths of these influences, including the quid-pro-quo effect. It is also interesting to note LIWC is to some extent able to capture linguistic distance between authors.

As a next step in this analysis, I hope to look more closely into alternative measures of textual similarity and the extent to which they can predict the number of comments exchanged in an author-commenter pair. One interesting candidate would be Latent Semantic Analysis, possibly using only words from certain semantic fields. It would also be interesting to investigate the differences between commenters in sensitivity to the discussed variables. The numbers given here reflect average behaviour, and it seems very likely that some commenters will care more than others for e.g. linguistic agreement between their own works and the works that they comment on. A follow-up question would then be whether it is possible to cluster commenters based on their preferences.

3. Discussion

To some extent a paper such as this one is unusual in the Digital Humanities context. Work in Digital Humanities has tended to focus on texts, their stylistics, authors, their sources, and the vehicles (editions) that bring texts to their modern students. Reader response has been largely ignored in our field. The discussion above shows some of the influence of contextual factors on literary evaluation, but many other sorts of analysis could and should be undertaken on the basis of this material. For one thing, I have not yet investigated the effects of 'power' (i.e. the influence of third persons on commenters' behaviour). Similarly, many analyses at the level of the individual work have yet to be performed. The existence of digital environments such as online writing communities has made these analyses feasible and can thus contribute to broadening the scope of Digital Humanities.

References

Boot, P. (2011a). Literary evaluation in online communities of writers and readers [to appear]. *Scholarly and Research Communication*.

Boot, P. (2011b). Predicting long-term activity in online writing communities: A Quantitative Analysis of Amateur Writing. Paper presented at Supporting Digital Humanities 2011, Copenhagen. http://crdo.up.univ-aix.fr/SLDRdata/doc/show/copenhagen/SDH-2011/submissions/sdh2011_submission_1.pdf (accessed 23 March 2012).

Fishelov, D. (2010). *Dialogues with/and great books: the dynamics of canon formation*. Eastbourne: Sussex Academic Press.

Janssen, S. (1998). Side-roads to success: The effect of sideline activities on the status of writers. *Poetics* 25(5): 265-280.

Miall, D. S. (2006). *Literary reading: empirical & theoretical studies*. New York: Peter Lang.

Pennebaker, J. W., and M. E. Ireland (2011). Using literature to understand authors: The case for computerized text analysis. *Scientific Study of Literature* 1(1): 34-48.

Rosengren, K. E. (1987) Literary criticism: Future invented. *Poetics* 16(3-4): 295-325.

Van Rees, K., and J. Vermunt (1996). Event history analysis of authors' reputation: Effects of critics' attention on debutants' careers. *Poetics* 23(5): 317-333.

Violence and the Digital Humanities Text as Pharmakon

Bradley, Adam James
adam.bradley@uwaterloo.ca
The University of Waterloo, Canada

There has long been a tendency of tools in the digital humanities to be objects through which literary scholars are intended to perform meta-analyses of texts. One of the inconsistencies with this act is that in the process of displaying a data visualization we destroy the actual text that is being studied. This act is a type of pharmakon, creative because it shifts the aesthetic of the text for re-interpretation and destructive because the original artifact is lost in this process of creation. By appropriating Rene Girard's definition of violence – namely that the attempt to represent texts visually and digitally is a form of mimetic rivalry that leads to violence in terms of the text itself, and using Denis Diderot's three fold vision of creation ('enthousiasme'), I will show that the mimetic nature of digital humanities tools creates an allowing condition for a type of violence that is both destructive (Girard's 'violence') and creative (Diderot's 'enthousiasme'). By juxtaposing the work of Girard and Diderot in this way it becomes possible to see digital humanities projects (specifically data visualization projects) as acts of violence. It was this line of reasoning that brought me to my primary research question: Is it possible to develop a visualization technique that does not destroy the original text in the process, one in which the digital humanist can be a creator in terms of Diderot's enthusiasme?

My paper will suggest a new method for visualizing texts that specifically addresses this question. The proposed method uses rigorous mathematics – developed with the aid of an interdisciplinary team of mathematicians and computer scientists at the University of Waterloo – that addresses this problem of violence and suggests that through interdisciplinary study and mathematical precision we can produce digital humanities tools that encode the original texts in their entirety without any violence towards the text. This project is an attempt to 'form a new kind of literary study absolutely comfortable with scientific methods yet completely suffused with the values of the humanities' as suggested by R. G. Potter in *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric.*

Using the resources of the University of Waterloo's Digital Media Lab, I created a data visualization technique that was originally intended to study form in poetry but has proven to have wide reaching applications. What differentiates this project from other similar work, such as the TextArc project, is that the algorithm created to model the visualization creates a one-to-one relationship between the visualization and the original text. What this means specifically is that unlike a meta-analysis tool, this project encodes the original layout and punctuation of a text within the algorithm that produces the visualization making it both a tool for meta-analysis and a keeper of the original text. This removes the need for an analyst that intimately understands the visualization and allows for a transition between the two states (aesthetic visualization and original text) without any hindrances. This effectively solves the problem of losing the text in the analysis, or more specifically, of traditional literary scholars complaints that they do not know how to interpret the visualization created by the digital humanities model. By including a one-to-one correspondence between the text and the visualization, the intuition that is being engaged in the criticism spawned by the visualization can take on a second dimension between the original text and the literary critic.

The actual program was written in python and uses an aesthetic inspired by avant-garde art, namely the shifting of perspectives to try to gain insight into the original text. This is accomplished by creating 3D graphs in an infinitely countable box that represents all of the possible words from a null string to an infinitely long word using the alphabet. Each word of the original text is then encoded with an algorithm that uses a combination of mathematics, developed by Georg Cantor, and a method of my own devising to create a unique set of co-ordinates for each word in the English language and graph them in 3d space. This creates a unique 3d object for every poem and engages a shifted aesthetic to try to find new insights into the arrangement of the original words. In terms of the actual aesthetics of the visualization the program claims nothing new, but the focus of this study was to try to devise a method for maintaining the text within the mathematics of data visualization and that is what I have done. I envision this project as axiomatic in nature and call in my paper for all data visualizations to live up to the standard of non-violence towards texts. I believe this exercise in mathematical rigor will inspire the digital humanist to consider the use of data visualization in a new way that will ultimately lead to better and more thoughtful tools for use in literary study.

Towards a bibliographic model of illustrations in the early modern illustrated book

Bradley, John
john.bradley@kcl.ac.uk
Department of Digital Humanities, King's College London, UK

Pigney, Stephen
s.pigney@gold.ac.uk
Goldsmith's College, London, UK

James Knapp, in his book *Illustrating the Past in Early Modern England* (Knapp 2003) points out that the History of the Book research has been 're-evaluating the relation of books to texts' (p 9). However, he then goes on to say that an 'overwhelmingly large amount of visual material' could be found 'on the pages of early modern English books' (p 37), and that these books tell a complex story that interconnects books not only with their texts but with their illustrations too. In this paper we will take up Knapp's observation about visual material in books and will present a bibliographic conceptual model of the relationship of images, texts, and their makers that could support the enhancement of our understanding of the creation of books from the early modern period.

We come by an interest in this topic honestly, since we were collaborators, with the project's Principal Investigator Professor Michael Hunter (Birkbeck) on a project called *British Printed Images until 1700* (BPI1700) that has resulted in what the front page of its website calls 'a database of thousands of prints and book illustrations from early modern Britain in fully-searchable form'. At present, bpi1700's materials come from the extensive collection of primarily single-sheet prints held by the British Museum's Department of Prints and Drawings and the Victoria and Albert Museum. Clearly, although the focus of bpi1700 has been on single-sheet prints up to now, the project is anxious to progress into the question of illustrations in printed books.

To understand the scale of such an undertaking, it is useful to realise that *many* early printed books contained illustrations. James Knapp claims that English printers in the mid 16th century commonly used many illustrations for books on historical subjects (Knapp 2003: 2). Luborsky and Imgram catalogued more than 5,000 woodcuts and engravings in English books printed between 1536

and 1603 alone, and a further study published by bpi1700 reveals thousands more for the period 1604-1640. There is plenty of material to work with. However, the numbers tell only a part of the story: what is the nature of the interaction between an intellectual study of the illustrations as prints with the intellectual study of the books as books, so that their parallel, but interconnected, histories can be more clearly recognised?

To help us explore these issues from a bibliographic perspective, we took up the approach described in the Functional Requirements for Bibliographic Records (*FRBR*) (Tillet 2004) to identifying the nature of the published illustrations and the books themselves. FRBR was developed primarily to clarify the bibliographic analysis of modern printed books, and is developed out of four different senses in which the word 'book' can be used. The base sense of book, as a physical item that takes up space in a library, is called an 'item'. A particular publication is called a 'manifestation', and roughly corresponds to an edition in modern publishing. FRBR identifies a higher level of structure in its entity 'expression', which, as Tillett describes it, organises a particular text by a specific language or media of delivery. Finally, the fourth and top level of abstraction is the 'work', the 'conceptual content that underlies all the linguistic versions', and represents things like the story being told, or the ideas in the author's head (Tillett 2004: 3).

Although bpi1700 was about printed images rather than modern printed books, FRBR had already proven to be useful in our growing understanding of the nature of prints. Bpi1700's 'impressions' were the actual paper objects held by the museums, and represented by digital surrogates in bpi1700, and these seemed to be modelled best by FRBR's 'item'. Metadata about them were provided for bpi1700 by our partner museums. Of course, many impressions would be printed from a single plate that had been created by the print maker. Thus, among all the holdings in the museums it was not unusual for the same print to turn up more than once. Out of this realisation came the concept of bpi1700's 'work' that fit FRBR's 'work' category: a bpi1700 work represented a single plate or woodcut and captured the creative act of the print maker when she or he created it. Finally, in early modern times a plate was often modified over time, and these modifications would then be witnessed in surviving prints. A print of Charles I might first show him as a boy, but another impression from what was manifestly the same plate would show him as a young man instead. Between the production of these two impressions, the image of the king on the plate had been modified. Bpi1700 called each of these surviving versions of the plate a 'state' (using terminology already in use by print cataloguers), and concluded that this most closely corresponded to FRBR's 'manifestation' category.

Early Modern printed books (in the time of the hand-press), on the other hand, although apparently closer in nature to what FRBR was developed for than print images might be, did not turn out to be entirely aligned to modern print practice for which FRBR had been designed. Works and editions were much more fluid than they are today and early modern books have therefore presented a challenge to the conventional cataloguing rules developed to handle more modern material (see the discussion of this in the context of 'rare books' in Moriarty 2004). We chose to examine the British Library's *English Short Title Catalogue* (ESTC 2011) for information about our books. It operates at two levels. Its top level consists of records that correspond broadly to book printings – this, in turn, maps broadly to FRBR's 'manifestation' category, since Moriarty (Moriarty 2004: 40) reminds us that in the hand-press era a reprinting of a book was really often like a new edition because it required the hand re-assembly of the type to print the pages again. Within each *printing* entry ESTC stores information about particular copies (which broadly correspond to FRBR's 'Items') which are grouped by holding institution. There is no explicit structural object corresponding to FRBR's 'Work' or 'Expression' in ESTC, although one can use the search forms provide to filter data by author and title, which conceptually allows for materials to be selected by something similar to FRBR's Work.

So, if we consider both the books' texts and their illustrations as both having an intellectual history that interests us, we have, then, two parallel and interconnected FRBR hierarchies in operation in early modern printed books. The issue is made more interesting by the fact that in the early modern period plates and/or woodcuts for illustrations were owned by printers and were often reused in more than one book. So, the same illustration might appear accompanying two completely different texts. Although a printer might decide to reprint a book, he or she could at the same time decide to change the illustrations that appeared in it to, say, increase the market for it. Furthermore, illustrations from plates had to be printed by a process separate from that used for the text, and these illustrations had to be inserted into the text pages by hand – a process subject to variability and error. Finally, because book binding was often arranged by the book buyer to be applied after she or he purchased the printed pages, book buyers often added illustrations they liked into printed books – a process called *extra-illustration* and described in more detail in the Folger library's online exhibition 'Extending the Book' (2010). Thus, a bibliographic model that accommodates both the text and the illustrations – that represents the

intellectual history of these two intertwined objects needs to accommodate a complex set of reasons why a set of illustrations (with their creative history) appears with the text of a particular book (with its separate history too).

In our presentation, we will examine some examples of illustrations in early modern books and show how this intertwining operated in practice. We believe that the parallel and intertwined existence of books and illustrations reveals a new and interesting story in the history of the book. We will present our model that represents the parallel FRBR-like nature of the images and the book with their complex web of possible interconnections, and we will discuss some of the possible significance of this for those interested in exploring the role of illustration in these early printed volumes.

(This paper is based on a presentation given by the authors at the Digital Humanities Day at Sheffield Hallam University on 13 December 2010, but will be extended beyond what was presented there with a more detailed presentation of the model for the data, as is suitable for a DH audience.)

References

Bbi1700. *British Printed Images before 1700*. Online at http://www.bpi1700.org.uk

ESTC (2011). *English Short Title Catalogue*. British Library. Online at http://estc.bl.uk/

Folger Shakespeare Library (2010). Extending the Book: The Art of Extra-Illustration. An exhibition January 28-May 25, 2010. Online version available at http://www.folger.edu/template.cfm?cid=3346

Knapp, J. A. (2003). *Illustrating the Past in Early Modern England: the representation of History in Printed Books*. Aldershot: Ashgate Publishing.

Luborsky, R., and E. Ingram (1998). *Guide to English Illustrated Books 1536-1603*. Tempe, AZ: MRTS.

Moriarty, K. S. (2004). *Descriptive Cataloging of Rare Materials (Books) and Its Predecessors: A History of Rare Book Cataloging Practice in the United States*. A Master's paper for the M.S. in L.S. degree. November, 2004. 100 pages. Advisor: Jerry D. Saye.

Tillett, B. (2004). *FRBR? A Conceptual Model for the Bibliographic Universe*. Washington: Library of Congress Cataloguing Distribution Service. Online at http://www.loc.gov/cds/downloads/FRBR.PDF

Automatic Mining of Valence Compounds for German: A Corpus-Based Approach

Brock, Anne
anne.brock@uni-tuebingen.de
University of Tuebingen, Germany

Henrich, Verena
verena.henrich@uni-tuebingen.de
University of Tuebingen, Germany

Hinrichs, Erhard
erhard.hinrichs@uni-tuebingen.de
University of Tuebingen, Germany

Versley, Yannick
yannick.versley@uni-tuebingen.de
University of Tuebingen, Germany

1. Introduction

The availability of large-scale text corpora in digital form and the availability of sophisticated analysis and querying tools have profoundly influenced linguistic research over the past three decades. The present paper uses this eHumanities methodology in order to automatically detect and analyze valence compounds for German. Valence compounds (in German: *Rektionskomposita*) such as *Autofahrer* 'car driver' have been subject to extensive research in the German linguistics. They are composed of a deverbal head (*Fahrer* 'driver') and a nominal non-head (*Auto* 'car'). As the corresponding verb *fahren* 'to drive', from which *Fahrer* is derived, governs its accusative object *Auto*, the compound *Autofahrer* is considered a valence compound.

The automatic detection and semantic interpretation of compounds constitutes an important aspect of text understanding for a language like German where compounding is a particularly productive means of word formation and accordingly occurs with high frequency. Baroni et al. (2002) report that almost half (47%) of the word types in the APA German news corpus, which they used as training material for a word prediction model for German, are compounds.

Due to their productivity, compounds in German do not form a closed class of words that can be listed in its entirety in a lexicon. Rather, as Lemnitzer (2007) has shown, new German compounds are coined daily, and some of them attain sufficient frequency to be eventually included in print dictionaries such as the

Duden. Novel compounds that are not yet listed in a dictionary pose a particular challenge for natural language processing systems that rely exclusively on dictionaries as their underlying knowledge source for word recognition.

Since the analysis of compounds constitutes a major challenge for the understanding of natural language text, the structural analysis and the semantic interpretation of compounds have received considerable attention in both theoretical and computational linguistics. Syntactic analysis of compounds focuses on the correct (left- vs. right-branching) bracketing of the constituent parts of a given compound, e.g., [[rock music] singer] vs. [deputy [music director]]. Research on the semantic interpretation of compounds has focused on the semantic relations that hold between the constituent parts of a compound. The present paper focuses entirely on the semantic interpretation of compounds; however see Henrich and Hinrichs (2011) for previous research on the syntactic analysis of nominal compounds in German.

2. Corpus-Based Experiments

The aim is to determine whether corpus evidence can form the basis for reliably predicting whether a given complex noun is a valence compound or not. For example, if we want to determine whether the complex noun *Taxifahrer* is a valence compound, we inspect a large corpus of German and investigate whether there is sufficient evidence in the corpus that the noun *Taxi* can be the object of the verb. The question of what exactly constitutes sufficient corpus evidence is of crucial importance. Three different measures were applied to answer this question:

1. Relative frequency: The percentage of the verb-object pairs in the corpus among all co-occurrences of the two words in the same sentence with any dependency relations,
2. The association score of mutual information for the verb-object pairs, and
3. The Log-Likelihood ratio for the verb-object pairs.

The measure in (1) above constitutes a simplified variant of the data-driven approach that Lapata (2002) applied for the purposes of automatically retrieving English valence compounds from the British National Corpus.

The starting point of the corpus-based experiments was a list of 22,897 German complex nouns and the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z).[1] This corpus consists of 200 Mio. words of newspaper articles taken from the *taz* ('die tageszeitung') and is thus sufficiently large to provide a reliable data source for the experiments to be conducted. The TüPP corpus was automatically parsed by the dependency parser MaltParser (Hall et al. 2006).

Each of the 22,897 potential valence compounds has been split into its deverbal head and its nominal modifier with the help of the morphological analyzer SMOR (Schmid et al. 2004). For example, the compound *Autofahrer* receives the analysis Auto<NN>fahren<V>er<SUFF><+NN> in SMOR. From the TüPP corpus, all occurrences of those object-verb pairs are extracted from those corpus sentences where either the verb in the sentence matches the deverbal head of the complex noun (e.g., fahren) or the accusative object of the sentence matches the nominal modifier (e.g., Auto) of the compound.

Figure 1 gives an example of the type of dependency analysis of the MaltParser from which the verb-object pairs are extracted. The dependency analysis represents the lexical tokens of the input sentence as nodes in the graph and connects them with vertices which are labeled by dependency relations. Recall that the MaltParser annotation is performed automatically and thus not 100% accurate. In the case of the sentence *Aber dann würde doch niemand mehr Auto fahren.* ('But then, no one would drive cars anymore.') shown in Fig. 1, *mehr* is erroneously attached to the noun *Auto* instead of to the noun *niemand*.

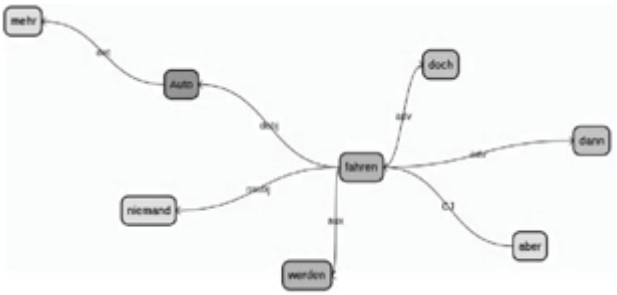


Figure 1: A MaltParser dependency graph for a TüPP corpus sentence Aber dann würde doch niemand mehr Auto fahren

Both the mutual information and the log-likelihood measures determine the association strength between two words by considering the relative co-occurrences shown in the contingency table (Table 1).

	Accusative object Auto	Accusative object ¬Auto
Verb fahren	*Auto fahren*	*Fahrrad fahren*
Verb ¬fahren	*Auto waschen*	*Wäsche waschen*

Table 1: Contingency table for verb-object pairs

The association strength increases for both measures the more the number of co-occurrences in the upper left corner of the contingency table outweighs the number of occurrences in the remaining cells.

3. Evaluation

From the list of 22,897 putative valence compounds, a balanced sample of 100 valence compounds and 100 non-valence compounds was randomly selected in order to be able to evaluate the effectiveness of the methods described above. Each entry in this sample was manually annotated as to whether they represent valence compounds or not. The sample as a whole serves as a gold standard for evaluation.[2]

For all three association measures described above recall, precision, and F-measure were computed. The results are shown in Fig. 2, 3, and 4, for log-likelihood, mutual information, and relative frequency, respectively. The first two measures yield a continuous scale of association strength values. The graphs in Fig. 2 and 3 plot, on the x-axis, association strength thresholds that correspond to a quantile between 100% and 10% of the values observed for these measures. The y-axis shows the corresponding effect on precision and recall for each measure.

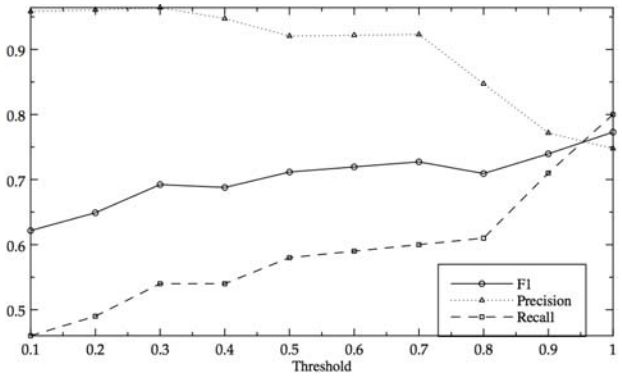


Figure 2: Precision, Recall, and F1 for Log-Likelihood

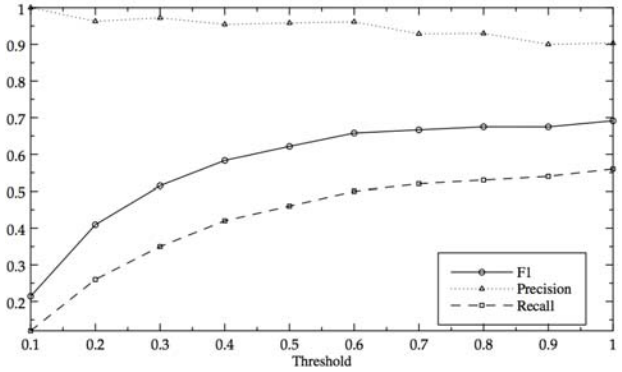


Figure 3: Precision, Recall, and F1 for Mutual Information

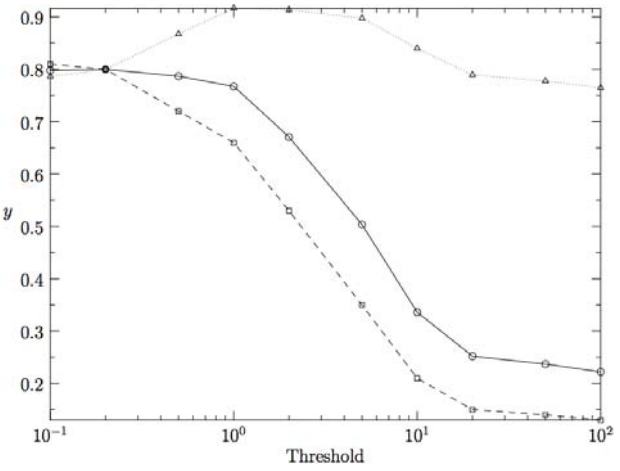


Figure 4: Precision, Recall, and F1 for Relative Frequency

For the relative frequency approach (Fig. 4) the decision to reject or accept a candidate pair is made by weighting occurrences as a verb-object pair against occurrences in other contexts. The weights can consist of any value between zero and positive infinity. Unlike the association measures (log-likelihood and mutual information), this approach does not yield a ranking of candidates; in consequence, the precision (shown in Fig. 4) does not decrease monotonically but shows an optimal parameter setting for values between 1.0 and 2.0.

The results show that all three measures are independently valuable in the corpus-based identification of valence compounds. Relative frequency and log-likelihood yield the best recall (up to 81%), while mutual information affords the best precision (up to 100%). Future research will address the effective methods for combining the complementary strengths of all three measures into an optimized classification approach.

In sum, the eHumanities method presented in this paper for the identification of valence compounds in German has proven effective and can thus nicely complement traditional methods of analysis which focus on the internal structure of valence compounds as such.

References

Baroni, M., J. Matiasek, and H. Trost (2002). Predicting the Components of German Nominal Compounds. In F. van Harmelen (ed.), *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI)*. Amsterdam: IOS Press, pp. 470-474.

Hall, J., J. Nivre, and J. Nilsson (2006). Discriminative Classifiers for Deterministic Dependency Parsing. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of*

the Association for Computational Linguistics (COLING-ACL) Main Conference Poster Sessions, pp. 316-323.

Henrich, V., and E. Hinrichs (2011). Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language* Processing (RANLP 2011), Hissar, Bulgaria, pp. 420-426.

Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics* 28(3): 357-388.

Lemnitzer, L. (2007). *Von Aldianer bis Zauselquote: Neue deutsche Wörter, woher sie kommen und wofür wir sie brauchen*. Tübingen: Narr.

Schmid, H., A. Fitschen, and U. Heid (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal, pp. 1263-1266.

Notes

1. See `http://www.sfs.uni-tuebingen.de/en/tuepp.shtml`
2. Precision measures the fraction of retrieved valence compounds that are correctly analyzed. Recall measures the fraction of actual valence compounds that are retrieved.

Networks of networks: a critical review of formal network methods in archaeology through citation network analysis and close reading

Brughmans, Tom
tb2g08@soton.ac.uk
Archaeological Computing Research Group, University of Southampton, UK

This paper will argue that archaeological network researchers are not well networked themselves, resulting in a limited and sometimes uncritical adoption of formal network methods within the archaeological discipline. This seems to have followed largely from a general unawareness of the historicity of network-based approaches which span at least eight decennia of multi-disciplinary research. Many network analytical techniques that would only find a broader use in the last 15 years were in fact introduced in the archaeological discipline as early as the 1970s. This paper does not aim to argue that every archaeological network study should include a historiography. It merely wishes to stress the need to explore the full range of existing network techniques and models. I will illustrate that knowledge of the diversity of archaeological and non-archaeological network methods is crucial to their critical application and modification within archaeological research contexts.

This paper will for the first time trace the academic traditions, network concepts, models and techniques that have been most influential to archaeologists. I will do this by combining a close reading of published archaeological network applications with citation network analysis techniques (Batagelj 2003; Hummon & Doreian 1989; White 2011), an approach that has not been applied to archaeological literature before. A citation network was created consisting of over 10,000 publications and their internal citations. This network consists of all archaeological network analysis applications, all publications cited by them and the citations between those publications. This data was extracted from Web of Knowledge (`http://wok.mimas.ac.uk/`) and manually when the publications were not included on Web of Knowledge.

The analysis revealed a number of issues surrounding the current use of network methods in archaeology,

as well as possible sources and explanations for these issues. They include the following: (1) Although many network techniques are rooted in graph theory, archaeological studies in graph theory were not influential at all to more recent archaeological network studies. The introduction of graph theory and social network analysis into the archaeological discipline happened largely independently and, unlike social network analysts, archaeologists did not collaborate with graph theorists to develop mathematical techniques tailored for their needs. (2) The potential of social network analysis techniques was explored (largely theoretically) by Cynthia Irwin-Williams no later than 1977. Many of the techniques she described were not applied in archaeological research until the last ten years, possibly because the limited availability of cheap and potent computing power and large digital datasets in the late 1970s. (3) Some social network analysis techniques (e.g. centrality measures) have received more attention than others (e.g. ego-networks). The popularity of these techniques seems to be related with their use by social network analysts rather than how archaeologists have used them before. (4) The archaeological use of complex network models is largely limited to the extremely popular small-world and scale-free models. Archaeologists have neglected to explore the potential use of alternative complex network models. (5) The availability of user-friendly software seems to determine the popularity of social network measures used by archaeologists. (6) On the other hand, simulations of complex network models by archaeologists are rare due to the required technological and mathematical skills.

The results of the citation analysis expose the insufficiently explored potential of formal network-based models and techniques and points out publications in other disciplines that might have interesting archaeological applications. The paper concludes that in order to move towards richer archaeological applications of formal network methods archaeological network analysts should become better networked both within and outside their discipline. The existing archaeological applications of network analysis show clear indications of methods with great potential for our discipline and methods that will remain largely fruitless, and archaeologists should become aware of these advances within their discipline. The development of original archaeological network methods should be driven by archaeological research problems and a broad knowledge of formal network methods developed in different disciplines.

References

Batagelj, V. (2003). Efficient Algorithms for Citation Network Analysis. *Arxiv preprint cs/03090* 23: 1-27.

Irwin-Williams, C. (1977). A network model for the analysis of Prehistoric trade. In T. K. Earle, and J. E. Ericson (eds.), *Exchange systems in Prehistory*. New York: Academic Press, pp. 141-151.

Hummon, N. P., and P. Dereian (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks* 11(1): 39-63.

White, H. D. (2011). Scientific and scholarly networks. In J. Scott, and P. J. Carrington (eds.), *The SAGE handbook of social network analysis*. London: Sage, pp. 272-285.

On the dual nature of written texts and its implications for the encoding of genetic manuscripts

Brüning, Gerrit
bruening@faustedition.de
Freies Deutsches Hochstift, Germany

Henzel, Katrin
henzel@faustedition.de
Klassik Stiftung Weimar, Germany

Pravida, Dietmar
pravida@faustedition.de
Johann-Wolfgang-Goethe-Universität Frankfurt am Main, Germany

1. Introduction

Discussions about a markup language for the genetic encoding of manuscripts have reached a crucial point. The endeavours undertaken by a couple of notable projects over the last years promise to yield a new encoding standard for the description of the genesis of texts. In our talk we will survey the past and current states of affairs and outline some problems of genetic encoding within and without the framework of the Text Encoding Initiative (TEI).

2. Crucial issues in genetic encoding

Early conceptions of a computer-aided genetic edition trace back to the 1970s (Gabler 1998). The absence of any established standard for projects that aim to record revised manuscripts and genetic relations across numerous witnesses was still felt at the end of the 20th century. Through the 1990s and early 2000s, the then available TEI guidelines did not meet the requirements of genetic textual criticism, for genetic manuscripts often lack textual structure in a more conventional sense and show complicated series of revisions. This incompletenesses gave rise to the construction of the HyperNietzsche Markup Language (Saller 2003, esp. note 3) which was a kind of spin-off, but no sustainable alternative.[1]Some of the problems that led to HNML resolved themselves with the previous version of the Guidelines (Burnard et al. 2008: 72-78, 335-374). Other expectations, however (Vanhoutte 2002; Pierazzo 2009), remained unfulfilled:

- A documentary or 'diplomatic' transcription which focuses less on the as yet privileged textual structure of the examined material.
- The possibility to record the temporal course of the genesis of texts.

The 'Encoding Model for Genetic Editions' (Workgroup on Genetic Editions, 2010) was to satisfy both requests. Now that it has been largely incorporated in chapter 11 of the second version of 'Proposal 5' (Burnard et al. 2011), successes and shortcomings can be discussed. As for the documentary transcription, a way of encoding has been made possible which could not be thought of in former versions of the TEI Guidelines up to the first version of P5 (Burnard et al. 2008). However, the suggested approach to transcribe texts, taking into account the spatial distribution of the inscription falls short of completion (Brüning et al., forthcoming). The conceptual tension between the documentary and the textual perspective which gave rise to the proposed encoding model has become even more pronounced. As for the markup of textual alterations, it is not always clear how the newly introduced elements relate to well-established practice.[2] The former stringency of chapter 11 of the Guidelines corresponded to the limited goal of providing 'methods for encoding as broad a variety of textual features as the consensus of the community permits' (Sperberg-McQueen et al. 1993). This consensus came to be known under the name of an 'agnostic' view on the issue of textuality. However, when chapter 11 was revised for the actual version, this consensus was tacitly and perhaps unawares abandoned. for the sake of much specific needs that parts of the new version of the chapter try to fulfill. Many of the needs which the newly introduced parts of the chapter try to fulfill are not covered by a broad consensus, but are based on very specific aims. Therefore the conceptional problems inherent in the genetic module cannot any longer be solved with reference to an alleged 'agnostic' basis. A discussion on the nature of text is not only of theoretical interest but also inevitable for practical reasons.

3. The dual nature of written texts

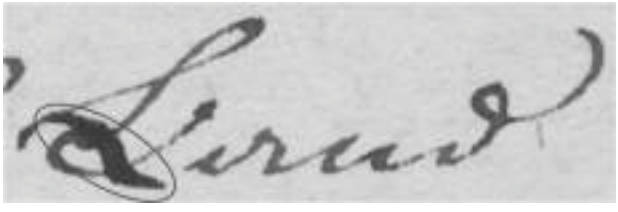
We would like to submit as a basic principle that any written text is, by virtue of being a linguistic entity, of a double-sided nature: First, it is by virtue of being a *written* text, a physical object that can be identified in space and time as a document or an inscription (the material or documentary dimension). Second, it is – by virtue of being a written *text* – an abstract object (the textual dimension), which has to be materialized in some way, not necessarily in one specific material form (Goodman 1977; Genette 1997). As can be learned from the epistemology of general linguistics

(Saussure 1916: introduction, ch. 3; cf. Bouquet 1997), it is imperative for sound methodology to keep irreducible points of view distinct, each of which allows a coherent strategy of inquiry. Any attempt to collapse two entirely different dimensions into one, will ultimately run up on a reef of methodological incoherence. Within the digital humanities, the use of the word 'text' is often inspired by the way texts are *represented* in the digital code: as an ordered hierarchy of content objects (Renear et al. 1996) or as a strictly linear sequence of characters (cf. Buzzetti & McGann 2006: 60). But the hierarchical and/or linear *representation* of a text in a computer file is not and cannot by principle be a text in a linguistically respectable sense of the word (Sperberg-McQueen 1991: 34).

4. Consequences

Adherence to our basic principle will lead to some important consequences for encoding written texts and to some proposals for the application and further development of the available TEI Guidelines.

1. The documentary perspective (1) takes the inscriptional point of view and focuses on the *materialization* of the text. The basic elements of this perspective are written letters and other inscriptional phenomena, such as cancellations and other sediments of the writing process. Inscriptional analyses can be focussed on units smaller than letters (e.g. allographic forms or elements below the niveau of the letter, see Feigs 1979). But due to the specific aims of our project, units smaller than letters are ignored. Letters can as well be regarded as linguistic units, and a linguistic understanding of what is written is a precondition for any manuscript to be deciphered. However, it is possible to disregard from this understanding for the purpose of a documentary edition. – The textual perspective (2) focuses on linguistic entities. Its basic elements are linguistic units. Depending on the perspective chosen, the following example has to be treated in two different ways:



i. From the documentary perspective, one letter (B) is changed into another (L). As said, even the fact that only a *part* of the original letter has been modified, might be of interest. But for the above reason, a replacement of letters is recorded. Therefore a documentary oriented encoding will usually look like this:[3]

```
<subst>
   <del>B</del>
   <add>L</add>
</subst>and
```

ii. From the textual perspective, a whole word is substituted by another, 'Band' (ligament) by 'Land' (land):

```
<subst>
   <del>Band</del>
   <add>Land</add>
</subst>
```

This way of encoding focuses on the textual impact of the overwriting of (a part of) one letter: the substitution of a whole word. It does not mean that the textual perspective is *confined* to the level of word forms. In fact, sometimes only the spelling is corrected, and the alteration can very well be recorded on the level of graphemic units alone – no matter how many letters of the word have been replaced, no matter if the complete word is rewritten for the sake of clarity in the manuscript. However, in the given case the substitution does *not* concern only single letters but the linguistic unit of the next level (the word), although the scribe carefully avoided to deface the manuscript more than was necessary.

How can the sediments of the writing process and the textual genesis be given their equal due?

As is clear from the above example, not only the markup is concerned by the difference of the two perspectives but also the marked up content ('BLand' vs. 'BandLand'). This is one of the reasons why a split-up of the encoding which we tried to avoid for a long time finally proved inevitable. We decided to introduce two separate transcripts: a documentary and a textual one (Bohnenkamp et al., forthcoming). The separate transcripts are

the source for different parts of the edition: the documentary for the diplomatic rendering and the textual for the reading text and the apparatuses. The disadvantages of the split-up will be dealt with by help of automatic collation (Brüning et al., forthcoming). Logically, it is possible to integrate both perspectives in a single transcript, under the only condition that the markup is dominated by one of the two perspectives (Renear et al. 1996). Practically, however, it is very difficult to give adequate information on the writing process and on the text with equal regard to both sides. Detailed information of one of both will inevitably get lost. Furthermore, the intermingling of both perspectives complicates the subsequent data processing.

2. Thanks to the split-up, the documentary transcript is kept free from the limitations inherent to the markup of the textual structure, so that all the information that is needed to generate a satisfying diplomatic rendering can be placed. Likewise, the textual transcript is kept free from descriptive information about the record. In turn, an aspect of the linguistic structure can be taken into account the importance of which for genetic encoding has not yet found sufficient perception. As a concatenation of linguistic entities is not a mere sequence but a structure, it is necessary to distinguish between syntagmatically defined positions and paradigmatically selected items occupying them.[4] In the above example, a position is initially occupied by the word 'Band', and subsequently by the word 'Land'. The surrounding <subst>-tag indicates the paradigmatic relation between both items, although this is not explicit in the definition (Burnard et al. 2011: 1348).[5] Problems arise in cases which deviate from the normal one-to-one-correspondence of items to positions, for example where the wording of the passage is held in abeyance as a consequence of an unfinished alteration.[6] Introducing the differentiation between positions and items into genetic encoding might help to spot, clarify, and solve these problems.

References

Bohnenkamp, A., G. Brüning, S. Henke, K. Henzel, F. Jannidis, G. Middell, D. Pravida, and M. Wissenbach (forthcoming). Perspektiven auf Goethes *Faust*. Zur historisch-kritischen Hybridedition des *Faust*. *Jahrbuch des Freien Deutschen Hochstifts 2011*.

Bouquet, S. (1997). *Introduction à la lecture de Saussure*. Paris: Payot & Rivages.

Brüning, G., K. Henzel, D. Pravida (forthcoming). Rationale of multiple encoding in the Genetic Faust Edition. *Journal of the Text Encoding Initiative* (submitted).

Burnard, L., and S. Bauman, eds. (2004). *TEI P4. Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition*. Oxford, Providence, Charlottesville and Bergen: The TEI Consortium.

Burnard, L., and S. Bauman, eds. (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange by the TEI Consortium*. Oxford, Providence, Charlottesville, Nancy: Tei Consortium.

Burnard, L., and S. Bauman, eds. (2011). *TEI P5: Guidelines for Electronic Text Encoding and Interchange by the TEI Consortium*. Charlottesville: Text Encoding Initiative Consortium.

Buzzetti, D., and J. McGann (2006). Electronic Textual Editing: Critical Editing in a Digital Horizon. In L. Burnard, K. O'Brien O'Keeffe, and J. Unsworth (eds.), *Electronic Textual Editing*. New York: MLA, pp. 51-71.

Feigs, W. (1979). *Deskriptive Edition auf Allograph-. Wort- und Satzniveau, demonstriert an handschriftlich überlieferten, deutschsprachigen Briefen von H. Steffens. Teil 1: Methode*. Bern: Peter Lang.

Gabler, H. W. (1998). Computergestütztes Edieren und Computer-Edition. In H. Zeller and G. Martens (eds.), *Textgenetische Edition*. Tübingen: Niemeyer, pp. 315-328.

Genette, G. (1997). *The Work of Art. Immanence and Transcendence*. Ithaca, NY: Cornell UP.

Goodman, N. (1977). *Languages of Art. An Approach to a Theory of Symbols*. 2nd ed. Indianapolis, IN: Hackett.

Huitfeldt, C. (1998). MECS – a Multi-Element Code System (Version October 1998), *Working Papers from the Wittgenstein Archives at the University of Bergen*, 3. `http://www.hit.uib.no/claus/mecs/mecs.htm` (accessed 9 March 2012).

Owens, J. (1988). *The Foundations of Grammar. An Introduction to Medieval Arabic Grammatical Theory*. Amsterdam, Philadelphia: Benjamins.

Pierazzo, E. (2009). Digital Genetic Editions: The Encoding of Time in Manuscript Transcription, In M. Deegan and K. Sutherland (eds.), *Text Editing, Print and the Digital World*. Farnham and Burlington, VT: Ashgate, pp. 169-186.

Renear, A., E. Mylonas, and D. Durand (1996). Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. *Research in Humanities Computing* 4:

263-280. http://www.stg.brown.edu/resources/stg/monographs/ohco.html (accessed 9 March 2012).

Saller, H. (2003). HNML – HyperNietzsche Markup Language. *Jahrbuch für Computerphilologie* 5: 185-192. http://computerphilologie.uni-muenchen.de/jg03/saller.html (accessed 9 March 2012).

Saussure, F. de (1916). *Cours de linguistique générale*. Lausanne: Payot.

Sperberg-McQueen, C. M. (1991). Text in the Electronic Age: Textual Study and Text Encoding with Examples from Medieval Texts. *Literary and Linguistic Computing* 6: 34-46.

Sperberg-McQueen, C. M., and L. Burnard, eds. (1993). *Guidelines for Electronic Text Encoding and Interchange* (P2). Chicago, Oxford: Text Encoding Initiative. http://www.tei-c.org/Vault/Vault-GL.html (accessed 9 March 2012).

Sperberg-McQueen, C. M., and L. Burnard, eds. (1999). *Guidelines for Electronic Text Encoding and Interchange*. 1994. Revised Reprint. Chicago, Oxford: Text Encoding Initiative.

Vanhoutte, E. (2002). *Putting Time back in Manuscripts: Textual Study and Text Encoding, with Examples from Modern Manuscripts*, paper presented at the ALLC/ACH 2002, Tübingen, 25 July 2002. http://www.edwardvanhoutte.org/pub/2002/allc02abstr.htm (accessed 9 March 2012).

Workgroup on Genetic Editions 2010. *An Encoding Model for Genetic Editions*. http://www.tei-c.org/Activities/Council/Working/tcw19.html (accessed 9 March 2012).

Notes

1. 'HyperNietzsche' nowadays operates under the name of 'Nietzsche Source', where the HNML based contents are regrettably cut off. They are still available by the following address: http://www.hypernietzsche.org/surf_page.php?type=scholarly . To become affiliated with the group of 'Source' projects, the 'Bergen Text Edition' of the Wittgenstein papers was converted from the specifically developed MECS to XML-TEI (P5) (http://www.wittgensteinsource.org/ ; Huitfeldt 1998). There are other ways to mark up transcriptions of modern manuscripts (see 'Les manuscrits de Madame Bovary', http://www.bovary.fr/ ; 'Les Manuscrits de Stendhal', http://manuscrits-de-stendhal.org/). But the set of tools provided by the TEI is clearly the one that is likely to become standard.
2. See esp. sect. 11.3.4.4 of the second version of P5 (Burnard et al. 367 sq.).
3. In fact, we do not use <subst> to record overwritings. Especially in the case of discarded starts overwriting letters do not positively substitute their predecessors. But we will not address this issue here.
4. he terminology of position-in-structure and of items-occupying-positions are very common in various schools of linguistics (cf. Owens 1988: 31-35) and in the computer sciences as well.
5. It is doubtful, therefore, if items which occupy 'different positions' (ibid., 352) can enter such a relation, as is, perhaps inadvertently, implied in the name of the newly introduced <substJoin> (ibid., 351 sq.).
6. For an example of the further, see sect. 11.3.4.6 of the TEI Guidelines (Burnard et al. 2011: 369). The suggested encoding (ibid., 370) neglects the symmetrical relation between the items 'before' and 'beside'. From a textual point of view, the latter can be considered *added* only insofar as the former is considered *deleted*.

Automatic recognition of speech, thought and writing representation in German narrative texts

Brunner, Annelen
annelen_brunner@gmx.de
Institut für deutsche Sprache, Mannheim, Germany

This paper presents a subset of the results of a larger project which explores ways to recognize and classify a narrative feature – speech, thought and writing representation (ST&WR) – automatically, using surface information and methods of computational linguistics.

Speech, thought and writing can be represented in various ways. Common categories in narratology are direct representation (*He thought: 'I am hungry.'*), free indirect representation, which takes characteristics of the character's voice as well as the narrator's (*Well, where would he get something to eat now?*), indirect representation (*He said that he was hungry.*) and reported representation, which can be a mere mentioning of a speech, thought or writing act (*They talked about lunch.*). ST&WR is a feature central to narrative theory, as it is important for constructing a fictional character and sheds light on the narrator-character relationship and the narrator's stance. The favored techniques not only vary between authors and genres, but have changed and developed over the course of literary history. Therefore, an automated annotation of this phenomenon would be valuable, as it could quickly deal with a large number of texts and allow a narratologist to study regularities and differences between different time periods, genres or authors.

The approach presented here specifically aims at applying digital methods to the recognition of features conceptualized in narrative theory. This sets it apart from other digital approaches to literary texts which often operate purely on a vocabulary level and are more focussed on thematic issues or on author or group specific stylistics. Recent approaches to the goal of automatically identifying ST&WR are either not concerned with narrativity at all, like Krestel et al. who developed a recognizer of direct and indirect speech representation in newspaper texts to identify second hand information, or not interested in the techniques themselves, like Elson et al. who use recognition of direct speech representation to extract a network of interrelations of fictional characters. Also, both approaches are for the English language and can therefore not be used in this project.

Basis for the research is a corpus containing 13 short narratives in German written between 1786 and 1917 (about 57 000 tokens). The corpus has been manually annotated with a set of ST&WR categories adapted from narratological theory. This step is comparable to the annotation project conducted by Semino and Short for a corpus of English literary, autobiographical and newspaper texts, but is something that has never been done for German literary texts before. The manual annotation gives empirical insight into the surface structures and the complex nature of ST&WR, but also serves as training material for machine learning approaches and, most importantly, as reference for evaluation of the automatic recognizer.

The main focus of this paper is the automatic recognition. Rule-based as well as machine learning approaches have been tested for the task of detecting instances of the narratological categories. In the scope of this paper, a subset of these strategies is presented and compared.

For the rule-based approach, simple and robust methods are favored, which do not require advanced syntactic or semantic preprocessing, automatic reasoning or complex knowledge bases. The modules make use of conventions like punctuation, as well as lexical and structural characteristics for different types of ST&WR. A central feature is a list of words that signal ST&WR, e.g. *to say* (*sagen*), *to whisper* (*flüstern*). For the recognition of indirect ST&WR specifically, patterns of surfaces and morphological categories are used to match the dependent propositions (e.g. *E r sagte, dass er hungrig sei.* [*He said that he was hungry.*]: signal word – followed by comma – followed by a conjunction – followed by a verb in subjunctive mode). This methods achieves F1 scores of up to 0.71 in a sentence-based evaluation. Direct representation can be detected with an F1 score of 0.84 by searching for quotation patterns and framing phrases (e.g. *he said*). Annotating reported representation, which is quite diverse, achieves an F1 score of up to 0.57.

The machine learning approach uses a random forest learning algorithm trained on the manually annotated corpus. Features like sentence length, number of certain word types and types of punctuation are used as attributes. The advantage of this approach lies in the fact that it can also be used to handle types of ST&WR which do have less obvious structural or lexical characteristics, like free indirect representation, for which an F1 score of 0.43 can be achieved when performing sentence-based cross validation on the corpus. The F1 score for detecting direct representation is 0.81, for indirect representation 0.57 and for reported representation 0.51.

The components of the automatic recognizer are modular and realized as working prototypes in the framework GATE (General Architecture for Text Engineering) (http://gate.ac.uk). When the project is finished, it is intended to publish these components as GATE plugins and make them available as a free download.

In the paper, the advantages and disadvantages of rule-based and machine learning approaches as well as possibilities for combination are discussed. For example, rules can be used for ST&WR strategies with clear patterns and conventions, like direct and – to an extent – indirect representation, but machine learning for the more elusive types like free indirect representation. It is also possible to get results for the same ST&WR category from different modules and use those to calculate scores. E.g. merging the results of rule-based and ML methods improves the overall F1 score for recognizing direct representation in the corpus.

Though the figures above give a rough idea of expected success rates, evaluation is in fact an analytic task itself: Results are not only quite different for different types of ST&WR and dependent on the exact configuration of the recognizer modules. There is also the question of what kind of results should be prioritized for narratological research and how to deal with cases which are problematic even for a human annotator. However, the modular structure of the recognizer is designed to allow for customization and the main goal of the project is to shed light on the relationship between the manual annotation, generated with narratological concepts in mind, and the possibilities and limitations of ultimately surface-based automatic annotation.

References

Cunningham, H., et al. (2011). Text Processing with GATE (Version 6): `http://www.tinyurl/gatebook` (accessed 14 March 2012).

Elson, D. K., N. Dames, and K. R. McKeown (2010). Extracting Social Networks from Literary Fiction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics 2010,* pp. 138-147.

Krestel, R., S. Bergler, and R. Witte (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008),* Marrakech, Morocco, May 28-30 2008.

Semino, E., and M. Short (2004). *Corpus stylistics. Speech, writing and thought presentation in a corpus of English writing*. London, New York: Routledge.

Bringing Modern Spell Checking Approaches to Ancient Texts – Automated Suggestions for Incomplete Words

Büchler, Marco
mbuechler@e-humanities.net
Leipzig University, Germany

Kruse, Sebastian
skruse@eaqua.net
Leipzig University, Germany

Eckart, Thomas
teckart@e-humanities.net
Leipzig University, Germany

One of the most challenging tasks for scholars working with ancient data is the completion of texts that have only been partially preserved. In the current situation, a great deal of scholarly experience and the use of dictionaries such as *Liddell Scott Jones* or *Lewis & Short* are necessary to perform the task of text reconstruction manually. Even though text search tools such as Diogenes or papyri.info exist, scholars still have to work through the results manually and require a very good knowledge about the text, its cultural background and its documentary form in order to be able to decide about the correct reconstitution of the damaged text. Therefore, a 'selective and relatively small scope' especially of younger scholars restricts the set of potential candidates.

To overcome these barriers an unsupervised approach from the field of machine learning is introduced to form a word prediction system based on several classes of spell checking (Kukich 1992; Schierle et al. 2008) and text mining algorithms.

Both spell checking and text completion can be separated into two main tasks: identification of incorrect or incomplete words and the generation of suggestions. While the identification of misspelled words can be a very difficult task when working with modern texts (such as with spell checking support provided by modern word processing suites), existing sigla of the Leiden Conventions (Bodard et al. 2009) can be used when dealing with ancient texts. The second step of the process is then to generate likely suggestions using methods such as:

- **Semantic approaches**: *Sentence co-occurrences* (Buechler 2008) and *document co-occurrences* (Heyer et al. 2008) are used to identify candidates based on different contextual windows (Bordag 2008). The basic idea behind this type of classification is motivated by Firth's famous statement about a word's meaning: '*You shall know a word by the company it keeps*' (Firth 1957).
- **Syntactical approaches**: *Word bi- and trigrams* (Heyer et al. 2008): With this method, the immediate neighbourhood of a word is observed and likely candidates are identified based on a selected reference corpus.
- **Morphological dependencies**: Similar to the *Latin and Greek Treebank of Perseus* (Crane et al. 2009) morphological dependencies are used to suggest words by using an expected morphological code.
- **String based approaches**: The most common class of algorithms for modern texts compares words by their word similarity on letter level. Different approaches like the *Levenshtein distance* (Ottmann & Widmayer 1996) or faster approaches such as *FastSS* (Bocek et al. 2007) are used to compare a fragmentary word with all candidates.
- **Named Entity lists**: With a focus on deletions of inscriptions, existing and extended named entity lists for person names, cities or demonyms like the *Lexicon of Greek Personal Names* (Fraser et al. 1987-2008) or the *Wörterlisten* of Dieter Hagedorn are used to look for names of persons and places and give them a higher probability.
- **Word properties**: When focusing on Stoichedon texts, word length is a relevant property. For this reason the candidate list can be restricted by both *exact length* as well as by *min-max thresholds*.

From a global perspective, every found word in a vocabulary is a potential suggestion candidate. To reduce this list of anywhere from several hundred thousand to several million words to a more reasonable size, the results of all selected algorithms are combined to a normalised score between 0 and 1 (Kruse 2009). In the last working step of this process, the candidates list (ordered by score in descending order) is then provided to the user.

Based on the aforementioned approaches the full paper will explain three different completion strategies:

1. Using only known information about a word (word length and some preserved characters),
2. using only contextual information such as word bigrams, co-occurrences, and classification data,
3. using all available information (combination of strategy a) and b)) of a word.

The main objective of this step by step explanation is to highlight both strengths and weaknesses of such a completely automatized system.

A video demonstration of the current implementation can be viewed at

`http://www.e-humanities.net/lectures/SS2011/2011-DigClassSeminar/THATCamp_DevChallenge_BuechlerEckart_TextCompletion.ogv`

References

Bocek, T., E. Hunt, and B. Stiller (2007). *Fast Similarity Search in Large Dictionaries*. Department of Informatics, University of Zurich.

Bodard, G., et al. (2009). *EpiDoc Cheat Sheet: Krummrey-Panciera sigla & EpiDoc tags*, 2006-2009. Version 1085, last accessed: Nov., 10th, 2009 [date] URL: http://epidoc.svn.sourceforge.net/viewvc/epidoc/trunk/guidelines/msword/cheatsheet.doc.

Bordag, St. (2008). *A Comparison of Co-occurrence and Similarity Measures as Simulations of Context*, 2008. In *CICLing*, Vol. 4919. Berlin: Springer (Lecture Notes in Computer Science).

Büchler, M. (2008). *Medusa. Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung*. Saarbrücken: Vdm Verlag Dr. Müller.

Crane, G., and D. Bamman (2009). *The Latin and Ancient Greek Dependency Treebanks*, 2009. URL: http://nlp.perseus.tufts.edu/syntax/treebank/ last accessed: Nov., 10th 2009.

Firth, J. R., *A Synopsis of Linguistic Theory*. Oxford.

Fraser, P. M. E. Matthews, and M. J. Osborne (1987-2008). *A Lexicon of Greek Personal Names*. (In Greek and English), Vol. 1-5, Suppl. Oxford: Clarendon Press.

Heyer, G., U. Quasthoff, and T. Wittig (2008). *Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse*. 2nd edition. Herdecke: W3L-Verlag.

Kruse, S. (2009). *Textvervollständigung auf antiken Texten*. University of Leipzig, Bachelor Thesis. pp 48-49. URL http://www.eaqua.net/~skruse/bachelor, last accessed on Nov., 10th 2009.

Kukich, K. (1992). Technique for Automatically Correcting Words in Text. *ACM Computing Surveys* 24(4).

Ottmann,T., and P. Widmayer (1996). *Algorithmen und Datenstrukturen*. Heidelberg: Spektrum Verlag.

Schierle, M., S. Schulz, and M. Ackermann (2008). From Spelling Correction to Text Cleaning – Using Context Information. In *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V.*

Designing a national 'Virtual Laboratory' for the humanities: the Australian HuNI project

Burrows, Toby Nicolas
toby.burrows@uwa.edu.au
University of Western Australia, Australia

1. Context

This paper looks at the design and architecture of the Humanities Networked Infrastructure (HuNI), a national 'Virtual Laboratory' which is being developed as part of the Australian government's NeCTAR (National e-Research Collaboration Tools and Resources) programme. One of NeCTAR's main goals is the development of up to ten discipline-based 'Virtual Laboratories'. The aims of this programme are to integrate existing capabilities (tools, data and resources), support data-centred research workflows, and build virtual research communities to address existing well-defined research problems.

Beginning in May 2012, HuNI has been funded until the end of 2013. It is being developed by a consortium of thirteen institutions, led by Deakin University in Melbourne.

2. Design Framework

HuNI is specifically designed to cover the whole of the humanities (defined as the disciplines covered by the Australian Academy of the Humanities). It uses scientific e-Research consortia as its model: large-scale, multi-institutional, interdisciplinary groups with an e-Research framework covering the entire field of research.

This approach has a sound academic basis. It emphasizes the interdisciplinary and trans-disciplinary reach of the e-Research services and tools which are included in the 'Virtual Laboratory' environment – and its value to researchers across the full range of humanities disciplines.

HuNI aims to join together the various digital services and tools which have already been developed for specific humanities disciplines, both by collecting institutions (libraries, archives, museums and galleries) and by academic research groups. It builds on these services and strengthens them, rather than superseding them.

Data-Centred Workflows

The very concept of 'data' can be problematic for the humanities. Nevertheless, a distinctive type of humanities data can be identified, different from the quantitative and qualitative data of the social sciences. This 'humanities data' consists of the various annotations, tags, links, associations, ratings, reviews and comments produced during the humanities research process, together with the entities to which these annotations refer: concepts, persons, places and events.

It is important to draw a distinction between 'data' in this sense and primary source materials, particularly in digitized form (Borgman 2007: 215-217 fails to make this distinction). Primary materials – even in the form of digital objects – are sources of data, rather than data per se.

A data-centred virtual laboratory for the humanities needs to include services for identifying these semantic entities and their relationships (using the Linked Open Data technical framework), and for capturing and sharing the annotations and other scholarly outputs which refer to them. While the Linked Open Data framework is a relatively recent development, there are already a sufficient number of projects and services underway in Europe and North America which can serve as case studies to demonstrate clearly the viability and value of this approach (Bizer, Heath & Berners-Lee 2009).

Integration of Existing Capabilities

Many significant Australian collections of digital content relevant to humanities research already exist. Some of these are descriptions of physical objects (e.g., books, museum objects and art works) and entities (e.g., people and places), some are collections of digital objects, and others are a mixture of the two. In most cases, these collections were not connected with each other except to the extent that they could be searched by services like Google. Working effectively across such a disparate range of sources has been a major challenge for humanities researchers.

The production-level tools for working with these collections of digital content are relatively limited. Most tools are designed to work with a single service or a single type of content, such as the visualization tools developed by AusStage[1], the user tagging developed by the Powerhouse Museum in Sydney[2], and the Heurist software[3] developed by the University of Sydney for archaeology. The LORE tool (for annotation and the construction of virtual collections) works mainly with AustLit[4], though its federated search also covers some other content services.

To meet NeCTAR's Virtual Laboratory criteria, content sources need to be integrated (or at least inter-linked), and tools need to be usable across as

many *sources* and *types* of content as possible. It is neither practical nor desirable to merge content from multiple disciplines into a single enormous database, given the extensive variations in standards and approaches. Federated searching across many services, on the other hand, will not build the data-centred platform required to support the other functions of the Virtual Laboratory. The only feasible solution for data integration is to deploy a Linked Open Data environment on a national scale.

Architecture and Services

The project has defined a data-centred workflow for the humanities, with three main stages:

- Discovery (search and browse services);
- Analysis (annotation, collecting, visualization and mapping);
- Sharing (collaborating, publishing, citing and referencing).

For the Analysis and Sharing functions, a suite of existing Open Source tools developed in Australia are being used and adapted as part of the project. These include:

- LORE – developed by AustLit for annotation, federated searching, visualization, aggregation and sharing of compound digital objects (Gerber, Hyland & Hunter 2010);
- Visualization tools developed by AusStage (Bollen et al. 2009);
- OHRM[5] – developed by the University of Melbourne to model entity relationships and publish information about collections into aggregated frameworks;
- Heurist and FieldHelper – developed by the University of Sydney to aggregate data, model entity relationships and publish collections of data to the Web (including maps and timelines).

The main adaptation required is to extend their functionality to work with Linked Data URIs and to be hospitable to cloud-based hosting. Where no Australian tool can be used or adapted, international Open Source tools will be used.

The annotations, compound objects and tags created by researchers using the Analysis tools will be stored in RDF as part of HuNI's Linked Data Service. Descriptions of these data collections will also be made available for harvesting in RIF-CS format by the Australian National Data Service (ANDS) for its Research Data Australia service.

HuNI's Discovery environment builds on the technologies used by a variety of Australian and international services to provide sophisticated searching and browsing across data extracted from heterogeneous data sets and combined into a Linked Data Service. These models include the Atlas of Living Australia[6], as well as general humanities-related services like SOCH (Swedish Open Cultural Heritage) and discipline-specific services like CLAROS. The British Museum's new ResearchSpace will also serve as a key exemplar.

The Discovery environment is underpinned by the Linked Data Service. The preferred solution for supporting and presenting Linked Data is the Vitro software developed by Cornell University, which serves as the basis for the VIVO research profiling service. VIVO has been implemented at two of the HuNI partner institutions: the University of Melbourne and the University of Western Australia.

The outputs from the Discovery environment will be produced in formats which can be consumed by tools performing Analysis and Sharing functions. The Linked Data Service will support an API for exposing data in RDF/XML and JSON formats. This API will form the basis for reuse of the data by service providers other than HuNI, as well as enabling the custodians of data sets which contribute to HuNI to build workflows for pulling new content from the Linked Data Store into their own data sets.

A necessary prerequisite for assembling heterogeneous data from different data sets into a Linked Data format is a Semantic Mediation and Mapping Service. There will be two main components to this service:

- Tools for extracting, exposing and transforming entity data contained in existing cultural data sets and in digital objects;
- An environment for harvesting, ingesting, matching, aligning, and linking the entity data.

For extracting and exposing entity data, two different approaches will be supported. The major content providers will develop and provide RESTful Web APIs for their data sets. A service will also be established to allow smaller data providers to expose their data for transformation and ingest without the need to develop an API. This service will use the harvest and ingest functionality of the Vitro software. Tools for converting static databases and spreadsheets to RDF will also be deployed. Data ingest from entity identification using text mining will follow at a later stage.

The Semantic Mediation and Mapping Service will draw on a range of vocabulary and identifier services, which will be managed through a controlled vocabulary registry with links to the ANDS identifier and vocabulary services. HuNI's initial focus will be on matching, aligning and linking data relating to people, places and objects, using several high-level vocabularies and ontologies. These

include Australian vocabularies like the Gazetteer of Australia[7] and the PeopleAustralia service (Dewhurst 2008).

The HuNI service will enable researchers to find and analyse data across a range of humanities disciplines, and to save the outputs of their analysis in a variety of forms, including compound digital objects, annotations, maps, timelines, and graphs. They will be able to share their results and outputs with other researchers.

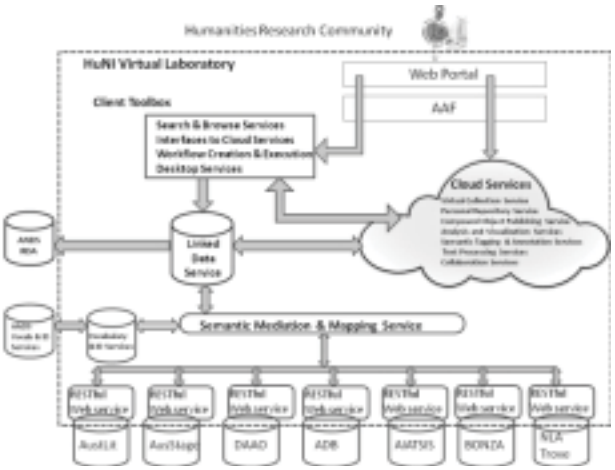


Figure 1: HuNI Architecture

References

Bizer, C., T. Heath, and T. Berners-Lee (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems* 5(3): 1-22.

Bollen, J., N. Harvey, J. Holledge, and G. McGillivray (2009). AusStage: e-Research in the Performing Arts. *Australasian Drama Studies* 54: 178-194.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, Mass.: MIT Press.

Dewhurst, B. (2008). People Australia: a Topic-Based Approach to Resource Discovery. In *VALA2008 Conference proceedings*. Melbourne: VALA. `http://www.valaconf.org.au/vala2008/papers2008/116_Dewhurst_Final.pdf` (accessed 30 March 2012).

Gerber, A., A. Hyland, and J. Hunter (2010). A Collaborative Scholarly Annotation System for Dynamic Web Documents – A Literary Case Study. In *The Role of Digital Libraries in a Time of Global Change* (Lecture Notes in Computer Science 6102). Berlin: Springer, pp. 29-39.

Notes

1. AusStage is the Australian performing arts service (www.ausstage.edu.au (`www.ausstage.edu.au`))
2. www.powerhousemuseum.com (`www.powerhousemuseum.com`)
3. www.heuristscholar.org (`www.heuristscholar.org`)
4. AustLit is the Australian literature service (www.austlit.edu.au (`www.austlit.edu.au`))
5. www.esrc.unimelb.edu.au/ohrm/ (`www.esrc.unimelb.edu.au/ohrm/`)
6. www.ala.org.au (`www.ala.org.au`)
7. www.ga.gov.au/place-names/ (`www.ga.gov.au/place-names/`)

Beyond Embedded Markup

Buzzetti, Dino
dino.buzzetti@gmail.com
formerly University of Bologna, Italy

Thaller, Manfred
manfred.thaller@uni-koeln.de
University at Cologne, Germany

1. Introduction (Manfred Thaller)

The unquestionable success of embedded markup methods and practice for the encoding of texts in the humanities is indisputably confirmed by the several projects and the authoritative collections of encoded texts now made available to the scholarly community.

The reasons of this success are many and diverse. An important one, however, consists in the influence that technological innovations have had on the accepted methodological principles of humanities computing. As John Unsworth has shown, we have been witnessing different phases of prevailing concerns in humanities computing projects and research: the chief orientation of interests has shifted from 'tools,' in the '50s, '60s, and '70s, to 'sources,' in the '80s and '90s, and now seems to be turning back from sources to tools (Unsworth 2004). From a computational point of view, what this change in orientation entailed was a shift of the attention focus from *processing* to *representation*, from developing algorithms applicable to information contents, to producing digital surrogates to replicate and visualise primary surce materials. The introduction of graphic interfaces, the development of WYSIWYG word processing systems on PCs, and the astounding explosion of the World Wide Web have undoubtedly favoured the expansion of this process.

The original purpose of the Web was actually to allow remote access to documents and to visualise them, and the languages developed to produce Web resources, HTML and now increasingly XML, are data representation languages and not data processing languages. Processing Web resources is heavily dependent on the structure these languages assign them, i.e. on hierarchical tree structures. XSLT, the language used to process XML data, 'takes a tree structure as its input, and generates another tree structure as its output' (Kay 2005). The point of view of the so-called 'document community' as opposed to that of the 'data processing' or 'database community' – i.e. 'to standardize the representation of data' *vs* 'to standardize the semantics of data' – was heavily influential on the decisions of the scholarly community, where 'attempts to define semantics [...] met with resistance,' and 'most notably' so in the Text Encoding Initiative. Thus, 'the route proposed by SGML,' and later XML, was to them 'a reasonable one' and embedded markup established itself as a standard for the encoding of texts in the humanities (Raymond 1996: 27-28).

However, from the original surmise that what text 'really is,' is nothing but an 'ordered hierarchy of content objects,' – the so-called OHCO thesis (De Rose et al. 1990) – the inadequacies of embedded markup have soon come to the fore, just for the sake of *representing* textual variation, not to mention *processing* textual information content. The need to overcome these difficulties has prompted several attempts to propose different approaches to text encoding. Among them we may mention the Layered Markup and Annotation Language (LMNL), (Piez s.d.), the eCommentary Machine web application (eComma), (Brown s.d.), the 'extended string' model, (Thaller 2006), and the Computer Aided Textual Markup and Analysis (CATMA) desktop application (CATMA s.d.).

One of the most efficient implementations of similar attempts to meet the difficulties of embedded markup is the proposal of 'standoff properties' as introduced by Desmond Schmidt. Properties can be assigned to given ranges of text that may nest and overlap, and can be stored separately from the text and in different formats suitable to specific needs. Standoff properties can be used to represent different textual features, for either interpretation or rendition purposes, and they can be organised in different sets, that can be easily merged, thus allowing for different encoding perspectives. The same technique is applicable to semantic annotation needs and stands as a viable and efficient alternative to it. But most importantly, as Desmond Schmidt has pointed out, whereas 'it is virtually impossible to freely exchange and interoperate with TEI-encoded texts,' with standoff properties 'interoperability and interchange are enhanced because the components are small and can be freely recombined and reused' (Schmidt, forthcoming). Moreover, the afforded flexibility of the standoff representation of property sets allows its 'textualisation': it can be expressed as a simple string of characters and its variations can be represented in the Multi-Version Document (MVD) format, i.e. as an oriented graph with a start-node and an end-node. Possible interpretative variants and encodings can then be easily compared and analysed.

In sum, approaches of this kind seem to be affording viable solutions to the challenge of putting to good use the invaluable wealth of data now made accessible and encoded, by 'building tools' that would enable us to proceed 'beyond

representation' (Unsworth 2004) and to process their information content. Standoff solutions can provide suitable means to deal with the different kinds of information conveyed by textual data structures and to assign them adequate data models for purposeful processing.

As it happens, however, the basic distinction between *data* and *information* is often overlooked, just as the clear severing of the two basic components of a text, its *expression* and *content*. This lack of distinction often leads to technical and conceptual shortcomings, sometimes intrinsic to the use of embedded markup. In this respect, standoff solutions can usefully complete and supplement embedded markup techniques with additional contrivances to distinguish between rendition and content features and to treat them appropriately. Specific case studies (see Buzzetti 2012, and Thaller 2012) can better exemplify the kind of problems that would require solutions that go beyond the mere representational scope of embedded markup and heed basic conceptual distinctions, such as those between data and information, or interpretation and rendition. To what extent these solution may also converge with the new technologies developed in the context of the Semantic Web would deserve a careful and more documented enquiry.

1.1. References

Brown, T. (s.d.). eComma: A commentary machine. http://ecomma.cwrl.utexas.edu/e392k/ (accessed 23 March 2012).

Buzzetti, D. (2012). Bringing Together Markup and Semantic Annotation. In this volume.

CATMA (s.d.). CATMA – Computer Aided Textual Markup & Analysis. http://www.catma.de/ (accessed 23 March 2012).

DeRose, St., et al. (1990). What Is Text, Really? *Journal of Computing in Higher Education* 1(2): 3-26.

Kay, M. (2005). What Kind of Language is XSLT? An analysis and overview. http://www.ibm.com/developerworks/library/x-xslt (accessed 23 March 2012).

Piez, W. (s.d.). LMNL Activity. http://www.piez.org/wendell/LMNL/lmnl-page.html (accessed 23 March 2012).

Schmidt, D. (forthcoming). Standoff Properties in HRIT.

Thaller, M. (2006). Strings, Texts and Meaning. In *Digital Humanities 2006: Conference Abstracts*. Paris: CATI – Université Paris-Sorbonne, pp. 212-214.

Thaller, M. (2012). What Is a Text within the Digital Humanities, or Some of Them, at Least? In this volume.

Unsworth, J. (2004). Forms of Attention: Digital humanities beyond representation. Paper delivered at *The Face of Text: Computer-Assisted Text Analysis in the Humanities*, the third conference of the Canadian Symposium on Text Analysis (CaSTA), McMaster University, November 19-21, 2004. http://people.lis.illinois.edu/~unsworth/FOA/ (accessed 13 March 2012).

2. What is a text within the Digital Humanities, or some of them, at least? (Manfred Thaller)

(i) The Humanities are a very broad field. The following ideas relate to those Humanities disciplines, which are dealing with 'historical texts' – or at least they started from them. 'Historical' in this context defines any text, which has been created by actors, which we cannot consult any more. This creates a complication when we understand an existing text as a message from a sender to a recipient – an understanding which is absolutely fundamental to modern information technology, as it is the model which has been used within Shannon's article of 1948, one of the corner stones of modern information theory and for most computer scientist, *the* corner stone of Computer Science upon which the later has been built. All of the measures Shannon proposes require an understanding, what the message that has been transmitted by the sender contained before transmission. Another important restriction Shannon acknowledges himself:

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.*

(Shannon 1948: 379)

The fact that information processing systems start with a model which ignores semantics from page one. is ultimately the reason, why meaning has to be added to the signal stream in ways, which allow the transmission (or processing) of that information as an integral part of the signal stream – today usually as embedded markup. Embedded into a signal stream, which has been created by a sender; so embedding anything into it would, according to the model of Shannon, require the markup being part of the transmitted message. This is indeed, what SGML has been created for: To enter the intentions of the

producer of a document about the formatting (and, up to a degree the meaning) of a data stream in such a way, that they would be independent of the requirements of specific devices.

When we are not able to check the meaning of a message with the sender we have to distinguish between the message, even if we do not understand it, and our assumptions about interpreting them. As we do not know the intent of the sender, the result of the 'transmission' of a historical text across time cannot be determined conclusively.

(ii) That data – as transmitted in signal streams – and information, as handled by humans, are not identical is a truism. They have long been seen as separate strata in information theory. (For a recent overview of the discussion see Rowley 2007.) A main difference between Shannon and the 'data – information – knowledge – wisdom' hierarchy has always been, that the former leads directly to an intuitive understanding of systems which can be realized by software engineering, while the later cannot. This is also true of attempts to use a similar scheme to understand information systems, notably Langefors (1995) *infological equation*.

(1) $I = i (D, S, t)$

Information (I) is understood here as the result of a process of interpretation (i) that is applied to data (D), applying previous knowledge (S) within the time available (t). The great attraction of this model is that – unlike Shannon's – it explicitly promises to model the meaning of messages, which are explicitly excluded from consideration by Shannon. To emphasize the difference between the models, we could say that Shannon assumes information to exist *statically*, therefore it can be broken into discrete units, independent of any process dealing with it, while Langefors understands information to be the result of a *dynamic* process, which, having a relationship to time, goes through different stages: So the amount of information existing at tn is not – or not necessarily – equal to the amount of information at tn-1, the ongoing process i having had the chance to produce more of it in the meantime.

The previous knowledge – S – can of course be easily seen as embodied in the interpreting scholar, who looks at the data. For the creation of systems of information processing Thaller (2009a: 228) has shown that Langefors original equation can be developed further. When we assume that knowledge is transformed from a static entity into a dynamic process, as Langefors has proposed for information, we can – via a few steps omitted in this abstract – reach

(2) $Ix = i (I_{x-\alpha}, s(I_{x-\beta}, t), t)$

Roughly: Information at point *x* is the result of the interpretation of an earlier level of information, in the light of knowledge generated from earlier knowledge, at a point of time *t*. As this allows the interpretation of data – e.g. a 'transmission' of a sender not living any more – as a process, which does not have to terminate, it is a better model for the handling of Humanities' texts as Shannon's.

(iii) This abstract model can be turned into an architecture for a representation of information, which can be processed by software. Thaller (2009b) has lead a project team within the digital preservation project PLANETS (cf. http://www.planets-project.eu/), which used this abstract model for the development of tools, which work on the comparison of the information contained within two different representations of an item according to two different technical formats. (Roughly: Does a PDF document contain exactly the same 'text' as a Word document.) For this purpose it is assumed, that all information represented in persistent form on a computer consists of a set of tokens carrying information, which exists within an *n*-dimensional interpretative space, each dimension of that space describing one 'meaning' to be assigned to it. Such a meaning can be a request directed at the rendering system processing the data to render a byte sequence in a specific way, or a connection to a semantic label empowering an information retrieval system. As such a representation is fully recursive, the requirements of formalism (2) above are fulfilled. For texts this can be simplified to an introductory example, where a text is seen as a chain of characters, each of which can be described by arbitrarily many *orthogonal* properties. (Whether the string *Biggin* within a text describes a person or an airfield is independent of whether that string is represented as italics or not; whether the string 'To be or not to be' is assigned to the speaker *Hamlet* is independent of whether it appears on page 13 or 367 of a book.)

(iv) Returning to the argument of section (i) we can see, that there is a direct correspondence between the two arguments. On the one hand the necessity to keep (a) the symbols transmitted within a 'message' from a sender who is irrevocably in the past and (b) our intellectual interpretations of them cleanly and unmistakably separate. On the other hand the necessity to distinguish clearly between (a) the tokens which transmit the data contained within a byte stream and (b) the technical information necessary to interpret that byte stream within a rendering system. If it is useful to transfer information transported within files with different formats into a representation, where the transmitted data are kept completely separate from the technical data needed to interpret them on a technical level, it is highly plausible, that that is even more the case,

when we are discussing interpretations of texts left to us by authors we can not consult any more.

This in turn is highly compatible to an architecture for virtual research environments for manuscript related work, where Humanities' work on historical texts is understood to consist of adding layers of changing and potentially conflicting interpretation unto a set of images of the manuscript to be interpreted. Ebner et al. (2011) have recently described an architecture for a virtual research environment for medieval manuscripts which implements this overall architecture, though using embedded markup for some of the layers for the time being.

To summarize the argument: (1) All texts, for which we cannot consult the producer, should be understood as a sequence of tokens, where we should keep the representation of the tokens and the representation of our interpretation thereof completely separate. (2) Such representations can be grounded in information theory. (3) These representations are useful as blueprints for software on highly divergent levels of abstraction.

2.1. References

Ebner, D., J. Graf, and M. Thaller (2011). A Virtual Research Environment for the Handling of Medieval Charters. Paper presented at the conference *Supporting Digital Humanities: Answering the unaskable*, Copenhagen, November 17-18, 2011 (forthcoming).

Langefors, B. (1995). *Essays on Infology*. Lund: Studentliteratur.

Rowley, J. (2007). The Wisdom Hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 33(2): 163-180.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3)**:** 379-423 and 27(4): 623-656.

Thaller, M. (2009a). The Cologne Information Model: Representing information persistently. In M. Thaller (ed.), *The eXtensible Characterisation Languages – XCL*. Hamburg: Kovač, pp. 223-240.

Thaller, M., ed. (2009b). *The eXtensible Characterisation Languages – XCL*. Hamburg: Kovač.

3. Bringing together markup and semantic annotation (Dino Buzzetti)

Far from having been convincingly clarified, the relation between *markup* and *semantics* still appears to be a perplexing one. The BECHAMEL project, a consistent and systematic attempt to provide a semantics for document markup (Renear et al. 2002), aimed at introducing mechanisms and rules for mapping syntactic markup structures into semantic domains of objects, properties and relations. In a convincing article, however, Dubin and Birnbaum (2004) acknowledge that 'all the distinctions that we're able to explicate using BECHAMEL' could either 'guide the re-tagging of documents with richer markup' or 'be serialized in the form of RDF or a topic map.' (p. 8) But, in the first case, is the prospect of expressing all semantic information through the markup a viable solution? As it has been pointed out, semantic and interpretative encoding prevents interoperability, and since any 'attempt to make a document interoperational' is 'likely to result in decreased expressiveness,' markup scholars are ready to admit that interoperability is 'the wrong goal for scholarly humanities text encoding.' (Bauman 2011) On the other hand, a purely semantic description is clearly incomplete, for it might disregard equally possible and semantically equivalent textual variants.

Keeping inline markup and semantic information distinctly severed proves to be a more appropriate approach. In the case of scholarly digital editions, the sole concern with markup has left us with 'a problem that still exists,' for 'we need (we still need) to demonstrate the usefulness of all the stuff we have digitized over the last decade and more – and usefulness not just in the form of increased access, but specifically, in what we can do with the stuff once we get it' (Unsworth 2003). How can we proceed 'beyond representation' and build tools that shall 'put us into new relationships with our texts' and enable us to process their information content? (Unsworth 2004) Embedded markup can best serve as a comprehensive information carrier for textual information content, but we need further solutions to process content in an efficient and functional way. For embedded markup provides a data structure, but it does not beget a suitable data model. (Raymond 1992, 1996) It defines a format, not a semantics to process its information content. Whether Semantic Web technologies do provide satisfactory data models for humanities research is still an open question, but the problem yet remains how markup and semantic description techniques can be suitably related, by heeding carefully the basic distinction between data and information content. TEI extensions on the one side and the RDFa syntax on the other, do not seem to provide an adequate approach, failing as they do to keep format and content concerns duly severed. The apparent markup overload they produce carries with it a dubious Ptolemaic flavour.

The relation between embedded markup and semantic description languages is an indetermination relationship. Dubin and Birnbaum (2004) fully recognise its very nature: 'the *same markup* can convey different meanings in different contexts,' and 'markup can communicate the *same meaning* in different ways using very different syntax.' It is, on both sides, a one-to-many relation. If you fix the syntax, the semantics may vary in various contexts, and vice versa, if you fix the semantics, you can use a different syntax to express the same content. Contrary to the tenets of hard artificial intelligence – 'if you take care of the syntax, the semantics will take care of itself' (Haugeland 1985: 106) – and of current analytic philosophy of language – 'to give the logical form of a sentence' is to 'bring it within the scope of a semantic theory' (Davidson 1980: 144) – there is no one-to-one correspondence between the logical form of a phrase and the structure of its semantic content. We should not take for granted that by processing a string of characters representing a text, we process its information content, for we can, and often do, process a string without processing the content. And far from being a drawback, this circumstance is actually an advantage, for by dealing with indetermination we can effectively chart variation.

Both the *expression* and the *content* (Hjelmslev 1961) of the text are open to variation. Dealing with textual variants is the task of textual criticism, just as dealing with interpretative variants is that of the literary critic. But we are not at loss in tackling these problems with computational means. We can exploit the ambivalent status of markup to represent the dynamics of variation. (Buzzetti 2009) As a diacritical mark, the markup can be construed either as belonging to the text or as providing an external description of its structure. We may therefore attribute to the markup both a descriptive and a performative function. (Renear 2000) Assumed in its performative capacity, the markup can be seen as an instruction, or as a rule, to determine the semantic structure of the text, whereas taken declaratively it can be equated to a variant of the string of characters that constitutes the text. Referring to a stand-off metadata representation of the textual information content, or to a stand-off markup of sorts with semantic import, we can all the same assume their structural marks in both acceptations, declarative and performative, and get an overall dynamic model of textual and interpretative variation, as shown in Figure 1:

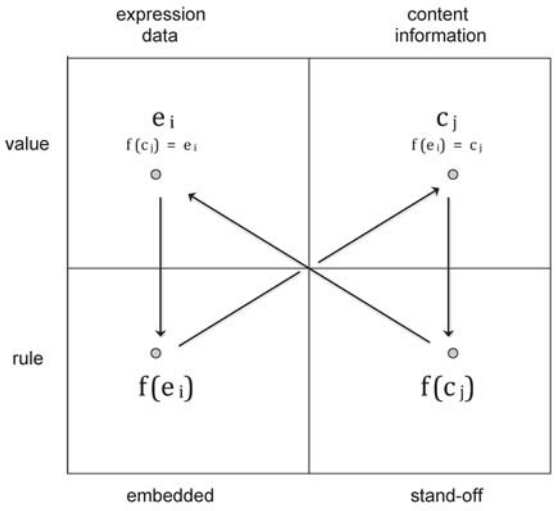


Figure 1

In this diagram, *ei* represents a specific element or construct of the *expression* of the text, conceived of as the set of all tokens that compose it, or E = { *e1* , *e2* , … , *en* }. In its performative capacity that element assumes a different logical status, and can be construed as a function *f (ei) = ci* mapping into the set of all tokens of a given content representation C = { *c1* , *c2* , … , *cn* }, whose specific elements *ci* act in a similar way as a function *f (ci) = ei* mapping into the set E of all the elements of the expression of the text.

Both kinds of variants, textual and interpretative, can be collectively represented, as a kind of 'logical sum' (Thaller 1993: 64), by means of an MVD (Multi-Version Document) graph, as shown by Schmidt and Colomb (2009). Each path of an MVD graph – a directed graph with a start-node and an end-node – represents a different version of the text. A totally isomorphic graph can be obtained also for interpretative variants. In the case of a Topic Maps representation of textual content, an MVD graph was obtained by collating textualized XTM representations of different maps referring to the same text (Isolani et al. 2009).

A comprehensive representation of this kind, of both textual and interpretative variants through MVD graphs, aims at finding efficient ways to determine which paths of the one graph, or which versions of the text, are compatible with specific paths of the other, or with different interpretations of its information content. Both graphs can be used to process the information they represent: the textual variants graph in order to visualise and display different views and versions of the text of a digital edition; the interpretative variants graph in order to process its information content. Promising and different approaches to that end have been proposed by Schmidt (forthcoming), in the context of the HRIT (Humanities Resources, Infrastructure and Tools) project, and by Thaller

(2009), through the development of the XCL (eXtensible Characterisation Language) language. The two methods can offer different implementations of the model here described for specific tasks of the editorial practice, and the pursuit of interoperability between them sets the goal for further development and research.

References

Bauman, S. (2011). Interchange vs. Interoperability. In *Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies*, vol. 7, doi:10.4242/BalisageVol7.Bauman01 (accessed 13 March 2012).

Buzzetti, D. (2009). Digital Editions and Text Processing. In M. Deegan and K. Sutherland (eds.), *Text Editing, Print, and the Digital World*. Aldershot: Ashgate, pp. 45-62.

Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Dubin, D., and D. Birnbaum (2004). Interpretation Beyond Markup. In B. T. Usdin (ed.), *Proceedings of the Extreme Markup Languages 2004 Conference*, Montreal, Quebec, August 2-6, 2004 http://www.ideals.illinois.edu/bitstream/handle/2142/11838/EML2004Dubin01.pdf (accessed 13 March 2012).

Haugeland, J. (1985). *Artificial Intelligence: The very idea*. Cambridge, MA: MIT Press.

Hjelmslev, L. (1961). *Prolegomena to a Theory of Language*. Madison, WI: U of Wisconsin P.

Isolani, A., C. Lorito, Ch. Genovesi, D. Marotta, M. Matteoli, and C. Tozzini (2009). Topic Maps and MVD for the Representations of Interpretative Variants. In *Digital Humanities 2009: Conference Abstracts*, Proceedings of the 2nd ALLC, ACH and SDH-SEMI Joint International Conference (University of Maryland, College Park, June 22-25, 2009), College Park, The Maryland Institute for Technology in the Humanities (MITH), pp. 8-11.

Raymond, D. R., et al. (1992). Markup Reconsidered. Paper presented at the *First International Workshop on Principles of Document Processing*, Washington, DC, 22-23 October 1992. http://www.cs.uwaterloo.ca/~fwtompa/.papers/markup.ps (accessed 13 March 2012).

Raymond, D. R., F. Tompa, and D. Wood (1996). From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML. *Computer Standards and Interfaces* 18(1): 25-36.

Renear, A. (2000). The descriptive/procedural distinction is flawed. *Markup Languages: Theory & Practice* 2(4): 411-20.

Renear, A., M. Sperberg-McQueen, and C. Huitfeldt (2002). Towards a semantics for XML markup. In R. Furuta, J. I. Maletic, and E. Munson (eds.), *DocEng'02: Proceedings of the 2002 ACM Symposium on Document Engineering*, McLean, VA, November 8-9, 2002, New York, NY: ACM Press, pp. 119-126.

Schmidt, D., and R. Colomb (2009). A Data Structure for Representing Multi-version Texts Online. *International Journal of Human Computer Studies* 67(6): 497-514.

Thaller, M. (1993). Historical Information Science: Is There Such a Thing? New Comments on an Old Idea. In T. Orlandi (ed), *Discipline umanistiche e informatica: Il problema dell'integrazione*. Roma: Accademia Nazionale dei Lincei, pp. 51-86.

Thaller, M., ed. (2009). *The eXtensible Characterisation Language: XCL*. Hamburg: Kovač.

Unsworth, J. (2003). Tool-Time, or 'Haven't We Been Here Already?' Ten Years in Humanities Computing. Paper presented at the conference *Transforming Disciplines: The Humanities and Computer Science*, Washington, DC, 17-18 Jnuary 2003. http://people.lis.illinois.edu/~unsworth/carnegie-ninch.03.html (accessed 13 March 2012).

Unsworth, J. (2004). Forms of Attention: Digital humanities beyond representation. Paper delivered at *The Face of Text: Computer-Assisted Text Analysis in the Humanities*, the third conference of the Canadian Symposium on Text Analysis (CaSTA), McMaster University, November 19-21, 2004. http://people.lis.illinois.edu/~unsworth/FOA/ (accessed 13 March 2012).

Myopia: A Visualization Tool in Support of Close Reading

Chaturvedi, Manish
chaturm@muohio.edu
Miami University, USA

Gannod, Gerald
gannodg@muohio.edu
Miami University, USA

Mandell, Laura
Mandell@tamu.edu
Texas A&M University, USA

Armstrong, Helen
armstrh@muohio.edu
Miami University, USA

Hodgson, Eric
eric.hodgson@muohio.edu
Miami University, USA

Although the term 'myopia' typically has negative connotations, focusing on what is nearby is not always a bad idea. Literary critics value what they call 'close reading', the process of carefully reading a poem or other literary work word-by-word and line-by-line in order to analyze how different features of the text – sound, syntax and rhythm work together to create meaning. Through close reading, literary critics hope, the reader might eventually succeed in uncovering hidden connotations. Typically, multiple readings are necessary to understand the structure of a literary work and to unearth multiple meanings, especially in the case of poetry. Only then can the reader perceive the poem's prosody. A non-abstract, enhanced visualization of the metrical, syntactical, and sonic structures of a poem can greatly ease this laborious process.

There has been a recurrent effort to develop text visualization and visual statistical tools to better understand the underlying structures of chosen textual resources. Some of these tool suites provide opportunities for exploration and collaboration in area of textual analysis [5,6,9,10]; others are more narrowly focused on comparison of plain or encoded text, across versions [3, 7]. Despite attempts to do so these digital tools have not resulted in significantly easing the process of close reading [2]. The side-by-side method of highlighting differences is not conducive to spot connotations and analyses, and thus does not help in fully exploiting potential of digital tools in literature.

At Miami University, we are developing the Myopia Poetry Visualization tool to facilitate display of a multidimensional representation of TEI [8] encoded poetry and text. Presenting an interactive visual representation of differently encoded versions of text, the Myopia tool seeks to amplify understanding and uncover new knowledge in the context of close reading. The tool currently is used to visualize text from The Poetess Archive currently moving to Texas A&M University. The Poetess Archive includes a database of electronic documents encoded using the TEI (Text Encoding Initiative) schemas [8] extended with the Poetess Archive tag-set that has been derived from widely used terms in literary analysis and criticism. Even though one schema is used, a schema extended from the TEI schema, the elements of a poem that are interesting to literary scholars have been formalized as tags and attributes based on XML standards. But the tags relevant to any single poem must be spread across multiple documents. XML by its very nature is hierarchical, and these hierarchies cannot overlap within an XML document. The elements of prosodic structures of poetry however do overlap: a metaphor might begin in the middle of one poetic line and end in the middle of another. Different features thus need to be encoded in separate XML files making a composite analysis difficult. Moreover, the XML representation of literary texts is suitable for machine processing and electronic exchange of information, but it is not intuitively comprehensible even by those scholars who know about poetry and would be most likely to adopt an XML encoding system. A tool which fuses TEI encodings spread over separate files into a single visually integrated document could help overcome the limitations of encoding literary texts in XML. We are developing a visualization tool that seeks to integrate multiple encodings while allowing comparative analysis of multiple poems encoded using the extended TEI tag-set. The proposed solution is an interactive visual representation of differently encoded versions of text that can enhance cognition, and aid in uncovering of new knowledge. This approach facilitates identification of frequently changing hotspots in encoded text and aids in the process of close reading.

Close reading means reading a poem word by word and then repeating the process line by line in order to explore sound, syntax, tropological figures, and the meter or underlying rhythm of the work. The reader might eventually succeed in uncovering hidden connotations. Multiple iterations are necessary to understand the structure and thus to unearth meanings from a poem. Only then is the reader in a position to describe underlying prosodic elements in literary text. A non-abstract, enhanced visualization of syntax and sound structures of a poem can greatly ease this laborious process.

The brain's capacity to rapidly process visual information, and discern visual patterns forms the basis for any visualization. Pirolli and Card developed a theory of information access to help designing interfaces to visual cognitive tools. People follow a scent as they forage for information, making just-in-time decisions about which path to follow [11]. An ideal cognitive loop between a computer and human would require the computer providing exactly the information needed at any given time. Thus only relevant information should be on screen, allowing users to forage as they explore the interface. In the Myopia tool, the cost of visual queries is reduced by the minimal layering of visual information, a layering that users can control.

The effectiveness of a given visualization is directly related to the properties of the mapping between the raw data and its rendering. In general any mapping should allow the user to access the underlying data, so that he or she can be assured that the visualization actually relates to the source data being rendered. In our visualization tool the textual content of the encoded poetry being visualized is always available to the user. The ability to revert to the source text at any stage of the visualization from abstract graphical mappings is a conscious design choice that we made in building our tool, and we believe that allowing users to see the text enhances the tool's effectiveness [12]. Myopia employs a direct representation of textual content overlaid with graphical metaphors to aid in literary analysis of poetry.

The User Interface of the *Myopia* tool is split into three distinct visual regions

- The Main Visualization Area that displays and allows users to interact with the rendered text.
- A Key/Legend at bottom of the visualization panel that holds a tabbed display of keys associated with the loaded visualization.
- GUI Panel at the right of the Visualization panel. This allows loading and manipulating source TEI documents and the resulting3D visual elements.

Figure 1: Myopia Tool – Principal Characteristics and Layout

The *Myopia* Poetry Visualization tool has been developed in Python and utilizes open-source Python libraries for rendering multidimensional graphics integrated with an intuitive user interface. The framework used for multidimensional graphics is called Panda3D, which was developed at Carnegie Mellon University [4]. It is useful to mix two-dimensional interface design elements with higher dimension graphical design elements. Navigational controls should always be visible and accessible to the user. In the Poetry Visualization tool the two-dimensional Graphical User Interface (GUI) elements are created using the wxPython framework.

We are conducting user studies to evaluate the effectiveness of our approach in helping the process of close reading. To achieve this objective we utilize the Pre-Experimental Design methodology known as Pretest-Posttest Experiment design ex- plored by Campbell, Stanley et. al. [1]. The experiment seeks to measure the effective- ness, adequacy and usability of the *Myopia* Poetry Visualization tool. The research questions of the study seek to determine the three aspects of the tool: its effectiveness, its adequacy, and its usability.

Reading literature closely is a time-consuming activity. Often combining the results of close reading can be difficult. By using Information Visualization, new knowledge about text can be synthesized. We are presenting a method to support close reading of literary texts using multi-dimensional composite visualizations. Offering an interactive visual representation of a poem's literary qualities, the Poetry Visualization tool al lows for a multidimensional representation of TEI-encoded poetry and text. Presenting an interactive visual representation of differently encoded

versions of text, the tool seeks to amplify understanding and uncover new knowledge. This tool will facilitate identifying hotspots in an encoded text and will visualize the attendant changes in areas of interest across various TEI documents: a poem may be most interesting metaphorically in one line, but metrically it may be more interesting in another. The goal is to help scholars analyze the original documents in new ways and explore information visualization as a rhetorical device in literary criticism. The contributions of this research are in the areas of Information Visualization and Digital Humanities. The purpose of the research is to support reading literature closely using information visualization. This will facilitate abstract analysis of literary texts by fusing multiple perspectives.

Figure 2: Myopia Tool – Multiple Perspectives

References

[1] **Campbell, D. T., and J. C. Stanley** (1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

[2] **Bulger, M., E. Lagresa, et al.** (2008). Close reading re-visited. http://english236-w2008.pbworks.com/Visualization-Project (accessed 11 March 2012).

[3] **Mittelbach, A.** (2009). Tei-comparator in my element. http://blogs.oucs.ox.ac.uk/jamesc/2009/09/04/tei-comparator (accessed 12 March 2012).

[4] **Carnegie Mellon ETC** (2010). Panda 3d - free 3d game engine. http://www.panda3d.org/ (accessed 12 march 2012).

[5] **UC BerkeleyVisualization Lab** (2010). Flare data visualization for the web. http://flare.prefuse.org/ (accessed 12 march 2012). August 2010.

[6] **IBM** (2011). Many eyes. http://manyeyes.alphaworks.ibm.com/manyeyes/ (ac- cessed 12 March 2012).

[7] **Juxta** (2011). Juxta collation software for scholars. http://www.juxtasoftware.org/about.html (accessed 12 March 2012).

[8] **TEI** (2011). TEI: Text encoding initiative. http://www.tei-c.org/index.xml (accessed 12 March 2012).

[9] **University of Nebraska** (2011). Tokenx: a text visualization, analysis, and play tool. http://jetson.unl.edu:8080/cocoon/tokenx/index.html?file=../xml/base.xml (accessed 12 March 2012).

[10] **Bradfford Paley, W.** (2011). Textarc.org home. http://textarc.org/ (accessed 12 March 2012).

[11] **Pirolli, P., S. K. Card, and M. M. van der Wege (**2001). Visual information foraging in a focus + context visualization. *Proceedings of the SIGCHI*

Translation Arrays: Exploring Cultural Heritage Texts Across Languages

Cheesman, Tom
t.cheesman@swansea.ac.uk
Swansea University, UK

Thiel, Stephan
mail@stephanthiel.com
Fachhochschule Potsdam, Germany

Flanagan, Kevin
K.FLANAGAN.667644@swansea.ac.uk
Swansea University, UK

Zhao, Geng
cszg@swansea.ac.uk
Swansea University, UK

Ehrmann, Alison
alison.ehrmann@t-online.de
Swansea University, UK

Laramee, Robert S.
R.S.Laramee@swansea.ac.uk
Swansea University, UK

Hope, Jonathan
jonathan.r.hope@strath.ac.uk
University of Strathclyde, UK

Berry, David M.
d.m.berry@swansea.ac.uk
Swansea University, UK

Astronomers use telescope arrays to create high-resolution images of distant objects. In a Translation Array, by aligning multiple divergent translations (target texts) we can produce new images of the historical and contemporary dynamics of translating cultures, and also of the originals (source texts), by showing how these are refracted through translators' varying interpretations.

Text variation – multiple documentary and curatorial witnesses – is a familiar challenge in digital edition-making (Price 2006; Schmidt & Colomb 2009). But in DH work so far, the 'story' which is told about a text or oeuvre is normally confined to its original language. (An exception is Rybicki 2003, 2006; and cf. Altintas et al. (2007) for a historical linguistics application of a 'time separated' parallel translation corpus.) In fact, the stories of 'cultural heritage texts' (Schmidt 2010) are normally multilingual and multicultural. Can we make it practically possible (and preferably enjoyable as well as informative) to explore, share, and debate the rich knowledge of texts and the world that is encoded in multiple translations?

The fact that multiple translations have remained, until now, an unsuspected resource for DH comes as no surprise from the perspective of Critical Translation Studies (CTS). *The Scandals of Translation* and *The Translator's Invisibility* are two titles by the doyen of the field, Venuti (1995, 1998). Few outside it appreciate the scale of divergence, of multiple kinds, caused by multiple factors, which is normal among multiple translations. The causes of divergence include varying translator competence and individual creativity, framed within target-context-specific norms: linguistic, cultural, political, legal, and ideological constraints and compulsions, which all affect the acts and outputs of translation or retranslation (Gürcaglar 2009). Multiple translations are therefore a very rich resource for transcultural study. They are tailor-made for digital approaches, because each one is implicitly aligned with each other one, and with the original – at least partially.

A Translation Array is a database (a multiple translation text corpus with rich metadata including rich segmental alignments to a source text) with a browser-based interface. It exploits existing and custom-built visualization tools for multi-scale viewing, exploration and creative interaction. Arrays will capture user generated content – both data (texts, text corpora) and metadata: catalogue information, annotations, commentaries, interpretations, algorithmic analyses, micro-alignments, and interlingual re- or back-translations. Data and metadata can be read at various scales. By supporting distant readings (à la Moretti 2005), algorithmic readings (Ramsay 2007; Hope & Witmore 2010), and close, critical and contextual readings, Translation Arrays can be powerful instruments for humanities researchers in diverse fields.

Multiply translated texts include the sacred scriptures of all religions, as well as ancient and modern literature and philosophy, political and historical texts, and so forth. Indeed, (re)translation corpora are coterminous with recorded cultural history. Beyond research applications, therefore, Translation Arrays can become valuable tools for presenting Digital Humanists' work to wider audiences in beguiling, interactive ways, and encouraging creative interaction with cultural heritage, across global cultural and linguistic barriers. They will be welcome in worldwide classrooms – especially multilingual and cross-culturally networked classrooms. Game-like and creative engagement with cultural heritage material is a key feature of translation itself. This will

be facilitated by Array interfaces, alongside more 'serious' modes of engagement. Artists and writers will find this type of resource irresistible. In short, Translation Arrays may become a signature Digital Humanities format, coupling screen-based exploration and interaction with truly global, multilingual and multicultural content – content which has (by definition) high international recognition value.

We will present at DH 2012 our work in progress towards building a prototype. We have an experimental corpus of over 50 different German translations of Shakespeare's *Othello*, 1766-2006, plus samples from and metadata about hundreds of other translations of the play, in (so far) over 40 languages, crowd-sourced at `http://www.delightedbeauty.org/` www.delightedbeauty.org. This site demonstrates the existence of worldwide communities interested in engaging with this type of work. As well as documenting existing translations, some users are also creating their own, ad hoc translations. The site also carries outputs of our project so far, such as: a survey of text visualisation tools, and prototype corpus overview and comparison tools (Geng et al. 2011a, 2011b); discursive text sample analyses (Cheesman 2011, 2012); a Web Science presentation (Wilson et al. 2011); and a paper on using stylometrics to sort multiple translation segments by 'distinctiveness', thereby also enabling a source text to be read in a new way, through its translations' variability (Cheesman et al. 2012).

By summer 2012 we aim to demonstrate a working Array prototype, with an interface designed by Stephan Thiel (see Thiel 2009 for related work). It will deploy Shakespeare's *Othello* and c. 20 German translations, which each reflect cultural-historical and idiosyncratic differences, and which collectively re-interpret the original text in ways which, within the Array interface, can even excite the interest of users who know no German.

The technical and conceptual challenges are fascinating. What is a translation? What is a good or interesting translation? What makes some translations last? Why do people keep retranslating the same texts? Are they in fact 'the same' texts? What do terms in Translation Theory like 'equivalence', 'adequacy' or 'appropriateness' mean in practice? How can machine- and user-generated 'back-translation' best be deployed in an Array?

From a Critical Translation Studies perspective, our 'source-text-centric' approach is questionable. The most interesting translations for CTS are often scarcely source-focused (i.e. not very 'faithful', 'literal', 'close', even 'equivalent'). Often, this means that they cannot easily be aligned with the source. Where diverse 'translations', 'adaptations', 'versions', and 'rewritings' shade into texts which are more loosely 'inspired by' or 'answer back to' source texts and previous reworkings of them, then our approach arguably reaches limits of vaildity. On the other hand, by juxtaposing multiple translations, and by reviving the cultural memory of obscure ones, a Translation Array creates a new and rich context in which radically 'unfaithful' 'un-translations', too, can be read and interpreted in new ways.

Funding

This work was supported in Feb-July 2011 by Swansea University, College of Arts and Humanities, Research Innovation Fund. The work is currently (Feb-Sept 2012) supported by the UK's Arts and Humanities Research Council, Digital Transformations Research Development Fund [AH/J012483/1]. The AHRC-funded team consists of: Principal Investigator, Dr Tom Cheesman (Swansea U); Co-Investigators, Dr Robert S. Laramee (Swansea U) and Dr Jonathan Hope (U Strathclyde); Research Assistant, Kevin Flanagan (Swansea U); Design Consultant, Stephan Thiel (FH Potsdam / Studio Nand, Berlin).

References

Altintas, K., et al. (2007). Language Change Quantification Using Time-separated Parallel Translations. *Literary and Linguistic Computing* 22(4): 375-93.

Cheesman, T. (2011). Thirty Times More Fair Than Black: Shakespeare Retranslation as Political Restatement. *Angermion* 4: 1-51.

Cheesman, T. (2012; forthcoming). "Far More Fair Than Black": Mutations of a Difficult Couplet. In B. Smith and K. Rowe (eds), *Cambridge World Shakespeare Encyclopaedia*, vol. 2: *The World's Shakespeare*. Cambridge: Cambridge UP.

Cheesman, T., and the Version Variation Visualization Team (2012; forthcoming). Translation Sorting: Eddy and Viv in Translation Arrays. In B. Wiggin (ed.), *Un/Translatables*. Evanstone: Northwestern UP.

Geng, Z., T. Cheesman, D. M. Berry, A. Ehrmann, R. S. Laramee, and A. J. Rothwell (2011a). Visualizing Translation Variation of *Othello*: A Survey of Text Visualization and Analysis Tools. `http://cs.swan.ac.uk/~cszg/text/textSurvey.pdf` (accessed February 29, 2012).

Geng, Z., T. Cheesman, D. M. Berry, A.Ehrmann, R. S. Laramee, and A. J. Rothwell, . (2011b). Visualizing Translation Variation: Shakespeare's Othello. In Bebis, G. et al. (eds.), *Advances in Visual Computing: 7th*

International Symposium, ISVC 2011, Part I, LNCS 6938, pp. 657–667. http://www.cs.swan.ac.uk/~cszg/text/isvc.pdf (accessed February 29, 2012).

Gürcaglar, S. T. (2009). Retranslation. In M. Baker and G. Saldanha (eds.), *Encyclopedia of Translation Studies*. Abingdon and New York: Routledge, pp. 232-36.

Hope, J., and M. Witmore (2010). "The Hundredth Psalm to the Tune of 'Green Sleeves'": Digital Approaches to Shakespeare's Language of Genre. *Shakespeare Quarterly* 61(3): 357-90.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Price, K. M. (2008). Electronic Scholarly Editions. In Schreibman, S. and Siemens, R. (eds.), *A Companion to Digital Literary Studies*. Oxford: Blackwell. http://www.digitalhumanities.org/companionDLS/ (accessed February 29, 2012).

Ramsay, S. (2007). Algorithmic Criticism. In S. Schreibman, and R. Siemens (eds.), *A Companion to Digital Literary Studies*. Oxford: Blackwell. http://www.digitalhumanities.org/companionDLS/ (accessed February 29, 2012).

Rybicki, J. (2003). Original Characters. http://www.cyf-kr.edu.pl/~strybick/original_characters.htm (accessed February 29, 2012).

Rybicki, J. (2006). Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. *Literary and Linguistic Computing* 21(1): 91-103.

Schmidt, D. (2010). The Inadequacy of Embedded Markup for Cultural Heritage Texts. *Literary and Linguist Computing* 25(3): 337-56.

Schmidt, D., and R. Colomb (2009). A Data Structure for Representing Multi-version Texts Online. http://www.sciencedirect.com/science/journal/10715819 *International Journal of Human-Computer Studies* 67/(6): 497-514.

Thiel, S. (2009). Understanding Shakespeare. http://www.understanding-shakespeare.com (accessed February 29, 2012).

Venuti, L. (1995). *The Translator's Invisibility: A History of Translation*. London: Routledge.

Venuti, L. (1998), *The Scandals of Translation: Towards an Ethics of Difference*. London: Routledge.

Wilson, M. L., Z. Geng, R. S. Laramee, T. Cheesman, A. J. Rothwell, D. M. Berry, and A. Ehrmann (2011). Studying Variations in Culture and Literature: Visualizing Translation Variations in Shakespeare's *Othello*. Poster paper at the ACM Web Science 2011 Conference, Koblenz, Germany, 14-17 June 2011. Poster attached to http://www.delightedbeauty.org Outputs.

Constructing a Chinese as Second Language Learner Corpus for Language Learning and Research

Chen, Howard
hjchen@ntnu.edu.tw
National Taiwan Normal University, Taiwan

Many researchers and language teachers around the world believe that language corpora have great potentials for improving second/foreign language learning and teaching. Among various types of corpora resources, learner corpora in particular have received much attention recently. One of the most influential projects is the ICLE (International Corpus of Learner English) project led by Professor Granger in University of Louvain, Belgium. The new version of ICLE corpus contains 3.7 million words of EFL writing from learners representing 16 mother tongue backgrounds. The ICLE corpus has helped to produce many research papers and pedagogical materials within the past decade. Although useful English learner corpora such as ICLE are widely available, very few learner corpora for other target languages are available.

Recently, because of the rapid economic growth in China, an increasing number of students are learning Chinese as a second language. Although the number of Chinese as second language (CSL) learners is increasing rapidly around the world; very few CSL learner corpora are available for teaching, learning, and research. For CSL research, learner corpus can play an important role. CSL teachers and researchers can conduct research on learners' interlanguage development and gain insights about leaner's difficulties and needs. Material writers can further use the results of error analysis to produce useful pedagogical materials. CSL leaner corpus might also be used to better understand the differences among learners at different proficiency levels (cf. Cambridge English Profile Project).

This paper will introduce a new Chinese as second language learner corpus and related corpus search tools developed by MTC (Mandarin Teaching Center) and SC-TOP (Steering Committee of Test of Proficiency) in Taiwan. MTC is located at National Taiwan Normal University and it is the largest Chinese teaching centers in Taiwan. There are more than 1600 students enrolled in each quarter, and there are more than 150 teachers in this center. Students from more than 70 countries are studying in this center. SC-TOP is a language testing research center sponsored by Ministry of Education for developing various Chinese as a second language proficiency tests. Based on the data provided by these two centers, a 3 million word Chinese as a second language learner corpus has been developed. The MTC-TOP learner corpus includes the following two different types of learner data:

1. CSL learners' short essays written in various TOP tests.
2. CSL students' writing assignments at MTC

To facilitate corpus search, the learner corpus was further automatically tagged with a Chinese tagger called CKIP (Chinese Knowledge Information Processing) tagger developed by Academia Sinica, Taiwan. The POS-tagged CSL corpus is useful for research and teaching. In addition to the learner corpus, a web concordancer which has several different search options was also developed. This web concordancer allows users retrieve specific words and phrases from CSL learner corpus. Thus, various CSL learners' usage can be retrieved and studied more easily and systematically. Furthermore, the POS-tagged learner corpus can be used to search for collocates used by CSL learners.

In addition, it is also important to further analyze various errors made by CSL learners. Since it is not possible for computers to identify errors, native speakers were asked to tag this learner corpus. 25 types of major errors were first identified, and about 800000 words were tagged so far. These manually tagged errors can also be searched via a web interface. Because these learner data were produced by learners from various L1 language backgrounds, teachers and researchers can also find errors and patterns produced by CSL learners from different native language backgrounds.

To illustrate how this MTC-TOP learner corpus can benefit CSL research, we used this learner corpus to conduct a study on the acquisition of Chinese classifiers by various CSL learners. Chinese classifier is a notoriously difficult language feature for many CSL learners. In the past, most studies on Chinese classifiers learning often involved few subjects and were based on very limited number of learner errors. With the help of this 3 million words CSL corpus, the classifiers errors made by many learners from various first language backgrounds can be found more easily and examined more systematically. The analysis on classifier errors produced by different CSL learners also helps researchers to better identify the common Chinese classifiers errors. In addition to classifiers, other common errors made by CSL learners can also be investigated with the help of computerized corpus. Several problematic areas (Chinese synonyms and Chinese particle le) will be further discussed in this paper. The CSL learner

corpus and the web concordancer should be able to help more researchers uncover CSL interlanguage patterns and conduct various types of research. It is evident that this CSL learner corpus can make significant impact on CSL teaching, learning, and research.

Social Curation of large multimedia collections on the cloud

Chong, Dazhi
dchong@odu.edu
Old Dominion University, USA

Coppage, Samuel
scoppage@odu.edu
Old Dominion University, USA

Gu, Xiangyi
xxxgu001@odu.edu
Old Dominion University, USA

Maly, Kurt
maly@cs.odu.edu
Old Dominion University, USA

Wu, Harris
hwu@odu.edu
Old Dominion University, USA

Zubair, Mohammad
zubair@cs.odu.edu
Old Dominion University, USA

1. Introduction and Prior Work

In previous Digital Humanities conferences we presented the Facet System, a system that improves browsing and searching of a large collection by supporting users to build a faceted (multi-perspective) classification schema collaboratively (Georges et al. 2008; Fu et al. 2010). The system is targeted for multimedia collections that have little textual metadata. In summary, our system (a) allows users to build and maintain a faceted classification collaboratively, (b) enriches the user-created facet schema systematically, and (c) classifies documents into an evolving facet schema automatically. It is hoped that through users' collective efforts the faceted classification schema will evolve along with the users' interests and thus help them navigate the collection. For the browsing and classification interfaces of the Facet System see Fig. 1 and Fig. 2. This paper discusses scalability challenges and improvements made to the system, cloud deployment, user studies and evaluation results.

Figure 1: Browsing of the system enabled by user-evolved faceted classification

Figure 2: The click-and-drag classification screen

2. Scalability Issues and System Improvement

Our Facet System allows users to collaboratively evolve a common, global schema that contains facets (perspectives such as *colour*) and categories (such as red and green in the *colour* facet) for a large multimedia collection, using a wiki-inspired interface. The global schema can then used by any users for browsing or searching the collection. The variety of user perspectives increases with the number of users. Even though we have taken measures to cure and reduce the global schema, we find that the global schema alone faces limitations in supporting a large number of users. Many users prefer to focus on a smaller subset of the schema or collection. To address this issue, we have improved the user interface to allow users switch between global and personal schemas. The personal schema allows users to organize digital documents in their own way. A group of people can share the same personal schema which applies to a subset of the large digital collection. On the backend, we developed a 'merging' algorithm that automatically enriches the global schema by merging items, categories and facets from personal schema into the global schema. The algorithm applies statistical techniques to item-category associations, category-subcategory and category-facet associations, and Wordnet-based similarities. Figure 3 shows both global and personal schemas that can be utilized by individual users or user groups. More technical details will be presented at the conference.

Figure 3: Global and personal (or local) schemas

3. Cloud Deployment

A large repository with millions of multimedia items may consume terabytes of storage. To support millions of users, the system needs to be partitioned, mirrored, or load balanced across an array of servers. When user traffic and growth trends are not perfectly predicable, cloud computing is not only the most cost effective but also likely the only viable option.

After evaluating several cloud service providers, we find that Microsoft Azure Cloud provides the best cost-effectiveness for us to host large digital repositories that require scalable database support as well as network bandwidth. Our Facet System was built on an open source technology stack including Joomla, PHP and MySQL. Joomla is a leading open source content management system implemented in PHP. Initially we deployed the system on two server layers, Web (Joomla/PHP) and Database (MySQL) layers, utilizing Web and Worker roles in the Azure cloud (see Fig. 4). The scalability and manageability of the MySQL database turned out to be prohibitive. We found that true database scalability can be accomplished only by cloud-tailored solutions such as Microsoft SQL Azure. We spent significant efforts in porting Joomla from MySQL to Microsoft SQL Azure. In addition, we moved Joomla's session management from the default storage in operating system files to Azure storage, to enable stateless load balancing across many servers on the Web layer. During the conference we will present more technical details which may benefit users of Joomla and other open source content management systems.

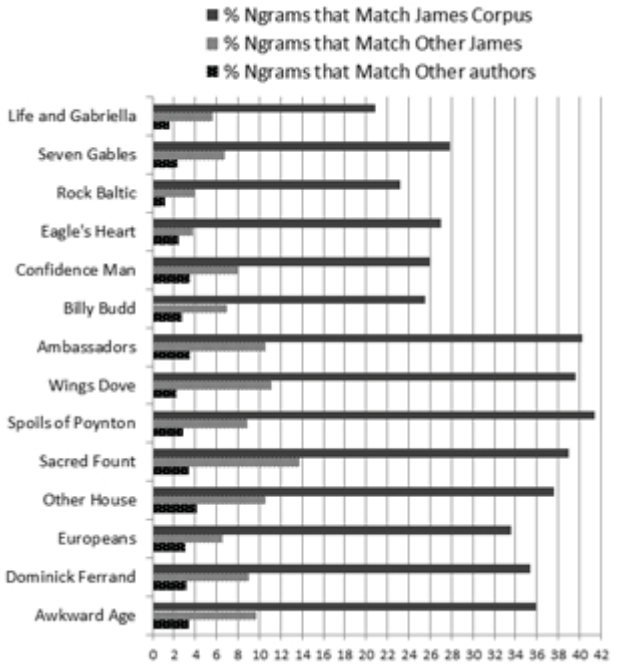


Figure 4: The deployment architecture in Microsoft Azure

4. Evaluation and user studies

We evaluated the system using 15 classes (different sessions of a university-wide undergraduate introductory information technology class) during Spring 2011, and continue the evaluation with 22 classes during Fall 2011 semester using the cloud deployment of our system. Each class has around 30 students. The system was used to teach the digital humanities and social computing component of the class, by asking students to collaboratively curate a special library collection of un-cataloged local newspaper clippings and images on African American activism in United States. We captured detailed usage information as users completed a series of tasks involving searching, classification, tagging and voting (providing feedback on digital items and metadata). Each class of students work as a group using a single personal schema, which may be better labeled as 'local schema' in the future to avoid confusion. A backend algorithm periodically merges categories and facets from these local schemas into the global schema. While global schema is useful, we find that the students spend a lot of efforts on curating the collection through building the local schemas for their own classes. Therefore we refer to our current system a Social Curation System. The performance of the system and cloud deployment proves to be satisfactory and actually exceeded our expectation in certain aspects.

A common challenge to curating digital collections is the lack of expertise necessary to classify, catalog and create metadata structures for such collections. The evaluation shows the feasibility of social curation by non-expert users supplemented by suitable algorithms for reinforcing, combining and judging user created associations and tags. From a social science perspective, we are interested in the relationship between student's prior experience and attitudes about social computing and their current behavior in a social curation environment. We used a pre-usage instrument to assess the users' prior experience with common social sites, then a post-usage instrument to measure post-usage attitudes and conceptual understanding of the resulting metadata among the users. We also measured students' knowledge in digital humanities and African American activism before and after the system usage. The change in perceived usefulness of the archive from the beginning of the experiment to the end was also measured. We will present the result of the data analysis at the conference and hope that it will help future designs and implementations of socially curated archives.

5. Related Work and Conclusion

There are many large scale multimedia collections on the Internet. Few of them, however, contain library-grade metadata for digital items. Researchers have proposed or studied how to build large, curated collections on the Internet. Recent Digital Humanities conferences have showcased efforts in trying to utilize multiple organizations' collaborative efforts in building federated or related collections (Timothy et al. 2011). Our project utilizes cloud computing to explore the scalability limitations of collaborative curation, and implement algorithmic approaches to synthesize individual users' or groups' curation efforts. Our work on backend algorithms are build upon recent advancements in ontology building, text categorization and link mining (e.g. Wu et al., 2006a, 2006b; Heymann et al. 2006; Joachims et al. 1998;Schmitz et al. 2006). Evaluation results show the promise of our social curation approach.

Funding

This work was supported by the National Science Foundation [grant number 0713290].

References

Fu, L., K. Maly, H. Wu, and M. Zubair (2010). Building Dynamic Image Collections from Internet. *Digital Humanities 2010*, London, UK, 2010.

Georges, A., K. Maly, M. Milena, H. Wu, and M. Zubair (2008). Exploring Historical Image Collections with Collaborative Faceted Classification. *Digital Humanities 2008*, Oulu, Finland, 2008.

Heymann, P., and H. Garcia-Molina (2006).Collaborative Creation of Communal

Hierarchical Taxonomies in Social Tagging Systems. *Stanford Technical Report InfoLab*.

Joachims, T. (1998) Text categorization with support vector machines. *10th European Conference on Machine Learning*. London, 1998.

Schmitz, P., and Yahoo! Research (2006). Inducing Ontology from Flickr Tags. *Workshop in Collaborative Web Tagging, WWW 2006, 2006,* Edinburgh, UK.

Cole, T., N. Fraistat, D. Greenbaum, D. Lester, and E. Millon (2011). Bamboo Technology Project: Building Cyberinfrastructure for the Arts and Humanities. *Digital Humanities 2011*, Stanford, CA, 2011.

Wu, H., M. Zubair, and K. Maly (2006). Harvesting Social Knowledge from Folksonomies. *ACM 17th Conference on Hypertext and Hypermedia*, 111-114, Odense, Demark, August 20-25, 2006.

Wu, H., M. Gordon, K. DeMaagd, and W. Fan (2006). Mining Web Navigations for Intelligence. *Decision Support Systems* 41(3): 574-591.

Sounding for Meaning: Analyzing Aural Patterns Across Large Digital Collections

Clement, Tanya
tclement@umd.edu
University of Texas at Austin, USA

Auvil, Loretta
lauvil@illinois.edu
University of Illinois, Urbana-Champaign, USA

Tcheng, David
davidtcheng@gmail.com
University of Illinois, Urbana-Champaign, USA

Capitanu, Boris
capitanu@ncsa.uiuc.edu
University of Illinois, Urbana-Champaign

Monroe, Megan
madey.j@gmail.com
University of Maryland, USA

Goel, Ankita
gargankita@gmail.com
University of Texas at Austin, USA

1. Introduction

This paper will discuss a case study that uses theories of knowledge representation and research on phonetic symbolism to develop analytics and visualizations that help users examine aural and prosodic patterns in text. This use case is supported by the Andrew W. Mellon Foundation through a grant titled 'SEASR Services.' Other collaborators include Humanities professors from Stanford, George Mason University, and University of Illinois Urbana-Champaign. All of the use cases within the project include research on how humanities scholars can use textual analytics and predictive modeling with visualizations in order to help scholars interpret large digital collections. In particular, while this paper will briefly describe a new reading of *Tender Buttons* (1914) by Gertrude Stein that was facilitated by text analysis, predictive modeling, and visualization, but this discussion is primarily focused on how the theories, research, and methodologies that underpin this work facilitate digital diversity.

2. The logic and ontologies of aurality

Relying on the research of literary scholars and psychologists concerning how humanists make meaning with sound, we have chosen to use the logics and ontologies of aurality to model and analyze sound in text. If we are unclear about *what* we mean and *how* we mean when we seek to represent 'sound,' we discourage diverse readings from those who wish to understand the results of the computational analytics we apply to that model. In order to create a methodological approach that supports digital diversity, we seek to be transparent about the logics and ontologies of sound we are using. To this end, we have listed below each of the five roles that knowledge representation plays in a project according to Davis et al. (qtd. in Unsworth) and how these roles play out in the methodologies we have chosen to use in our use case.

(1) 'A knowledge representation is most fundamentally a surrogate, a substitute for the thing itself' (Davis et al. 1993).

In this project, we are defining sound as *the pre-speech promise of sound* as it is signified within the structure and syntax of text. Charles Bernstein focuses on the 'aurality' of text, which he calls the 'sounding of the writing' (Bernstein 13). 'Orality,' Bernstein writes, 'can be understood as a stylistic or even ideological marker or a reading style; in contrast, the audiotext might more usefully be understood as aural – what the ear hears . . . *Aurality precedes orality*, just as language precedes speech' (Bernstein 13). Using textual analytics to create a surrogate of *potential* sound is an essential aspect of the theory of aurality underpinning the methods in this project.

(2) Knowledge Representation 'is a set of ontological commitments, i.e., an answer to the question: In what terms should I think about the world?' (Davis et al. 1993).

In this project, we are committed to ontologies of sound that define sound as a meaningful aspect of text. This debate concerning whether or not sound makes meaning and how has a long history.[1] These ideas, that the meaning of sound correlates to the structure of the text underpins the ontologies we have chosen for this project. Ultimately, the act of reading requires digital diversity since readers create their own sense making systems by making connections and understanding patterns. The same is true of reading sound: computers do an excellent job of modeling the features with which readers make these kinds of connections.

(3) Knowledge Representation 'is a fragmentary theory of intelligent reasoning, expressed in terms of three components: (i) the representation's fundamental conception of intelligent reasoning; (ii) the set of inferences the representation sanctions; and (iii) the set of inferences it recommends' (Davis et al. 1993).

The above theories in aurality and research in phonetic and prosodic symbolism underpin our choice to use OpenMary, an open-source text-to-speech system, as a fundamental first step to generate the features that correspond to sound in texts. Developed as a collaborative project of Das Deutsche Forschungszentrum für Künstliche Intelligenz (GmbH) (German Research Center for Artificial Intelligence) Language Technology Lab and the Institute of Phonetics at Saarland University, OpenMary's rule set or algorithm for 'intelligent reasoning' is based on the research of both linguists and computer scientists. Created to generate audio files of spoken text from written transcripts, OpenMary captures information about the structure of the text (features) that make it possible for a computer to read and create speech that is comprehensible to humans in multiple languages (German, British and American English, Telugu, Turkish, and Russian; more languages are in preparation). Specifically, OpenMary is a system that accepts text input and creates an XML document (MaryXML) that includes 'tokens along with their textual form, part of speech, phonological transcription, pitch accents etc., as well as prosodic phrasing,' all of which are important indicators of how the text could be 'heard' by a reader ('Adding support for a new language to MARY TTS') (see Figure 1)

(4) Knowledge Representation 'is a medium for pragmatically efficient computation' (Davis et al. 1993).

SEASR's Meandre provides basic infrastructure for data-intensive computation, and, among others, tools for creating components and flows, a high-level language to describe flows, and a multicore and distributed execution environment based on a service-oriented paradigm. The workflow we created in Meandre processes each document in our collection through the OpenMary web service at a paragraph level to create a tabular representation of the data that also includes features that maintain the structure of the documents (chapter id, section id, paragraph id, sentence id, phrase id). Each word is parsed out into accent, phoneme and part-of-speech in a way that keeps the word's association with the phrase, sentence, and paragraph boundaries. The ultimate benefit to creating this flow in Meandre is digital diversity: it is meant to be accessible to future

users who wish to produce or tweak similar results (see Figure 2 and Figure 3).

In order to make comparisons evident with predictive modeling, we have also developed an instance-based, machine-learning algorithm for predictive analysis. We use an instance-based learning algorithm because it was thought to make similar mistakes as humans do (Gonzalez, Lerch, Lebiere). To focus the machine learning on prosody, we only used features that research has shown reflects prosody including part-of-speech, accent, stress, tone, and break index (marking the boundaries of syntactic units) (Bolinger). To further bias the system towards human performance, we selected a window size of fourteen phonemes, because it was the average phrase length and research has shown that human prosody comparisons are made at the phrase level (Soderstrom et al. 2003).

(5) Knowledge Representation 'is a medium of human expression, i.e., a language in which we say things about the world' (Davis et al. 1993).

An essential aspect of this project is ProseVis, a visualization tool we developed that allows a user to map the results of these analytics to the 'original' text in human readable form. This allows for the simultaneous consideration of multiple representations of knowledge or digital diversity. Digital Humanities scholars have also used phonetic symbolic research in the past to create tools that mark and analyze sound in poetry (Plamondon 2006; Smolinsky & Sokoloff 2006). With ProseVis, we also allow users to identify patterns in the analyzed texts by facilitating their ability to highlight different features within the OpenMary data (see Figure 4) as well as the predictive modeling results (see Figure 5) mapped onto the original text. The visualization renders a given document as a series of rows that correspond to any level in the document hierarchy (chapter/stanza/group, section, paragraph/line, sentence, phrase). In these ways, the reader can examine prosodic patterns as they occur at the beginning or end of these text divisions. Further, ProseVis represents the OpenMary data[2] within ProseVis by breaking down the syllables into consonant and vowel components such as:

Word	Sound	Lead Consonant	Vowel Sound	End Consonant
Strike	s tr I ke	s	I	Ke

This breakdown provides the reader with a finer-grained level of analysis, as well as a simplified coloring scheme. See Figure 6 (beginning sound), Figure 7 (vowel sound), and Figure 8 (end sound).

3. Re-reading Tender Buttons by Gertrude Stein

Our predictive modeling experiment was to predict *Tender Buttons* out of a collection of nine books. The prediction was based on comparing each moving window of fourteen phonemes, with each phoneme represented solely by its aural features listed above. The hypothesis was that Gertrude Stein's *Tender Buttons* (1914) would be most confused with *The New England Cook Book* (Turner 1905). Margueritte S. Murphy hypothesizes that *Tender Buttons* 'takes the form of domestic guides to living: cookbooks, housekeeping guides, books of etiquette, guides to entertaining, maxims of interior design, fashion advice' (Murphy 1991, p. 389). Figure 9 visualizes the results of our predictive analysis by showing the percentage of times the computer confuses its predictions for each of the nine books. In the visualization, row two shows the results for *The New England Cook Book* with *Tender Buttons* as the text that the computer 'confuses' with the highest percentage when trying to predict the cookbook. By modeling the *possibility* of sound with pre-speech features marking aurality, we are inviting diverse modes of discovery in textual sound.

```
<prosody pitch="+1%" range="+6%">
<phrase>
<t accent="L+H*" g2p_method="lexicon" ph="r I - ' p i - t I N" pos="NN">
Repeating
</t>
<t g2p_method="lexicon" ph="' D E n" pos="RB">
then
</t>
<t g2p_method="lexicon" ph="' I z" pos="VBZ">
is
</t>
<t g2p_method="lexicon" ph="' I n" pos="IN">
in
</t>
<t g2p_method="lexicon" ph="' E - v r= - i" pos="DT">
every
</t>
<t g2p_method="lexicon" ph="' w V n" pos="CD">
one
</t>
<t pos=",">
,
</t>
<boundary breakindex="4" tone="H-L%"/>
</phrase>
</prosody>
```

Figure 1: Shows the phrase 'Repeating then is in everyone,' from Gertrude Stein's text The Making of Americans

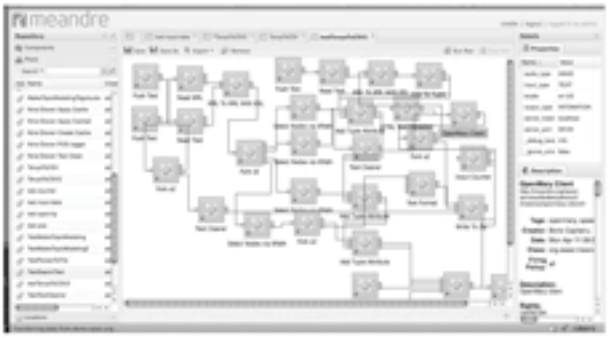


Figure 2: Shows the flow with the components that are used for executing OpenMary and post-processing the data to create the database tables

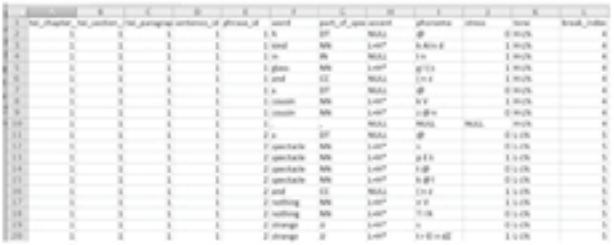


Figure 3: This is a sample of the tabular output of Tender Buttons by Gertrude Stein created within Meandre

Figure 4: Sounds of each syllable in Tender Buttons by Gertrude Stein visualized using ProseVis

Figure 5: Predictive Modeling data comparing Tender Buttons to seven other texts visualized using ProseVis

Figure 6: Beginning Sound of each syllable in Tender Buttons by Gertrude Stein visualized using ProseVis

Figure 7: Vowel sound of each syllable in Tender Buttons by Gertrude Stein visualized using ProseVis

Figure 8: End sound of each syllable in Tender Buttons visualized using ProseVis

Figure 9: These are the results of a predictive modeling algorithm that was run on nine texts. Each color represents the percentage of times the algorithm predicts one of the nine in its attempt to pick a given text

References

Adding support for a new language to MARY TTS. MARY Text To Speech. Web. 8 Sept. 2011.

Bernstein, C. (1998). *Close Listening: Poetry and the Performed Word*. Oxford: Oxford UP.

Bolinger, D. (1986). *Intonation and Its Parts: Melody in Spoken English*. Stanford, CA: Stanford UP.

Clement, T. (2008). *'A thing not beginning or ending': Using Digital Tools to Distant-Read Gertrude Stein's* The Making of Americans. *Literary and Linguistic Computing* 23(3): 361-82.

Notes

1. See Bolinger, Cole, Newman, Ong, Plato Sapir, Saussure Shrum and Lowrey, and Tsur.
2. These sounds are documented here `http://mary.opendfki.de/wiki/USEnglishSAMPA` . However, over the course of testing ProseVis, we uncovered two additional vowel components, @U and EI, which are now included in the implementation.

The Programming Historian 2: A Participatory Textbook

Crymble, Adam H.
acrymbl@uwo.ca
Network in Canadian History & Environment, UK

MacEachern, Alan
amaceach@uwo.ca
Network in Canadian History & Environment, UK

Turkel, William J.
wturkel@uwo.ca
Network in Canadian History & Environment, UK

The second edition of *The Programming Historian (PH2)* is an open access and open source introductory text designed to teach the fundamentals of programming to humanists. The first edition (Turkel & MacEachern 2007), while popular, is basically a traditional text presented online. It was limited to the lessons that we provided and followed a single narrative arc. The second edition, also online, has been restructured to invite ongoing community participation and authorship.

The project is designed with a core sequence of lessons that get readers up to speed with basic programming concepts. These show how to programmatically download sources from the Internet, process them for keywords, and visualize some textual attributes. Contributors can introduce new lessons at any point by branching off from a previous lesson. They might, for example, introduce a related technique, or show how to implement something in another programming language, on another platform, or using different kinds of sources. The branching structure allows members of the digital humanities community to add lessons based on their own specialties, or to create new sequences for use in the classroom, while still guaranteeing that readers will have the background required to understand a particular branch if their study of *PH2* has taken them to that point. Each lesson is intended to take two hours or less to complete.

For contributors, *PH2* provides a peer-reviewed platform for methodology. Both literary and technical reviewers vet all submissions with quality emphasized over quantity. Lessons are presented in a custom WordPress installation and code snippets are stored in GitHub for easy maintenance and sharing. To make the site as easy as possible to use, code snippets are pulled directly into the WordPress posts making GitHub essentially invisible for both

authors and readers not familiar with or interested in its features. Individual lessons are attributed to their authors, and reviewers and other contributors receive public credit for their labour. For learners, *PH2* focuses on the use of programming to solve problems common to academic research, such as obtaining, analyzing, processing, and mashing up sources. While there are hundreds of programming books on the market, very few teach these kinds of hands-on tasks that are crucial to the practice of the digital humanities.

With the first edition of the *Programming Historian*, we showed that there is a great demand for programming introduction that focuses on the day-to-day work of digital humanists both historians and otherwise. We have commissioned a number of new lessons from guest authors for the second edition and hope to open our submission process to the wider community shortly. From our past experience we know that people well beyond our target audience of digital humanists can and will use the project as a way to delve into programming. When *PH2* launches publicly, we believe that the digital humanities community will embrace and expand it and we look forward to working with them.

Funding

This work was supported by *The Network in Canadian History & Environment,* and the *Roy Rosenzweig Center for History and New Media*.

References

Github . http://github.com

Wordpress. http://wordpress.org (accessed 1 March 2012).

Turkel, W. J. and A. MacEachern (2007). *The Programming Historian 1st Edition.* http://niche-canada.org/programming-historian (accessed 1 March 2012).

Turkel, W. J., A. Crymble, J. Boggs, M. Posner, et al., eds. (2012). *The Programming Historian 2.* http://chnm.gmu.edu/press/programminghistorian/ (accessed 1 March 2012).

Multilingual and Semantic Extension of Folk Tale Catalogues

Declerck, Thierry
declerck@dfki.de
DFKI GmbH, Germany, ICLTT, Austrian Academy of Sciences, Austria

Lendvai, Piroska
lendvai.piroska@nytud.hu
Research Institute for Linguistics, Hungary

Darányi, Sándor
sandor.daranyi@hb.se
Swedish School of Library and Information Science, Sweden

1. Introduction

The modelling of phenomena related to higher-order knowledge mechanisms poses significant challenges to research in any domain, this being also valid for narratives, which 'are an important form of knowledge representation', as (Tuffield et al. 2006) state. In the context of the AMICUS network[1] we are more specifically interested in the representation of motifs in narratives. Motifs are recurring conceptual, textual, audio or visual units appearing in artefacts – in folk tales, they can be seen as cognitively complex notions (e.g. 'Rescue of princess', 'Helpful animal', 'Cruel stepmother', etc.), expressed by lexically and syntactically variable narrative structures. The modelling of such motifs, including their typical realizations in natural language form, can help in supporting the automatic motif detection in large (multilingual) text collections[2]. But this kind of formal representation of motifs is so far unresolved, and thus a large amount of cultural heritage collections of which motifs are typical constructive units can still only be manually indexed, which significantly limits access to these resources and to this type of knowledge. Recently, we started to investigate the utility of linguistic and semantic analysis and mark-up of motifs (Lendvai et al. 2010), which we would like to extend and apply to studies on motif sequencing (Darányi, Wittek & Forró 2012).

In the current study, we address two priorities of the Digital Humanities discipline: devising procedures that integrate semantic enhancement of legacy folk tale indexes, classification systems and taxonomies (Declerck & Lendvai 2011) and their automatized translation (Mörth et al. 2011). For this study we are dealing with two extended catalogues that

hold conceptual schemes of narrative elements from folktales, ballads, myths, and related genres: the *Thompson's Motif-Index of Folk-Literature*, TMI (Thompson 1955) and *The Types of International Folktales*, ATU (Uther 2004). TMI has an available on-line version[3], only in English language, but the digitized ATU is not yet available on-line in its entirety, only some of its subsections are reproduced in Wikipedia, in various languages[4].

TMI indexes and ATU types are both combined with extensive labels, and a novel approach to those resources is that this combination can be linguistically processed (Declerck & Lendvai 2011), semantically represented, and, ultimately, turned into domain-specific ontology classes that can be interlinked. The upgrade leads to semantically enriched catalogues that qualify as interoperable language technology resources that can be harnessed to assign text units to the folktale domain classes, creating in an automatized way indexed folktale corpora. As part of such a normalization process, catalogues have to be made interoperable with each other, and stored in a semantically harmonized representation, like the SKOS[5] standard, which is using RDF[6] as its formal representation language.

2. Towards Semantic and Multilingual Extensions of TMI

The Thompson's Motif-Index of Folktale-Literature is organized by alphanumeric indexes, which resembles a taxonomy structure of motifs, but does not express hierarchy or inheritance properties. i.e. it is not made explicit that some elements of the taxonomy introduce mere classification information over a span of labels *('A0-A99. Creator'*, split into finer-grained subclasses, e.g. *'A20. Origin of the creator.')*, that some elements are abstractions of motifs *('A21. /Creator from above./')*, and that some elements are summaries of a concrete motif, supplying source information as well *('A21.1. / Woman who fell from the sky./--Daughter of the sky-chief falls from the sky, is caught by birds, and lowered to the surface of the water. She becomes the creator.--*Iroquois: Thompson Tales n.27.--Cf. Finnish: Kalevala rune 1.')*.

We prepared a program that converts TMI to an XML representation and marks such properties explicitly by designated tags, as exemplified below:

<label class="TMI_A0" span="A0-A99" type="*abstract*" lang="en">Creator</label>

<label class="TMI_A20" span="A21-A27" type="*abstract*" Property_Of="A0" lang="en">Origin of the Creator</label>

<label class="TMI_A21" span="A21.1-A21.2" type="*abstract*" SubClassOf="A20" lang="en">Creator from Above</label>

<label class="TMI_A21.1" span="A21.1-A21.1" type="*concrete*" SubClassOf="A21" lang="en">Woman who fell from the sky</label>

This representation makes explicit the fact that the natural language expressions (like 'Creator from Above') are labels of classes that we explicitly organise in a class hierarchy. The feature 'type' can take two values: 'abstract' or 'concrete'. The latter is indicating that the index (for example: A21.1) is pointing to concrete examples in selected tales. The 'span' feature is indicating the number (from 1 to many) of subsumed indexes. We noticed that indexes ('classes' in our terminology), which are in fact leaves (spanning only over their own number) point all to concrete tales, and can thus be considered as first level motifs, whereas the other classes have more a classification role.

Ongoing work is dedicated to upgrading the XML representation to SKOS-RDF, to provide adequate means for differentiating between hierarchical realizations and properties associated with classes, and the possibility to compute inheritance properties o of the class hierarchy. SKOS-RDF is also appropriated for publishing the enriched TMI resource on the LOD. We display below a simplified example (where 'TMI_A*' are shortcuts for URIs):

<TMI_A0> rdf:type skos:Concept ;

skos:prefLabel "Creator"@en.

<TMI_A20> rdf:type skos:Concept ;

skos:prefLabel "Origin of the Creator"@en

skos:related <TMI_A0>.

<TMI_A21> rdf:type skos:Concept ;

skos:prefLabel "Creator from Above"@en.

<TMI_A21.1> rdf:type skos:Concept ;

skos:prefLabel "Woman who fell from the sky"@en ;

skos:broader <TMI_A21>

In parallel, we target the extension of motifs listed in TMI in English into a multilingual version. This is carried out by accessing the *Wiktionary* lexicon[7], via the LOD-compliant *lexvo* [8] service. Actual work (Declerck et al. 2012) is aiming at extracting from Wiktionary a multilingual lexical semantics network for the labels included in TMI.

3. Towards the Harmonization of Multilingual ATU On-line Data

As we mentioned above, only segments of ATU are available on-line, in the context of Wikipedia articles, in different languages. We note the following discrepancies in these:

- (EN) 310 Rapunzel (Italian,Italian, Greek, Italian)
- (DE) AaTh 310 Jungfrau im Turm KHM 12
- (FR) AT 310: 'La Fille dans la tour'(*The Maiden in the Tower*): version allemande

The English Wikipedia links *Rapunzel* to four Wikipedia pages on tales belonging to the same ATU type. Rhe Wikipedia page for the original Rapunzel tale is reached if the reader clicks on 'Rapunzel'. The German page links additionally to the German classification KHM (*Kinder- und Hausmärchen*). The French Wikipedia page gives an English translation of the French naming, while the French title of the Rapunzel tale is accessible only if the user clicks on the link 'version allemande', leading to the French Wikipedia page 'Raiponce'). There is a clear need to structure this disparate information in one representation format. We turned the basic information from the Wikipedia pages into an integrated SKOS representation:

<ATU_310> rdf:type skos:Concept;

skos:prefLabel "Rapunzel"@en;

skos:altLabel "The Maiden in the Tower"@en;

skos:prefLabel "Jungfrau im Turm"@de;

skos:altLabel "Rapunzel"@de;

skos:prefLabel "La Fille dans la Tour"@fr;

skos:altLabel "Raiponce"@fr.

On the basis of a small fragment of such aligned information from Wikipedia pages, a representative multilingual terminology of ATU terms can be aggregated, and this terminology can be re-used for supporting the translation of other labels of ATU or of TMI. We investigate for this terminology alignment techniques used in the Machine Translation field, adapting them to the short terms that are employed in the catalogues.

4. Relating the Upgraded Representations of TMI and ATU

The SKOS vocabulary allows to establish matching relations between the upgraded TMI and ATU catalogues, so that for example the motif TMI_A2223.1. 'Cat helps man build house: may occupy chimney corner.' can be linked to ATU_545 'The Cat as Helper':

<TMI_A2223.1> rdf:type skos:Concept ;

skos:prefLabel "Cat helps man build house: may occupy chimney corner"@en.

<ATU_545> rdf:type skos:Concept;

skos:prefLabel "The Cat as Helper"@en.

<TMI_AA2223.1> skos:relatedMatch <ATU_545>

Whereas we still do not take decisions on the type of matching relation, which could be 'broader', 'narrower' or 'close'. Additionally we currently investigate the concrete linguistic and semantic markup of the tokens used in labels, using another RDF-based formalism: the *lemon* [9] representation scheme. Precise linguistic mark-up allows establishing in an automated way this kind of SKOS matching, since the noun 'Helper' can be marked as a derivation of the verb 'help'. This work also supports the building of a lexical semantics network for folktales, based on the labels used in TMI and ATU.

5. Towards the Automated Analysis of Motif Sequences in Folktale Plots

The semantically upgraded TMI and ATU resources incorporate information about class hierarchy, which allows potential users to query for motifs that are connected on the same or on different levels of the taxonomy. This is feature that could be exploited for the analysis or generation of storytelling scenarios, since plot graphs of folktales are like watersheds, progressing by connecting motifs at different hierarchical levels as points at various heights. Thus the semantically upgraded TMI could help to learn about the concatenation patterns of motif categories used in tales In turn, one could use such category information as a source of probabilities for network construction. Such probabilistic networks, prominently Hidden Markov Models, are a highly significant research topic in bioinformatics, a field as much utilizing the concept of motifs as narratology, therefore methodological cross-pollination is a totally realistic option. Darányi et al. (2012) have recently shown that ATU tale types as motif strings exemplify the basic recombination patterns of chromosome mutations already on a limited sample of types, and more new findings can be expected by an extension of the quest for 'narrative DNA'. Further, the manual tagging of tale types by motifs in ATU is known to be incomplete, so to repeat this tagging by automatic means by annealing motif

definitions and NLP-based term or phrase extraction is an ultimately fruitful endeavour.

6. Conclusion

In this study, we focus on converting TMI and ATU into adequate semantic resources, compliant with linguistic and terminological requirements, to enable multilingual, content-level indexing of folktale texts. TMI and ATU in their paper form have mainly been used for manual assignment of English-language metadata. To upgrade these established classification systems of folk narratives, we specify a development program that enhances and links them using language processing and semantic technologies. Recent work was dedicated to establish terminological relationships between both catalogues, using for this the SKOS framework establishing semantic links between specific labels of both semantically enriched catalogues. Our work will lead to the deployment of such resources in the LOD framework, complementing the work of national libraries that already ported their bibliographic data to the LOD. This may not only facilitate automated text indexing, but also allow for more detailed analysis of motif sequencing.

Acknowledgements

The contribution of DFKI to the work reported in this paper has been partly supported by the R&D project 'Monnet', which is co-funded by the European Union under Grant No. 248458.

References

Darányi, S., P. Wittek, and L. Forró (2012). Toward Sequencing 'Narrative DNA': Tale Types, Motif Strings and Memetic Pathways. *Proceedings of the Workshop on Computational Models of Narrative*, Istanbul, May 2012.

Declerck, T., and P. Lendvai (2011). Linguistic and Semantic Representation of the Thompson's Motif-Index of Folk-Literature. *Proceedings of the International Conference on Theory and Practice of Digital Libraries*, Berlin, Germany, Springer, 9/2011.

Declerck, T., and N. Koleva (2012). An Iterative Ontology-Based Text Processing Strategy for Detecting and Recognizing Characters in Folktales. *Proceedings of Digital Humanities -2012*.

Declerck, T., K. Mörth, and P. Lendvai (2012). Accessing and standardizing Wiktionary Lexical Entries for supporting the Translation of Labels in Taxonomies for Digital Humanities. *Proceedings of LREC-2012*.

Lendvai, P., T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado (2010). Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. *Proceedings of the Seventh International conference on Language Resources and Evaluation*, Pages 1996-2001, Valetta, Malta, European Language Resources Association (ELRA).

McCrae, J., E. Montiel-Ponsoda, and P. Cimiano (2012). *Integrating WordNet and Wiktionary with lemon. Proceedings of LDL 2012.*

Mörth, K., T. Declerck, P. Lendvai, and T. Váradi (2011). Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts. *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web , Bonn, Germany, Springer, 10/2011.*

Thompson, S. (1955). *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends . Revised and enlarged edition.* Bloomington: Indiana UP, 1955-58.

Tuffield, M. M., D. E. Millard, and N. R. Shadbolt (2006) Ontological Approaches to Modelling Narrative. In: *2nd AKT DTA Symposium*, January 2006, AKT, Aberdeen University.

Uther, H.-J. (2004). *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. FF Communications no. 284–286. Helsinki: Suomalainen Tiedeakatemia.

Zöllner-Weber, A. (2008). *Noctua literaria : a computer-aided approach for the formal description of literary characters using an ontology*. Ph.D. Thesis, Bielefeld University.

Notes

1. AMICUS (Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts), `http://ilk.uvt.nl/amicus`.
2. The interconnection of Knowledge Representation systems and automated textual analysis is being successfully applied to various domains. One of our goal is to adapt and apply such technologies to the automated analysis of folk narratives, using for example the approach presented in (Declerck & Koleva 2012), to be extended among others by ontology elements described in (Zöllner-Weber 2008).
3. See `http://www.ruthenia.ru/folklore/thompson/index.htm`.
4. See `http://en.wikipedia.org/wiki/Aarne%E2%80%93Thompson_classification_system` for the English version, `http://de.wikipedia.org/wiki/Aarne-Thompson-Index` for the German version, and for the French version `http://fr.wikipedia.org/wiki/Classification_Aarne-Thompson`. Note that the online ATU data do not reflect the layout of the original catalogue, as was the case for TMI.

5. SKOS stands for 'Simple Knowledge Organization System'. See for more details: http://www.w3.org/2004/02/skos/.
6. RDF stands for 'Resource Description Framework'. See http://www.w3.org/RDF.
7. See http://en.wiktionary.org/wiki/Wiktionary:Main_Page
8. 'Lexvo.org brings information about languages, words, characters, and other human language-related entities to the Linked Data Web and Semantic Web.' See http://www.lexvo.org.
9. 'lemon (LExicon Model for ONtologies) is an RDF model that allows to specify lexica for ontologies and allows to publish these lexica on the Web' (see McCrae et al. 2012)). This model was developed within the European R&D project "Monnet" (see www.monnet-project.eu (www.monnet-project.eu)

Digital Language Archives and Less-Networked Speaker Communities

Dobrin, Lise M.
dobrin@virginia.edu
University of Virginia, USA

During the past decade, digital language archives have flourished as the field of language documentation has entered mainstream linguistics. Major international funding bodies, both public and private, support research on endangered languages, and it is now possible to publish practical and theoretical work on endangered languages and documentary methods in dedicated journals and book series. One of the themes that has emerged most clearly in this literature is that of *collaboration*. It is now widely agreed that language documentation should be equally responsive to both the technical questions posed by linguists and the more immediate practical interests of speakers and their communities. Issues of rights and access are no longer anxious afterthoughts; they are fundamental matters for negotiation between researchers and speakers, mandatorily addressed in research agreements and funding proposals, and threaded through documentation projects from their very conception (see, e.g., Yamada 2007; Czaykowska-Higgins 2009).

Yet paradoxically, this intense focus on collaborative methods seems to stop where the digital archiving of language data begins. Language archives are doing their best to ensure that source communities can in principle gain *access* to the language materials they produce, but they are only just beginning to consider possibilities for formally integrating speaker communities into the process of archive curation. Such involvement could potentially transform language archives from repositories of static objects into sites for ongoing dialogue and exchange with living communities.

The obstacles to collaborative archiving are particularly acute in the case of source communities located in hard-to-reach places – often the third world – where many of today's small, minor, and endangered languages are spoken. As might be expected, a disproportionate percentage of documentary projects are being carried out in such locations (see, e.g., http://www.paradisec.org.au/blog/2010/09/where-does-the-dosh-go), with their outputs being deposited into first world archives. It is imperative that western scholars

and institutions be cognizant of the impact their documentation projects have on such communities. In Melanesia, which is arguably the world's most linguistically diverse area, local value is validated through relationships with outsiders. If community input ceases once the fieldwork phase of a western-sponsored project is over, this can reinforce the feelings of marginalization that motivate language shift in the first place (Dobrin 2008).

In April 2012, the University of Virginia's Institute for Advanced Technology in the Humanities (IATH) is hosting an international group of scholars, technical experts, and community members for an intensive two-day meeting, sponsored by the National Endowment for the Humanities, to explore appropriate social and technical models for facilitating the ongoing involvement of less-networked source communities in the digital archiving of their endangered languages. Participants will present on the current state of the art in language archiving, the cultural and infrastructural situations of representative world regions (Papua New Guinea and Cameroon), and promising 'bridging' technologies. *This paper will report on the results of the meeting, describing their implications for language archiving and for the digital humanities more generally.*

There are a number of reasons why digital language archives must begin to find ways to support direct, ongoing relationships with speakers and community members, and not just with data depositors or researchers. One of these is the unfolding nature and experience-dependence of informed consent. When speakers are first recorded, they may consent to certain levels of access for the resulting materials (full access, access with conditions, researcher-only, etc.; see `http://elar.soas.ac.uk/node/32620` and `http://www.mpi.nl/DOBES/archive_access/access_procedures`). But they cannot possibly foresee all potential future uses of the recordings. And where communities have little or no familiarity with the medium of distribution, how can they be expected to grasp even the most basic uses to which their material will be put? Given the opportunity to make their wishes known, judgments made by communities at a given time may be modified, refined, or even reversed later.

Developing direct lines of communication between archives and communities will also improve the quality and discoverability of archived data. Changing circumstances may make possible the engagement of knowledgeable stakeholders who were formerly reluctant or inaccessible. Also, though linguists may be diligent about collecting detailed metadata when making recordings in the field, this process inevitably leaves gaps. After fieldwork is over, missing information, such as the identity of a speaker or dialect, might only be known by difficult-to-reach community members. Whole categories of metadata that once seemed irrelevant (details of local history, relationships between consultants, and so on) often take on new significance as projects evolve. By enabling mechanisms by which community members can identify and attach metadata to recordings that concern them or their communities, endangered language resources will be enriched, making them easier to find and more useful for all users.

Finally, there is a practical need for the kind of direct relationships we are proposing to facilitate, as archives are now receiving expressions of interest from individuals who speak the languages of their collections. At the University of London's School of Oriental and African Studies (SOAS), the Endangered Languages Archive (ELAR) has registered speakers of Bena and Pite Saami as users; IATH occasionally receives queries from town-dwelling Arapesh people interested in the Arapesh Grammar and Digital Language Archive (AGDLA). ELAR depositors sometimes specify that access to their deposit requires community approval; the Archive of the Indigenous Languages of Latin America (AILLA) sets up 'community controlled' as a systematic level of access (though at present this must be mediated through the institution). Increasingly, we can expect archives to be routinely receiving user account applications from individuals who were either directly involved as research participants or who have ancestors, family members, friends, or community associates who were involved in language documentation projects. But at present there is a clear disconnect between what depositors are able to achieve using the tools and systems provided by archives, and what community members are able to do – especially those with limited internet access.

Some digital language archives recognize these problems and are seeking to overcome them. Edward Garrett of ELAR has been experimenting with social networking technologies such as the open-source Drupal Content Management System in order to make their archive more user-centered; a desire to extend the ELAR system to include community members is one of the motivations for our meeting. The project BOLD-PNG (see `http://www.boldpng.info/home`) has begun putting digital recording equipment into the hands of community members in Papua New Guinea and training them in techniques of basic oral language documentation, giving people the resources to become data collectors, if not depositors. Kimberly Christen's Mukurtu platform was designed to facilitate community-controlled archiving, digitally instituting traditional cultural protocols to manage access to archived objects (see, e.g., Christen 2008). But none of these projects

develops a generalizable approach to integrating communities without internet access in the ongoing curation of digital language materials they have produced.

We are particularly interested in the possibilities afforded by the increasingly common presence of mobile communication technologies in areas where basic infrastructure such as electricity, running water, and even roads remain absent. In such areas, new social institutions involving cell phones have begun to evolve, for example, conventions for exchanging cell-phone minutes, or new gathering spaces defined by signal access. Exploiting this new form of technology, the non-profit organization Open Mind has developed a system called 'Question Box' (see http://questionbox.org/about-mission) which allows remote communities to access information in critical domains such as health, agriculture, and business. Google's voice-based social media platform SayNow is being used to allow cell phone users in Egypt and elsewhere to leave voicemail messages that appear online immediately as Twitter audio feeds (http://www.nytimes.com/2011/02/02/world/middleeast/02twitter.html). We believe that similar methods are worth exploring as a means to creatively connect less-networked language communities with researchers and archives.

References

Christen, K. (2008). Archival Challenges and Digital Solutions in Aboriginal Australia. *SAA Archeological Recorder* 8(2): 21-24.

Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. *Language Documentation and Conservation* 3(1): 15-50.

Dobrin, L. M. (2008). From Linguistic Elicitation to Eliciting the Linguist. *Language* 84: 300-324.

Yamada, R.-M. (2007). Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language documentation and conservati on* 1(2): 257-282.

Language Documentation and Digital Humanities: The (DoBeS) Language Archive

Drude, Sebastian
Sebastian.Drude@mpi.nl
Max-Planck-Institute for Psycholinguistics, The Netherlands

Trilsbeek, Paul
Paul.Trilsbeek@mpi.nl
Max-Planck-Institute for Psycholinguistics, The Netherlands

Broeder, Daan
Daan.Broeder@mpi.nl
Max-Planck-Institute for Psycholinguistics, The Netherlands

1. Overview

Since the early nineties, the on-going dramatic loss of the world's linguistic diversity has gained attention, first by the linguists and increasingly also by the general public. As a response, the new field of language documentation emerged from around 2000 on, starting with the funding initiative 'Dokumentation Bedrohter Sprachen' (DoBeS, funded by the Volkswagen foundation, Germany), soon to be followed by others such as the 'Endangered Languages Documentation Programme' (ELDP, at SOAS, London), or, in the USA, 'Electronic Meta-structure for Endangered Languages Documentation' (EMELD, led by the LinguistList) and 'Documenting Endangered Languages' (DEL, by the NSF). From its very beginning, the new field focused on digital technologies not only for recording in audio and video, but also for annotation, lexical databases, corpus building and archiving, among others. This development not just coincides but is intrinsically interconnected with the increasing focus on digital data, technology and methods in all sciences, in particular in the humanities.

As one result, a number of worldwide and regional specialized language archives have been established, devoted to a new type of corpora: digital, multimedia, multi-purpose. The DoBeS archive alone contains data on more than 60 languages; it is hosted at the Max-Planck-Institute for Psycholinguistics (MPI-PL) in Nijmegen, where it is combined with data from other field research projects, in total covering around

two hundred languages. The Technical Group of MPI-PL (now as a new unit, 'The Language Archive', TLA) has not only been developing new tools such as ELAN and the language archiving technology suite, but is also active in building regional archives offering the same technology around the globe, and in networking with other archives, for instance in DELAMAN.

Furthermore, The Language Archive (TLA) also participates, based on our experience with the DoBeS archive, in international initiatives such as CLARIN, which have a much broader scope than endangered languages. Still, the new type of corpora, including the DOBES archive, have only incipiently been explored with novel research questions and methods, and the full potentials of cooperation between language documentation and, e.g., computational linguistics and automatic data processing methods have only begun to be exploited. This also holds for topics such as community member participation and representation, engaging wider audiences, access methods, ethical issues around privacy and 'publicness', documentation and archives' relations to social networking and other mass participation platforms. In general it seems obvious that neighbouring disciplines can benefit from language documentations, and that also language documentation and archiving can gain much from cooperation with similar activities and research in other fields, which may even change the way language corpora are designed and future documentation projects are carried out. Nevertheless, the concrete details have still to be determined in the quickly growing environment of 'digital humanities'.

2. Outline of the DoBeS programme and its relation to the digital humanities

The DoBeS programme emerged from long discussions between the Volkswagen foundation and a small group of linguists around the recently founded German 'society for endangered languages', concerned with language diversity and linguistic field research. First result of this interaction was a successful summer school about 'language description and field research' in Cologne in 1993, a very early response to the first public statements urging to get active about the problem of 'endangered languages', i.e. the imminent loss of humanities linguistic (and cultural) diversity. To that time the very first uses for the internet started to be developed, such as the world wide web; email was rarely used in business and academia, university term papers were still often written on a typewriter or by hand. Although the beginning of 'computing and the humanities', 'linguistic computing' etc. actually dates back long before the arrival of personal computers (at least to the 1970ies, when todays major DH associations were founded), the digital humanities (DH) as we know them today existed in the early nineties maximally in an embryonic state, for neither sufficient digital data sets nor the needed computational tools and methods existed.

The first call for applications for individual projects of what would soon become to be called the DoBeS programme was made in 1999. Despite the long period of intense debate and planning, it was still all but clear how language documentation (LD) exactly would be carried out. Himmelmann had already published his seminal paper 'Documentary and descriptive linguistics' (1998), but still the exact goals of such a research program and especially its methodologies had to be decided, tested and established, and the needed technology had to be identified, designed and developed. Thus, the first year of DoBeS was designated to be a pilot phase, where a consortium of linguists, leading eight different documentation projects on a variety of languages all over the globe,[1] together with the technicians of the technical project at the MPI-PL, would discuss and eventually decide on fundamental, methodological, technical and legal/ethical issues, giving clear guidelines to the next generations of projects and establishing the fundaments of a technical infrastructure to support the building of digital language corpora of a new kind – focussed on actual oral language use in a natural cultural context.

By the beginning of the programme in 2000, the general technological setting had changed dramatically. Semi-professional audio and video equipment with satisfactory quality for serious documentation work were now available to affordable prices (although to that time, video recordings were often still analogue, e.g. in HI8 format, and audio recordings were often done in a compressed format, e.g. ATRAC). Most importantly, digital storage capacities had grown to a point where even video recordings with a reasonable resolution and bit-rate could be stored on usual hard discs and even on removable media such as DVDs. This allowed LD to be a fundamentally digital enterprise, aiming at the building of large digital language archives, i.e., multimedia corpora with natural speech data.

At the same time, the technical group (TG) at the MPI-PL had started to tackle the problem of the increasing amount of digital data produced and used at that institute. Some of the data obviously were of relevance for the future – for re-use with different research goals, or just for being able to check and reproduce the results, making the research more accountable. The TG had already started to work on a digital archive of research data (including

data from linguistic field research) and was thus in the best position to function as the technical centre for the DOBES programme, which in turn for several years boosted the development at MPI-PL. Most importantly, a meta-data-schema, IMDI, was developed with decisive input from the DOBES consortium of the pilot phase and in the first years of the main phase (most of the 8 projects of the one-year pilot phase also participated in the first years of the main phase of the DOBES programme). Also, the development of a multi-media annotation tool, ELAN, started. Other tools and infrastructure elements were added over the years to what by around 2008 came to be called the 'Language Archiving Technology' suite.

These developments occurred basically without connection to the first developments in the mainstream 'digital humanities' or 'E-Humanities'. In the 1990ies, when larger data collections became available, when individual computers became part of every university department, and when the development of tools tailored to specific needs was easily done, research techniques and tools such as text mining, quantitative text analysis, complex databases were increasingly often employed by technophile humanities practitioners and computer linguists. Language documentation, however, was developed by a completely different community of field researchers (linguists, anthropologists, music-ethnologists etc.) which usually were not akin to digital technologies and computational methods. The object of study had little overlap, too: computer linguistics and the emerging digital humanities in general were (and continue to be) mainly concerned with major languages and predominantly in their written form, whereas LD by definition is concerned mainly with small and understudied languages, most of which are only occasionally written.

So for about almost a decade, the two areas developed mostly independently one from another. The DoBeS program grew (each year between five and ten new projects started, each with a duration of usually three to four years) and had followers (see above) and became more mature, as the basic standard methodologies were clear and new research questions were introduced and a stronger interdisciplinary approach consolidated. The necessary tools became available, more stable and increasingly easy to use. In particular, Language Archiving Technology with the web-based programs LAMUS for the upload and archive integration, AMS for archive management, and other IMDI or ELAN related tools were built, so that the DoBeS archive (at the core of the larger digital archive with language resources at MPI-PL, 'TLA') became an example for digital archives.

In the last years, the two communities and research traditions (DH & LD) have begun to come closer. In linguistics, language documentation has contributed to raise the interest in linguistic diversity, linguistic typology and language description. This general movement also affects computational linguistics which now increasingly shows also interest in small languages. At the same time, language documentation has from its beginning been concerned with digital data and methods, even if mostly for data management and archiving and less for linguistic analysis. Still, teams for LD projects nowadays usually have quite good computational expertise, and field workers are less distant from digital tools and data than they used to be, and than many other linguists. Furthermore, soon it became obvious that the time was ripe for novel research questions and topics that would make use of these linguistic corpora of small languages, which is where computational linguistics and other statistical methods come into play (see below).

In the opposite direction, the digital humanities grow and consolidated to the point of engaging in constructing major research infrastructures for their needs, integrating the numerous individual data sets and tools 'out there' in the many departments and individual computers, and allowing humanities research to be carried out on a completely new higher level. The DH projects such as DARIAH and in particular CLARIN count with the experiences at the MPI-PL, in particular with the DoBeS archive. The Language Archive is now one of the backbone centres in CLARIN and participates actively in developing this DH infrastructure in Germany, the Netherlands and in a European and international level.

With a total of about 70 major project funded, the DoBeS programme is in its final phase now (the last call for projects was in 2011), and already it has been one of the most successful programs of the Volkswagen Foundation, in terms of impact and public awareness. Internationally, LD has grown into a respected sub-discipline of linguistics and neighbouring fields on its own right, as is witnessed by a successful on-line peer-reviewed journal (Language Documentation and Conservation, LD&C) and a bi-annual international conference (ICLDC, 2009 and 2011 in Hawaii) that is attended by many hundreds of participants from all over the world. In this community, the DoBeS programme is generally recognized as trendsetter and in several aspects as a model, and there are numerous personal and technical links between DoBeS and other LD initiatives (ELDP, DEL, EMELD, PARADISEC, etc.).

3. Some key issues of Language Documentation and Digital Humanities

One distinctive feature of The Language Archive is diversity, on different levels. First, it is concerned with the very linguistic diversity inner- and cross-language, uniting data from many different languages with unique features and a broad variety of communicative settings. But the data, produced by many different teams with different background and research interests, are also diverse in their formats and contents – what is annotated, and how it is annotated. While with respect to metadata and archiving, DoBeS was successful to create agreement and consensus among the researchers, the same does not hold with respect to levels and conventions of annotation, from the labelling of 'tiers' in ELAN or other annotation tools to the abbreviations used for grammatical glosses and labels. This now constitutes a major obstacle for advanced cross-corpora research, and even with the general ISOcat data category registry, much manual work has to be done before different corpora are interoperable. The same holds for lexical data, as the work in the RELISH project has shown which created interfaces and conversions between different standards for lexical databases (LMF & Lexus and LIFT & LEGO).

Another issue is the question of sustainability. Still, too often one finds great initiatives that produce wonderful tools and/or data sets, but without any long term plan – when the funding ends and/or the developer leaves, the resources are abandoned and not rarely unusable after some years, when hardware and software changes. There are different aspects to sustainability – one is the sheer preservation of the bit stream, which is threatened by eroding media such as hard disks or optical discs. The necessary automatic copying of several backups to different locations and the constant replacement of out-phased hardware can only be done by data centres, with which smaller archives should cooperate. The Language Archive was lucky enough to be able to negotiate a 50 year guarantee for bit stream preservation by the Max Planck Gesellschaft already around 2006. For data format accessibility, one needs to rely on a manageable number of open standards (such as XML and UNICODE for text data, or widely used open and preferably not compressed codecs for audio and video data), and has to be prepared to migrate the whole corpus from one format to another if new standards supersede the ones used in the archives (although the original files should always be preserved, too). Finally, the problem of maintenance of tools is generally not satisfactorily solved. Due to changing hardware, platforms, drivers, standards, most tools are bound to need constant maintenance even if no new features are to be implemented (which usually happens if a tool is well received by a large user community), and few funds are available for this kind of activity. One has to be constantly considering how and with how many resources the currently offered software can and should be maintained or further developed. All these questions are by now well known in the DH and now addressed by the infrastructure projects such as DARIAH and CLARIN, but have been addressed at a comparatively early stage at The Language Archive.

It is interesting to notice that one of the strengths of The DoBeS Language Archive is its connecting character. Not only are the data in the archive relevant for many different disciplines, not just linguistics, but also anthropology, history, psychology, music-ethnology, speech technology etc. Also some of the tools developed at The Language Archive are now more widely used, in particular ELAN, which is now used not just by descriptive and documentary linguists, but fostered multi-modality (gesture) and sign language research and is even employed in completely unrelated fields. Also ISOcat has the potential to bridge the gap between different traditions of labelling entities in areas much broader than linguistics, and ARBIL, the successor of the IMDI metadata-editor, is by now being prepared to be the major tool for the creation of modular and flexible CMDI metadata as used in CLARIN. Finally, other archives using Language Archiving Technology have been and continue to be set up at different locations in the whole world, constituting a network of regional archives which soon can interact, exchange their data for backup and ease of access purposes, and hence strengthen and consolidate not just LD but generally the archiving of valuable digital research data.

There are several big challenges ahead for language documentation and the corpora it produces. Some, as the tension between the general movement to open access to research data and the need to protect the individual and intellectual property rights of speakers and researchers are in principle solved by providing different levels of (controlled and possibly restricted) access and employing codes of conducts and other ethical and legal agreements. Still, this state of affairs has constantly to be re-thought and discussed. The same holds for new insights and methods for language strengthening and revitalization, promoting multilinguality and the fruitful interaction between coexisting cultures, and the digital inclusion of linguistic, social and cultural minorities. Obviously, these are questions that cannot be solved by science (alone).

Others are only beginning to be properly addressed, such as the mobilization language data: the future shape of language documentation and archives

will be radically different from its current state, in view of new opportunities, needs and goals beyond data gathering, language description and revitalization. How can be ensured that many users, from researchers via the speakers themselves and the general public, make the best use of the data, that novel research questions are addressed and answered with the support from language documentation corpora? How can the process of creating, annotating and archiving new, high-quality documentation data be made easier even for community members without the presence and support of a LD project and researchers from abroad?

The closer interaction of LD and computational techniques as being developed in the context of the DH will certainly help to improve the situation with respect to one of the major impediments in LD: the high costs for annotating the rich corpora. This is currently done mostly by hand, albeit in some cases with semi-automatic support, such as in the case of morphological glossing based on string-matching with a lexical database. Still, segmenting recordings into utterances, identifying the speaker, transcribing the original utterance and providing further annotation is a very time-intensive work which is mostly done by experts and only to a smaller part can be delegated for instance to well-trained native speakers. Here new computational methods which work reasonably well for written major languages can be generalized or adapted for other languages, and statistical methods can be used to segment and label audio and video data based on speaker and/or gesture recognition. The team at The Language Archive is working on the inclusion of such methods into their tools. The better this integration is, the broader are the possible uses in many humanities disciplines way beyond the field of language documentation.

Notes

1. Three projects on geographically close languages in the Upper Xingu area in central Brazil formed a collaborative project within the consortium.

The potential of using crowd-sourced data to re-explore the demography of Victorian Britain

Duke-Williams, Oliver William
o.duke-williams@ucl.ac.uk
University of Leeds, UK

This paper describes a new project which aims to explore the potential of a set of crowd-sourced data based on the returns from decennial censuses in nineteenth century Britain. Using data created by an existing volunteer based effort, it is hoped to extract and make available sets of historical demographic data.

Crowd-sourcing is a term generally used to refer to the generation and collation of data by a group of people, sometimes paid, sometimes interested volunteers (Howe 2007). Whilst not dependent on Web technologies, the ease of communication and ability to both gather and re-distribute digital data in standard formats mean that the Web is a very significant enabling technology for distributed data generation tasks. In some cases – the most obvious being Wikipedia – crowd-sourcing has involved the direct production of original material, whilst in other cases, crowd-sourcing has been applied to the transcription (or proof-reading) of existing non-digitised material. An early example of this – predating the Web – was Project Gutenberg (Hart 1992), which was established in 1971 and continues to digitize and make available texts for which copyright has expired. A more recent example in the domain of Digital Humanities is the Transcribe Bentham project (Terras 2010), which harnesses international volunteer efforts to digitize the manuscripts of Jeremy Bentham, including many previously unpublished papers; in contrast to Project Gutenberg, the sources are hand-written rather than printed, and thus might require considerable human interpretative effort as part of the transcription process. Furthermore, the Transcribe Bentham project aims to produce TEI-encoded outputs rather than generic ASCII text, potentially imposing greater barriers to entry for novice transcribers.

FreeCEN[1] is a project which aims to deliver a crowd-sourced set of records from the decennial British censuses of 1841 to 1891. The data are being assembled through a distributed transcription project, based on previously assembled volumes of enumerator's returns, which exist in physical form and on microfiche. FreeCEN is part of

FreeUKGEN, an initiative aiming to create freely accessible databases of genealogical records, for use by family historians; other member projects being FreeREG, a collection of Parish Registers, and FreeBMD a collection of vital events registers. The database has grown steadily since its inception in 1999, but coverage is variable with differing levels of completeness between censuses, and for individual counties within each census. The current delivery of data from the FreeCEN project is targeted explicitly at genealogists, and is already of benefit to family historians, by delivery freely accessible data. However, the data (and associated interfaces) are not designed for demographic analysis of the content. Two main approaches may be of interest to demographers, historical geographers and historians in this regard: a series of cross-sectional observations, and a linked sample (such that an individual and hir/her circumstances at decennial intervals might be observed).

The first aim of the project reported in this paper is to extract aggregate counts of individuals grouped by their county of enumeration (that is, their place of residence at the time of the census) and their county of birth. Together, these form a 'lifetime net migration' matrix, which can be disaggregated by age and sex to allow exploration of differential patterns of mobility for a variety of sub-populations in nineteenth century Britain. An associated requirement will be the aggregation of area level population counts (disaggregated by the same age and sex categories to be used in the migration analysis) in order to permit systematic calculation of migration rates (as opposed to absolute volumes). A second phase of analysis will be to use individual and household level records in order to classify households into 'household types' (i.e. persons living alone, married couples with and without dependent or co-resident children etc, households with domestic staff present, households with lodgers present etc).

As usable lifetime migration data sets are extracted, they will be made available via CIDER[2], the Centre for Interaction Data Estimation and Research (Stillwell & Duke-Williams 2003). CIDER is funded as part of the ESRC Census Programme 2006-2011, and currently provides access to interaction (migration and commuting) data sets from the 1981, 1991 and 2001 Censuses, together with an increasing number of flow data sets from a variety of administrative sources. The addition of data from historic censuses will require the modification and extension of existing metadata structures, in order to effectively document the data.

The FreeCEN project aims to transcribe data from the 1841, 1851, 1861, 1871, 1881 and 1891 Censuses of Britain. The 1841 Census can be seen as a transitional census. Whereas the first British censuses - from 1801 onwards - had been completed at an area level by local officials and members of the clergy, the 1841 Census was based around a household schedule which listed each individual together with basic demographic details (rounded age, sex, profession and details of place of birth). However a variety of organisation weaknesses (Higgs 1989) were reflected in idiosyncratic results. A much more robust administrative structure was developed for the 1851 Censuses onwards, with improved instructions given to both enumerators and householders. The number and range of questions asked of householder changed over the course of the nineteenth century, with new questions introduced regarding marital status, relationship to head of family, whether the individual was blind, deaf and dumb, an 'imbecile or idiot', or 'lunatic' and whether the individual was an employer or employee. The aggregation work will concentrate in the first instance on results from the 1891 Census, as these are expected to be of higher quality than the earlier returns. A second strand will then examine data from the 1841 Census, precisely because the quality of the enumerator's returns was poorer, and thus systemic problems are likely to be exposed. How do transcribers cope with poorly recorded or ambiguous original data? How widely do transcribers vary in their notation style?

As noted above, the data are far from complete, and so comprehensive national analysis is not currently possible. However, there are some counties such as Cornwall for which enumeration is complete or close to completion for all censuses in the period, and county or region based analyses should be possible. The overriding aim of the work reported in this paper is to explore the potential for use of these data, and to ascertain the ease with which aggregate observations can be extracted from the FreeCEN database as it grows in the future.

More significantly, there are more subtle issues with the records contained within the FreeCEN data, and there are a variety of issues to be explored. Primarily, it is necessary to examine the accuracy of transcriptions. Over the course of the 19th century the administrative prowess of census taking increased, and improved instructions to enumerators (probably in combination with improved literacy rates in the general population) caused an overall improvement in the quality of the original data; however the accuracy of current transcriptions must also be assessed. It must also be noted that both the quality and comprehensiveness of records transcribed will vary by area and transcriber: the initial impetus for many will be transcription of their own ancestors, and the sample may not therefore be complete or representative. The degree to which the sample and representative can be gauged by

comparison to published totals, although regional and small area totals are less readily available than national totals, constraining the ability to make local level assessments of quality. Mapping of the data will demonstrate the extent to which there are spatial biases (at both local and national levels) in the records selected for transcription.

The lifetime net migration tables to be produced from this project will permit new analysis of mobility in Victorian Britain. Working with the results of the 1871 and 1881 Censuses, Ravenstein (1885) derived a number of frequently cited and debated 'laws of migration', including that migrants tend to move in a series of small steps towards a 'centre of absorption', that counter-flows exist, and that migration propensities reduce with distance. The actual 'laws' are not enumerated as a clear list, but described at separate locations within more than one paper; for a fuller discussion see Grigg (1977). In a discussion of Ravenstein's work, Tobler (1995) states that the most interesting of a set of maps included in the 1885 paper is not referenced or described in the text, noting that such an omission would cause most modern editors to remove the map. Such a map could perhaps be re-created if migration matrices from the period were made available. Many subsequent studies have been made using the originally published results of these censuses (see for example, Lawton (1968) and Friedlander and Roshier (1966)). However, source data are hard to find, and it is now difficult for the general researcher to re-examine such work: thus, an aim of the work reported in this paper is to facilitate future research in this area.

A second approach to the study of these data is to develop a life course analysis, and find individuals linked across censuses, and examine the ways in which their lives have developed at decennial intervals. Whilst appealing, this approach is significantly harder due to difficulties in linking individuals between censuses. Anderson (1972) studying a sample of 475 people in consecutive censuses (1851, 1861) for example, found that 14% of the sample had inconsistent birthplace recording. Some inconsistencies are down to minor misspellings, but this underlines the difficulty of matching individuals.

This project is currently at its outset, and the paper will report on progress and demonstrate how the data can be accessed and used.

References

Anderson, M. (1972). The study of family structure. In E. Wrigley (ed.), *Nineteenth-century society*. Cambridge: Cambridge UP.

Friedlander, D., and R. Roshier (1966). A Study of Internal Migration in England and Wales: Part I. *Population Studies* 19(3): 239-279.

Grigg, D. (1977). E. G. Ravenstein and the "laws of migration". *Journal of Historical Geography* 3(1): 41-54.

Hart, M. (1992). *Gutenberg: the history and philosophy of Project Gutenberg* http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart , retrieved 31 October 2011

Higgs, E. (1989). *Making Sense of the Census, Public Record Office Handbooks No 23*. London: HMSO.

Lawton, R. (1968). Population Changes in England and Wales in the Later Nineteenth Century: an Analysis of Trends by Registration Districts. *Transactions of the Institute of British Geographers* 44: 55-74.

Howe, J. (2007). The Rise of Crowdsourcing. *Wired* 14(6), http://www.wired.com/wired/archive/14.06/crowds.html , retrieved 31 October 2011

Ravenstein, E. (1885). The Laws of Migration. *Journal of the Statistical Society* 46: 167-235.

Stillwell, J., and O. Duke-Williams (2003). A new web-based interface to British census of population origin-destination statistics. *Environment and Planning A*, 35(1): 113-132.

Terras, M. (2010). Crowdsourcing cultural heritage: UCL's Transcribe Bentham project. Presented at: Seeing Is Believing: New Technologies For Cultural Heritage. International Society for Knowledge Organization, UCL (University College London).

Tobler, W. (1995) Ravenstein, Thornthwaite, and beyond. *Urban Geography* 16(4): 327-343.

Notes

1. http://www.freecen.org.uk
2. http://cider.census.ac.uk http://cider.census.ac.uk

Sharing Ancient Wisdoms: developing structures for tracking cultural dynamics by linking moral and philosophical anthologies with their source and recipient texts

Dunn, Stuart
stuart.dunn@kcl.ac.uk
King's College London, UK

Hedges, Mark
mark.hedges@kcl.ac.uk
King's College London, UK

Jordanous, Anna
anna.jordanous@kcl.ac.uk
King's College London, UK

Lawrence, Faith
faith.lawrence@kcl.ac.uk
King's College London, UK

Roueche, Charlotte
charlotte.roueche@kcl.ac.uk
King's College London, UK

Tupman, Charlotte
charlotte.tupman@kcl.ac.uk
King's College London, UK

Wakelnig, Elvira
wakelnig@yahoo.de
University of Vienna, Austria

1. Introduction

The Sharing Ancient Wisdoms (SAWS)[1] project explores and analyses the tradition of wisdom literatures in ancient Greek, Arabic and other languages, by presenting the texts digitally in a manner that enables linking and comparisons within and between anthologies, their source texts, and the texts that draw upon them. We are also creating a framework through which other projects can link their own materials to these texts via the Semantic Web, thus providing a 'hub' for future scholarship on these texts and in related areas. The project is funded by HERA (Humanities in the European Research Area) as part of a programme to investigate cultural dynamics in Europe, and is composed of teams at the Department of Digital Humanities and the Centre for e-Research at King's College London, The Newman Institute Uppsala in Sweden, and the University of Vienna.

2. Historical background

Throughout antiquity and the Middle Ages, anthologies of extracts from larger texts containing wise or useful sayings were created and circulated widely, as a practical response to the cost and inaccessibility of full texts in an age when these existed only in manuscript form (Rodríguez Adrados 2009: 91-97 on Greek models; Gutas 1981). SAWS focuses on *gnomologia* (also known as *florilegia*), which are manuscripts that collected moral or social advice, and philosophical ideas, although the methods and tools developed are applicable to other manuscripts of an analogous form (e.g. medieval scientific or medical texts; Richard 1962).

The key characteristics of these manuscripts are that they are collections of smaller extracts of earlier works, and that, when new collections were created, they were rarely straightforward copies. Rather, sayings were reselected from various other manuscripts, reorganised or reordered, and subtly (or not so subtly) modified or reattributed. The genre also crossed linguistic barriers, in particular being translated into Arabic, and again these were rarely a matter of straightforward translations; they tend to be variations. In later centuries, these collections were translated into western European languages, and their significance is underlined by the fact that Caxton's first imprint (the first book ever published in England) was one such collection (Craxton [1477] 1877]. Thus the corpus of material can be regarded as a very complex directed network or graph of manuscripts and individual sayings that are interrelated in a great variety of ways, an analysis of which can reveal a great deal about the dynamics of the cultures that created and used these texts.

3. Methods: publishing, linking, and creating tools for other related projects

The SAWS project therefore has three main aspects:

1. The encoding and publication of a digital archive of editions of a selected number of these texts;
2. The identification, publication and display of the links between the anthologies, their source texts, and their recipient texts;
3. The building of a framework for scholars outside the SAWS projects to link their texts to ours and to other relevant resources in the Semantic Web.

3.1. TEI XML markup of editions

Each of the texts is being marked up in TEI-conformant XML and validated to a customised schema designed at King's College London for the encoding of *gnomologia*. Our structural markup reflects as closely as possible the way in which the scribe laid out the manuscript. The TEI schema uses the **<seg>** element to mark up base units of intellectual interest (not necessarily identified as single units by the scribe), such as a saying (statement) together with its surrounding story (narrative). For example:

Alexander, asked whom he loved more, Philip or Aristotle, said: 'Both equally, for one gave me the gift of life, the other taught me to live the virtuous life.' [2]

This contains both a statement and a narrative:

<seg type="contentItem">

<seg type="narrative">

Alexander, asked whom he loved more, Philip or Aristotle, said:

</seg>

<seg type="statement">

Both equally, for one gave me the gift of life, the other taught me to live the virtuous life.

</seg>

</seg>

Each of these **<seg>** elements can be given an **@xml:id** to provide a unique identifier (which can be automatically generated) that differentiates them from all other examples of **<seg>**, for instance **<seg type="statement" xml:id="K.al-Haraka_ci_s1">**. In other words, it allows each intellectually interesting unit (as identified by our team's scholars) to be distinguished from each other unit, thus providing the means of referring to a specific, often very brief, section of the text.

3.2. RDF linking within and between manuscripts

Several types of relationships have been identified within and between the manuscripts. These manuscript relationships exist at many different levels of granularity, from links between individual sayings to interconnections in families of manuscripts. These relationships are represented using an ontology that extends the FRBR-oo model (Doerr & LeBoeuf 2007) (the harmonisation of the FRBR model of bibliographic records (Tillett 2004) and the CIDOC (Doerr 2003) Conceptual Reference Model (CIDOC-CRM)). The SAWS[3] ontology, developed through collaboration between domain experts and technical observers, models the classes and links in the SAWS manuscripts. Basing the SAWS ontology around FRBR-oo provides most vocabulary for both the bibliographic (FRBR) and cultural heritage (CIDOC) aspects being modelled.

Using this underlying ontology as a basis, links between (or within) manuscripts can be added to the TEI documents using RDF markup. RDF[4] is a specification to represent one-way links (known as triples) from an object entity to a subject entity. The use of RDF to represent relationships in markup has to date primarily been seen in XHTML documents containing RDFa[5] (RDF in attributes); however it is desirable to extend the scope of RDF in markup, so that RDF links can be added directly into documents such as the TEI XML documents used in SAWS (Jewell 2010).

To include RDF triples in TEI documents, three entities have to be represented for each triple: the subject being linked from, the object being linked to, and a description of the link between them. The subject and object entities in the RDF triple are represented by the **@xml:id** that has been given to each of the TEI sections of interest. We use the TEI element **<relation/>** (recently added to TEI) to place RDF markup in the SAWS documents, with four attributes as follows:

- The value of **@active** is the **@xml:id** of the subject being linked from;
- The value of **@passive** is the **@xml:id** or **URI** (Uniform Resource Identifier) of the object being linked to;
- The value of **@ref** is the description of the relationship, which is drawn from the list of relationships in the ontology;
- The value of **@resp** is the name or identifier of a particular individual or resource (such as a bibliographic reference). Many of the links being highlighted are subjectively identified and are a matter of expert opinion, so it is important to record the identity of the person(s) responsible.

Example:

```
<seg type="statement" xml:id="K._al-Haraka_ci_s5">
برهان ثالث لكل متحرك لذاته فهو راجع إلى ذاته
</seg>

<seg type="contentItem" xml:id="Proclus_ET_Prop.17_ci1">
Πᾶν τὸ ἑαυτὸ κινοῦν πρώτως πρὸς ἑαυτό ἐστιν ἐπιστρεπτικόν.
</seg>

<relation
active="K._al-Haraka_ci_s5"
ref="saws:isCloseRenderingOf"
passive="Proclus_ET_Prop.17_ci1"
resp="Wakelnig2012"
/>
```

This is equivalent to stating that the Arabic segment identified as 'K._al-Haraka_ci_s5' is a close rendering of the Greek segment identified as 'Proclus_ET_Prop.17_ci1', and that this relationship has been asserted by Elvira Wakelnig, 2012.

The **<relation/>** element can be placed anywhere within the XML document, or indeed in a separate XML document if required: for our own purposes we have found it useful to place it immediately after the closing tag of the **<seg>** identified as the 'active' entity.

3.3. The SAWS 'hub' for enabling related projects to annotate and link their own texts

The project is thus producing a framework for representing these relationships, using an RDF-based semantic web approach, as well as tools for creating these complex resources, and for visualising, analysing, exploring and publishing them (Heath & Bizer 2011). We are engaging with scholars working on related projects in order to establish an agreed set of predicates; the version currently being evaluated and deployed by scholars is available as an ontology at `http://purl.org/saws/ontology` . The number of manuscripts of this type is large, and the project is creating the kernel of an envisaged much larger corpus of interrelated material. Many of the subsequent contributions will be made by others; therefore we are creating a framework of tools and methods that will enable researchers to add texts and relationships of their own, which will be managed in distributed fashion. We are also linking to existing Linked Data sources about the ancient world, most notably the Pleiades gazetteer[6] of ancient places and the Prosography of the Byzantine World[7] which aims to document all the individuals mentioned in textual Byzantine sources from the seventh to thirteenth centuries.

Thus we will create an interactive environment that enables researchers not only to search or browse this material in a variety of ways, but also to process, analyse and build upon the material. This environment will provide tools to browse, search and query the information in the manuscripts, as well as making available SAWS-specific TEI schema and ontology files and XSLTs used to extract semantic information from the structural markup and metadata within the

TEI markup. We also hope to make available tools for adding and editing manuscripts within the SAWS environment. The ultimate aim is to create a network of information, comprising a collection of marked-up texts and textual excerpts, which are linked together to allow researchers to represent, identify and analyse the flow of knowledge and transmission of ideas through time and across cultures.

References

Caxton, W. ([1477] 1877). *The Dictes and Wise Sayings of the Philosophers* (originally published London, 1477), reprinted 1877 (London: Elliot Stock).

Doerr, M. (2003). The CIDOC CRM – an Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine* 24(3).

Doerr, M., and P. LeBoeuf (2007). Modelling Intellectual Processes: The FRBR – CRM Harmonization. *Digital Libraries: Research and Development, Vol.* 4877. Berlin: Springer, pp. 114-123.

Gutas, D. (1981). Classical Arabic Wisdom Literature: Nature and Scope. *Journal of the American Oriental Society* 101(1): 49-86.

Heath, T., and C. Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*. San Rafael, CA: Morgan & Claypool, pp.1-136.

Jewell, M. O. (2010). Semantic Screenplays: Preparing TEI for Linked Data `http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-878.pdf/` *Digital Humanities 2010*, Friday 9 July, London, UK.

Pleiades gazetteer: `http://pleiades.stoa.org/` Last accessed 9th March 2012.

Prosopography of the Byzantine World: `http://www.pbw.kcl.ac.uk/` . Last accessed 9th March 2012.

RDF/XML Syntax Specification (Revised) `http://www.w3.org/TR/rdf-syntax-grammar/` . Last accessed 31st October 2011.

RDFa in XHTML: Syntax and Processing. http://www.w3.org/TR/rdfa-syntax/ . Last accessed 31st October 2011.

Richard, M. (1962). Florilèges grecs. *Dictionnaire de Spiritualité* V, cols. 475-512.

Rodríguez Adrados, F. (2009). *Greek wisdom literature and the Middle Ages: the lost Greek models and their Arabic and Castilian*. Bern: Peter Lang.

Tillett, B. (2004). What is FRBR? A Conceptual Model for the Bibliographic Universe. *Library of Congress Cataloging Distribution Service, Library of Congress* 25, pp. 1-8.

Notes

1. http://www.ancientwisdoms.ac.uk
2. Gnomologium Vaticanum
3. SAWS ontology: http://purl.org/saws/ontology ontology. Last accessed 16th March 2012
4. http://www.w3.org/TR/rdf-syntax-grammar/ Last accessed 31st October 2011.
5. RDFa in XHTML: Syntax and Processing http://www.w3.org/TR/rdfa-syntax/ . Last accessed 31st October 2011.
6. Pleiades gazetteer: http://pleiades.stoa.org/ . Last accessed 9th March 2012
7. Prosopography of the Byzantine World: http://www.pbw.kcl.ac.uk/ . Last accessed 9th March 2012

Recovering the Recovered Text: Diversity, Canon Building, and Digital Studies

Earhart, Amy
aearhart@tamu.edu
Texas A&M University, USA

This paper examines the state of the current digital humanities canon, provides a historical overview of the decline of early digitally recovered texts designed to expand the literary canon, and offers suggestions for ways that the field might work toward expansion of the digital canon. The early wave of small recovery projects has slowed and, even more troubling, the extant projects have begun to disappear, as is apparent from looking through the vast number of projects that are if we are lucky, ghosts on the wayback machine. Alan Liu's *Voice of the Shuttle* provides a good measure of the huge number of lost early recovery projects. A quick perusal of 'The Minority Studies' section reveals that of the six sites listed in 'General Resources in Minority Literature,' half cannot be located with a careful search, suggesting that they have been removed. The same trend is found with other projects listed on the site. While the 'General Resources in Chicano/ Latino Literature' section, only 50% of projects are still online, other areas, such as Asian American literature, have a higher percentage of active projects. Some projects, such as Judith Fetterley's *19th Century Women's Bibliography*, are the living dead and have not been updated for years. Similarities exist among the extinct projects. Most were produced in the late 1990s by single scholars at institutions that did not have an etext center or a digital humanities center, never attracted external support, and never upgraded to SGML or TEI. The canon problems are driven by the dual issues of preservation and the stagnation of digitization recovery efforts.

We should find it troubling that the digital canon is losing the very texts that mirrored the revised literary canon of the 1980s. If we lose a large volume of these texts, and traditional texts such as Whitman, Rossetti, and Shakespeare are the highlighted digital literary texts, we will be returning to a new critical canon that is incompatible with current understandings of literature. The early digital recovery work I am discussing is not easily available. Few of these texts are in print. A few more are available on for-profit databases or on microfilm, but most are available only with a return to the one or

two libraries that own the original physical copy of the book, journal, or newspaper.

Some have posited that a structural problem in the emergence of digital humanities contributes to the selection of materials for digitization and preservation. Martha Nell Smith contends that digital humanities developed as a space to which practitioners hoped to flee from the shifts in the profession that arose out of the cultural studies movement. Smith highlights the digital humanities' retreat into modes of analytics, objective approaches as 'safe' alternatives to the messy fluidities found in literary studies. If Smith is correct, then there should be no surprise that recovery of messy lost texts has not been a priority for the field. Others, such as Kenneth Price, view the funding mechanism as a contributing factor to the increasingly traditional canon of digitized texts. Price notes that the criteria of impact, crucial to the granting decision, favor the canonical text. The National Endowment of Humanities (NEH) awarded 141 start-up grants from 2007 through 2010. Of those grants, only twenty-nine were focused on diverse communities and sixteen on the preservation or recovery of diverse community texts. It is striking to examine the authors and historical figures individually cited in the list of funding: Shakespeare, Petrach, Melville, and Jefferson. While there are grants to support work on indigenous populations, African and African American materials, and Asian American materials, in addition to other populations, the funding of named great men of history deserve scrutiny.

In addition, the turn to increased standardization (TEI) and big data troubles our efforts at small-scale recovery project, as DIY scholars, outside the DH community, have difficulty gaining access to required technical skills for small projects, leading to a decline in small-scale digital recovery projects. The poor literary data sets impact digital humanities efforts to experiment with visualization and data mining techniques. For example, the UC Berkley WordSeer tool is remarkably useful, but the data set used to test the tool and the conclusions drawn about the texts are problematic. The team chose to examine the slave narratives housed at *Documenting the American South*. They ran an analysis on the set to see if the literary conventions of the text corresponded with critic James Olney's claim that autobiographical slave narratives follow a list of tropes such as 'I was born' or 'Cruel slavemaster.' However, the chosen data set was not appropriate for the research question. The 300 narratives labeled slave narratives at this site are actually a mixed bag of first person narratives that are fictional and non-fictional, black authored and white authored, pre and post Civil War, some autobiographical, some biographical and some anti and some pro slavery, or at least apologetic. Given the narratives, it is hard to believe that Olney's criteria, which he says are to be applied to autobiography, would be able to be proven by the test set of data. Literary datamining is still in its infancy, but, to date, the most successful literary results are those that utilize smaller, curated data sets, such as Tanya Clement's fascinating Stein project.

Dan Cohen writes, 'Instead of worrying about long-term preservation, most of us should focus on acquiring the materials in jeopardy in the first place and on shorter-term preservation horizons, 5 to 10 years, through well-known and effective techniques such as frequent backups stored in multiple locations and transferring files regularly to new storage media, such as from aging floppy discs to DVD-ROMs. If we do not have the artifacts to begin with, we will never be able to transfer them to one of the more permanent digital archives being created by the technologists.' I want to spin out the two crucial points that I take from Dan's argument. 1) we must continue to focus on acquiring artifacts and 2) we must work with short term preservation strategies to stop immediate loss.

If Matt Kirschenbaum is correct and preservation is not a technical problem but a social problem, then it follows that the digital humanities community should be able to address the lack of diversity in the digital canon by attention to acquisition and preservation of particular types of texts. We need a renewed effort in digitizing texts that occurs in tandem with experimental approaches in data mining, visualization and geospatial representations. Recent critiques have tended to separate the two, a mistake that will damage both efforts. We can create infrastructure to support scholars interested in digital recovery. Leveraging existing structures to support new, small scale scholarship, continued outreach and training for scholars not interested in become digital humanists but interested in digital recovery is crucial.

Preservation of existing digital recovery projects needs to begin immediately. We need a targeted triage focused on digital recovery projects that allow access to materials that are not accessible in print or other digital form. Using existing structures of support, like NEH's Access and Preservation Grant and Brown's newly launched TAPAS, we might build an infrastructure to ingest selected digital recovery projects. While it is unlikely that we can preserve the entirety of the project we might, as Dan Cohen has argued, preserve simply. We may not preserve the interface, but if we can preserve the recovery core – the text – then we will not lose the valuable work that the early digital pioneers completed.

We need to examine the canon that we, as digital humanists, are constructing, a canon that skews

toward traditional texts and excludes crucial work by women, people of color, and the GLBTQ community. We need to reinvigorate the spirit of previous scholars who believed that textual recovery was crucial to their work, who saw the digital as a way to enact changes in the canon. If, as Jerome McGann suggests, 'the entirety of our cultural inheritance will be transformed and reedited in digital forms,' (72), then we must ensure that our representation of culture does not exclude work by diverse peoples.

References

Clement, T. E. (2008). 'A Thing Not Beginning and Not Ending': Using Digital Tools to Distant-read Gertrude Stein's The Making of Americans. *Literary and Linguistic Computing* 23(3): 361-381.

Cohen, D. J. (2005). The Future of Preserving the Past. *CRM: The Journal of Heritage Stewardship* 2(2): viewpoint.

Cole, J. L. (2004). Newly Recovered Works by Onoto Watanna (Winnifred Eaton): A Prospectus and Checklist. *Legacy: A Journal of American Women Writers* 21(2): 229-234.

Cole, J. L. Winnifred Eaton Digital Archive. Formerly `http://etext.virginia.edu/eaton/`

Liu, A. Voice of the Shuttle `http://vos.ucsb.edu/` .

McGann, J. (2005). Culture and Technology: The Way we Live Now, What is to be Done? *New Literary History* 36(1): 71-82.

Olney, R. (1984). I Was Born: Slave Narratives, Their Status as Autobiography and as Literature. *Callaloo* (20): 46-73.

Price, K. M. (2009). Digital Scholarship, Economics, and the American Literary Canon. *Literature Compass* 6(2): 274-290.

Smith, M. N. (2007) The Human Touch: Software of the Highest Order: Revisiting Editing as Interpretation. *Textual Cultures* 2(1): 1-15.

University of Virginia Library. Electronic Text Center `http://www2.lib.virginia.edu/digital curation/etext.html` .

Wordseer. The University of Berkley, California wordseer.berkeley.edu.

Mind your corpus: systematic errors in authorship attribution

Eder, Maciej
maciejeder@gmail.com
Pedagogical University, Krakow, Poland

1. Introduction

Non-traditional authorship attribution relies on advanced statistical procedures to distil significant markers of authorial style from a large pool of stylistic features that are not distinctive enough to provide reliable information about authorial uniqueness. In other words, the goal is to find as much order in 'randomness' as possible. The better the method applied, the more regularities can be extracted from a population that seems to contain nothing but noise. However, it does not mean that one can overcome the impact of randomness: noise is an inherent feature of all natural languages. In particular, word frequencies in a corpus are *random variables*; the same can be said about any written authorial text, like a novel or poem.

Although dealing with this unavoidable noise is *crème de la crème* of computational stylistics, any other influence of additional randomness – e.g. caused by an untidily-prepared corpus – might lead to biased or false results. Relying on contaminated data is quite similar to using dirty test tubes in laboratory: it inescapably means falling into systematic error. Certainly, quite an important question is what degree of nonchalance is acceptable to obtain sufficiently reliable results.

The problem of systematic errors in stylometry has already been discussed. Rudman (1998a, 1998b, 2003) has formulated a number of caveats concerning different issues in non-traditional authorship attribution, including possible pitfalls in corpus preparation. Noecker et al. (2008), in their attempt to test the impact of OCR errors on attribution accuracy, have observed that a moderate damage of input texts does not affect the results significantly. Similarly, Eder (2011) has noticed that a faultily prepared corpus of Greek epic poems displayed an unexpectedly good performance. In another study, a strong correlation between the length of input samples and attribution performance has been shown (Eder 2010). In these and many other studies, however, the problem of systematic errors has not been addressed systematically.

The nature of noise affecting the results is quite complex. On the one hand, a machine-readable text might be contaminated by a poor OCR, mismatched codepages, improperly removed XML tags; by including non-authorial textual additions, such as prefaces, footnotes, commentaries, disclaimers, etc. On the other hand, there are some types of unwanted noise that can by no means be referred to as systematic errors; they include scribal textual variants (*variae lectiones*), omissions (*lacunae*), interpolations, hidden plagiarism, editorial decisions for uniform spelling, modernizing the punctuation, and so on.

To verify the impact of unwanted noise, a series of experiments has been conducted on several corpora of English, German, Polish and Latin prose texts, the corpora being roughly similar in length and number of authors tested. In 100 iterations, a given corpus was gradually damaged and controlled tests for authorship have been applied (the procedure has been inspired by the study of Noecker et al. 2008). It can be obviously assumed that heavy damage will spoil the results substantially. The aim of this study, however, is to test whether this decrease of performance is linear. On theoretical grounds, one can expect either a linear regression (the more errors, the worse the results), or some initial resistance to small errors, followed by a steep drop of performance.

In all the experiments, the Delta method has been used (Burrows 2002). It is a very intuitive procedure that performs considerably well as compared with much more sophisticated techniques. Like other machine-learning methods, Delta is very sensitive to the choice of number of features to be analyzed (Jockers & Witten 2010). For that reason, each experiment has been approached in a series of 30 independent tests for attribution, increasing the number of MFWs analyzed: 100, 200, 300, and so on, all the way to 3,000. The obtained results should be – by extension – valid for other methods that rely on multidimensional comparison of frequencies of MFWs.

It should be stressed that one can deal with two general types of systematic errors. First, when *all* the samples are damaged to a similar degree (e.g. if the corpus was not cleaned of markup tags); second, when, in a carefully collected corpus, *some* of the samples (e.g. one sample) are of poor quality. The latter case will not be discussed in the present study.

2. Misspelled characters

The first experiment addresses a very trivial, yet severe, type of damage – the situation where single characters are misspelled due to transmission disturbance, imperfect typing, poor quality of scanned documents, using untrained OCR software, etc. This type of error is distributed randomly in a string of characters; however, some letters are more likely to be damaged than others. Especially, white spaces are rarely misspelled.

The experiment, then, was designed as follows. In each of 100 iterations, an increasing percentage of letters (excluding spaces) were replaced with other randomly chosen letters. To give an example: in the 15th iteration, every letter of the input text was intended to be damaged with a 15% probability; in consequence, the corpus contained roughly 15% of independently damaged letters.

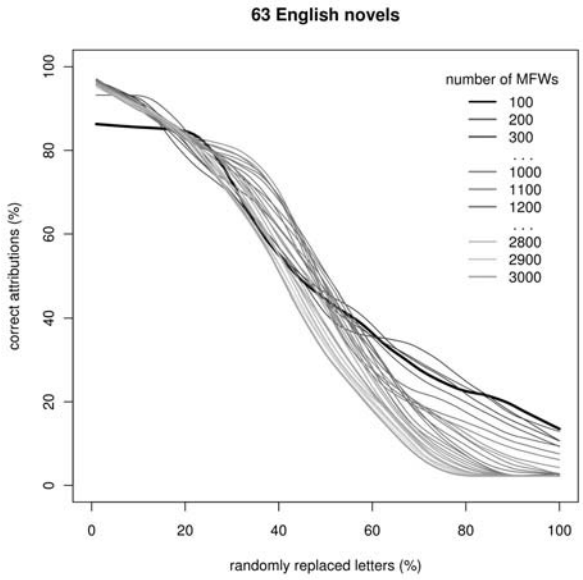


Figure 1: Simulation of poor OCR quality in the corpus of English novels: in 100 iterations, increasing percentage of intentionally misspelled characters has been tested for 30 different MFW vectors

The results were quite similar for most of the languages tested. As shown in Fig. 1 and 2, short vectors of MFWs (up to 500 words) usually provide no significant decrease of performance despite a considerably large amount of noise added. Even 20% of damaged letters will not affect the results in some cases! However, longer MFW vectors are *very sensitive* to misspelled characters: any additional noise means a steep decrease of performance. Intuitively, this could be expected, as the top of frequency lists is usually occupied by short words, which are less likely to contain randomly misspelled characters. This phenomenon is quite clearly evidenced in the German corpus (Fig. 2): i.e. in a language with words usually longer than those in other languages.

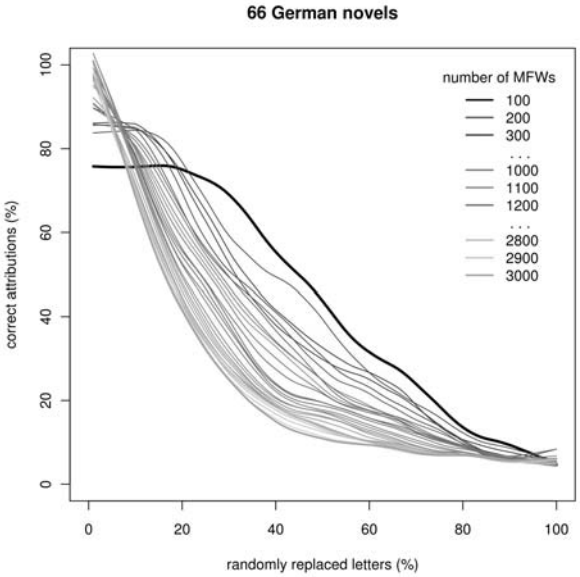


Figure 2: Simulation of poor OCR quality in the corpus of German novels

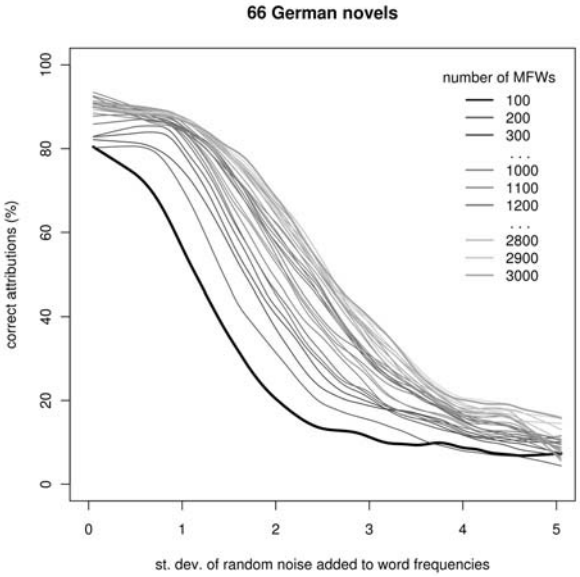


Figure 3: Simulation of editorial modifications in the corpus of German novels

This is very important to stress, however, that using short MFW vectors – despite their considerable resistance to damaged texts – still provides *worse* performance than relying on a large number of MFWs. This means that the 'garbage in, gospel out' optimism is in fact illusory.

3. Noise in word frequencies

The aim of the second experiment is to explore the impact of scribal and editorial modifications of the literary texts. These include orthographic variants, scribal interpolations, editorial textual adjustments, punctuation introduced by modern scholars, etc. A corpus that contains such texts is not merely *damaged*, i.e. it is clean of misspelled characters. However, due to different spellings used, the obtained word frequencies are likely to be biased. This bias might be used to find unique scribal idiolects (Kestemont & Dalen-Oskam 2009); it can be also subjected to automatic disambiguation of spelling variants (Craig & Whipp 2010). In most approaches, however, there is no sufficient awareness of potential systematic error (Rudman 1998a).

The potential bias in word counts can be simulated by adding pure random noise – gradually increasing its standard deviation – to the computed tables of word frequencies. Thus, in the first of 100 iterations, the added noise would have as little variance as 0.05 σ *i* and the last iteration would include a huge noise of 5 σ *i* variance (which means that the noise is 5 times stronger than the variance of a given word frequencies it is added to).

The results seem to be quite similar to those obtained in the previous experiment – but it is worth to note that the pictures are in fact *symmetrical* (Fig. 1-2 vs. 3-4). Here, short MFW vectors are significantly sensitive to randomness in frequency tables, while the longest vectors can survive a moderate earthquake: even very strong noise – its strength comparable with the variance of the words it infects – has a rather weak influence on attribution effectiveness. The results were roughly similar in each corpus tested.

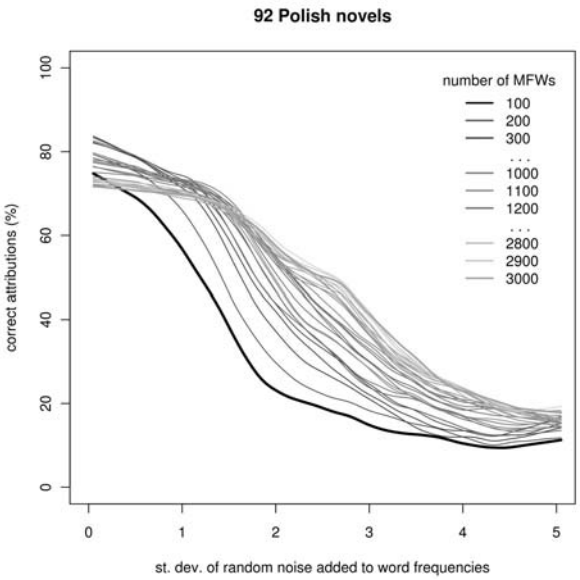


Figure 4: Simulation of editorial modifications in the corpus of Polish novel

4. Impact of literary tradition

The last type of noise can hardly be called systematic error. Namely, the aim of this experiment is to simulate the impact of literary inspirations (plagiarism, imitations, intertextuality, etc.) on attribution effectiveness. In authorship studies, there is always a tacit – and somewhat naïve – assumption that texts in a corpus are purely 'individual' in terms of being written solely by one author and not influenced by other writers – as if any text in the world could be created without references to the author's predecessors and to the whole literary tradition. The problem of collaborative nature of early modern texts has been discussed by traditional literary criticism (Hirschfeld 2001; Love 2002), but it is hardly reported in computational stylistics. In authorship attribution, this feature of written texts is certainly a pitfall; however, the same feature makes it possible to use stylometric techniques to trace stylistic imitations or unconscious inspirations between different authors.

The experiment is designed as follows. In each of 100 iterations, for each text, a consecutive percentage of original words are replaced with words randomly chosen *from the entire corpus*. Thus, a simulation of increasing intertextual dependence between the texts in a corpus is obtained.

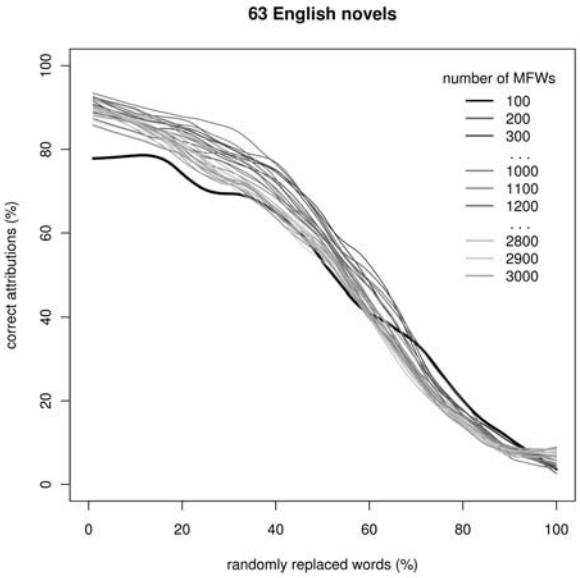


Figure 5: Simulation of extreme intertextuality in the corpus of English novels

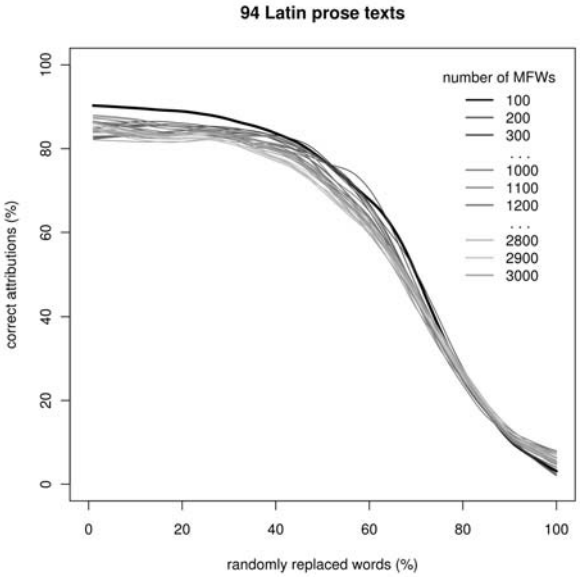


Figure 6: Simulation of extreme intertextuality, in the corpus of Latin prose texts

The obtained results are quite interesting. The corpora of modern literatures, i.e. English (Fig. 5), German and Polish, displayed a gentle decrease of performance (despite the number of MFWs analyzed) in correlation with the amount of 'intertextuality' added. The Latin corpus (Fig. 6) behaved as if the authorial uniqueness could be traced through a mass of external quotations: a considerably good performance was achieved despite 40% of original words replaced. This deserves further investigation.

References

Craig, H., and R. Whipp (2010). Old spellings, new methods: automated procedures for indeterminate linguistic data. *Literary and Linguistic Computing* 25(1): 37-52.

Burrows, J. F. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17(3): 267-287.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. *Digital Humanities 2010: Conference Abstracts*. King's College London, pp. 132-135.

Eder, M. (2011). Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics* 6 (in press).

Hirschfeld, H. (2001). Early modern collaboration and theories of authorship. *PMLA* 116(3): 609-622.

Jockers, M. L., and D. M. Witten (2010). A comparative study of machine learning methods

for authorship attribution. *Literary and Linguistic Computing* 25(2): 215-223.

Kestemont, M., and K. van Dalen-Oskam (2009). Predicting the past: memory based copyist and author discrimination in Medieval epics. *Proceedings of the 21st Benelux Conference on Artificial Intelligence (BNAIC) 2009*. Eindhoven, pp. 121-128.

Love, H. (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge UP.

Noecker, J., M. Ryan, P. Juola, A. Sgroi, S. Levine, and B. Wells (2009). Close only counts in horseshoes and... authorship attribution? *Digital Humanities 2009: Conference Abstracts*. University of Maryland, College Park, MD, pp. 380-381.

Rudman, J. (1998a). Non-traditional Authorship Attribution Studies in the 'Historia Augusta': Some Caveats. *Literary and Linguistic Computing* 13(3): 151-157.

Rudman, J. (1998b). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities* 31: 351-365.

Rudman, J. (2003). Cherry picking in nontraditional authorship attribution studies. *Chance* 16(2): 26-32.

Underspecified, Ambiguous or Formal. Problems in Creating Maps Based on Texts

Eide, Øyvind
oyvind.eide@edd.uio.no
University of Oslo, Unit for Digital Documentation, Norway

1. Introduction

How can a reading of a textual description of a landscape be expressed as a map? Maps form a medium different from verbal texts, and the differences have consequences not only for *how* things are said, but also for *what* can be said at all using maps. Where are these limitations to be found?

In this abstract, I will discuss the relationship between verbal and map based geographical communication. I have created a model of the geographical information read from a source text, then tried to express the contents of the model as maps. I will show that types of geographical information exist that can be stored in and read from verbal texts, but which are impossible to express as geographical maps without significant loss of meaning.

2. Object of study

I used a set of Scandinavian border protocols from the eighteenth century (Schnitler 1962) as source material for this research. The text is based on interrogations about geography with more than 100 different persons, of whom many presumably did not use maps very much if at all. It was created in a society, or a set of societies, on the brink of the transformation from oral to written cultures, where some were firmly placed within the written culture, while others had only been exposed to the activity of reading texts for a few decades. The voices in the text represent persons coming from different ethnic and professional backgrounds, e.g., Sami reindeer herders, Norwegian farmers, and Danish military officers, thus bringing a set of different perspectives into the geographical conversation.

3. Method

Modelling is to create and manipulate models in order to learn from the experiences we gain through the process. Models are in this context

representations of something which is created for the purpose of studying what is modelled more closely (McCarty 2005: 24). In the work reported on here, a computer based model was constructed based on the source text. Through experiments on this model, new knowledge about the text was gained. The creation of the model itself also led to problems giving new insights, so it was part of the series of experiments. The same can be said of the development and tailoring of the software used.

The whole experiment was performed in a qualitative manner. Occurrences in the text were not counted and compared, they were rather evaluated as individual cases, in accordance with traditional humanities research where seeking local knowledge about specific cases is the main activity (Galey 2010: 95-100).

The point of a modelling exercise lies in the process, not in the model as a product. Modelling mediates epistemologically between modeller and the modelled, that is, between the researcher and the data. The model was built through a combination between computer based tools and a thorough close reading in which all geographic information I was able to extract from the text was included. The model was stored as a set of connected facts in a computer application I developed in Java for this research.

In the application the results of the close readings are stored as triplets in a fact database. Then the 'raw' triplets go through several steps of interpretation, leading to more and more formal structures. These steps eventually result in the standard format RDF[1], then further interpretation leads to geographical vector data in a simplified version of the GML format[2] which can be imported to GIS software to be viewed as maps.

The modelling process thus involves several steps, each formalising the data in a stricter sense. For each step, what 'falls off,' what is difficult to avoid loosing, is the interesting part. This 'falling off' will include things that cannot survive a transfer from one system to another, which misses the added level of formality. Through this process I attempted to translate geographical information from one medium to the other. The 'fall-offs' show us what was lost in the process; this was not always because of differences between the media, but candidates for further examination were found in the 'fall-offs'. Examples from this modelling work will be shown in the paper, including some situations where 'fall-off' occurs.

As the goal is to understand how texts express geographical information, I made an effort to base the maps on what I read from the text only. While it is impossible to exclude all context information, I tried hard to exclude any local context based on geographical knowledge of the area. This makes the process different from mapping place name information in literary texts, as done by e.g. Smith and Crane (2001). The latter process includes adding a significant amount of contextual knowledge to the process through what is learned from the pre-existing map.

4. Main results

The issues faced in the modelling experiments led to the development of the following typology of textual expressions:

1. **Under-specified texts.** *Based on such a text, more than one map can be drawn, and at least two of these maps are significantly different.*

This situation occurs when the geographical information in a text can be expressed as more than one significantly different map. This happens when directions such as 'east' or measurements such as '1-2 miles' are used: given lack of other evidence in the text, a number of different geometrical interpretations of the statements are possible.

2. **Fully specified textual descriptions.** *Only one map can be drawn based on the description. If the text mentions something, it is fully specified geometrically.*

Texts in formal languages, such as GML, are typical cases. I have not found examples of this type in natural language texts.

3. **Ambiguity and negation.** *The text includes expressions in the forms 'A or B is located at C' (ambiguity) or 'There are no A's in B' (negation).*

The spatial information read from the text cannot be represented as one single map; one will need two or more maps, a dynamic map or a map with its geometry corrected by a textual description.

The choice of the words 'significantly different' as opposed to merely 'different' in the description of type 1 is necessary and highlights how the border to type 2 is fuzzy. Two maps can always be made which are slightly different, but which will occur as similar to the reader. Small adjustments in location are routinely made in cartographical work in order to improve the readability of maps. The choice of symbols can also change without its leading to significantly different maps. Any text, also of type 2, could be represented by two slightly different maps.

In the case of Schnitler, type 1 situations are abundant, we see a handful of type 3, and none of type 2. The groups are partly exclusive: 1 and 3 can exist in the same textual description, but is hard to see how a

textual description can be both 2 and 1 or both 2 and 3.

5. Discussion

The discussion of the relationship between geographical maps and verbal texts fall into the long tradition of inter-art and intermedia studies. Two important oppositions established by Lessing (1893) can be summarised this way:

1. Actions in time should be applied in poetry, and bodies in space in painting.

2. What is hidden is not seen in painting, while things hidden can still be seen in poetry.

Whereas Lessing's dichotomy between poetry and painting has been questioned, e.g., by Frank (1963) and Mitchell (1980) in connection to their discussions of spatial form in literature, it has never been eliminated. A recent way of formalising such distinctions is the model of Elleström (2010). Instead of starting from a set of different media or art forms, he takes a bottom-up approach, starting from a set of media modalities. His set includes four, namely, the material, sensorial, spatiotemporal and semiotic modalities. The differences between texts and maps fall mainly in the latter two categories.

The two problem areas found in the modelling of Schnitler, *under-specified texts* and *ambiguity and negation*, are both connected to Lessing's thinking. When 'east' in a text can mean a number of different things, potentially covering half of the possible directions from a place, the direction between two features on a map is measurable, with only a limited inaccuracy. Space is expressed differently in the two media. And when disjunction is hard to put on a map, it is also connected to the fact that the sign systems used to refer to the external reality are different. This will be discussed in light of Lessing's opposition and Elleström's system in the paper. Some comments will also be made as to how this links to narratology.

References

Elleström, L. (2010). The Modalities of Media: A Model for Understanding Intermedial Relations. In L. Elleström (ed.), *Media borders, multimodality and intermediality*. Basingstoke: Palgrave McMillan, pp. 11-48.

Frank, J. (1963). Spatial Form in Modern Literature. In *The widening gyre: crisis and mastery in modern literature*. (First published: 1945). New Brunswick, N.J.: Rutgers UP, pp. 3-62

Galey, A. (2010). The Human Presence in Digital Artefacts. In W. McCarty (ed.), *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Open Book Publishers, pp. 93-117.

Lessing, G. E. (1893). Laokoon: oder über die Grenzen der Mahlerey und Poesie. Erster Theil. In K. Lachmann and F. Muncker (eds.), *Gotthold Ephraim Lessings sämtliche Schriften. Neunter Band* (3., aufs neue durchges. und verm. Aufl.). Stuttgart: Göschen. Photographic reprint. Berlin : de Gruyter 1968, pp. 1-177. First published: 1766.

McCarty, W. (2005). *Humanities computing*. Basingstoke: Palgrave Macmillan.

Mitchell, W. J. T. (1980). Spatial Form in Literature: Toward a General Theory. *Critical Inquiry* 6 (3): 539-567.

Schnitler, P. (1962). *Major Peter Schnitlers grenseeksaminasjonsprotokoller 1742 - 1745*, Volume 1. Oslo: Norsk historisk kjeldeskrift-institutt.

Smith, D. A., and G. Crane (2001). Disambiguating Geographic Names in a Historical Digital Library. In *Research and Advanced Technology for Digital Libraries: 5th European conference, ECDL 2001*. Darmstadt: Springer, pp. 127-136.

Notes

1. Resource Description Framework: `http://www.w3.org/RDF/` (checked 2011-10-31)
2. Geography Markup Language: `http://www.opengeospatial.org/standards/gml/` (checked 2011-10-31)

A Frequency Dictionary of Modern Written and Oral Media Arabic

Elmaz, Orhan
orhan.elmaz@univie.ac.at
University of Vienna, Austria

1. Learning Arabic

Having about 500 million speakers, Arabic is not just the only Semitic world language but also one of the six UN-languages. The interest in Media Arabic more specifically increased with the popularity of the satellite station al-Jazeera and the modern necessity for reflecting the Arab perspective of what is going on in the world. People interested in the Arab opinions about what is going on in the Middle East as an example should read Arabic newspapers since in 'Western' media – as Mol (2006: 309f.) states: 'often only information on the Arab world gathered by people who have not mastered the Arabic language [...] often only interpretations of the Arab world from a Western view are found.'

2. Methods & Results

On the base of newspaper texts (Written Media Arabic) and – up to date disregarded – transcripts (Oral Media Arabic), a learner's vocabulary was developed and evaluated in a computer-assisted way. The corpora put together comprise 23.5 million tokens (cp. Routledge Frequency Dictionaries) and were edited in Perl (a.o. CP-1256 incompatible ligatures were broken down and data read wrongly [about 2%] was deleted).

```
INPUT STRING: القطاع
LOOK-UP WORD: AlqTAE
SOLUTION 1: (AlqiTAE) [qiTAE_1] Al/DET+qiTAE/NOUN
(GLOSS): the + sector/section +
SOLUTION 2: (AlqiTAE) [qiTAE_1] Al/DET+qiTAE/NOUN_PROP
(GLOSS): the + Strip (Gaza) +
SOLUTION 3: (AlqaT~AE) [qaT~AE_1] Al/DET+qaT~AE/NOUN
(GLOSS): the + stone-cutter/wood-cutter +
SOLUTION 4: (AlqiTAE) [qaTiyE_1] Al/DET+qiTAE/NOUN
(GLOSS): the + groups/herds/flocks +
SOLUTION 5: (AlquT~AE) [qATiE_3] Al/DET+quT~AE/NOUN
(GLOSS): the + cutters +
```

The corpora were analysed using Tim Buckwalter's word-stem based Buckwalter Arabic Morphologic Analyzer (BAMA 1.0) the output of which was changed in such a way that only possible lemmas per token could be saved (without any further information). Then, line duplicates (semantic ambiguity) were deleted and the line resp. lemma set frequencies were calculated. The latter were broken up and analysed fully forming our primary knowledge base together with each lemma's estimated frequency in Google News. The material was analysed manually in a top-down manner (average morphological ambiguity rate: 3 possible analyses per token) and a frequency vocabulary of just above 3,500 lemmas (estimated to correspond to CEFR B2) sorted by their roots resp. word families, was extracted. It was demonstrated that these cover 95% of non-trivial, recent media texts which were lemmatised manually! Further, the number of look-ups in dictionaries is reduced to 24 ± 10 times for texts with an average length of about 550 ± 60 words.

In order to be on line with the language of media, neologisms (extracted by word length, patterns and concordances) and generally disregarded multi word items (extracted by n-gram analyses, finding collocations [after Dunning 1993]) which constitute integral elements of it, were dealt with separately.

A discursive glossary including dialectal variants which textbooks published up to now are lacking tops the study off.

virostatic, anti-viral (neolog.)	مضاد لفيروس
in response to a question about (journ.)	ردا على سؤال حول
cease-fire resolution (dipl.)	قرار وقف اطلاق النار
header after a cross (sports)	رأسية اثر تمريرة عرضية
's true and [is]n't (dial.)	صح ولا
This isn't our topic! (dial.)	مش هاد موضوعنا

The present material can be useful for reading courses in which students of different levels and background come together in order to level differences in vocabulary. Moreover, the corpora can be used for learning collocations with the help of concordances in an explorative manner and make idiomatic corrections in writing assignments.

3. Corpora

The largest corpus is made up of texts published between 1 June 2008 and 31 May 2009 in the pan-Arab newspaper al-Quds al-ʿArabī (London) consisting of about 20 million tokens. In order to augment the sections of economy and sports a corpus for each was built using texts from the Qatari newspaper al-Šarq. Oral Media Arabic was regarded by creating a corpus of about 1,5 million tokens out of al-Jazeera transcripts.

4. Text coverage

The following table informs about the contents and length of ten texts (disregarding digits), for which the dictionary's coverage was tested. The first six are from Lahlali 2008. One finds the number of unknown

words (NF), the text coverage rate, the number of words which can be understood by morphological and generative knowledge (NWB) and the number of therein subsumed proper nouns (PN) as well as the therewith corrected coverage rate and the number of look-ups (LU) in dictionaries.

Text	Tokens	NF	TC_{NF}	NWB (PN)	TC_{NWB}	LU
Diplomacy: Darfur	489	50	89.78%	28 (19)	95.23%	15
Elections: Yemen 2006	625	65	90.40%	41.5 (33)	96.03%	27
Violence: France 2005	498	84	83.13%	41 (26)	90.81%	35
War: Lebanon 2006	540	58	89.28%	40 (17)	96.41%	12
Economy: Egypt-China	683	80	88.29%	57 (17)	96.33%	21
Disasters: Katrina	509	87	83.10%	55 (42)	93.17%	40
Journalism: MENA	524	39	92.56%	29 (16)	95.99%	10
Integration: Germany	555	77	85.92%	41 (13)	93.42%	30
Human Rights: Headscarf Ban	527	80	84.67%	46 (25)	93.68%	21
Speech: Arab Summit '09	534	58	89.06%	23 (7)	93.40%	30

Therefore, one can relativise the FSI classification of Arabic – ex aequo with Chinese, Japanese and Korean – as super hard language and the opinion of some natives who – as Belnap puts it ironically (2008: 59) – 'almost universally revere Arabic as the most difficult language in the world and therefore essentially unlearnable by mortals.'

References

Belnap, R. K. (2008). If you build it, they will come. In Z. Ibrahim and S. A. M. Makhlouf (eds.), *Linguistics in an Age of Globalization: Perspectives on Arabic Language and Teaching*. Cairo: AUC Press, pp. 53-66.

Buckwalter, T. (2002). *Buckwalter Arabic Morphological Analyzer Version 1.0*. Linguistic Data Consortium.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-74.

Lahlali, El Mustapha (2008). *Advanced Media Arabic*. Edinburgh: Edinburgh UP.

Mol, M. van (2006). Arabic Receptive Language Teaching: a new CALL approach. In K. M. Wahba et al. (eds.), *Handbook for Arabic language teaching professionals*. Mahwah: Lawrence Erlbaum, pp. 305-314.

Texts in Motion – Rethinking Reader Annotations in Online Literary Texts

Fendt, Kurt E.
fendt@mit.edu
Massachusetts Institute of Technology, USA

Kelley, Wyn
wkelley@mit.edu
Massachusetts Institute of Technology, USA

Zhang, Jia
zhangjia@mit.edu
Massachusetts Institute of Technology, USA

Della Costa, Dave
dfd@mit.edu
Massachusetts Institute of Technology, USA

Taking Reader's Response Theory as the basis, this short paper discusses approaches to enhancing reader interaction with literary texts through social annotation and visualization tools. The integration of multimedia repositories, collaborative annotation mechanisms, and advanced visualization techniques creates a digital environment for students and scholars alike and provides readers with rich tools for both interpreting literary texts and visualizing reader interactions with these texts. Examples will be drawn from Mary Shelley's *Frankenstein* and Herman Melville's *Moby-Dick*.

Notes in margins of books fulfill several functions for the individual reader: recording thoughts and associations while reading, marking important passages or references to other parts within the same text, noting possible links to other texts and documents, or simply creating a reference point to come back to in a subsequent reading. From a cognitive point of view, marginalia offer insights into the interaction of a reader with a given text and offer hints towards understanding the process of reading and sense making. Reading a book that already contains notes by another reader can be both enlightening or distracting, yet it allows us to encounter another person's reading experience, forming another layer of meaning that might influence our own interpretation. It's this interactive nature of the reading experience, the constant interaction between the text itself and the reader's response to it that, according to Wolfgang Iser in the *Implied Reader* (1974)[1], allows us to interpret a literary text. The result is a 'virtual text' that is formed by a text and the reader's interaction with it.

For centuries, marginalia have served as instantiations of a rich reader engagement with a text, often providing hints to interpretations from different times, languages, and cultures thus forming critical insights into texts that would otherwise be lost. With more and more texts being digitized or already born digital texts available online or on electronic readers, the notion of marginalia in a digital space poses a number of interesting questions. How can we preserve, enhance, and expand this critical interaction especially with literary texts, particularly when we consider the social dimension offered through digital media? How can we make use of large digital text, image, video, and audio repositories to help us represent a text as a 'multi-dimensional space,' as a 'tissue of quotations drawn from innumerable centers of culture' (Roland Barthes)[2]? What if we had digital tools that allowed us to even visualize how readers experience a text by following their interactions, for example by graphically representing their annotations across the whole text? How do readers discover connections within a text, across different texts, to source texts, to adaptations in other media, or derivative texts? What if we could finally visualize the 'act of reading' (Iser) and analyze a literary text from the perspective of both the author and the readers?

A multidisciplinary research team at HyperStudio, the Massachusetts Institute of Technology's Digital Humanities Lab within Comparative Media Studies, has been exploring these questions, developing digital tools that allow readers to create multimedia annotations to literary texts, and testing them in the literature classroom. HyperStudio has created specific visualization tools that allow for a visual representation of the structural features of a text and, more importantly, offer graphs that show the reader interactions with a given text both on page and whole book levels. At the same time, these visualizations also serve as navigational mechanisms through the texts and annotations, both fine-grained and on a whole-text basis.

Pedagogically this work is informed by a wide range of theoretical and practical approaches: concepts of media literacy, work on interactive and collaborative text editing, a pedagogy of close reading and critical writing devised to place literary texts in relation to their multimedia sources and adaptations; and new Digital Humanities tools for visualization, data mining, and social networking.

At the heart of these different approaches lie theories and technologies for understanding the mix as a central feature of new media 'literacies' or 'competencies,' in Henry Jenkins' terminology[3].

The notion of remix borrowed from popular musical forms has come to define new modes of artistic production and consumption that inspire active creators. Similarly in academic scholarship and pedagogy, an appreciation for how artists borrow and rework cultural materials has energized the study of creative processes. In a literature classroom, traditional methods of close reading, source study, and literary analysis can merge with newer interpretive models to view texts in creative flux: as fluid vessels for recombining older forms of inspiration and engendering new ones in different media adaptations. Reading literary works as textual remixes reinforces strong close reading and critical skills while rejuvenating source study. We see text annotation at the center of this notion of remix.

The visualization approaches have been informed by a range of current online tools for visualizing texts but also by artistic representations of text corpora such as David Small's work[4], timeline tools such as HyperStudio's *Chronos* [5], and playful approaches such as the ones by John Maeda[6]. Existing visualization concepts such as word clouds, word trees, phrase nets are only a few of the online visual tools that help us graph occurrences of single words or the relationship between words within texts. Popular visualization tools such as the ones available for experimentation at IBM's Many Eyes[7] or scholarly tools such as Voyeur[8] help us quickly discover word frequencies and usage patterns. Google's Ngram Viewer[9] makes the leap beyond the single text and searches across 'millions' of books and displays occurrences of single or multiple words as line graphs along a timeline that spans decades or even centuries. While these tools help us greatly analyze a range of linguistic aspects of the work of an author, they tell us little about how readers interact with literary texts. The tools that HyperStudio is developing allow us to visualize how readers interact with a text, how they discover connections within a text, across texts, to source texts, to adaptations in other media, or derivative texts. In the end, they aim at visualizing the 'act of reading' (Iser) and helping to analyze a literary text from the perspective of both the author and the readers.

From a software development point of view, we are using principles of agile development and co-design, involving students, faculty, designers, and assessment specialists from the outset. Building on HyperStudio's Ruby-on-Rails-based platform Repertoire, the text annotation tools feature shareable text repositories, easy role and group management, filtering of annotations via tags and search, and display of overlapping annotations from different users/groups. The visualization tools will allow a seamless transition from fine-grained display of annotations to a global view of annotations within a whole literary text.

We recognize that there are a number of text annotation tools available, both open source and commercial, that include some of the features we are describing here. However, the approach to annotating texts with multimedia materials, annotation across texts, a deep integration of visualization tools for both textual features and reader interaction, combined with a strong educational focus, is unique in the field.

Early feedback from students using a prototype of the annotation tools in two literature classes at MIT has shown a deeper engagement by students not only in processes of close reading and creating references to external sources but also in students' requests for private workspaces where they could upload their selection of literary texts and annotate them in the same way for their own research, not even related to any class assignments. In addition, the HyperStudio team has also received a number of feature requests from students. One feature in particular was in great demand (on our schedule for later implementation): the flexible comparison and joint annotation of two texts side-by-side. Each text can be independently defined as source, version, base, or adaptation, allowing for the display of referenced annotations even when the related text is not present. Additional features such as annotation of non-contiguous portions of text and the creation of multimedia essays based on reader annotations will be implemented in a next phase. By the time of DH2012, HyperStudio will have tested the annotation tools and visualizations in half a dozen courses and conducted extensive assessment.

Funding

This work was supported by the MIT Class Fund and the MIT SHASS Teaching and Learning Fund.

Notes

1. Iser, W. (1974). *The Implied Reader; Patterns of Communication in Prose Fiction from Bunyan to Beckett*. Baltimore: John Hopkins UP.
2. Barthes, R. (1997). The Death of the Author. In *Image – Music– Text*. New York: Hill & Wang, pp. 142-148.
3. Jenkins, J., et al. (2006). *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*. Chicago: The John D. and Catherine T. MacArthur Foundation. http://digitallearning.macfound.org/atf/cf/%7B7E45C7E0-A3E0-4B89-AC9C-E807E1B0AE4E%7D/JENKINS_WHITE_PAPER.PDF accessed 25 March 2012.
4. Small, D. (1999). *Talmud Project*. http://www.davidsmall.com/portfolio/talmud-project/ accessed 25 March 2012

5. MIT HyperStudio (2011). *Chronos Timeline*. http://hyperstudio.mit.edu/software/chronos-timeline/ accessed 25 March 2012.

6. http://www.maedastudio.com/index.php

7. http://www-958.ibm.com/software/data/cognos/manyeyes/ accessed 25 March 2012.

8. http://hermeneuti.ca/voyeur/ accessed 25 March 2012.

9. http://books.google.com/ngrams accessed 25 March 2012.

May Humanists Learn from Artists a New Way to Interact with Digital Technology?

Franchi, Stefano
stefano@tamu.edu
Texas A&M University, USA

1. Takeovers

The two best-known interactions between computational technologies and traditional Humanist pursuits are represented by the Artificial Intelligence/Cognitive science movement and the roughly contemporary Digital Humanities approach (although the label became popular only recently). Classic Artificial Intelligence saw itself as 'anti-philosophy' (Dupuy 2000; Agre 2005; Franchi 2006): it was the discipline that could take over philosophy's traditional questions about rationality, the mind/body problem, creative thinking, perception, and so on, and solve them with the help of a set of radically new synthetic, experimentally based techniques. The true meaning of the 'computational turn in philosophy' lies in its methodology, which allowed it to associate engineering techniques with age-old philosophical questions. This 'imperialist' tendency of cognitive science (Dupuy 2000) was present from the very beginning, even before the formalization of the field into well – defined theoretical approaches (McCulloch 1989 [1948]; Simon 1994).

The Digital Humanities represent the reverse modality of the encounter just described. The most common approach (Kirschenbau 2010) uses tools, techniques, and algorithms developed by computer scientists to address traditional questions about the meaning of texts, their accessibility and interpretation, and so on. Other approaches turn technology into the scholar's preferred object of study (Svensson 2010). The recent approach pioneered by the 'Philosophy of Information' (Floridi 1999, 2011) follows this pattern. Its focus on the much broader category of 'information' substantially increases the scope of its inquiries, while firmly keeping it within philosophy's standard reflective mode.

The common feature of these two classic encounters between the Humanities and computational theory and technology is their one-sidedness. In either case, one of the two partners took over some relevant aspects from the other participant and fit it within

its own field of inquiry (mostly questions, in AI's case; mostly tools, for the Digital Humanities). The appropriation, however, did not alter the theoretical features of either camp. For instance, AI and Cognitive Science researchers maintained that philosophy pre-scientific methodology had only produced mere speculation that made those problems unsolvable. Therefore, philosophy's accumulated wealth of reflection about the mind, rationality, perception, memory, emotions, and so forth could not be used by the computational approach. In McCulloch's famous phrase, the 'den of the metaphysician is strewn with the bones of researchers past.' In the Digital Humanities' case, the takeover happens at the level of tools. In most cases, however, this appropriation does not become an opportunity for a critical reflection on the role of the canon on liberal education, or for a reappraisal of the role of the text and the social, political, and moral roles it plays in society at large.

2. Artistic Digital Practices

Meetings between artists and computational technology show the possibility of a different paradigm. In many cases, making music, painting, producing installations, and writing with a computer changes the concepts artists work with, and, at the same time, forces computer sciences to change theirs as well. There are many examples in the rich history of 'digital art,' broadly understood (OuLiPo 1973; *ALAMO (Atelier de Littérature Assistée par la Mathématique et les Ordinateurs)*; Schaeffer 1952). I will illustrate their general features with reference to two more recent projects: the 'microsound' approach to musical composition (Roads 2004) and the T-Garden approach to agency (Sha 2002).

'Microsounds' are sonic objects whose timescale lies between that of *notes* – the smallest traditional music objects, whose duration is measured in seconds or fractions thereof – and *samples* – the smallest bit, measured in microseconds (10^{-6}). The manipulation of microsounds broadens substantially the composer's palette, but it is impossible without the help of technological devices of various kinds, from granular synthesis software to high-level mixing interfaces. Composers wishing to 'sculpt' sounds at the microlevel face a double challenge that translates into a mutual collaboration between compositional and algorithmic techniques. On the one hand, they need to broaden the syntax an grammar of music's language to allow the manipulation and aesthetic assessment of previously unheard of objects (Vaggione 2001). On the other hand, they need computer scientists and mathematicians to develop alternative analytic and synthetic models of sound (in addition to Fourier-transforms and similar methods) capable of capturing the features of sonic events lasting only a few milliseconds (Vaggione 1996).

The T-Garden environment produced at the Topological Media Lab follows a structurally similar paradigm. It is a closed room within which people wearing wireless sensors on their chest and limbs can project images and sounds on the room's walls and ceiling. As the participants move around, their gestures across the floor perturb the preexisting sonic and visual fields, thereby modifying it and introducing new aural and visual patterns. As Sha describes it:

[The environment] was built to explore how people can improvise gestures out of dense, evolving fields of media. In ordinary informal conversation, you can spontaneously drag or pitch your speech to express irony, sympathy, and so forth. Similarly, in a T-Garden, by waving your arm you write video or leave a trace in sound, and by moving about the space, solo or in concert with other people, you construct a voice for yourself out of a sound field that is summed from all the instrumental voices in the room. […] as you play, your continuous motions create an aural and visual 'voice' for yourself out of the ambient perceptual field (Sha 2002: 441-442).

Whereas Roads explored a new sonic object lying between samples and notes, Sha investigated a segment of human agency that is neither fully free nor fully constrained. He took human gestures as a proxy for agency in general and conceived a project that demanded theoretical and technical work on two fronts. On the artistic side, it mandated the translation of a theoretical reflection upon the status of semiotic structures into a concrete installation. It required the construction of an event that forced the participants to reassess their conception of communication and 'freedom of speech.' On the technical sides, it forced the computer scientists to develop real-time systems capable of interpreting human gestures and translating them into sonic and visual equivalents the participants could reintegrate into their communicative actions.

These two examples points to a pattern of cooperation between work in computational and non-computational disciplines that is deeply at odds with the AI/CogSci and DigHum patterns discussed above. Instead of a takeover, the artistic model produces a true encounter that changes both partners' technical and theoretical apparatus.

3. Digital Theoretical *Poiesis*

Could the interaction pattern most favored by artists be generalized to the Humanities? With the help of Aristotle's classification of human activity as *poiesis*, *praxis*, and *theoria* (Aristotle 1984, 1025b19-30, p.

1619), it could be objected that artistic practices are examples of *poiesis*. They produce objects on the basis of materials and manipulating tools. Anything (including digital tools) that provides new materials (material substrate) or new ways of organizing them will necessarily change artistic practices. And artists have always been extremely keen on expanding their material and formal palettes, long before the advent of digital techniques. The daily work of Humanists, it could be claimed, is essentially *theoretical*, since their task is to critically reflect upon a previously produced reality.

Paradoxically, however, a serious critical engagement with contemporary reality shows that the Aristotelian tripartite distinction is untenable in the current landscape. How can Humanities' traditional inquiries about human nature and human cultural production still be relevant in a landscape in which some of the communicating agents may not be human, partially or entirely? Can they go on in the same way? And *vice versa*: are science and technology fully aware that the new digital artifacts they are shepherding into the world may change its landscape and transform worldly action at the pragmatic as well as at the theoretical level? Or are they still relying upon a pre-digital universe in which technological artifacts were always to be used as mere tools deployed by humans, an assumption that seems increasingly questionable? The genuine challenge Humanists face when confronting the 'computational turn' demands a truly poietic effort. Humanists will have to engage in the constructions of new concepts in a close peer-to-peer interaction with the 'sciences of the artificial' (Cordeschi 2002) that may result in a form of 'digital theoretical *poiesis*' much closer to artistic practice than to their traditional forms of inquiry.

This suggestion does not pretend to exhaust the theoretical options we have at our disposal when reflecting upon the computational turn. There are certainly other views that may legitimately claim to be seeking the same goal. My contention, however, is that artistic practices in all forms of 'digital art' can serve as an inspiration to all of the Humanities disciplines. We can follow their path toward a new mode of digital encounter that does not fall into the well-worn path of hostile takeovers by either partner.

References

Agre, P. E. (2005). The Soul Gained and Lost: Artificial Intelligence as Philosophical Project. In S. Franchi and G. Güzeldere (eds.), *Mechanical Bodies, Computational Minds*. Cambridge, Mass.: MIT Press.

ALAMO (Atelier de Littérature Assistée par la Mathématique et les Ordinateurs). `http://alamo.mshparisnord.org/index.html` (accessed 24 March 2012).

Aristotle (1984). *The complete works of Aristotle*. Ed. by J. Barnes. Princeton, NJ: Princeton UP.

Cordeschi, R. (2002). *Discovery of the Artificial: Behavior, Mind, and Machines Before and Beyond Cybernetics*. Dordrecht: Kluwer.

Dupuy, J.-P. (2000). *The Mechanization of the Mind: On the Origins of Cognitive Science*. Princeton, N.J.: Princeton UP.

Floridi, L. (1999). *Philosophy and Computing: An Introduction*. London, New York: Routledge.

Floridi, L. (2011). *The Philosophy of Information*. Oxford: Oxford UP.

Franchi, S. (2006). Herbert Simon, Anti-Philosopher. In L. Magnani (ed.), *Computing and Philosophy,* Pavia: Associated International Academic Publishers, pp. 27-40.

Kirschenbau, M. G. (2010). What Is Digital humanities and What's It Doing in English Departments? *ADE Bullettin* 150: 1-7.

McCulloch, W. S. (1989[1948]). Through the Den of the Metaphysician. In W. S. McCulloch, *Embodiments of Mind*. Cambridge, Mass.: MIT Press, pp. 142-156.

OuLiPo (1973). *La littérature potentielle*. Paris: Gallimard.

Roads, C. (2004). *Microsound*. Cambridge, Mass.: MIT Press.

Schaeffer, P. (1952). *À la recherche d'une musique concrète*. Paris: Seuil.

Sha, X. (2002). Resistance Is Fertile: Gesture and Agency in the Field of Responsive Media. *Configurations* 10(3): 439-472.

Simon, H. (1994). Literary Criticism: a Cognitive Approach. In S. Franchi and G. Güzeldere (eds.), *Bridging the Gap,* Vol. 4. 1. *Stanford Humanities Review*, Special Supplement, pp. 1-26.

Svensson, P. (Summer 2010). The Landscape of Digital Humanities. *Digital Humanities Quarterly* 4(1).

Vaggione, H. (1996). Articulating Microtime *Computer Music Journal* 20(2): 33-38.

Vaggione, H. (2001). Some Ontological Remarks about Music Composition Processes. *Computer Music Journal* 25(1): 54-61.

A flexible model for the collaborative annotation of digitized literary works

Gayoso-Cabada, Joaquin
gayoxo@gmail.com
Universidad Complutense de Madrid, Spain

Ruiz, Cesar
cruiz85@gmail.com
Universidad Complutense de Madrid, Spain

Pablo-Nuñez, Luis
lpnunez@filol.ucm.es
Universidad Complutense de Madrid, Spain

Sarasa-Cabezuelo, Antonio
asarasa@fdi.ucm.es
Universidad Complutense de Madrid, Spain

Goicoechea-de-Jorge, Maria
mgoico@filol.ucm.es
Universidad Complutense de Madrid, Spain

Sanz-Cabrerizo, Amelia
amsanz@filol.ucm.es
Universidad Complutense de Madrid, Spain

Sierra-Rodriguez, Jose-Luis
jlsierra@fdi.ucm.es
Universidad Complutense de Madrid, Spain

1. Introduction

The Complutense University has been one of the first European universities that has collaborated with Google's project[1] by putting on the Web 100,000 volumes from its ancient fund. However scholars notice that these digitized texts are often of no much use to professors-researchers-students in literature unless additional tools are provided, to enhance the educational and research value of this material. In particular, the ability of making annotations on these texts has been largely recognized as a basic mean of adding value to this kind of digitized resources (Rios da Rocha et al. 2009). In this paper we present the annotation model used in @Note 1.0, a system developed at UCM funded by the Google's 2010 Digital Humanities Award program.

@Note 1.0 allows us to retrieve digitized works from Google Books collection and add annotations to enrich the texts with research and learning purposes: critical editions, reading activities, e-learning tasks, etc. One of the main features of @Note annotation model, which distinguishes it from similar approaches (Azouaou & Desmoulins 2006; Bechhofer at al. 2002; Koivunen 2005; Rios da Rocha et al. 2009; Schroeter et al. 2006; Tazi et al. 2003), is to promote the collaborative creation of annotation schemas by communities of researchers, teachers and students, and the use of these schemas in the definition of annotation activities on literary works. It results in a very flexible and adaptive model, able to be used by many different communities of experts in literature defending different critical literary theories and for different annotation tasks. In this paper we present this annotation model.

2. The @Note Annotation Model

2.1. Structure of the model

The structure of the @Note annotation model is summarized in the UML class diagram (Booch et al. 2005) of Fig. 1. In this model:

- *Annotation management communities* are groups of *annotation managers*, experts in literature (teachers, researchers, etc) who act as administrators to create activities, to select works and to organize activity groups.
- *Annotation communities*, in their turn, are groups of *annotators*, students / pupils interested in literature who perform proposed annotation activities.
- Each a*nnotation activity* comprises (i) a *digitized work*, (ii) a *metalevel-oriented annotation schema*, (iii) a *work-oriented annotation schema*.
- In this context, the *works* are the literary texts that can be annotated during the annotation activities. *Annotations*, in their turn, are characterized by: (i) an *annotation anchor* (the region of the work to which the annotation refers), (ii) an *annotation content* (a free rich-text piece that actually configure the annotation), (iii) a set of *annotation types* (semantic qualifiers for annotations) chosen from the annotation schemas attached to the annotation activity (at least one from the metalevel-oriented annotation schema).

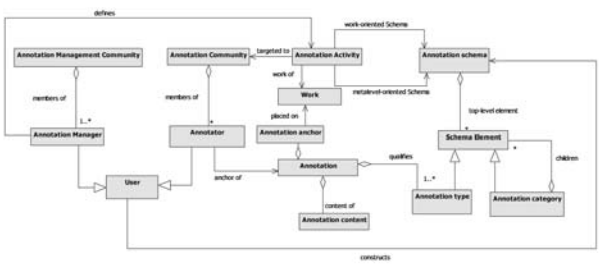


Figure 1: @Note information model

- The *annotation schemas* are explicit formalization of the types of annotations that can be carried out on works. In @Note, annotation schemas are hierarchies formed by annotation types and *annotation categories* (sets of annotation types and/or others, more specifically, annotation categories). In their turn, they can be *metalevel-oriented annotation schemas* (schemas which usually comprise concepts concerning particular literary theories around which the annotation activities are articulated), or *work-oriented annotation schemas* (schemas that capture aspects relative to the relationships between annotations and their anchors). While schemas of the first type are created by annotation managers, schemas of the second type are usually created by annotators.

In the context of the annotation management community, an annotation schema can be *public* or *private*. A private schema is only accessible for the annotation manager who created it. On the contrary, a public schema is accessible for all the annotation managers. Annotation managers have unlimited privileges on all the schemas to which they can access (i.e., they can create new annotation types and categories, they can blend two different types/categories in a single one, they have renaming and erasing privileges, etc), with the exception of modifying the *public/private* character (it only can be done by the schema's creator). In addition, when annotation managers create annotation activities, they only can choose those schemas to which they have access grants. Concerning annotators, they can add new types and categories to the book-oriented annotation schema, but they can't perform any other modification.

Figure 2: Example of rules governing the annotation process (informally described using natural language)

2.2. The annotation process

The @Note annotation process governs how to create the different types of information elements envisioned in the annotation model. For this purpose, @Note introduces a set of rules governing aspects like information visibility, creation and modification privileges of annotations and annotation schemas, etc. Although, by lack of space, these rules will not be detailed here, in Fig. 2 we include an example concerning an informal description of some of the rules governing the management of annotation schemas.

2.3. Annotation browsing and recovering

Annotation schemas in @Note are seen as T-*boxes* of description logic theories (Brachman & Levesque 2004). For instance, Fig 3a shows, edited in @Note, a fragment of the annotation schema used at UCM in an English Literature introductory course, while Fig 3b despites the description logic's counterpart. This simple interpretation is still powerful-enough to enable powerful *annotation browsing* and *annotation recovering* behavior. Indeed:

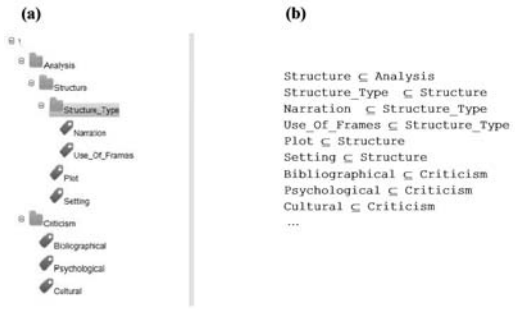


Figure 3: (a) A fragment of annotation schema (b) description logic counterpart

- Annotations can be browsed using annotation schemas, in a similar way to a folder explorer in a filesystem. In each step, there is a category or a type selected, and the user will *see* all the annotations entailed by such a selected element.
- Queries consist of arbitrary Boolean formulae involving annotation types and categories, being the outcomes the annotations entailed by such formulae.

In both cases, since entailment will be performed according to the description logic principles, the process will take into account the *is-a* relationship made explicit by the annotation schema.

2.4. Some technical details

The system has been entirely developed using Google technologies for the development of Rich-Internet Applications (RIAs) (Fraternali et al. 2010): GWT in the client side and the Google App Engine's facilities in the server side (Unruh 2010). Fig. 4 shows some snapshots of the system. The current version runs on the fully free-access books integrated in Google Books, and, in particular, on the UCM-Google collection. The works retrieval is achieved by the use of the Google Books API through REST (Richardson & Ruby 2007), and then presented to clients in an asynchronous way to keep them responsive to their events.

Figure 4: Some snapshots of @Note

3. Conclusions and Future Works

@Note promotes a fully collaborative annotation process, in which not only literary works are collaboratively annotated, but also annotation schemas are collaboratively created. The @Note system has been evaluated at UCM by several researchers and students in literature. They highlighted the flexibility of the annotation model, and, in particular, the ability to create and share annotation schemas tailored according to different critical perspectives and annotation activities. Additionally, they appreciate a sufficient expressive power from a browsing and recovering point of view. They also remarked the educational potential of the tool, although some advanced features could add some conceptual difficulties for students.

Currently we are working to adapt the annotation tool in order to facilitate its connection to a repository of learning objects, so as to allow the storage of literary texts' annotations as learning objects (Polsani 2003), and to make possible the recovery of those annotations and move about them according to the associated metadata. Thus, we are developing a communal working space for the creation of written compositions in different traditions and languages. We are also experimenting with the students' capability for developing their own catalogues, annotating the literary texts according to them and reusing their annotations in the production of critical essays. Additionally, we are working on connecting our system with other digital libraries (in particular, with Hathi Trust[2]). Finally, we are planning to address interoperability issues, in order to enable the interchange of annotations according to some of the emerging standards proposed by the digital humanities community (e.g., OAC[3]).

Acknowledgements

This work has been funded by Google with a grant of the Google's 2010 Digital Humanities award program entitled *Collaborative annotation of digitalized literary texts*. Additionally, this work has been performed in the context of the project grants of the Spanish Ministry for Research and Innovation (FF12008-06924-C02-01 and TIN2010-21288-C02-01), UCM (PIMCDs 2010/177 and 2011/313) and Santander-UCM (GR 42/10 - 962022).

References

Azouaou, F., and C. Desmoulins (2006). MemoNote, a context-aware annotation tool for teachers. *Proceedings of the 7th International Conference on Information Technology Based Higher Education and Training* ITHET'06, Sidney, Australia, July 2006.

Bechhofer, S., L. Carr, C. Goble, S. Kampa, and T. Miles-Board (2002). The Semantics of Semantic Annotation. *Proceedings of the First International Conference on Ontologies, DataBases, and Applications of Semantics for Large Scale Information Systems* ODBASE'02, Irvine, CA, USA, October 2002.

Booch, G., J. Rumbaugh, and I. Jacobson (2005). *The Unified Modeling Language User Guide* (*2nd Edition*). Boston: Adisson-Wesley.

Brachman, R., and H. Levesque (2004). *Knowledge Representation and Reasoning*. Amsterdam: Morgan-Kauffmann.

Fraternali, P., R. Gustavo, and F. Sánchez-Figueroa (2010). Rich Internet Applications. IEEE *Internet Computing* 14(3): 9-12.

Koivunen, M.-R. (2005). Annotea and Semantic Web Supported Collaboration. *Proceedings of the UserSWeb Workshop – 2nd European Semantic Web Conference*, Heraklion, Grece, June 2005.

Polsani, P. (2003). Use and abuse of reusable learning objects. *Journal of Digital Information* 3(4).

Richardson, L., and S. Ruby (2007). *Restful web services*. Beijing: O'Reilly.

Rios da Rocha, T., R. Willrich, R. Fileto, and S. Tazi (2009). Supporting Collaborative Learning Activities with a Digital Library and Annotations. *Proceedings of the 9th IFIP World Conference on Computers in Education* WCCE'09, Bento Gonçalves, Brazil, July 2009.

Schroeter, R., J. Hunter, J. Guerin, I. Khan, I., and M. A. Henderson (2006). Synchronous Multimedia Annotation System for

Secure Collaboratories. *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing* e-Science '06, Amsterdam, Netherlands, December 2006.

Tazi, S., Y. Al-Tawki, and K. Drira K. (2003). Editing pedagogical intentions for document reuse. *Proceedings of the 4th IEEE Technology Based Higher Education and Training* ITHET'03, Marrakesh, Morocco, July 7-9, 2003.

Unruh, A. (2010). *Google App Engine Java and GWT Application Development*. Birmingham: Packt Publishing.

Notes

1. http://www.ucm.es/BUCM/atencion/25403.php
2. http://www.hathitrust.org/
3. http://www.openannotation.org/

HyperMachiavel: a translation comparison tool

Gedzelman, Séverine
severine.gedzelman@ens-lyon.fr
ENS de Lyon, France

Zancarini, Jean-Claude
jean-claude.zancarini@ens-lyon.fr
ENS de Lyon, France

1. Introduction

The HyperMachiavel project started with the idea of a tool that would aid research communities comparing several editions of one text and in particular comparing translations.

The Italian studies department (Triangle laboratory) at ENS de Lyon has been working for many years on fundamental texts, from Machiavelli, Guicciardini and other contemporary followers, that put forward new political concepts throughout Europe in the 16th century. The question addressed in the project was mainly about the transfer of these concepts from one language to another, and especially their reception in France. The first aligned corpora tested in our tool gathers different editions of Machiavelli's *Il Principe*, the *princeps edito* from Blado in 1532 and the first four French translations of the 16th century.

Inspired by machine translation and lexicographic domains, the system presented in this paper proposes an annotation environment dedicated to the edition of lexical correspondences and offers different views to assist humanities researchers in their interpretations of the quality and the specificities of translator's work.

2. Viewing and Searching in Aligned Corpora

2.1. Synoptic View

To be able to identify lexical correspondences, machine translation tools usually propose a frame of two panels, one for the source text and the other for the target text. The visualized interface is meant for annotators to easily revise the results obtained from automatic word alignment. In general it only considers a pair of texts at a time.

In the world of digital editions, text comparison has always been of great interest and the request to view diplomatic vs normalized transcriptions, or simply

different editions of one text increases all the more that digital data become truly available. Systems such as the Versioning Machine (Schreibman et al. 2007) already offers parallel views for TEI aligned corpora with no constraint on the number of displayed texts. Being able to take into consideration several target versions was an important starting point when designing **HyperMachiavel**. Therefore the main panel offers a 'Parallel view' (see Fig. 1) to interact with processes that need global context visualization (e.g. concordancer results).

2.2. Alignment

To deal with text alignment, many tools working with pair of texts (bitext2tmx, alinea or mkAlign[1], Fleury et al. 2008) uses statistical measures of co-occurrence or combines word distribution algorithm with cognates detection. Although original investigation on automatic alignment has been proceeded for non modern texts like in the Holinshed Project (Cummings & Mittelbach 2010), our focus was to fulfill the need for editing and controlling lexical correspondences. Exploring these techniques will be done in a second development phase but for the moment, HyperMachiavel imports aligned corpora encoded in XML-TMX or XML-TEI and proposes a manual alignment feature for additional text files.

Because old edition texts usually present great variability in the lexical forms, we favored a computer-aided system for bilingual search and put efforts in the lexicography work environment.

2.3. Lexicography and Lexical Tagging

Linguistic annotation tools make use of external dictionaries to tag texts but cannot be applied on old texts per se. Combining this process with the user's analysis of the text to build its own lexical resource, either by bringing corrections to the model or proposing partial tagging, is a necessary condition for our study. Recent work conducted by Lay et al. 2010 have also emphasized the user's interaction in the lexical exploration and offers customization of lexical resources. HyperMachiavel follows that path by presenting tokenizer's choice (among which TreeTagger java implementation by Helmud Schmid) and a general framework to consolidate endogenous or external lexical resources that would work in harmony with the defined corpora.

Information obtained from tokenization like category, lemma, word onset and offset localisation in the texts are registered in XML files. A full-text search engine, specially designed for the tool, uses this information, so does the 'Vocabulary view' listing all the forms for each language in the corpora. This view displays frequency information by text (see Fig. 2) and brings a first impression on lexical distribution among versions, or authors in our case.

The use of concordancer (Key Word In Context) that classically links dictionary entries to corpora occurrences is central in the system (results of monolingual search or after a selection from the 'Vocabulary view', e.g. Fig. 3). As for equivalences detection are concerned, some authors have shown that bilingual concordancers (ParaConc; Barlow et al. 2004; ConcQuest; Kraif 2008) are very helpful. However viewing results of bilingual search with concordancers in our case is not sufficient. As we want to compare numerous source and target texts, a 'Parallel Indexes view' has been designed and is described hereafter.

3. Exploring Equivalences

3.1. Equivalences Detection and Validation

Annotating equivalences is a long process and a semi-automatic scenario has been introduced with the bilingual search. Results from that search are displayed in the 'Parallel indexes view' (see Fig. 4). Some results will show co-presence in an aligned segment of at least one source and target searched items, but source or target occurrences could be found alone.

When the user is faced with empty results (source or target), the 'Parallel view' helps detect other potential equivalence expression. In some cases, it could mean that no translation has been given or that the occurrence belong to another semantic group (e.g. homonyms, polysemes).

Validation of real equivalences versus suggested equivalences from the bilingual search, is done by checking the lines. However manual completion is often required, especially when source or target refer to specific expressions or to other lexical terms that were not used in the search.

3.2. Graph Visualizations

Exploring edited equivalences can be pursued in other environments like excel or any data analysis programs but HyperMachiavel has developed some feature to highlight thematic groups of equivalences and show stylistic differences between translations.

For each dictionary entry, one can ask to see validated target equivalences. Information about their frequency and distribution between versions of text help the user evaluate translator's appropriation

of the source text (see Fig. 5). Graph visualizations illustrate the same information but the user can go further and ask to see equivalents of equivalents until no new information is provided. Groups of equivalences usually appear and interrogate the user on the conceptual themes exposed in the corpora.

In the current aligned corpora, we have noticed that most French translators of *Il Principe* depict the polysemic and complex reality of new political concepts at that time by giving various lexical equivalences. Yet one of the translators prefers to steadily deliver concepts with almost unique translation (e.g. 'stato' with only 'estat(s)') perhaps believing that the rest of the text would do the rest.

4. Conclusion

This paper presents the state of development of HyperMachiavel[2], a tool distributed under a french open-source licence (Cecill-B). Although the demonstration was done with only french-italian texts the tool was defined to support manual edition of equivalences for aligned multilingual corpora and lexicography work.

Providing import and export formats recommended by the digital humanity community ensure reuse of corpora in other environment such as text mining platform using the full complexity of TEI encoded corpora (e.g. TXM[3] , Heiden 2010). Use of automatic alignment algorithms will be tested with another aligned corpora based on Machiavelli's *The Art of War*-translations. The idea is to extend our tool features for comparable corpora (different texts of authors highly influenced by Machiavelli's work).

Funding

The work presented here is carried out within the research project (2008-2011) 'Naissance, formes et développements d'une pensée de la guerre, des guerres d'Italie à la paix de Westphalie (1494-1648)', funded by the French Agence Nationale Recherche.

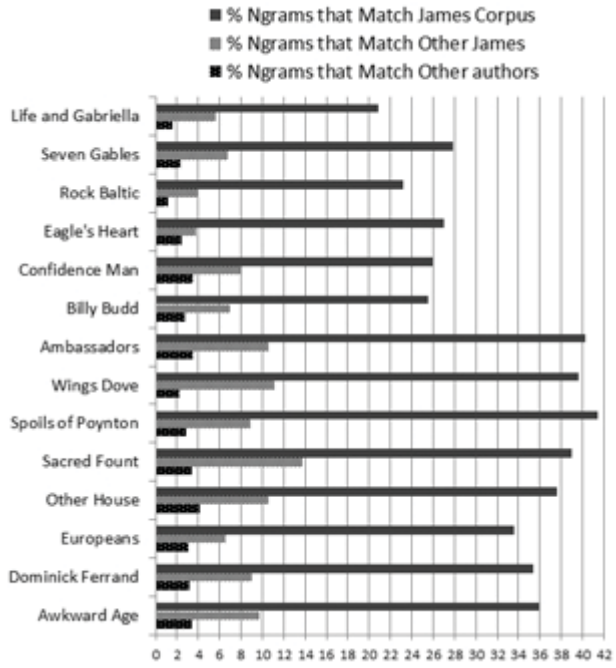


Figure 1: 'Parallel view' of the aligned texts (right panel) and 'Corpus view' are showing the text structure common to all versions (left panel)

Figure 2: The 'Vocabulary view' presents lexical forms and their frequency (number of occurrences) in the different aligned texts

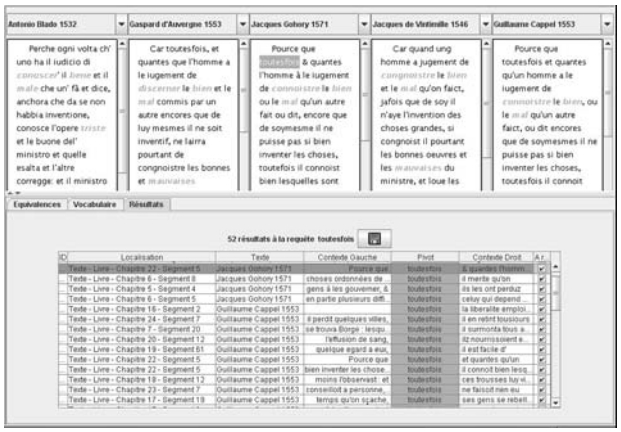


Figure 3: The 'KWIC view' shows occurrences of the french old form 'toutesfois'

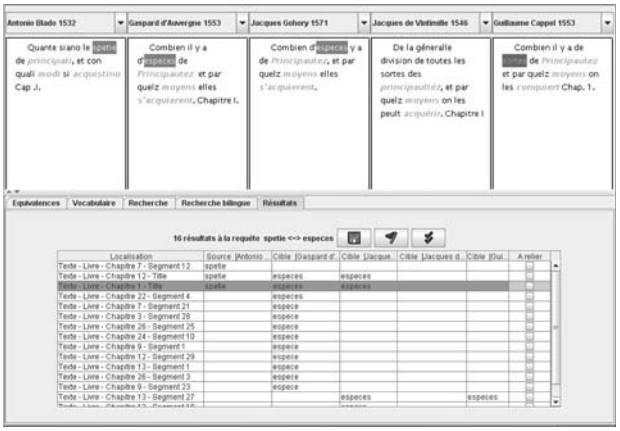


Figure 4: The 'Parallel Indexes view' shows searched source and target expressions that appear in the same localization. The bilingual search example was performed with 'spetie' as source expression and 'espece(s)' as target expressions

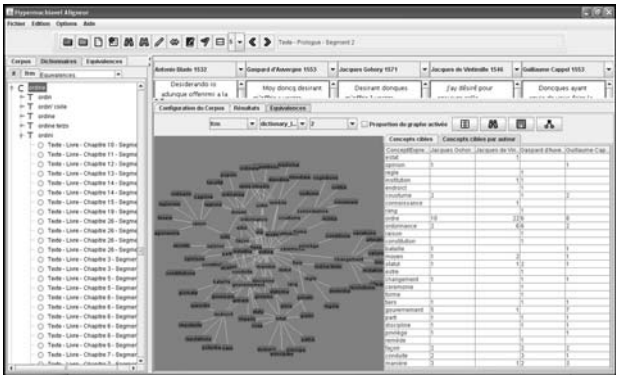


Figure 5: Views of French equivalents for the Italian concept 'ordine'

References

Bowker, L., and M. Barlow (2004). Bilingual concordancers and translation memories: A comparative evaluation. In E. Y. Rodrigo (ed.), *Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training*. Geneva, Switzerland, August 28, 2004, pp. 70-83

Cummings, J., and A. Mittelbach (2010). The Holinshed Project: Comparing and linking two editions of Holinshed's Chronicle. In *Citation Information*. International Journal of Humanities and Arts Computing 4: 39-53 DOI 10.3366/ijhac.2011.0006, ISSN 1753-8548, October 2010.

Fleury, S., and M. Zimina (2008). Utilisations de mkAlign pour la traduction philologique. *Actes JADT 2008, Journées Internationales d'Analyse Statistiques des Données Textuelles*, Lyon, 2008.

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japan 2010.

Kraif, O. (2008). Comment allier la puissance du TAL et la simplicité d'utilisation ? L'exemple du concordancier bilingue ConcQuest. In *Actes JADT 2008, Journées Internationales d'Analyse Statistiques des Données Textuelles*, Lyon, 2008.

Lay, M.-H., and B. Pincemin (2010). Pour une exploration humaniste des textes : AnaLog. In *Actes JADT 2010, Journées Internationales d'Analyse Statistiques des Données Textuelles*, Roma, 2010.

Schreibman, S., A. Hanlon, S. Daugherty, and T. Ross (2007). The Versioning Machine v3.1: A Tool for Displaying and Comparing Different Versions of Literary Texts. In *Digital Humanities 2007*. Urbana, Illinois. Jun. 2007.

TEI Consortium. (2008). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.1.0.,

4th July. L. Burnard and S. Bauman, eds., TEI Consortium. `http://www.teic.org/Guidelines/P5`

Notes

1. Software 'alinea' is developed by O. Kraif (`http://w3.u-grenoble3.fr/kraif/index.php?Itemid=43&id=27&option=com_content&task=view`), 'bitext2tmx' (`http://bitext2tmx.sourceforge.net`) by S. Santos, S. Ortiz-Rojas, M. L. Forcada and now R. Martin, and 'mkAlign' by S. Fleury (`http://www.tal.univ-paris3.fr/mkAlign`)
2. The software can be downloaded at `http://hypermachiavel.ens-lyon.fr`
3. TXM is a collaborative, open-souce software initiated by S.Heiden and the textometrie project members (`http://textometrie.ens-lyon.fr/`)

Discrimination sémantique par la traduction automatique, expériences sur le dictionnaire français de Littré

Glorieux, Frédéric
frederic.glorieux@enc.sorbonne.fr
École nationale des chartes, France

Jolivet, Vincent
vincent.jolivet@enc.sorbonne.fr
École nationale des chartes, France

L'exploration d'un corpus sur un champ sémantique, avec plusieurs mots clés, produit de longues concordances fastidieuses à dépouiller. Les outils proposent de trier les résultats par les mots du contexte, ce qui permet rarement de dégager des classes sémantiques éclairantes quand le vocabulaire est varié, sans mots spécialement fréquents. Des regroupements peuvent être opérés par référence à un système de traits sémantiques, ou 'sèmes', attribués aux mots du contexte. Maurice Gross[1] a ainsi montré comment quelques traits peu équivoques tels que *humain–non humain* pouvaient distinguer des acceptions: par exemple, le verbe *permettre* signifie généralement 'autoriser' lorsque son sujet est une personne, ou bien, il signifie 'rendre possible' lorsque le sujet est une chose. La sémantique syntaxique n'est cependant pas suffisante pour dégager toutes les significations. François Rastier[2] a montré que l'interprétation du sens dans un texte, et non seulement dans une phrase, s'appuyait sur l'isotopie de bien d'autres sèmes, par exemple */noirceur/*, */rapidité/*, et tout le spectre des valeurs et des sentiments. Il a par ailleurs insisté sur les limites des ontologies classificatoires de concepts, comme WordNet[3]. Si un lexique sémique informatisé serait souhaitable, il n'est malheureusement pas encore constitué[4]. Notre communication explore les potentialités et les limites de la traduction automatique comme instrument de discrimination sémantique, en observant d'abord son application à un dictionnaire.

L'informatisation de la désambiguïsation sémantique a très tôt (1949) été liée à la traduction automatique, notamment par Warren Weaver[5]. Lorsqu'un traducteur humain rencontre en contexte le mot français 'société', il a l'expertise pour choisir en anglais entre '*society*' et '*company*'. Comment résoudre informatiquement cette ambiguïté? Weaver appelait à la constitution d'un lexique inventoriant les différents sens de chaque mot, avec des critères repérables dans le contexte permettant d'opérer la distinction. Une telle ressource suppose l'existence d'universaux linguistiques partagés entre les langues, (*language invarients* selon les termes de Weaver). À cette époque, c'était visionnaire, la désambiguïsation sémantique se présentait comme un préalable à la traduction automatique.

Actuellement, les traducteurs automatiques en ligne sont de plus en plus performants, pourtant, la désambiguïsation sémantique ne présente pas de résultats aussi convaincants. Par exemple les moteurs de recherche ne distinguent pas encore strictement plusieurs sens d'un même mot, comme *société*: 'communauté organisée' ou 'entreprise'. Des résultats ont été obtenus, mais les progrès butent sur un obstacle humain. Jean Véronis[6] a ainsi montré qu'avec un même dictionnaire et les même textes, des personnes n'affectaient pas le même sens à un mot équivoque (*barrage*). Les nuances proposées par les lexicographes diffèrent selon les dictionnaires et sont comprises différemment par les lecteurs. Si la traduction automatique a progressé, ce n'est pas par la désambiguïsation sémantique, mais par la loi de Moore qui a augmenté exponentiellement la capacité des ordinateurs, qui désormais mémorisent et traitent une grande quantité de traductions alignées.

Même si elle n'est pas explicitement formalisée, la traduction automatisée effectue pourtant bien des distinctions sémantiques. Dans votre traducteur préféré, proposez par exemple ces deux phrases: 'La **société** doit protéger les faibles. La **société** protège ses biens.'; vous pouvez obtenir les traductions: '**Society** must protect the weak. The **company** protects its assets.', 'Die **Gesellschaft** muss den Schutz der Schwachen. Das **Unternehmen** schützt sein Vermögen.', 'La **sociedad** debe proteger a los débiles. La **empresa** protege sus activos.' Pourrait-on mobiliser cette information sémantique pour opérer des distinctions dans une même langue?

Pour observer le phénomène plus précisément, nous avons établi une édition électronique du *Dictionnaire de la langue française* d'Émile Littré (1863-1872)[7]. Comme tout dictionnaire, le *Littré* distingue les différentes significations d'un même mot vedette, mais il a aussi l'avantage de comporter de nombreuses citations et exemples d'emplois. L'auteur revendique un ordre historique des significations, ou du moins, fidèle à une reconstruction génétique selon ses convictions positivistes. Si la finesse des définitions et des distinctions de Littré impressionne encore, leur ordre convient beaucoup moins au lecteur d'aujourd'hui. Nous avons fait l'expérience d'utiliser la traduction automatique pour réordonner l'article

selon l'équivalent traductionnel proposé dans une langue cible. Soit par exemple l'article *SOCIÉTÉ*, toutes les occurrences de *société* dans les exemples et les citations sont en gras. La traduction automatique en anglais renvoie l'article traduit, avec en gras des occurrences de *society*, *company*, voire *corporation*, *venture*, ou *partnership*. Ces mots servent ensuite de clés pour regrouper différemment les paragraphes de Littré, et proposer un nouvel ordre de lecture, selon le réseau sémantique de la langue cible. Avec une langue qu'il connait, le lecteur constate évidemment des erreurs de traduction, notamment dans les textes anciens ou les phrases tronquées, mais il est aussi surpris par la reconnaissance de locutions rares (*Petites-Maisons*: 'madhouse'). Sur des articles longs, avec assez d'exemples par significations, le classement par équivalent traductionnel dégage généralement des distinctions éclairantes, en pondérant mieux les acceptions fréquentes, en regroupant des usages que la doctrine historique de Littré avait séparé. Au fond, le procédé retrouve la pratique ancienne dans les langues européennes de se référer à une autre langue (en général le latin) pour distinguer les significations dans une langue vivante. La traduction automatique n'a évidemment pas l'exactitude d'un jugement humain, mais elle permet de projeter les textes sur des dizaines de langues, qui sont autant d'espaces sémantiques originaux. La comparaison et la combinaison de plusieurs langues produisent des ordres de lecture parfois difficiles à interpréter, mais rarement dénués de sens.

La traduction automatique articule des distinctions sémantiques, certes biaisées par les limites de l'informatique, les intérêts pratiques des utilisateurs, ou le lexique de l'anglais qui sert généralement de langue pivot, mais pourtant révélatrices de l'expérience singulière de nombreuses langues. Cette approche convient aux *digital humanities* parce qu'elle ne suppose pas a priori des universaux de signification et tente plutôt d'instrumentaliser l'interculturalité afin de mieux comprendre sa propre culture. La procédure peut être imitée par des dizaines d'autres langues, sans constitution coûteuse de lexiques sémantiques ou de corpus annotés. Le lexicographe, d'abord, sera invité à revoir le plan de ses articles, à se dégager de ses présupposés logiques, par comparaison au découpage lexical d'autres langues. Nous utilisons ce type de procédures dans des interfaces interactives de recherche, afin d'élargir le bouquet de mots d'une requête, puis pour trier les concordances.

Notes

1. Les bases empiriques de la notion de prédicat sémantique. *Langages* 63, Paris 1981, pp. 7-52.
2. *Arts et sciences du texte*. Paris: PUF 2001.
3. `http://wordnet.princeton.edu/`
4. Mathieu Valette et al. (2006). Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens. In *TALN'06*, 2006.
5. Le *Translation* memorandum, 1949.
6. Sense tagging: does it make sense? In A. Wilson, P. Rayson et T. McEnery (dir.), *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*. Frankfurt: Peter Lang 2003.
7. `http://corpus.enc.sorbonne.fr/littre/`

The Myth of the New: Mass Digitization, Distant Reading and the Future of the Book

Gooding, Paul Matthew
paul.gooding.10@ucl.ac.uk
University College London, UK

Warwick, Claire
c.warwick@ucl.ac.uk
University College London, UK

Terras, Melissa
m.terras@ucl.ac.uk
University College London, UK

1. Introduction

This paper proposes the importance of developing a realistic theoretical framework for the rise of mass digitization: the importance of the diffusion of innovations in shaping public opinion on technology; hyperbole and the myth of the 'digital sublime' (Mosco 2004); and the role of different media in shaping human understanding. It proposes that mass digitization can only be viewed as a significant deviation from the print medium when it emerges from the shadow of the Gutenberg Press and exerts a similarly powerful paradigm of its own, and that as a result there is a disconnect between the reality of the medium and the discourse with which it is surrounded.

2. Research context

In recent years companies such as Google have generated huge interest in their attempts to digitize the world's books. Described by Crane as 'vast libraries of digital books' (2006), they provide a powerful future vision where books exist within a digital version of the mythical universal library. Yet we know very little about what to do with these books once they are digitized, or even how they are being used by researchers and the wider public. The contemporary debate is defined by hyperbole and the belief that digital texts will inevitably destroy the print paradigm, a discourse that reflects the rhetorical direction of many historical reactions to technological change.

At the same time, a new research method which Moretti labels 'distant reading' (2007) allows researchers to undertake quantitative analysis of these massive literary corpora using tools such as the Google Ngrams Viewer[1]. Important work on these corpora has been carried out within the Digital Humanities community, such as projects by Jockers (2011) and Cohen (2006), and further research is being made possible by Digging into Data[2] funding from NSF[3], NEH[4], SSHRC[5] and JISC[6]. However, as we shall argue below, it is vital that we understand the theoretical background of such research, and its potential effects on scholarly methodology in both DH and traditional literary analysis.

3. Quantitative methods and theories of technology

Rogers (2003) tells us that far from being a deterministic inevitability, the success of a new technology relies in no small part upon external human factors. The role of opinion leaders is essential to this, and in the case of mass digitization the literature is filled with examples of the fetishization of both print and digital technologies. Gurus and promoters tell us that digitization will improve the world, and render obsolete what has gone before; others express concern about the cataclysmic effects that new technologies may have on humanity (Mosco 2004: 24). The result is a powerful narrative of ruptures and dramatic change, a myth that is historically common yet rarely reflected in reality.

This narrative has been shaped by a debate that has spanned decades (Benjamin 1936; McLuhan 1962; Barthes 1977; Baudrillard 1994; Lanier 2011), and so it is essential to consider the influence that theorists have exerted on the contemporary debate, and the misconceptions that have arisen as a result. For example, McLuhan's conception of a global village, a 'single constricted space resonant with tribal drums' (1962), is at odds with our experience of the contemporary World Wide Web. The village operates in a semi-closed system with clear boundaries, whereas the Web is unprecedented in its scale and openness. It more accurately resembles the loose structure of a modern city, and therefore mirrors the behavioural changes that have been noted in residents of large cities. With increased access to information comes an increased opportunity cost, the sense that whatever a person does will necessarily involve missing out on something else potentially useful (Deutsch 1961: 102). Accordingly, we can see that both city dwellers and web users exhibit similar behaviour in filtering and sorting the mass of information to which they are exposed. As a result of this promiscuous, diverse reading style (Nicholas et al. 2004) the authorial voice has been side-lined, and the growth of a specific cultural movement that

hyperbolises quantitative analysis of literary corpora threatens to hasten this process.

Many of the quantitative methods we have already seen hold great potential as research methods (Cohen 2006; Jockers 2011; Michel & Shen 2010; Moretti 2007), but there is a wider social discourse that appears to have taken quantitative analysis as a sign that traditional methods have, as with print, become an unnecessary distraction. This culture has redefined the content of texts as information, in an overly literal interpretation of the death of the author that undermines both authority and context. Information, viewed in this manner, is an abstract entity analogous to computerized data and therefore open to the same methods of interrogation (Nunberg 2010). There is no doubt that quantitative analysis holds great potential when combined with close reading, but we have witnessed a rhetoric that argues that close reading becomes unnecessary when large data becomes available (Anderson 2008). What appears to be a revolutionary faith in mass digitization, though, disguises a deeply conservative technological determinism that is reflected in the manner that digitization still so closely remediates the print medium.

Instead of acting as a disruptive force, mass digitization borrows greatly from the print medium that it remediates, a feature common to technologies in the early stages of their development (Bolter & Grusin 1996: 356-357). Even quantitative analysis can operate as a confirmatory method; rather than creating new knowledge it often acts only to confirm what humanities scholars already knew. The digital form, then, still operates 'not as a radical break but as a process of reformulating, recycling, returning and even remembering other media' (Garde-Hansen et al. 2009: 14). The reality is that quantitative methods are most effective when used alongside the close textual reading that allows us to contextualize the current glut of information. Yet the high speed processing of huge numbers of texts brings with it a potential negative impact upon qualitative forms of research, with digitisation projects optimised for speed rather than quality, and many existing resources neglected in the race to digitise ever-increasing numbers of texts.

4. Conclusion

We must therefore learn more about the potential, and limitations, of large-scale digitization in order to ensure that a focus on quantity rather than quality does not override the research needs of the wider community. If accepted, therefore, this paper will present some of the results of Gooding's on-going doctoral research into the use and impact of mass digitization projects, which will include case studies from institutions such as the British Library. It will argue that we must look beyond the noise that characterises the theoretical framework proposed above, and learn more about how the true impact of mass digitization.

References

Anderson, C. (2008). The end of theory: the data deluge that makes the scientific method obsolete. *Wired*. Available at: `http://www.wired.com/science/discoveries/magazine/16-07/pb_theory` [Accessed May 31, 2011].

Barthes, R. (1977). The death of the author. In *Image-Music-Text*. London: Fontana Press.

Baudrillard, J. (1994). *Simulacra and simulation*. Ann Arbor: U of Michigan P.

Benjamin, W. (1936). The work of art in the age of mechanical reproduction. In *Illuminations*. New York: Schocken Books.

Bolter, J. D., and R. A. Grusin (1996). Remediation. *Configurations* 4(3): 311-358.

Cohen, D. (2006). From Babel to knowledge: data mining large digital collections. *D-Lib Magazine* 12(3). Available at: `http://www.dlib.org/dlib/march06/cohen/03cohen.html` [Accessed February 7, 2011].

Crane, G. (2006). What do you do with a million books? *D-Lib Magazine* 12(3). Available at: `http://www.dlib.org/dlib/march06/crane/03crane.html` [Accessed January 7, 2011].

Deutsch, K. W. (1961). On social communication and the metropolis. *Da* 90(1): 99-110.

Garde-Hansen, J., A. Hoskins, and A. Reading, eds. (2009). *Save as...* Basingstoke: Palgrave MacMillan.

Jockers, M. (2011). Detecting and characterizing national style in the 19th Century novel. In *Digital Humanities 2011*. Stanford University, Palo Alto.

Lanier, J. (2011). *You are not a gadget*. London: Penguin.

McLuhan, M. (1962). *The Gutenberg galaxy: the making of the typographic*. Toronto: U of Toronto P.

Michel, J.-B., and Y. K. Shen (2010). Quantitative analysis of culture using millions of digitized books. *Science Magazine*.

Moretti, F. (2007). *Graphs, maps, trees: abstract models for literary history*. London and New York: Verso.

Mosco, V. (2004). *The digital sublime: myth, power, and cyberspace*. Cambridge Mass.: MIT Press.

Nicholas, D., et al. (2004). Re-appraising information seeking behaviour in a digital environment: bouncers, checkers, returnees and the like. *Journal of Documentation* 60(1): 24-39.

Nunberg, G. (2010). Counting on Google Books. *The Chronicle of Higher Education*. Available at: http://chronicle.com/article/Counting-on-Google-Books/125735 [Accessed September 15, 2011].

Rogers, E. M. (2003). *Diffusion of innovations*. Fifth Edition. New York: Simon & Schuster.

Notes

1. http://books.google.com/ngrams
2. http://www.diggingintodata.org
3. http://www.nsf.gov/
4. http://www.neh.gov/grants/guidelines/diggingintodata.html
5. http://www.sshrc-crsh.gc.ca/home-accueil-eng.aspx
6. http://www.jisc.ac.uk/whatwedo/programmes/digitisation/diggingintodata.aspx

Designing Navigation Tools for an Environmental Humanities Portal: Considerations and Critical Assessments

Graf von Hardenberg, Wilko
wilko.hardenberg@carsoncenter.lmu.de
Rachel Carson Center for Environment and Society - LMU, Germany

Coulter, Kimberly
kimberly.coulter@carsoncenter.lmu.de
Rachel Carson Center for Environment and Society - LMU, Germany

As digital humanities projects bring together more data than ever before, ways of presenting and accessing this data are transforming rapidly: text mining becomes more powerful and 'searching' and 'browsing' become more complex and interrelated. These tools not only help us find what we are looking for; they also shape our paths of inquiry. For this reason, it is important to carefully consider both the ontologies behind the tools (what exists and how can it be ordered?) as well as their epistemological effects (how do classification and representation affect our understanding?). How can we preserve, in a digital environment, the spirit of serendipity and discovery that characterizes humanities research? By asking such critical questions, we hope to forge tools that may generate unexpected results and new perspectives.

This paper addresses such considerations in the selection and design of navigation tools for humanities subject portals, as well as some current options and concerns related to these choices. As a concrete case study, we reflect on our experiences creating the Environment & Society Portal (http://www.environmentandsociety.org), a not-for-profit education, research, and outreach project at the Rachel Carson Center for Environment and Society in Munich, funded by the German Federal Ministry for Education and Research.

The Environment & Society Portal, which launched in early 2012, makes digital multimedia and interpretive materials in environmental humanities freely and openly accessible to academic communities and the interested public internationally. To enhance user engagement, in the coming year we will be building on features

that facilitate and encourage feedback, interaction, contribution and dissemination of information.

The Portal's multimedia content mixes retrodigitized environmental humanities materials with born-digital content, such as interpretive exhibitions, short descriptions of places and events, or the localized environmental histories related to broader issues found in the sub-project 'Arcadia: European Environmental Histories' (this collaboration with the European Society for Environmental History eventually plans to include all world regions). Content may be linked or clustered according to diverse research paths. It will be possible to add new entries on the same topics to provide a fuller picture or alternative perspectives.

The Portal employs three custom-designed interactive navigation tools (map viewer, timeline, and keyword explorer) that invite users to explore content of interest and to visualize chronological, spatial, and conceptual connections. While the Portal offers full-text indexing and searching as well, these three interconnected navigation tools represent its unique functionality. Similar tools exist elsewhere, of course. But rather than simply apply existing tools, we have reviewed our assumptions about categorization and representation and have either modified existing taxonomies or created our own controlled vocabulary; we have reviewed multiple benchmarks, considered the purpose of our representations, and have created our tools' designs from scratch. While we have done user testing and can discuss our Google Analytics data, our main objective is to reflect on how data selection, categorization, and representation create constraints and opportunities for inquiry. We offer an examination of our decisions regarding the creation of these three navigation tools, and reflect critically, six months after our launch, on their performance.

1. Map viewer

Navigation via the Portal's map viewer is possible by zooming and browsing on a map and/or searching for placenames. After considering several alternatives, we chose to use GeoNames' highly hierarchical, but open-source, gazetteer (adopted also by other major European digital humanities portals such as Europeana). We took this decision to reflect as accurately as possible the complexities and vagaries of geographical and administrative dependencies, to potentially ease the retrieval of nested data (e.g., showing results related to a certain village among the results for its parent country), and to allow for a greater degree of interoperability with other websites. The most difficult and important representational consideration was our decision to represent only point data, not lines or areas. This was not only a pragmatic technical decision, but serves the purpose of our map viewer as a content index, not as a thematic mapping of a coherent data set.

2. Interactive timeline

The Portal's timeline feature allows users to scroll forward and backward in time, to zoom in and out, and to plot the results of up to three searches on an interactive timeline canvas. One of the most difficult decisions was limiting the temporal metadata to years as points in time, to the exclusion of periods, eras, and specific or 'fuzzy' dates. Representational considerations included how to distinguish between search results for the same year and the use of a graphic scale for time. Furthermore it was important to us to present the Timeline as a content index, not as a 'comprehensive' collection of events. The decision to represent the quantity of relevant Portal results in the timeline slider/scale bar as connected 'peaks' may be one of the Portal's most problematic representations.

3. Keyword explorer

While the map and timeline allow users to compare results geographically and temporally, the keyword search tool offers thematic navigation functionally resembling the former Google 'wonder wheel.' Specifically, our keyword explorer lets users refine a search by choosing among the tags that occur most frequently in relation to a given keyword, using a narrowing search and, in future, also a more exploratory, expansive approach in which keywords are linked without limiting the results to previously identified keywords. To facilitate this, we chose to adopt – and constantly develop – a flat controlled vocabulary instead of imposing a predetermined hierarchy of themes and keywords. This means that search experiences may change substantially as new materials are added to the Portal.

As new content is added, users' search possibilities will grow exponentially. That is, it will be possible to use any tool to refine searches made with any of the other tools. The aim is to allow the users to create completely personalized paths through environmental scholarship and materials gathered on the Portal.

As the ways we organize and represent information have important consequences for the communication and use of knowledge, an enhanced navigational platform provides opportunities to rethink our usual approaches. We hope these tools will offer multiple ways to find and compare information, inspire alternative historical frames, and encourage the formation of new connections across disciplinary and political boundaries.

By examining the complex ontologies behind navigation tools and the epistemological considerations concerning their application and use, we hope our experiences and critical reflections may offer new perspectives for our colleagues working with digital humanities data and databases.

Processing Email Archives in Special Collections

Hangal, Sudheendra
hangal@cs.stanford.edu
Stanford, USA

Chan, Peter
pchan3@stanford.edu
Stanford, USA

Lam, Monica S.
lam@cs.stanford.edu
Stanford, USA

Heer, Jeffrey
jheer@cs.stanford.edu
Stanford, USA

1. Introduction

Libraries and scholarly institutions often acquire the archives of well-known individuals whose work and life has significant historical or research value. Email collections are now an important part of these archives – the Academic Advisory Board Members in the Paradigm project ranked them the most valuable from among images, speeches, press releases, personal websites and weblogs, campaign materials, engagement diaries, presentations, etc. [3]. The detailed record embedded in email provides access to the donor's thoughts and actions at a level that has rarely been available in the past and enables researchers to probe questions like: What was the process the donor used to come up with a particular breakthrough? What were they reading at the time and how may it have influenced them? [4] Further, these archives are being accumulated not just by famous people; with about 2 billion users, email reaches just about every section of wired societies. Indeed, the British Library has collected sample email messages from ordinary Britons as a way of capturing a sense of life in the 21st century [5].

In this paper, we describe a new technique for processing email archives in special collections using MUSE (Memories USing Email), an email browsing and visualization system developed at Stanford University. The technical details of Muse are covered in a separate paper [1]. While Muse was initially designed for individuals to browse their own long-term email archives, we have added features that help archivists in processing email archives of others as well. To illustrate with a concrete example, we report our experiences with using Muse to process the email archives

of noted American poet Robert Creeley, whose archives are hosted at Stanford University Libraries. The video at http://mobisocial.stanford.edu/muse/creeley.mp4 demonstrates MUSE running on the Creeley archives and supplements the descriptions below. MUSE is publicly available at the URL: http://mobisocial.stanford.edu/muse.

2. Challenges in Processing Email

Today, email archives are being collected and preserved, but are rarely processed, let alone delivered to researchers and end-users. This is due to concerns about privacy and copyright as well as the relative difficulty of processing large-scale archives with conventional email tools. While paper records are scanned and processed manually by archivists, such a process is cumbersome for archives with tens of thousands of email messages. Hence the potential of email archives for research remains under-tapped and they are often listed as a single series or sub-series in a 'Finding Aid' in special collections, making it hard for researchers to make practical use of them. We elaborate on these challenges below.

Stakeholders

There are several stakeholders in the process of acquisition and use of email archives: the donor, the curator, the archivist who processes the collection, and the researcher who uses it. Each of these stakeholders has different requirements and expertise.

Donors are sometimes hesitant to turn over their email archives to curators as they may contain deeply personal information such as family or financial records, confidential letters of recommendation, health matters, etc. Donors are often busy people and may not have the time to perform a detailed assessment of their archives. Further, a donor may sometimes not be the creator, but say, a family member. Curators develop library collections and maintain relationships with donors.

Archivists are generally well versed in tools and archival processes, but may not be subject matter experts. While archivists want to provide broad access to the archives and encourage exploratory use, they also have to be cautious due to embargoes established by the donor, privacy considerations and copyrights restrictions.

Researchers may be familiar with the subject, but may not be experts with tools. Typically, they would like to gain a sense of the content in the email correspondence through the process of exploratory browsing. They may want to know if certain people or subjects are mentioned in the archives even before making a visit to the collection or raising funding for a project.

Data gathering and cleaning

It is common for digital archives to be acquired at different times, over several rounds of accession, and to be scattered across a variety of digital media including floppy disks, Zip drives, CDs, DVDs and hard drives. Email archives are no exception and we find that, over time, donors change computers, accounts, email clients etc, and store email in different formats (such as Eudora, Outlook and mbox). We have found tools like Emailchemy (http://www.emailchemy.com) useful to convert email in disparate formats to the mbox format that Muse can read. Individuals' email foldering practices tend to be inconsistent over time, and messages are frequently duplicated in various folders. Muse takes care of this problem by detecting and eliminating duplicates. Muse also organizes messages by automatically inferred (but manually editable) groupings of people in the archives, making the folder structure less critical. Further, email addresses and name spellings for the same person tend to change over time; therefore, Muse performs entity resolution to try and merge records for the same individual.

In the Creeley archives, there are about 80,000 emails; after removing duplicates, Muse is left with 40,038 messages. Of these, 14,770 are outgoing messages and 25,268 are incoming messages. Creeley corresponded with about 4,000 people in these archives.

Muse displays graphs of email communication activity, which show that most messages in these archives are from 1996 to 1998, and from 2001 to 2005 (when Creeley passed away), with a sudden dip at the beginning of 2002. This tells us that the archives are missing material from the years 1999 and 2000, and possibly for some period in early 2002. Such signals are useful to the archivist to know that some information may have been missed at some step in the archival process.

Capture and Authenticity

A major benefit of digital records is that they are easy to capture and store compared to paper records. Thus it is possible for a donor to retain access to his or her records for many years or decades, and for archivists to capture the archives of many more individuals and store them in a reasonable amount of space at reasonable cost.

Another benefit of email messages is that, while it is difficult to get access to letters sent by a donor, email copies frequently exist with both the sender and the receiver, leading to a more detailed record. Physical correspondence also has problems with completeness of information. For example, in the Republic of Letters project, many letters were not dated (http://republicofletters.stanford.edu/).

In contrast, email messages have an automatic timestamp. While both paper and digital formats can decay physically over time, it is easier to preserve a large volume of digital data.

The techniques to determine the authenticity of a paper document are well established. Paper or vellum can be appraised against a familiar set of physical characteristics, such as ink, handwriting, letterhead, paper quality and signs of tampering. However, there are new problems with electronic records. The Paradigm workbook cited above points out that the capture process itself can alter the perceived creation date, and that author metadata is often inaccurate or misleading. Further it notes that establishing intellectual property rights is a key concern for the digital curator who will need to determine who authored a photograph or article, whether they are still alive, whether they still hold copyright and how long that copyright will last.

3. Cues Provided by MUSE

We ran Muse on the Creeley archives, and found the following cues useful in gaining a quick overview of the archives.

1. Calendar view of terms. Muse displays a calendar view of the 30 most important terms per month based on statistical ranking, with a novel time-based TF-IDF metric. The terms scored are named entities extracted from the messages using the Stanford NLP toolkit (http://nlp.stanford.edu). We found this feature useful to give ourselves and potential researchers a high-level sense of the contents of the archives; at the same time, the small number of terms makes it easy for the archivist to manually ensure that they are appropriate for public distribution.

2. Sentiment Analysis. Muse uses sentiment analysis techniques to identify messages that may reflect certain categories of sentiments including emotions (such as love, grief, anger, etc), family events, vacations, congratulatory messages, etc. We have developed these categories and word lists for personal archives instead of relying on more general lexicons like LIWC [2]. See Fig. 1 for a graph of these sentiments over time in the Creeley archives. The Muse lexicon can be tuned by the user by adding or deleting words to a category, or entire categories themselves. One use we found of this feature was to add a category to identify potentially sensitive messages involving health, finances, recommendation letters, etc.

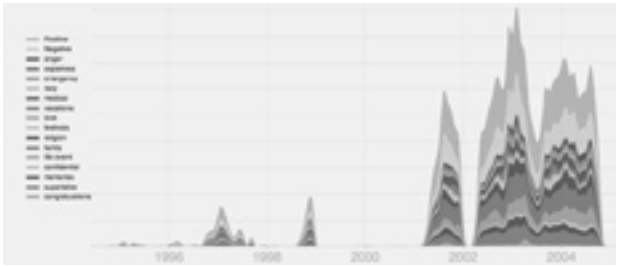


Figure 1: Sentiments over time in the Robert Creeley email archives

3. Attachment wall. To facilitate rapid scanning of picture attachments in email, Muse displays them on a 2.5D zoomable and draggable photo wall (Fig. 2). In the Creeley archives, there are 6,282 picture attachments, of which 4,769 are unique. The archives include many interesting pictures, for example, those of Creeley and his family, his home, trips he took, various forms of artwork, and scanned announcements of events. In general, pictures and documents have associated copyrights, so the archivist cannot publish these images and attachments publicly.

Browsing features

When the user is interested in following a cue, Muse launches a message view with all the messages related to that cue. These views can be fairly large and consist of hundreds of messages. To make it easy to rapidly skim a collection of messages, Muse provides a faceted browsing interface, where the facets are sentiments, groups, people, original folders, email direction (incoming vs. outgoing), and month or year. It also provides a jog dial interface that lets users rapidly flip through messages without the need for keypresses and mouse clicks. The jog dial is very popular with users of Muse.

Figure 2: Image attachments in the Robert Creeley email archives

Multiple views

While we initially thought that Muse would be used primarily by archivists, we realized that it can be useful to donors and researchers as well. To support these stakeholders, two distinct views of the archives are needed. The first is a full-

access view for donors and archivists to use when processing the archives. The same interface can also be made available to researchers in a reading room environment. The second, more limited interface can be made public and can provide enough detail for potential researchers to get an overall sense of the archives' contents. It can include a calendar view of important terms, and perhaps the overall patterns of communication with different groups and sentiment. However, the actual message contents are omitted. Since Muse stores message headers, bodies, and attachments separately for each folder in its own cache, we found it easy to support both views for the Creeley corpus; in the public view, we simply hide the message bodies and attachments.

Message selection and export

We envision that Muse can be used by donors themselves to screen their archives before turning them over to the library, with the help of features like automatic grouping of email and sentiment analysis. We have added a feature in Muse to allow users to tag messages and export all messages with a particular tag (thereby including only the selected messages), or without a particular tag (to redact certain messages). These features can also be used by an archivist to screen the archive for sensitive material.

4. Connecting the Archives to Web Browsing

While scanning the Creeley archives through Muse, we realized that it would be useful to look in it for terms for which Robert Creeley is most famous. This is a difference from the original purpose of Muse; the archivist or researcher is not expected to be intimately familiar with the life of the donor. For example, according to his Wikipedia page, Creeley is known as a Black Mountain poet; searching for this term in his archives returns 259 messages. We therefore implemented a browser plug-in that searches for named entities on the page being browsed and highlights those that are also present in the archives. Clicking on the highlighted text lets the user explore email messages that include the term. This lets a researcher bring the archives' lens into his normal browsing.

We hypothesize that researchers can find this feature useful to browse their own research. The browsing lens will automatically find terms in the archives that are relevant to the researcher's interest and highlight them.

5. Conclusion

Our overall experience of processing email archives using Muse was quite positive. Muse can help archivists by letting them spot missing or unclean data, performing quick scans of the contents for material that needs to be restricted, and make parts of the archives publicly available for researchers' use. Researchers benefit by gaining an overall sense of the material in the archives; when they need to drill down into the actual contents, an interactive browsing and navigation interface aids them explore the archives efficiently, and a browser plug-in lets them bring a lens from the archives into their normal browsing.

Using Muse, archivists can hope to process email archives quickly and make valuable information available for researchers. Further, we believe that making Muse even simpler to use will enable ordinary individuals to browse their own long-term email archives, or those of people close to them such as family members. This will allow the study of personal archives on a scale that has not been possible until now.

Acknowledgements

We thank Glynn Edwards, the Andew W. Mellon Foundation, NSF POMI 2020 Expedition Grant 0832820 and the Stanford Mobisocial lab for supporting this work.

References

1. **Hangal, S., M. S. Lam, and J. Heer** (2011). Muse: Reviving memories using email archives. *Proceedings of UIST 2011*. ACM, 2011.

2. **LIWC Inc**. *Linguistic Inquiry and Word Count*. http://www.liwc.net.

3. **Thomas, S.** (2005). *Paradigm Academic Advisory Board Report*. John Rylands University Library, Manchester, Dec. 12, 2005.

4. **Wright, M.** (2007). *Why the British Library archived 40,000 emails from poet Wendy Cope*. Wired, May 10, 2011.

5. **Zjawinski, S.** (2007). *British Library Puts Public's Emails on The Shelves*. Wired, May 29, 2007.

The Stylometry of Collaborative Translation

Heydel, Magda
magdalena.heydel@tww.pl
Jagiellonian University, Krakow, Poland

Rybicki, Jan
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland

1. The Problem

A translated work of literature is a collaborative effort even if performed by a single translator, always haunted by the ghost of the author of the original. The relationship between the two has been at the centre of mainstream translation studies and in the discipline's corpus-based and stylometric varieties, as evidenced by a growing body of scholarship (Olohan 2005; Oakes & Ji 2012). Stylometric problems multiply when the term 'collaborative translation' is taken to signify a joint rendering of a single author into a different language by two (or more) translators, or by translator and editor (Rybicki 2011). In general, stylometry based on multivariate analyses of word frequencies successfully detects the author of the original – rather than the translator – in translations (Rybicki 2010); sometimes, this success varies from translator to translator (Burrows 2002a). It is only when translations of the same author are compared is there any hope for stylometric machine-learning methods to tell translator from translator (Rybicki 2012).

This is exactly why this study focuses on a problem situated somewhat in the middle of the above, as collaborations between translators on a *single* literary work are a fact in the publishing industry. In the Polish market, this is perhaps most famously evidenced by Maria and Cezary Frąc, responsible for the third Polish translation of Tolkien's *The Lord of the Rings*. It is notoriously difficult to obtain information on the reasons and details of such translatorial collaborations from either the translators or their publishers; usually, looming deadlines for lengthy popular novels are blamed (Kozieł 2011).

But a collaborative translation can also be made for other reasons. When one of Poland's most eminent translators, Anna Kołyszko died of cancer during her work on Virginia Woolf's *Night and Day*, leaving a finished draft of much of the book and some notes on additional sections, the translation was taken over by Magda Heydel, a particular specialist in Woolf (*Jacob's Room*, *Between the Acts*, *A Moment's Liberty, On Being Ill*), who also performed some editing on the entire text. As Heydel stated in a TV interview, it was for the readers to see whether there was or there was not a rift in the middle of the book where one translator took over from the other; she hoped her editing made the narration coherent as far as the style was concerned. She also emphasized the uniqueness of the translator's experience to confront her own intuition of her voice in the text with that of another (Heydel 2011). Thanks to her previous work and research on Woolf, she had had quite a definite idea of what stylistic shape *Night and Day* should obtain in its Polish translation. Her linguistic image of Woolf's style, being, as it to an extent must, rooted in her own idiosyncratic 'feeling' of the language, was also informed by tangible evidence in Woolf scholarship. The technique of the changing point of view, to be elaborated in Woolf's mature work into the stream of consciousness, a very important achievement of the writer, is clearly visible already in this early novel. Recreating this aspect of her writing has always been one of Heydel's central concerns. Also, her particular translation technique is to a large extent based on the idea of the voice of the speaker, with the actual reading aloud for the effect of naturalness as the ultimate test for a successful rendition. In Woolf the recognizable 'voice' of the focalizer is central as it produces the point of view in narration (Rait 2010). Thus the changes Heydel introduced into Anna Kołyszko's text were not (or very rarely) lexical but mainly syntactical. She worked with the famously long and intricate Woolfian sentences, the more so that the Polish language, with its extremely flexible sentence structure, locates most of its rhetorical and pragmatic devices here. Also for this reason, most-frequent-word analysis was a well-suited approach to this experiment in translatorial attribution.

Indeed, this seems a translatorial counterpart of David Hoover's study on *The Tutor's Story*, a novel begun by Charles Kingsley and completed by his daughter Mary under her pen name Louis Malet, with some information available on who wrote what (Hoover 2011). In the Polish *Night and Day* case, this information is exact; in both cases, the early chapters have been written by the first, the final ones by the second translator. It is also reminiscent of an earlier study on the Middle Dutch epic *Walewein* (van Dalen-Oskam & van Zundert 2007). The main difference consists in the fact that Heydel is available to confirm or deny the findings of the quantitative analysis.

2. The Method and the Corpus

This study applies Cluster Analysis to Delta-normalized word frequencies in texts; as shown by

(to name but a few) Burrows (2002) and Hoover (2004, 2004a), and despite limitations discussed by Smith & Aldridge (2011), this is one of the most precise methods of 'stylistic dactyloscopy.' A script by Maciej Eder, written for the R statistical environment, converts the electronic texts to produce complete most-frequent-word (MFW) frequency lists, calculates their z-scores in each text according to the Burrows Delta procedure (Burrows 2002); selects words for analysis from various frequency ranges; performs additional procedures for better accuracy (Hoover's culling and pronoun deletion); compares the results for individual texts; produces Cluster Analysis tree diagrams that show the distances between the texts; and, finally, combines the tree diagrams made for various parameters (number of words, degree of culling) in a bootstrap consensus tree (Dunn et al. 2005, quoted in Baayen 2008: 143-47). The script was demonstrated at Digital Humanities 2011 (Eder & Rybicki 2011) and its ever-evolving versions are available online (Eder & Rybicki 2011a). The consensus tree approach, based as it is on numerous iterations of attribution tests at varying parameters, has already shown itself as a viable alternative to single-iteration analyses (Eder & Rybicki 2011b; Rybicki 2011).

The Woolf translation was analysed by comparing equal-sized fragments (at various iterations of fragment size) of the translation of *Night and Day* to determine the chapter where Heydel had taken over from Kołyszko; Heydel was consulted only after the initial determination had been made. At this point, the Kołyszko and the Heydel portions of the book were compared to other translations by Heydel (Woolf's *Jacob's Room*, *A Moment's Liberty* and *Between the Acts*, Graham Swift's *The Light of Day* and Conrad's *Heart of Darkness*) and by Kołyszko (McCarthy's *Child of God*, Miller's *Tropic of Capricorn*, Roth's *Portnoy's Complaint*, Rushdie's *Midnight's Children*), and then to an even more extended corpus of author-related translations.

3. Results

All iterations of different fragment sizes of the Kołyszko/Heydel translation pointed to the beginning of Chapter 27 as the place where Heydel took over the translation from Kołyszko. Figure 1 shows the attribution of medium-sized fragments (approximating, in this case, mean chapter length) of the translation. In fact, Kołyszko has completed 25 full chapters and left scattered notes on Chapter 26; these have been collected, organized and edited by Heydel. If we are to believe stylometric evidence, the latter made a very good job of preserving the style of the former, so that her own style is only visible in earnest in Chapter 27.

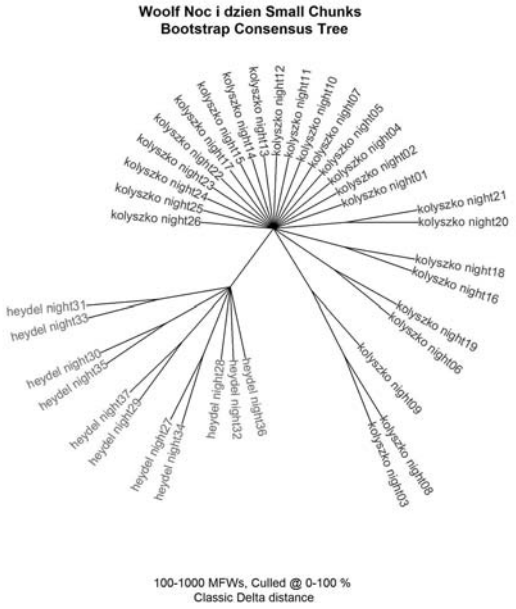


Figure 1: Consensus tree for equal-sized fragments of Kołyszko and Heydel's translation of Woolf's Night and Day, performed for 0-1000 MFWs with culling at 0-100%. The beginning of fragment 27 of the translation roughly coincided with the beginning of Chapter 27

Once the authorship of the translation of *Night and Day* became confirmed, it was possible to place this shared work in the context of other translations by Heydel and Kołyszko; unfortunately, they never translated different books by the same author. The consensus tree in Figure 2 is divided neatly between Kołyszko and Heydel. In this context, the overall editing by Heydel might be visible in that both parts of *Night and Day* remain in her section of the tree; this might equally be a result of the original authorship, as the Woolf novels are close neighbours here.

When more translations by the same authors but by other translators are added to the corpus, the balance between translatorial and authorial attribution is clearly shifted towards the latter. In Figure 3, five translations of Woolf (including *Night and Day*) by four different translators occupy the lower branches of the consensus tree, with Bieroń's *Orlando* being the only (relative) outsider. Kołyszko's and Heydel's translations in the upper half of the graph cluster with translations of respective authors by other translators.

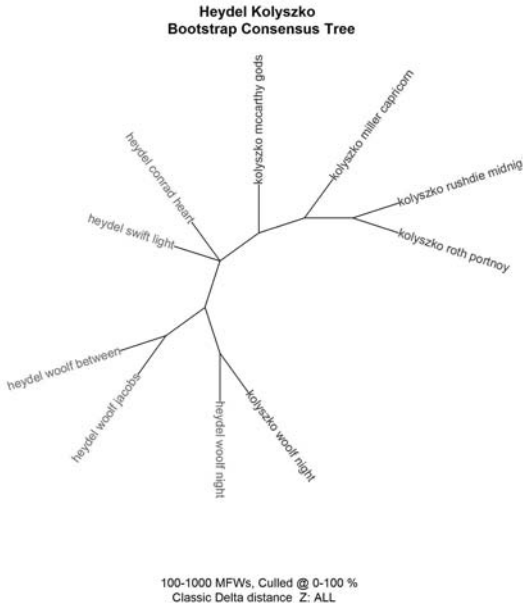


Figure 2: Consensus tree for translations by Kołyszko and Heydel, performed for 0-1000 MFWs with culling at 0-100%

4. Conclusions

So far, attempts at finding stylistic traces of the translator or the editor have been only partially successful. Word-frequency-based stylometric methods have shown that they are better at attributing the author of the original than the translator (Rybicki 2009, 2010; Eder 2010) – as has already been stated, unless translations of a single author are compared. In the latter case, Burrowsian stylometry is quite capable of telling translator from translator.

The translatorial attribution is greatly helped by the adoption of the bootstrap consensus tree approach, which minimizes attributive errors due to unlucky combinations of parameters as, simply speaking, Delta and similar distance measures are more often right than wrong, but the proportion between right and wrong might vary for a variety of reasons – especially language (Eder & Rybicki 2011b). This is particularly significant in a rare translatorial attribution case such as the Kołyszko/ Heydel translation of *Night and Day*, where the results of stylometric analysis can be confirmed or denied by the translator herself. Equally importantly, this study demonstrates that although stylometry seems to find traces of the author as well as of the translator, these traces can be disambiguated by placing the disputed translations in contexts of various corpora.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge UP.

Burrows, J. F. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17: 267-87.

Burrows, J. F. (2002a). The Englishing of Juvenal: Computational Stylistics and Translated Texts. *Style* 36: 677-99.

Dalen-Oskam, K. van, Zundert, J. van (2007). Delta for Middle Dutch – Author and Copyist Distinction in Walewein. *Literary and Linguistic Computing* 22: 345-62.

Dunn, M., A. Terrill, G. Reesink, R. A. Foley, and S. C. Levinson (2005). Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science* 309: 2072-075.

Eder, M., and J. Rybicki (2011). Stylometry with R. Poster. Stanford: Digital Humanities 2011.

Eder, M., and J. Rybicki (2011a). *Computational Stylistics* `http://https://sites.google.com/site/computationalstylistics/`.

Eder, M., and J. Rybicki (2011b). Do Birds of a Feather Really Flock Together, or How to Choose Test Samples for Authorship Attribution. Stanford: Digital Humanities 2011.

Heydel, M. (2011). Interview in 'Czytelnia.' *TVP Kultura*, 12 Feb.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing* 19: 453-75.

Hoover, D. L. (2004a). Delta Prime? *Literary and Linguistic Computing* 19: 477-95.

Kozieł, M. (2011). *The Translator's Stylistic Traces. A Quantitative Analysis of Polish Translations of The Lord of the Rings*. MA thesis, Uniwersytet Pedagogiczny, Kraków.

Oakes, M., and M. Ji (2012). *Quantitative Methods in Corpus-Based Translation Studies*, Amsterdam: Benjamins.

Olohan, M. (2004). *Introducing corpora in translation studies*. London: Routledge.

Rait, S. (2010). Virginia Woolf's early novels: Finding a voice. In Seller, S. (ed.), *The Cambridge Companion to Virginia Woolf*, Cambridge: Cambridge UP.

Rybicki, J. (2010). Translation and Delta Revisited: When We Read Translations, Is It the Author or

the Translator that We Really Read? London: Digital Humanities 2010.

Rybicki, J. (2011). Alma Cardell Curtin and Jeremiah Curtin: The Translator's Wife's Stylistic Fingerprint. Stanford: Digital Humanities 2011.

Rybicki, J. (2012). The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation. In M. Oakes and M. Ji (eds), *Quantitative Methods in Corpus-Based Translation Studies*, Amsterdam: Benjamins.

Smith, P., and W. Aldridge (2011). Improving Authorship Attribution: Optimizing Burrows's Delta Method. *Journal of Quantitative Linguistics* 18(1): 63-88.

Focus on Users in the Open Development of the National Digital Library of Finland

Hirvonen, Ville
ville.z.hirvonen@helsinki.fi
The National Library of Finland, Finland

Kautonen, Heli Johanna
heli.kautonen@helsinki.fi
The National Library of Finland, Finland

1. The Public Interface of the National Digital Library of Finland

The tools for discovering cultural heritage material are changing. For more than a decade there has been a demand for better services for acquiring and working with the ever increasing amount of digitized collections of archives, libraries, and museums. Almost quarterly we hear about new electronic search and research services, which promise better assistance to academic workers. These tools are brought out by software vendors, academic research and development projects, or publishers. Looking at these news we can estimate that the research infrastructure for scholars in the humanities, currently concentrating on resource databases, will be replaced by integrated, modular services, which utilize the potential of social metadata.

As a response to the increasing demand for novel research infrastructure services, we in Finland have started building our National Digital Library (NDL). The NDL project covers the digitization of Finnish archives, libraries, and museums; the long term preservation system; and a next generation search application called the Public Interface, which harvests the metadata from various data repositories and provides it for the end users. The NDL will also serve as the national aggregator to the European Digital Library, Europeana. The development project is funded by the Finnish Ministry of Education and Culture until 2014, after which there should be permanent funding for maintenance and support. The project started in 2008 and will reach the publication of the first end-user interfaces displaying the first set of material by the end of the year 2012. (Further information: www.kdk.fi/en/public-interface.)

The Public Interface facilitates access to the diverse digital repositories of Finnish libraries, archives and museums. The users can use it to find physical items, digital objects, and licensed electronic materials. The metadata for these are harvested from the organizations' back-end systems via an OAI-PMH interface. Access to restricted and licensed materials will be provided by integrating user authentication from the corresponding back-end system.

Within the service there will be one common user interface for all free contents within the system, called the National View. It will serve as the basis for those organizations who will not need specific system integrations and modifications. Organizations who wish to integrate external services, such as an online payment system or an ontology, to the core service can set their own interface instances and thus create so called Local Views to the NDL. All will still have the entire set of data available for their customers, via whichever user interface, if desired.

2. A Technically Challenging Project

From the technical point of view, the scope of the project is very ambitious. So far there has been a lot of hard work choosing and testing the satisfactory software for the Public Interface and defining various system interfaces. Potentially all Finnish archives, libraries, and museums are entitled to join the project and get their services within the NDL. While some of them will follow given standards and some have similar back-end systems, there will be those who will require completely individual solutions for systems' integration.

Scalability and additional service integrations also require thorough planning and good management. The amount of material in a country like Finland is not enormous and the number of possible service integrations can be controlled. However, the number of organizations and user interfaces is easily in the hundreds. If not compensated, these can have some effect on server loads and delays, which in turn will have impact on the end-user experience.

The greatest technical challenge so far has been the choice of the core software for constructing the Public Interface. When the available software solutions were initially compared, the information discovery and delivery system Primo by the Ex Libris Group was considered as the most suitable for the needs of the NDL. During the pilot testing, which took almost sixteen months, it became apparent that the customizations needed in Primo would require much more time and resources than was reserved for the project. Although the software procurement contract between the National Library of Finland and Ex Libris was cancelled in January 2012, the partnership between these organizations continues. The National Library still examines possibilities to provide access from the NDL Public Interface to the Primo Central Index.

In order to stay in the schedule of the project, a new solution was needed quickly. After a speedy market analysis, the National Library decided to turn to use an open source software solution. Having some experience of open source development, and avoiding a tedious start from the scratch, the National Library studied some alternatives to be used as the basis for the software development. The library resource portal VuFind was selected to serve as the starting point.

3. Focus on Usability

From the very beginning of the NDL project the importance of end-users and their expectations has been seriously regarded. As a means to realize the aim of good usability, a Usability Working Group and a reasonable budget has been assigned to meeting the goals and coordinating the needed tasks. The group has established the usability requirements and defined target users for the Public Interface service, and gradually completed a Usability Plan for the project and its outcome. Applicable design approaches (e.g. Meroni & Sangiori 2011) and best practices (*e.g.* Clark et al. 2010) have been consulted all along the process. The plan covers all project phases from planning to maintenance, but is still flexible enough to adjust to changes in technical solutions or the scope of the project.

Years 2010 and 2011 were mostly spent on testing the Primo software, but also on testing the service concept among prospect users. Eight organizations among Finnish archives, libraries, and museums volunteered, and started integrating their back-end systems to Primo. The concept testing was conducted by the University of Tampere in summer 2010 (Lavikainen 2010). Some months later a user study was conducted by an usability consultant company Adage on the first end-user interface layouts of the National View (Hirvonen 2011). The results of both studies indicated that some features of the future service might be difficult for end-users to comprehend, and therefore the final design of the end-user interfaces will be crucially important. However, the results gave no reason to change the scope of the service or the system requirements. These studies, on their part, supported the demand to chance the software solution.

4. Several Local Views, One National View

Although the core software had to be changed after a reasonably long period of work, the efforts put on testing one solution could be immediately utilized with the other option. Especially the concept of different views was considered applicable on VuFind, too.

During the year 2011 some Local Views were implemented with Primo and exposed to technical and functional tests. The first in line was the Turku City Library, which integrated Primo to their web content management and publication system. In their solution the main objective was to merge the Digital Library into a dynamic and locally oriented customer service. On the second track there was the National Board of Antiquities of Finland, which has invested considerable resources on establishing common technological solutions for Finnish museums. They implemented some varying user interface instances that displayed the museum repositories and utilized the features offered by Primo. Their focus was on experimentation of data visualization and service integration, since they intend to offer one Local View for all museums. The third track of the Local View testing concentrated on a portal-like application that leaves the external services to be run on the back-end system. The National Archives of Finland worked on mapping and normalizing aggregation levels of archival data. They also aim at producing solutions and standards for all archives in Finland. Such applications of Local Views will later provide excellent field for usage studies, since they reveal three entirely different viewpoints to the same service.

The National View, which will serve as the user interface for compiling all free material within the NDL, could not be thoroughly implemented and tested with Primo. Therefore there are now great expectations to the implementation of those features that support effective discovery and display of different types of content, i.e. content from archives, libraries, and museums at the same time. (See Fig. 1.) According to the recently updated Usability Plan the National View will be exposed to several usage tests. Before the summer 2013 there will be a comprehensive log analysis, some customer surveys, and other usability studies, which aim at further analyzing the service concept and giving input to the user interface design.

Along with the new software solution, the software development methods, the collaboration practices, and the user testing processes have been renewed. The usability tests planned for the implementation and production phases will still be valid and executed in time. In addition, a constant row of end-users will be involved in the agile development process to test freshly developed features immediately after they have been published.

Figure 1: The layout of the National View exposed to end-user tests in March 2012

5. More Challenges Ahead

With the exception of certain back-end a *pplication programming interfaces (APIs)*, the NDL Public Interface service will be mainly developed by the staff of the National Library. It will be crucially important to share knowledge and results with the open source community involved in VuFind. The national open source development community will also be communicated with. This sets extra demands on project management and collaboration management.

In future years the Public Interface of the NDL will replace several research services currently used by humanities' scholars. Within the National Library of Finland alone the user interface of the national bibliographic database, the portal of electronic publications in Finnish libraries, and the interface of digital material repository will be 'powered by' the Public Interface of NDL. It is difficult yet to foresee how the actual research activities will change due to the new tools. At least scholars themselves should be more involved in systems' and services' development than today (see *e.g.* Jeffreys 2010).

Nevertheless, the Usability Working Group has already made preliminary plans for examining the scholarly use of the NDL service.

As soon as a sufficient number of organizations have gone live, and there is a critical mass of coherent data within the service, we can start proper tracking and evaluation of the system as a research infrastructure. It is to be noticed, that the service is not just a library in a digital environment, since it can offer

and recommend content from archives and museums on single search request. It will be particularly interesting to study how students and scholars will be able to utilize such diverse and multifaceted information, and how that information can best be represented and laid out to the user. Therefore, it is yet to be seen, how well the National Digital Library of Finland will meet the great expectations.

References

Clark, D. J., D. Nicholas, and I. Rowlands (2011). *Europeana Log Analysis Report 1*. Europeana Connect. http://ciber-research.eu/download/20110821-M3.1.2_eConnect_LogAnalysis.pdf (accessed 14 March 2012).

Hirvonen, V. (2011). *Kansallisen näkymän käytettävyystesti*. Project report, The National Library of Finland.. http://kdk.fi/wiki/images/4/45/KDK-k%C3%A4ytett%C3%A4vyystesti_2011_01.pdf (accessed 14 March 2012).

Jeffreys, P. (2010). The Developing Conception of e-Research. In W. H. Dutton and P. W. Jeffreys (eds.), *World Wide Research. Reshaping the Sciences and Humanities*. Cambridge, Mass.: MIT Press, pp. 51-66.

Lavikainen, V. (2012). *Kansallisen digitaalisen kirjaston palvelukonseptin testaaminen*. Project report, University of Tampere. http://kdk.fi/wiki/images/9/96/KDKn_palvelukonseptin_testaaminen.pdf (accessed 14 March 2012).

Meroni, A., and D. Sangiori (2011). *Design for Services*. Surrey, UK: Gower.

The Rarer They Are, the More There Are, the Less They Matter

Hoover, David
david.hoover@nyu.edu
New York University, USA

In computational stylistics and authorship attribution recent work has usually concentrated on very frequent words, with a seemingly compelling logic: consistent and characteristic authorial habits seem most likely to exist among frequent function words not closely tied to plot, setting, and characterization, and not likely to be consciously manipulated. Analyses using frequent words have been very successful, and continue to be used (Craig & Kinney 2009). Some researchers have recently begun using many more words in analyses, often achieving excellent results by analyzing the 500-4,000 most frequent words (Hoover 2007). Yet even these words are not truly rare in a 200,000 word novel, where the 4,000th most frequent word typically has a frequency of 3-4.

Brian Vickers has recently attacked work based on common words, claiming that distinctive and unusual word Ngrams are better authorship markers than words for 16th century plays (Vickers 2008, 2009, 2011). He argues that rare Ngrams that are not found in a large reference corpus, but *are* found in one author's corpus and in an anonymous text, strongly suggest common authorship. Vickers tests whether Thomas Kyd might have written the anonymous *Arden of Faversham* by searching for 3- to 6-grams shared by *Arden* and the three plays known to be by Kyd, but not found in a 64-play reference corpus of texts of similar genre and date. He finds more than 70. This method, an improved method of parallel hunting, is essentially an extreme version of Iota (Burrows 2007), but based on Ngrams rather than words. As he puts it, 'One or two such close parallels might be dismissed as imitation, or plagiarism, but with over 70 identical word sequences, often quite banal phrases . . . that explanation falls away' (Vickers 2008: 14). In a later article he reports 95 such matches between *Arden* and Kyd's corpus (2009: 43). Does the existence of such rare matches constitute reliable evidence that Kyd wrote *Arden?* We can begin to answer this question by testing the method on texts of known authorship.

The texts Vickers works on are not ideal for testing: 16th century texts have substantial spelling

variation, making automated methods problematic, the authorship and dates of individual plays in the period are often uncertain, and many of them involve collaboration. I instead test the method on American fiction ca 1900 and modern American poetry. Because Ngrams have been used increasingly in authorship study (Clement & Sharp 2003; Grieve 2007; Juola 2008), it seems useful to test frequent word Ngrams before looking at rare ones. My American fiction corpus consists of a bit more than 2 million words of third person narration (10,000-39,000 words each) extracted from 83 texts by 41 authors, 20 of the authors represented by one or more additional texts, 36 test texts in all. A cluster analysis of all 83 texts correctly groups all texts by 16 of the 20 authors; in three cases, one of an author's texts fails to cluster with the others, and in the other case, the six texts by an author form two separate clusters. Delta correctly attributes 35 of the 36 possible texts in some analyses (on Delta, see Burrows 2002). Bigrams are less effective: a cluster analysis correctly groups only 14 of the 20 authors, and Delta correctly attributes a maximum of 32 of the 36 test texts. Trigrams are even less effective: cluster analysis is fairly chaotic, and Delta correctly attributes a maximum of 31 of the possible 36, with much weaker results overall. The same is true for American poetry, 2.7 million words of poetry by 29 authors, arranged in 23 primary composite samples (19,000-138,000 words), tested against 39 individual long poems (900-1,500 words). Word bigrams are much less effective than words and trigrams do a poor job. While word Ngrams have sometimes given better results than words alone, the results here do not support Vickers's basic premise that Ngrams are inherently superior to words in characterizing an author's style, even though they add syntactic information.

Rare Ngrams are a different matter, of course. In order to duplicate the scenario of Vickers's tests on Kyd, I used 64 of the sections of third person narration above (by 23 authors) as a reference corpus. I collected a three-text corpus for James and eight other James texts as test texts. I also held out six other texts for further testing. There are about 6.5 million 3- to 6-grams in the entire corpus; because the method requires at least one match between texts, I first removed all hapax Ngrams and those found in only one text, leaving about 100,000. I then removed all those found in the reference corpus, leaving about 9,600. I took each of the eight James test texts in turn, and compared it with the James corpus, with the seven other James texts, and with the six texts by other authors, with respect to the proportion of Ngrams in each text that match those in the James corpus or the other texts. The results are shown in Figure 1.

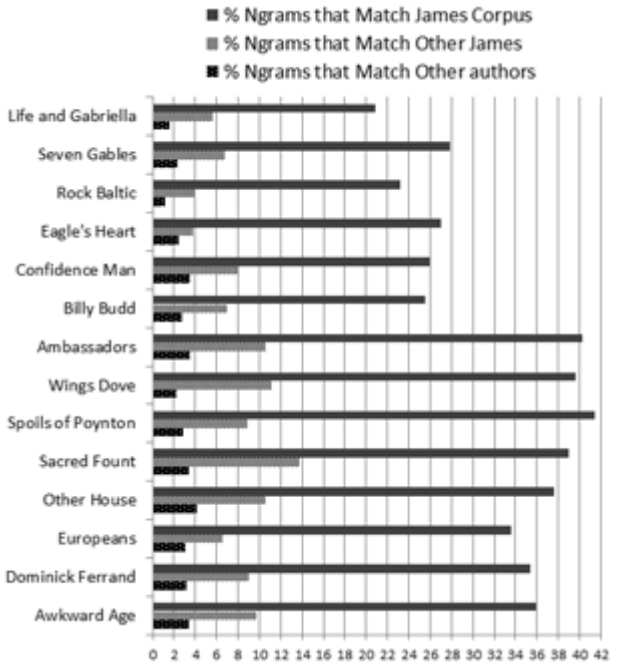


Figure 1: Rare Ngrams in James's Fiction

It is clear that the Ngrams in the James test texts have a higher proportion of matches in the James corpus than do the texts by others, ranging from about 33% to 42%, while the other texts range from about 21% to 28%. But the difference between the James texts and the others is disturbingly small. Also disturbing is that two of the six other authors have more matches to other James texts than one of James's own texts, and that all of the texts have about the same percentage of matches with texts by other authors. These facts are disturbing because Vickers bases authorship claims on the existence of ngrams found in an authorial corpus and an anonymous text that are not found elsewhere without doing the further tests above. Thus, if one were seriously testing The House of the Seven Gables for the possibility that Henry James wrote it, the fact that 28% (291) of the 1045 Ngrams found in Seven Gables have matches in the James Corpus but are found nowhere else in the reference corpus might seem like a strong argument. Remember that Vickers based an argument for Kyd's authorship on only about 70 matches between the Kyd corpus and his three anonymous plays that were not found in the reference corpus (or 95 in his later article). (Why there are so many more here needs further investigation, but the somewhat greater average length of my samples is almost certainly a factor.) For the poetry corpus, things are even worse: Kees shows a higher proportion of matches to Rukeyser's corpus than two sections of her own poetry, and Stafford is not far behind, as Fig. 2 shows.

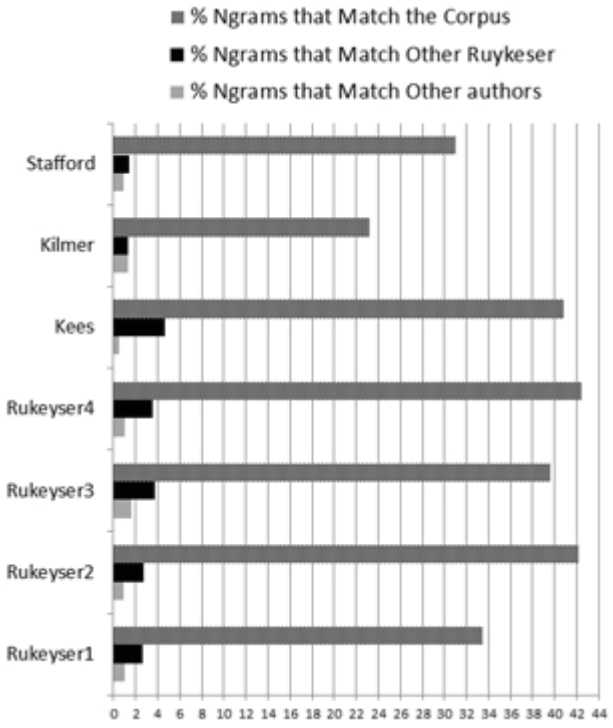


Figure 2: Rare Ngrams in Rukeyser's Poetry

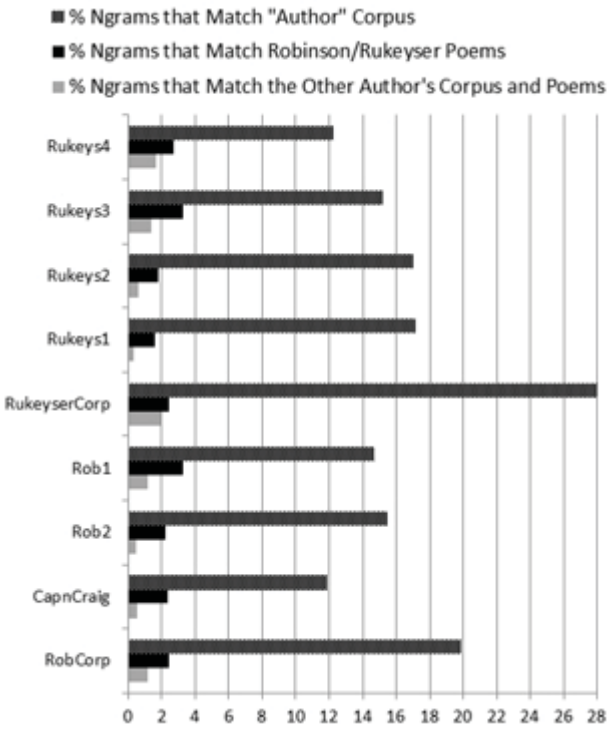


Figure 3: Rare Ngrams in a 'Composite' Author

Finally, I tested a bogus composite author corpus consisting of the samples by Stafford, Kilmer, and Kees. As Fig. 3 shows, even with an unreal author, there are many Ngram matches between the bogus authorial corpus and each other text that are not found in the reference corpus. For many of the samples, 15% or more of their Ngrams match the bogus corpus but are not found anywhere else. If these texts were anonymous, it would be easy to repeat Vickers's argument and claim that all of them must be by this illusory 'author.'

Finally, preliminary tests in which additional texts are added to the analysis show that many of the matches between an authorial corpus and any given test text that are not found in the original reference corpus also have matches in one or more of the new texts. The results presented above suggest that rare Ngrams are unlikely to be good indicators of authorship, in spite of their initial plausibility. There are about 1.3 million trigrams in the 83 narrative texts discussed above, but only about 150,000 of them occur more than once, and the highest frequency is 853. In contrast, there are about 49,000 different words, about 30,000 of which occur more than once, and the highest frequency is 136,000. One might say simply, and somewhat paradoxically, that, although Vickers's matching Ngrams are rare, there are so many of them that matches can be expected between almost any pair of texts, and some of these matches will also be absent from the reference corpus.

References

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *LLC* 22: 27-47.

Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *LLC* 17: 267-87.

Clement, R., and D. Sharp (2003). Ngram and Bayesian Classification of Documents. *LLC* 18: 423-47.

Craig, H., and A. Kinney, eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge U. Press.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *LLC* 22: 251-70.

Juola, P. (2008). Authorship Attribution. *Foundations and Trends in Information Retrieval* 1: 233-334.

Hoover, D. L. (2007). Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style* 41: 174-203.

Vickers, B. (2011). Shakespeare and Authorship Studies in the Twenty-First Century. *Shakespeare Quarterly* 62: 106-42.

Vickers, B. (2009). The Marriage of Philology and Informatics. *British Academy Review* 14: 41-4.

Vickers, B. (2008). Thomas Kyd, Secret Sharer. *Times Literary Supplement,* 18 April 2008: 13–5.

Experiments in Digital Philosophy – Putting new paradigms to the test in the Agora project

Hrachovec, Herbert
herbert.hrachovec@univie.ac.at
Faculty of Philosophy, University of Vienna, Austria

Carusi, Annamaria
annamaria.carusi@oerc.ox.ac.uk
Oxford e-Research Centre, University of Oxford, UK

Huentelmann, Raphael
info@ontosverlag.com
Ontos verlag, Heusenstamm nr. Frankfurt, Germany

Pichler, Alois
Alois.Pichler@fof.uib.no
Wittgenstein Archives at the University of Bergen, Norway

Antonio, Lamarra
antonio.lamarra@iliesi.cnr.it
Istituto per il Lessico Intellettuale Europeo e la Storia delle Idee (ILIESI), Consiglio Nazionale delle Ricerche, Italy

Cristina, Marras
cmarras@gmail.com
Istituto per il Lessico Intellettuale Europeo e la Storia delle Idee (ILIESI), Italy

Alessio, Piccioli
piccioli@netseven.it
Net7 srl, Internet Open Solutions, Pisa, Italy

Lou, Burnard
Lou.burnard@retired.ox.ac.uk
Oxford e-Research Centre, University of Oxford, UK

Agora extends an existing federation of semantically structured digital libraries named *Philosource*, a collection of high quality OA content from classical to contemporary European philosophy. Agora intends to transform this federation into a specialised and highly innovative OA archive and publishing venue for new scholarship, by carrying out five experiments: (1) Semantic linking; (2) Linked Open Data (LOD); (3) Advanced Scholarly Linking; (4) Open Peer review; and (5) OA business models in the field of European philosophy. The LOD portal created within experiment (2) will serve as unique access point to the philosophical content of the federation and will expose its metadata to the LOD

cloud. Agora's ultimate goal is to arrive at a model for establishing and accessing a growing open research archive for scholarly publications in the humanities.

The presentation will focus on the rationale for the Agora project's vision of innovative enhancement of Open Access resources in European Philosophy, and on the TEI framework. The experiments and evaluation strategies will not be presented in detail, and are briefly summarised below. More information can be found on the project website: `http://www.project-agora.eu`.

1. TEI Framework

The general vision that underlies the Agora project, as well as the mark-up challenges that arose during Agora's initial phase focused on content preparation for the experiments. It was decided to define a simple TEI-based interchange format which could be used both for existing and for newly published scientific journal articles, enabling different partners to combine these secondary resources seamlessly with other resources within the evolving AGORA framework. To define this format we surveyed about half a dozen different TEI-based publishing systems, as used by different journal publishers in Europe and North America, and identified a common subset of useful markup constructs. We documented this using the TEI's 'ODD' system, which is also used to generate the schemas used to validate the content produced by all partners. For output to readable PDF and HTML formats we rely on a customization of the TEI Stylesheet library. In a parallel experiment, we have used the same library to define an input mapping from 'docx' format into TEI, enabling partners to draft new documents in the familiar Microsoft Word environment using a specially customized Word stylesheet.

2. Experiments

Each of the Agora experiments is a trial of a mode of enhancing an aspect of philosophical scholarship through the possibilities offered by digital publications and repositories in philosophy.

The Semantic Linking Experiment has as the goal of enabling a novel way of building, querying, and browsing a knowledge network and to assess its suitability as a collaborative research tool as well as a learning device. A subset of the new content (secondary sources in the form of scholarly articles, monographs) will be interlinked via semantic annotations among each other and its underlying datasets (primary sources in the form of editions of texts, manuscripts) most of which are already available in the Philosource Federation. A subset of secondary sources will be connected with keywords that at the same time will make up domain concepts of the secondary sources ontology. The resulting ontology, in the form of RDF files, will be published on the web and the users will be offered the possibility to browse it, use the ontology to tag additional material, and add to the ontology generating their own notebooks. The semantically linked content will enhance browsing, querying, collaboration and learning.

The Linked Open Data (LOD) Experiment aims to offer users and machines a unique access point to the federation content and to augment the user search and browsing experience. The Philosource Federation is already a federation in the political sense by sharing common values and technological standards. The Federation Portal will mainly serve as a centralised aggregator of all the metadata (semantically structured and non-semantically structured) produced by the federation nodes. It will thereby interlink the federation nodes, and link the federation content with relevant external sources. It is hoped that the linking will make the content more interoperable, easier to discover and retrieve, and provide a richer context for the federation content. The metadata made available through the portal API will make the federation content usable to build other third party applications and aims to stimulate innovative and creative re-uses of our data. The portal will offer a unique SPARQL endpoint to access RDF metadata of all the federation content to third party data consumers; semiautomatically enrich the data to link certain type of resources (most notably books, people, places and events) via rdfs:sameAs properties, to other LOD providers such as DBPedia, Freebase, The Library of Congress, The Open Library, Geonames, Europeana; offering an augmented search facility by expanding queries with data extracted from third party LOD providers. Enhanced search will be offered in an appealing and easy to use interface, for example, by allowing visually browsing search results on a map and on a timeline.

The Open Collaborative Peer Review Experiment has as its goal to experiment in the field of open peer review in order to enhance and determine standards for peer review in the humanities and social sciences in the digital age. In this experiment authors submit their papers to two new online journals (Nordic Wittgenstein Review and Lexicon Philosophicum) via Open Journal System (OJS). Papers will undergo a general eligibility and quality check by the editors and editors-in-chief. Papers will undergo double blind review during which authors and reviewers will remain anonymous. Papers that are accepted for publication will go through one month of additional open review or commentary, during which registered users will be

asked to comment on and to discuss the accepted papers. Editors and editors-in-chief will moderate discussions. Editors will proceed with suggesting revisions on the basis of the open review. Authors will be able to use the suggestions, comments and discussions from the open review / comment stage in finalizing their paper for publication. The online publication of the articles will include functionality for readers to evaluate, rate and recommend articles in a comment stage. The user will then be able to select from different criteria of evaluation, such as highest rated, most read, and most cited.

The goal of the PPP OA Business Models Experiment carried out by the philosophy publishing house Ontos is to experiment with different OA Business Models in order to gain insight into their effects on commercial business. A commercial publisher will republish a selection of monographs in digital format and make them available in Open Access, under a Creative Commons Attribution-Non Commercial-No Derivative license. It will study the effects on sales figures over time of re-publishing online in OA existing inprint closed access monographs. It will further study the effects of publishing the Nordic Wittgenstein Review under hybrid and delayed business models.

3. Evaluation Strategy

The overarching evaluation strategy that is used has been adapted from Value Sensitive Design (VSD), a design methodology geared towards the values that are involved in technology customization and design, which consists of ongoing evaluation on three levels: conceptual, empirical and technical. VSD has been informed by philosophy from the outset of its development, and is therefore closer to a philosophical ethos, and also makes room for a philosophical as well as user-based evaluation of the technologies as they develop. Even though this methodology has been primarily used for ethical values of technology design and implementation, we will be extending it to epistemic values as well. This is particularly appropriate for technologies for scholarship where the two forms of values are very closely connected, and will also contribute to the further development of VSD as a methodology.

References

Bartscherer, T., and R. Coover, eds. (2011). *Switching Codes. Thinking through New Technology in the Humanities and the Arts*. Chicago: U of Chicago P.

Bizer, C., T. Heath, and M. Hepp, eds. (2009). Special Issue on Linked Data. *International Journal on Semantic Web and Information System* 5(3).

D'Iorio, P. (2008). L'île des savoirs choisis. De HyperNietzsche à Scholarsource: pour une infrastructure de recherche sur le Web. *Recherches & Travaux* 72.

D'Iorio, P., and M. Barbera (2011). Scholarsource: A Digital Infrastructure for the Humanities. In T. Bartscherer and R. Coover (eds.), *Switching Codes. Thinking through New Technology in the Humanities and the Arts*. Chicago: U of Chicago P.

Galey, A. (2011). Reading the Book of Mozilla: Web Browsers and the Materiality of Digital Texts. In R. Crone and S.Towheed, *The History of Reading, Vol. 3: Methods, Strategies, Tactics*. New York: Palgrave Macmillan.

Luke, T. W., and J. Hunsinger, eds. (2009). *Putting Knowledge to Work and Letting Information Play*. Blacksburg: Center for Digital Discourse and Culture.

Mika, P., C. Bizer, M. C. Schaefel, and L. Rutledge (2010). Semantic Web Challenge 2009: User Interaction in Semantic Web Research. Special Issues of *Journal of Web Semantics* 8(4).

Pichler, A. 2010. Towards the new Bergen Electronic Edition. In N. Venturinha (ed.), *Wittgenstein After His Nachlass*. Basingstoke: Palgrave Macmillan.

Pichler, A., and S. Säätelä, eds. (2008). *Philosophy of the Information Society: Proceedings of the 30th International Ludwig Wittgenstein Symposium, Kirchberg am Wechsel, Austria 2007*. Frankfurt: Ontos.

Information Discovery in the Chinese Recorder Index

Hsiang, Jieh
jhsiang@ntu.edu.tw
National Taiwan University, Taiwan

Kong, Jung-Wei
dino.kong@gmail.com
National Taiwan University, Taiwan

Sung, Allan
allan.sung@gmail.com
National Taiwan University, Taiwan

1. The Chinese Recorder

Chinese Recorder (CR) is a (first bi-monthly, then monthly) journal published by the Protestant missionaries in China between 1867, a few years after the 1860 treaty that allowed missionaries to enter China, and 1941, when the U.S. became engaged in the Pacific theater of the Second World War. Except for the nineteen-month interruption from June 1872 to the end of 1873, CR was the longest running English missionary journal in China. The period that CR covered was a tumultuous time in Chinese history, when the country went through the Taiping Rebellion, the Boxer Rebellion, the various wars with the foreign powers, the Republic Revolution of 1911, the civil wars among the war lords, the rise of communism, the invasion of Japan, and the great cultural and social transformation during the late 19th and early 20th century. Being based within China, CR provided a close look at all spectrum of the Chinese society, not only missionary affairs, but also issues such as Chinese civilization, healthcare, education, political situation, opium, and other social issues of the day. CR is unique in that the articles were written by missionaries in China for the benefit of their fellow missionaries. Being supported by missions and not sponsored by any government, Chinese or foreign, the views presented in CR were more candid and not blurred by political agenda. It, thus, provides an angle unlike any others. Although one cannot say that CR is not biased, at least it is biased in its own special way.

2. The Chinese Recorder Index

With its 73 volumes, one for each year, and over 50,000 pages in total, CR is difficult to use on its own. In 1986 Kathleen Lodwick published the 2-volume *The Chinese Recorder Index: a Guide to Christian Missions in Asia, 1867–1941* (CRI), which made it much easier for scholars to utilize CR in their research (Lodwick 1986).

CRI is more than an index. It consists of 3 indices and 6 special lists (of Persons by Affiliation, Persons by Location, List of Women, etc). The *Person Index* (*PI*) includes 8,391 names with 192,149 page records; the *Mission/Organization Index* (*MI*) 712 mission entries with 34,851 page records; and the *Subject Index* 4,691 entries with 83,636 page records. Altogether, there are 13,794 entries and 310,626 page records, averaging 22.5 page records per entry.

Unlike a conventional book index that only provides the pages in which an entry appears, CRI assigns tags to PI and MI, thus provides additional information about the nature of the occurrence of the entry on that page. Tables 1 and 2 give the names, nature, and numbers about the tags that we have tabulated.

There is a great wealth of information hidden in the tags. For example, if a certain page appears in a person name entry A, and if that page also appears in another person name entry B under an ART (article) tag, then we know that A appears in an article written by B. Using the same page number to check all entries, and we can find all person names, locations, missions, subjects, etc that appear in the same article. If the same page appears in a Subject Index entry indicating a certain event, then we may even 'guess' what the article is about without reading the article itself.

As another example, the ATT (attacks) tag tells how many attacks of missions and missionaries had been reported in CR, who were attacked, in what years did they occur and where. This information should be valuable to someone studying the attitude of the Chinese society towards Christianity when it was re-introduced to China in the late 19th century. However, since 'attack' is not an entry by itself, this information is scattered all over CRI. One has to pore through the 13,000 entries to collect every relevant bit of information.

Tag name of Person Index	Tag value	number/no value	explanation
AFF (Affiliation)	Name of mission	27,678 / 14	Affiliation
ARR (Arrival)	Year	8,705 / 6,283	Arrival year (if no value, then it's indicated by volume number)
ART (Articles)		11,125	Article written by person
ATT (Attacks)		1,164	Attack on the person
CHI (Children)		7,765	Children of the person
CON (Conferences)	Location, year	3,449 / 1,793	Conference attended
COR (Correspondence)		2,276	Correspondence to the editor
DAT (Dates in China)	Year	139 / 2	Dates in China
DEA (Death)	Precise year	2,459 / 2,035	Death (if no value, then it's indicated by volume number)
DEP (Departures)	Year of departure	5,496 / 5,000	Departure in the year(if no value, then it's indicated by volume number)
ITI (Itinerancy)	Location	445 / 413	Itinerancy
LOC (Location)	Location	39,744 / 8	Location
OTH (Other Publications)		10,095	Other publications
POS (Positions)	Title and mission	6658 / 2	Position of the person
SPO (Spouse)		13,212	Spouse of the person
UNS (Unspecified)		51,739	Unspecified

Table 1: Person Index

Tag name	Tag value	Number/no value	explanation
ATT (Attacks)		275	Attacks on the M/O
CON (Convert)		996	Converts
FIN (Finances)		1,585	Finances of the M/O
HIS (History)		650	History of the M/O
HOS (Hospitals)		639	Hospital
MEE (Meetings)		2,340	Meeting
LOC (Location)	Location	9,533 / 32	Location
OPR (Opium Refuges)		0	Opium refuge
ORD (Ordained Asians)		349	Asians ordained by mission
ORP (Orphanages)		32	Orphanages
PER (Personnel)		1,520	Personnel
PRE (Press)		1,554	Press
REP (Reports)		772	Reports by the M/O
SCH (Schools)		1,309	School run by the M/O
STA (Statistics)		2,510	Statistics about the M/O
UNS (Unspecified)		10,763	Unspecified

Table 2: Mission/Organization Index

3. Reindexing the CRI

To fully utilize the wealth of information embedded in CRI, we developed a system that incorporates the 3 indices of CRI into one uniform framework, under which the indices are fully integrated and cross referenced. This integration enables the user to explore the rich information hidden within CRI.

Our approach starts with a data structure that decomposes an index entry into a number of *page records*, each consists of the entry name (***n***), a volume number (***vol***) and page numbers (the start page and the end page, ***s*** and ***e***), a tag (***t***) and a tag value (***v***). (The tag part may not be present if it is not indicated.) Thus, a page record is a tuple (***n*** ; ***vol*** : ***s*** - ***e*** ; ***t*** : ***v***). If there is only one page indicated, then the start page and the end page will be the same. For instance, the first page record in the George Leslie Mackay entry (Figure 1) is (Mackay, George Leslie; 13:74-74; AFF: CPM) and the second page record is (Mackay, George Leslie; 13:312-312; AFF: CPM). (The George Leslie Mackay entry as shown in this example becomes 46 page records.) This process will decompose the 13,794 entries in

the three indices into 310,626 page records. We then designed algorithms to mine the relationships among the page records, mainly using the page numbers and the tags as reference points (Kung 2011). While the detail of the algorithms cannot be covered in this abstract, we will use two simple examples to demonstrate the outcome.

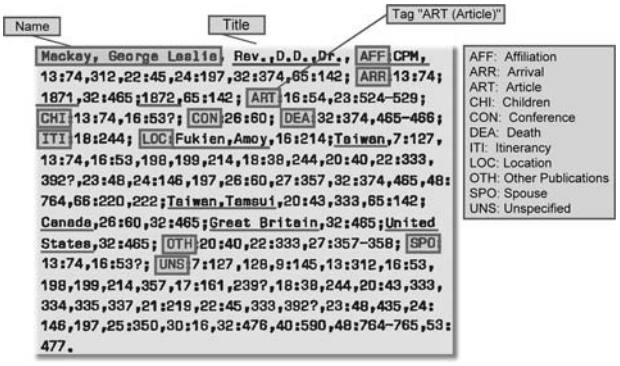


Figure 1: The 'George Mackay' entry in Person Index

One of the page records in the Person Index entry of George Mackay is (Mackay, George Leslie; 16:214-214; LOC: Fukien, Amoy). Since in the entry of Thomas Barclay, there is a page record (Barclay, Thomas; 16:214-215; ART), we know that Mackay was mentioned in an article written by Barclay. Other page records show that the same page appeared in the Subject Index of Sino-French War (1884–1885), the Person Index of James Maxwell, William Thow, etc. (Figure 2), plus others. Thus, when issuing George Mackay as a query, instead of simply returning Vol. 16, p. 214 as one of the pages in which Mackay appears, all information in CRI about that page will be organized and returned (Figure 3). (Indeed, it was an article written by Barclay about the returning of the missionaries from Amoy to Taiwan after the lifting of the blockade of Taiwan at the end of the Sino-French War.) Figure 4 shows the webpage resulting from the query 'George Mackay'. Note that a chronological distribution is presented and a foldable/expandable classification of the return is on the left.

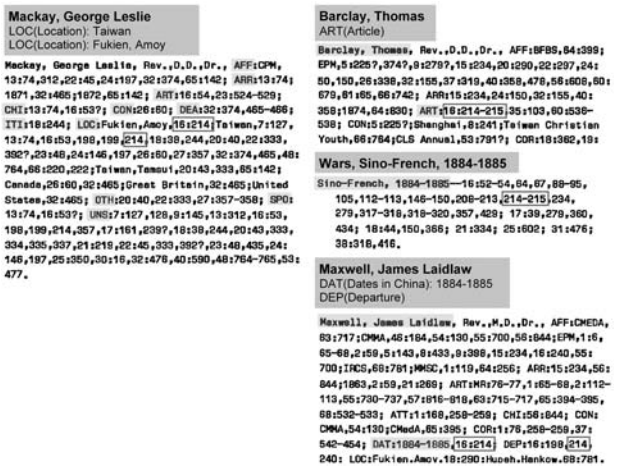


Figure 2: Some entries containing Vol. 16, p. 214

Figure 3: Information returned about Vol 16, p214 when given query 'George Mackay'

Figure 4: Query result of 'George Mackay'

Figure 5: summary of query 'attack'

Figure 5 shows the return of the query 'attack', for which a total of 574 different pages in the entire CR are retrieved. (There are all together 1,439 page records with the ATT tag. But some referred to the same pages.) The peak occurred in 1900, during the Boxer Rebellion, which may not be surprising. However, the post-classification of the query result (Figure 6), an important feature of our system, shows the lists of authors, persons, missions, which might provide interesting information for scholars to pursue further.

The method and system that we have developed provide a more global view of both CR and CRI. It integrates the three indices of CRI and thus reveals relations among them and information implicitly embedded. Our work should make CR and CRI more accessible to the research community. The approach that we have taken is also a general one that can be applied to any index with a similar structure.

Figure 6: Classifications of the query result of 'attack' according to Author, Person, Mission, and Subject

References

Kong, J.-W. (2011). *Design and Implementation of a Retrieval System for the Chinese Recorder Index.* MS Thesis, National Taiwan University.

Lodwick, K. (1986). *The Chinese Recorder Index: A Guide to Christian Missions in Asia, 1867-1941.* Scholarly Resources Inc. Washington D.C.

Complex Network Perspective on Graphic Form System of Hanzi

Hu, Jiajia
hjj81@126.com
Beihang University, China

Wang, Ning
niwangning@263.net
Beijign Normal University, China

According to Saussure, there are two systems of writing: (1) The system commonly known as 'phonetic' tries to respond to the succession of sounds that make up a word. Phonetic systems are sometimes syllabic, sometimes alphabetic i.e. based on the irreducible elements used in speaking. (2) In an ideographic system, each word is represented by a single sign that is unrelated to the sounds of the word itself. Each written sign stands for a whole word and, consequently for the idea expressed by the word. The classic example of an ideographic system of writing is Chinese. Thus there is a common misunderstanding that treated each Chinese character as a separate sign, especially in Chinese information processing field, and caused a major problem namely the mass encoding set. Hanzi, however, as a kind of common information carrier created and shared by the whole Chinese society, is impossible to be disorderly and unsystematic. There must be some intrinsic laws which make Hanzi the most mature ideograhic system in the world. This paper aims to use some analysis methods for complex network to identify these intrinsic laws of Hanzi system.

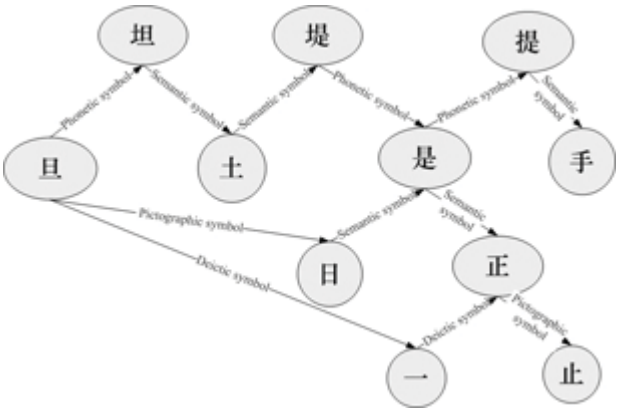
Being the representative of ideographic systems of writing, the most noticeable feature of Hanzi is that each character's graphic form was coined according to its original meaning. The elements of Hanzi graphic forms are called components which have certain functions when they were used to form characters. 梅, for example, is formed by 2 components 木 and 每, with 木 indicating 梅 is a name of tree and 每 indicating the pronunciation of 梅. A Chinese character may be formed only by one component, namely the character itself, and called single-element character; or may be formed by no less then 2 components and called compound-element character. The classical six principles for analyzing Hanzi graphic forms are know as 'Liushu'. Modern study on the Hanzi graphic forms, on the basis of 'Liushu', puts forward a new theory that uses 'component and component functions' to analyze Hanzi graphic forms. A component could play a quadruple role in forming a Chinese character: (1) as a pictographic symbol; (2) as a semantic symbol; (3) as a phonetic symbol; (4) as a deictic symbol. They are called component functions. All Chinese characters formed by different components with different functions could be divided into eleven categories, called the formation modes, as listed in Table 1.

MODE	DEFINITION	ie
0	Single-element character	止
Deictic-pictographic	composed of a pictographic component and a deictic component	正
Deictic-semantic	composed of a semantic component and a deictic component	本
Deictic-phonetic	composed of a deictic component and a phonetic component	百
Pictographic-phonetic	composed of a pictographic component and a phonetic component	琴
Semantic-phonetic	composed of a semantic component and a phonetic component	堤
Comprehensive composite with phonetic component	composed of a phonetic components and other components	寶
Pictographic composite	composed of 2 pictographic components	步
Semantic-pictographic	composed of a pictographic component and a semantic component	電
Semantic composite	composed of 2 semantic components	是
Comprehensive composite without phonetic component	composed of different components except phonetic component	羅

Table 1: Eleven Kinds of Hanzi Formation Modes

As a same component could be used to form different Chinese characters with other components, those characters having a common component (with the same function) could be linked to each other, and make the whole set of Hanzi a huge complex network, as Fig.1 shows. Modern Chinese graphology, based on comprehensive and detailed descriptions of Chineses characters' components and components' functions, has made some statistical analysis about the systematic features of Hanzi graphic forms, like the number of primitive components on Hanzi if different periods varied from 250 to 400; more

than 87% Chinese characters (since seal script) are composed of a semantic component and a phonentic component, or in other words, are of semantic-phonetic mode. These conclusions, however, can not give further explanations on the structure profile of Hanzi System, like how are hundreds of thousands of Chinese characters composed of no more than 400 primitive components; what is the average distance between 2 characters, are there any clusters, what are their densities, do they show a strict hierachy, are there some centers and so on.



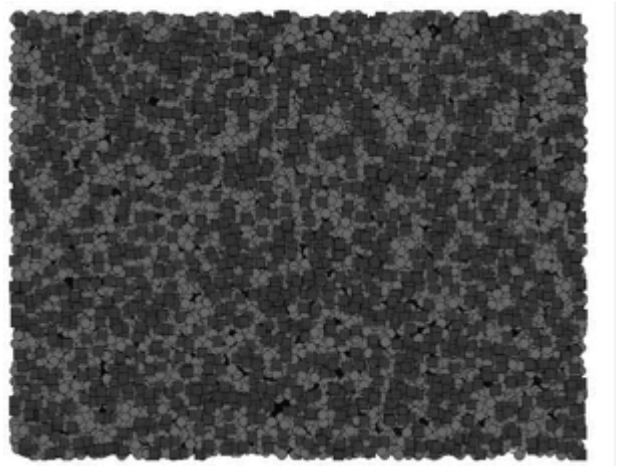


Figure 1: The Network of Hanzi
Graphic Forms (Local and Global)

This paper views the system of Hanzi graphic forms as a complex network with each Chinese character linked to its components by an edge indicating the component's function in the character (see Fig.1); and uses a series of network metics, such as the degree, the path length, the clustereing coefficient, the centrality, the coreness, the betweeness and so on, to analyze its topology features. This could introduce some more in-depth discussions on the structure and formation mechanism of Hanzi graphic form system and help us get more thorough understanding of the nature of Chinese. The innovation of this paper lies in the integration of the techniques of complex network theory and the scientific analysis of Hanzi graphic forms, which would bring some new insights in the study on Chinese graphology and provide some useful help in the teaching of Hanzi and Chinese information processing.

A Computer-Based Approach for Predicting the Translation Time Period of Early Chinese Buddhism Translation

Hung, Jen-Jou
jenjou.hung@gmail.com
Dharma Drum Buddhist College, Taiwan

Bingenheimer, Marcus
m.bingenheimer@gmail.com
Temple University, USA

Kwok, Jieli
guo.jieli@ddbc.edu.tw
Dharma Drum Buddhist College, Taiwan

Buddhism is a world-religion which has managed to take roots in cultures vastly different from that of its origin. Its transmission from India to China between the 2nd and the 10th centuries happened against all odds. The 'Buddhist conquest of China' can be partly attributed to the successful translation of a great number of texts translated into Chinese from Indian languages. The current standard edition of the Chinese Buddhist canon (*Taishō shinshū daizōkyō* (Abbr.: T.) 大正新修大蔵經, edited 1924-1934) contains 3053 works in 85 volumes, including about 1000 texts of Indian (or alleged Indian) provenance. However, ca. 150 of these texts are marked as *shiyi* 失譯, indicating that the name(s) of the translator(s) are unknown. Furthermore, for the texts that were translated between the 2nd and the late 6th century, many attributions are uncertain, problematic or simply incorrect. The issue of doubtful and wrong attributions has been debated in the field of Buddhist studies over the last few decades, e.g., Zürcher (1991), Harrison (1993), (and) Nattier (2008).

Over the years Buddhist scholars have leveraged traditional text-critical methods to corroborate or dispute traditional attributions yet like every method philology has its limits. Faced with a large number of texts in 'Buddhist Hybrid Chinese' of unknown provenance/origin, the long-established note-taking on the usage of characters and words quickly runs into problems. As with European languages, computational linguistics might offer new avenues of data collection and verification. The corpus of Buddhist Hybrid Chinese is available in a reliable digital format (XML/TEI) since the first 55 volumes of the Taishō edition were published freely by the Chinese Buddhist Electronic Texts Association (CBETA).

We are now able to apply statistical methods and artificial intelligence algorithms to the analysis of this corpus. This enables us to obtain new evidence bearing on translatorship attribution problems. The major advantage of quantitative methods for translatorship attribution is being able to analyze large amounts of data and to discover patterns which are not evident to the human reader.

Quantitative translatorship attribution is often considered to be a classification problem, that is, a text with uncertain or problematic authorship will be analyzed and compared with a corpus of texts by possible authors and then attributed to the author which whose works the texts shares most 'characteristics.' Recent years have seen renewed interest in many issues involved in optimizing quantitative authorship attribution. One of them is the effect of the size of possible candidate authors. As Luyckx and Daelemans (2010) have shown the accuracy of authorship analysis will decrease as the number of possible authors increases. It is therefore advisable to limit the number of possible authors in order to get a high accuracy analysis result. In our case, however, many of the early Chinese Buddhist translations are only rarely mentioned in historical records and canonical catalogues, and few have attracted the attention of philologists. For these translations, it is difficult to reduce the range of possible translators.

Therefore, as part of our attempt to establish a foundation for quantitative translatorship attribution for early Chinese Buddhist translations, we propose a classification mechanism based on predicting the translation time or period of a text. The advantage of this mechanism is twofold. First, within a given time bracket for the translation, the number of possible authors is limited, thereby improving the performance of the translatorship attribution. Second, by examining the result classification mechanism, we are able to identify possible and probable stylistic features of translations for different periods.

The time periods we focus on in the present study include three early Chinese dynasties: the Eastern Han (C.E. 25-220), the Three Kingdoms (C.E. 220-280) and the Western Jin (C.E. 266-316). These three dynasties constitute the earliest phase of Buddhist translation history and most of the translations from these periods present attribution problems. In this research, we build up classification mechanisms for each of the three dynasties. These can be used to test whether the translation style of a text is similar to the one *prevalent* during a certain period. We are aware of the fact that within

Buddhist Hybrid Chinese translation styles within a given period can vary greatly.

For the Eastern Han (C.E. 25-220) and the Three Kingdoms periods we build on recent philological scholarship (Nattier 2008), which has ascertained a number of attributions for this period. For the Western Jin textual corpus, we rely on contemporary research on traditional Buddhist sūtra catalogs, from which we exclude those texts for which current scholarship has not reached a consensus (Lancaster 2008; Lü 1981; Ren 1985; Yu 1993; Xu 1987). We then adopt the Variant Length N-gram algorithm (Hung et. al. 2009) to extract the stylometric features from the three corpora of ascertained texts. Variant Length N-gram is an extended form of the traditional n-gram algorithm. In the traditional n-gram algorithm, the length of grams *n* is fixed. Although the exploitation of n-gram algorithm has great impact on the performance of following analysis, deciding the best value of *n* is not straightforward. The Variant Length N-gram algorithm generates grams of all possible lengths, then removes those which are not significant. Thus, the importance of stylometric features is measured across grams of different length. This is crucial as there are no word boundaries in Buddhist Hybrid Chinese: gram-based analysis must therefore include grams of any length.

In the final stage, we use Fisher Linear Discriminant Analysis (FLDA) to analyze the stylometric features that have been extracted from the translations and to build up the classification mechanisms. The FLDA is a well-known dimension reducing and classification algorithm. It returns a linear function that transfers the high dimension source data of different groups into one-dimension points such that the ratio of total variances of projected points to the variances between groups of projected points is maximized. Since the FLDA's transformation is based on assigning weight to n-grams, the analysis is capable of yielding distinctive features, i.e. strings of Chinese characters, that are characteristic of the dynasties in question.

According to our experiments, the classification mechanisms for the three dynasties have all reached an accuracy rate higher than 90%. Moreover, when the three classification mechanisms are combined and usedto predict the translation time of an unknown translation, we can achieve an accuracy rate and a recall rate both above 80%. Besides, we are able to identify characteristic translation terms for different time periods.

References

Harrison, P. (1993). The Earliest Chinese Translations of Mahāyāna Buddhist Sūtras: Some Notes on the Works of Lokaksema. *Buddhist Studies Review* 10(2): 135-177.

Hung, J., M. Bingenheimer, and S. Wiles (2009). Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations. *Literary and Linguistic Computing* 25(1): 119-134.

Lancaster, L. (2008). *Catalogues in the Electronic Era: CBETA and The Korean Buddhist Canon: A Descriptive Catalogue*. CBETA, Taipei, 2008 (electronic publication). Retrieved from `http://jinglu.cbeta.org/lancaster.htm`.

Lu, Cheng 呂澂 (1981). *Xinbian hanwen dazangjing mulu* 新編漢文大藏經目錄. Jinan: Jilu shushe 齊魯書社.

Luyckx, K., and W. Daelemans (2010). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*. 26(1): 35-55.

Nattier, J. (2008). *A Guide to the Earliest Chinese Buddhist Translations: Texts from the Eastern Han* 東漢 *and Three Kingdoms* 三國 *Periods*. Tokyo: The International Research Institute for Advanced Buddhology, Soka University.

Ren Jiyu 任繼愈 (1985). *Zhongguo fojiao shi* 中國佛教史. Vol 1. Beijing: Zhongguo shehui kexue 中国社会科学出版社.

Xu Lihe 許理和 (1987). Zui zao de fojing yiwen zhong de donghan kouyu chengfen 最早的佛經譯文中的東漢口語成分, Yu yan xue lun cong 語言學論叢, Vol. 14. Beijing: Shangwu yinshuguan 商務印書館, pp. 197-225.

Yu Liming 俞理明 (1993). *Fojing wenxian yuyan* 佛経文献語言 [The Language of the Buddhist Scriptures]. Chengdu: Bashu shushe巴蜀書社, p. 206.

Zürcher, E. (1991). A New Look at the Earliest Chinese Buddhist Texts. In K. Shinohara et al. (eds.), *From Benares to Beijing: Essays on Buddhism and Chinese Religion in Honour of Prof. Jan Yün-hua*.Oakville, Ontario: Mosaic Press, pp. 277-304.

Bridging Multicultural Communities: Developing a Framework for a European Network of Museum, Libraries and Public Cultural Institutions

Innocenti, Perla
perla.innocenti@glasgow.ac.uk
University of Glasgow, UK

Richards, John
John.Richards@glasgow.ac.uk
University of Glasgow, UK

Wieber, Sabine
Sabine.Wieber@glasgow.ac.uk
University of Glasgow, UK

Global migration is here to stay

Knut Kjeldstadli[1]

1. Introduction

Traditionally, museums and libraries developed as historically separate institutional contexts and distinct cultures. Trant[2] noted how philosophies and policies of museums and libraries reflect their different approach to interpreting, collecting, preserving and providing access to objects in their care. Bishoff remarked that 'libraries believe in resource sharing, are committed to freely available information, value the preservation of collections, and focus on access to information. Museums believe in preservation of collections, often create their identity based on these collections, are committed to community education, and frequently operate in a strongly competitive environment'[3]. In the last century policy makers have attempted to group and bridge these communities of practices through 'their similar role as part of the informal educational structures supported by the public, and their common governance'[4]. Such commonalities are increasingly important to the sustainability of museums, libraries (and archives) in a globalised world. However a theoretical framework to scope and address such collaborative models still needs to be developed. This is particularly urgent in the specific context of our transnational and multicultural societies.

One of the goals of the recently funded EU FP7 SSH MeLa Project[5] is to fill this gap by investigating, identifying and proposing innovative coordination strategies between public European museums, libraries and public cultural institutions, for the benefit of multicultural audiences and towards European integration and European cultural commons [6]. The idea of laying the foundations for a European network of museums, libraries and public cultural institutions that address globalisation, migration and new media is particularly fitting for the configuration of migrant communities, which 'in the receiving countries can best be described from a structural perspective as a network of organizations.'[7]

2. Collaborations between Museums and Libraries: Potentialities and Challenges

In the first phase of our MeLa research, we have been focusing in particular on collaborations between museums and libraries. Some studies in this area[8] have highlighted the benefits of joining forces and resources in a variety of areas, including but not limited to:

- library activities and programmes related to museum exhibits
- travelling museums exhibitions hosted in libraries
- links established between web-based resources in library and museum websites
- library programs including passes to museums
- collaborative digitisation and digital library projects enhancing access to resources in both museums and libraries
- collaborative initiatives to bring in authors as speakers
- museums and libraries partnerships with cultural and education organisational for public programmes.

The overall opportunities for improving collections, increasing the number of users, leveraging experiences and funding also represent some of the main benefit of such partnerships. These studies also often included archives as a third player in museums and archives collaborations. The aims and objectives of collaboration projects between museums and libraries that have been investigated in previous studies, include: educational focus (e.g. learning about past civilisations, encourage families learning together etc.), cross-over visits between institutions, promoting resources to various target groups, improving coordination between

institutions, demonstrating joint working or training activities, providing models for working practices.

The International Federation of Libraries Association (IFLA) remarked that museums and libraries would indeed be natural partners for collaboration and cooperation[9]. In this context, a study in the United States observed that 'collaboration may enable [...] museums and libraries to strengthen their public standing, improve their services and programs, and better meet the needs of a larger and more diverse cross-sections of learners.'[10] The nature of this collaboration can be multifaceted and varied, and the terminology lends itself to diverse meanings, in particular regarding the degree of intensity of the collaboration and its transformational capacity. Gibson, Morris and Cleeve noted that "'*Library-museum collaboration*" can be defined as the cooperation between a library and a museum, possibly involving other partners. [...].'[11] Here, the authors use the term 'collaboration' with the meaning indicated by Diamant-Cohen and Sherman, as a 'more involved cooperation where there is a more in-depth sharing and pooling of resources.'[12]

Museums and libraries seem well positioned to synergically support and enable multicultural identities of migration societies. [13] As a result, museums are ideally placed to interpret and preserve culturally diverse heritage.[14] As centres for culture, information hubs, learning and gathering, libraries seemingly represent service providers for culturally diverse communities, enabling inter-cultural dialogue and education while supporting and promoting diversity.[15]

Nevertheless, the fruitful convergence between museums and libraries faces a number of challenges. Some authors[16] have highlighted the risks and obstacles encountered on the road to establish a successful collaboration between museums and libraries with respect to their different missions, cultures, organisational and funding structures. In terms of change management, Zorich, Waibel, Erway suggested that it is important to differentiate between coordination and cooperation, and pointing to the organisational changes required for a deep collaboration between libraries, museums and archives[17]. In particular for collaboration on digital libraries, Bishoff[18] and Innocenti *et alii* [19] remarked that interoperability is critical to the digital library community. Innocenti *et alii* further stressed the diverse organisational, semantic and technical interoperability levels that need to be addressed in a digital library, using the classification of the European Interoperability Framework for eGovernment services.[20] Achieving effective organisational interoperability between digital libraries can imply a radical change in the way that organisations work, manage and share their digital assets.

3. Towards a European Network of Museum, Libraries and Public Cultural Institutions

In the MeLA project, the ongoing research programme on a Network of Museums, Libraries and Public Cultural Institutions is articulated through a series of enquiries that intend to:

- Investigate the relations between transnational museums, library and public cultural institutions collaborations and the society of migration.
- Identify and describe how transnational museums, library and public cultural institutions collaborating together present themselves to various public communities.
- Understand and evaluate the effects (benefits and disadvantages) of transnational museums, library and public cultural institutions collaborations.

The results of this research will be made available in three books (a source book, proceedings of an international conference[21] and a reference book), including a coordination framework and policy briefs for the museum Community, policy makers and for European Commission. It is our wish that this coordination framework will contribute to new collaborative models of museums and libraries, testing the possibility of a European 'imagined community'[22] and the idea of European cultural commons, encompassing both cultural and scientific expressions and artefacts.

At the DH 2012 Conference we are presenting an overview of the desk and field investigation on selected case studies, organised in four thematic clusters: Collaboration models, European cultural and scientific heritage, Migration and mobility, Narratives for Europe. Each cluster includes case studies articulated in primary, secondary and tertiary level, with additional information from an online survey on museum collaborations with libraries and public cultural institutions.[23] Areas of collaboration explored in our research include the core activities of archiving, preserving and framing memory and the associated categories of hierarchies of cultural value and historical identity. The geographic coverage comprises trans-national and trans-local connections of museums and libraries, to allow more flexible and heterogenic connections to be considered, both within Europe – where for example public libraries are at the forefront of leading initiatives addressing multicultural diversity

– and outside its assumed confines (for example the Mediterranean), also in terms of European Union legitimacy and identity.[24] We will provide an overview of differences and current tension points between museums, libraries and public cultural institutions investigated in our research, and will discusssome initial suggestions on how they may overcome the challenges built into their infrastructure[25] and manage the change conveyed by collaborations and use of ICTs.

4. Conclusions and Outlook

Within the MeLa project, we are investigating innovative coordination strategies of transnational European museums, libraries and public cultural institutions, for the benefit of multicultural audiences and towards European integration and European cultural commons. Our research aims to provide evidence of collaborations and networks that can positively impact on the visibility of institutions involved, the improvement of the diffusion and accessibility of the collections, the effectiveness of an integrated organisational structure at European level and the coherence with European policies towards a common cultural and scientific heritage definition. We expect that the results of our investigation, which will continue through 2013, will contribute to the advancement of knowledge in this area, to provide museums, libraries and public cultural institutions with collaboration model framework, and to shaping European policies on multiculturalism and migration.

Funding

This work was supported by the European Commission within the MeLA - European Museums and Libraries in/of the Age of Migration, EU co-funded FP7 Collaborative Project, SSH-2010-5.2-2, 2011-2014 [Grant Agreement No. 266757].

Notes

1. Kjeldstadli, Knut (2010). Concepts of nation – and tasks of libraries, Keynote speech at IFLA Section Library Services to Multicultural Populations, IFLA Satellite Meeting, Copenhagen, 17-18 August 2010.
2. Trant, Jennife (2004). Emerging convergence? Thoughts on museums, archives, libraries, and professional training. Museum Management and Curatorship 24(4): 369-387.
3. Bishoff, Liz (2004). The Collaboration Imperative. Library Journal, 2004. Accessed August 16, 2011. http://www.libraryjournal.com/article/CA371048.html .
4. Trant, Jennifer (2004). Emerging convergence? Thoughts on museums, archives, libraries, and professional training. Museum Management and Curatorship 24(4): 369.
5. MeLa – European Museums in an Age of Migrations, http://www.mela-project.eu/ . The work presented in this paper is being conducted within 'MeLa Research Field 03 - Network of Museums, Libraries and Public Cultural Institutions.' http://wp3.mela-project.eu/ . Last accessed September 23, 2011. MeLa Research Field 03 is led by History of Art at the University of Glasgow (GU); the research team include staff members from Politecnico di Milano, Copenhagen Institute of Interaction Design, Museu d'Art Contemporani de Barcelona, Muséum National d'Histoire Naturelle/Musée de l'Homme, The Royal College of Art, L'Orientale University of Naples.
6. The definition 'European Cultural Commons' has been recently used in relation to digital content within the European Cultural Commons conference (europeanculturalcommons.eventbrite.com (europeanculturalcommons.eventbrite.com), videos at http://www.youtube.com/playlist?list=PLE244178708CBB62C) organised by Europeana on October 12, 2011. Europeana, an initiative endorsed by the European Commission, is a single access point to millions of books, paintings, films, museum objects and archival records that have been digitised throughout Europe. In this paper we are using the definition in a wider and more general meaning.
7. Faist, Thomas (1998). Transnational social spaces out of the international migration: evolution, significance and future prospects. *Archives Européennes de Sociologie* 39: 213-247.
8. See for example: Hannah Gibson, Anne Morris and Marigold Cleeve (2007). Links between Libraries and Museums: Investigating Museum-Library Collaboration in England and the USA. Libri (57): 53-64. Accessed August 16, 2011 www.librijournal.org/pdf/2007-2pp53-64.pdf (www.librijournal.org/pdf/2007-2pp53-64.pdf) ; Diane M. Zorich, Gunter Waibel, and Ricky Erway. Beyond the Silos of the LAMs: Collaboration Among Libraries, Archives and Museums. Report produced for OCLC Research, 2008. Accessed August 16, 2011 www.oclc.org/research/publications/library/2008/2008-05.pdf (www.oclc.org/research/publications/library/2008/2008-05.pdf) ; Alexandra Yarrow, Barbara Clubb, and Jennifer-Lynn Draper. Public Libraries, Archives and Museums: Trends in Collaboration and Cooperation. The Hague: IFLA Headquarters, 2008. Accessed August 16, 2011 www.ifla.org/VII/s8/pub/Profrep108.pdf (www.ifla.org/VII/s8/pub/Profrep108.pdf) . The WP3 team is preparing a selected bibliography for the purpose of the workpackage activities.
9. Yarrow, Alexandra, Barbara Clubb and Jennifer-Lynn Draper. Public Libraries, Archives and Museums: Trends in Collaboration and Cooperation. IFLA Professional Reports N. 108. The Hague: IFLA Headquarters, 2008. Accessed August 16, 2011 www.ifla.org/VII/s8/pub/Profrep108.pdf (www.ifla.org/VII/s8/pub/Profrep108.pdf) .
10. Charting the Landscape, Mapping New Paths: Museums, Libraries, and K-12 Learning. Institute of Museum and Library Services. Aug. 2004. Accessed August 16, 2011 http://www.imls.gov/assets/1/AssetManager/Charting_the_Landscape.pdf
11. Gibson, Hannah, Anne Morris and Marigold Cleeve. Links between Libraries and Museums: Investigating Museum-Library Collaboration in England and the USA. Libri (57) 2007: 53-64. Accessed August 16, 2011, p.53 www.librijournal.org/pdf/2007-2pp53-64.pdf (www.librijournal.org/pdf/2007-2pp53-64.pdf) .
12. Diamant-Cohen, Betsy, and Dina Sherman. Hand in Hand: Museums and Libraries Working Together. Public Libraries 42.2 (Mar./Apr. 2003): 102-105.
13. The definition of culture I am looking at can be found in the UNESCO Universal Declaration on Cultural Diversity:

'culture should be regarded as the set of distinctive spiritual, material, intellectual and emotional features of society or a social group, and that it encompasses, in addition to art and literature, lifestyles, ways of living together, value systems, traditions and beliefs' (UNESCO Universal Declaration on Cultural Diversity. UNESCO: 2002. Last accessed 18 August 2011, http://unesdoc.unesco.org/images/0012/001271/127160m.pdf). In this paper the terms "multicultural", "multiculturalism" and "cultural diversity" are considered synonymous.

14. See for example Barker, Emma, ed. Contemporary Cultures of Display. London: Yale UP 1999; Bennett, Tony, The Birth of the Museum. History, Theory, Politics. London & New York NY: Routledge, 2009. Gonzalez, Jennifer A. Subject to Display. Reframing Race in Contemporary Installation Art. Cambridge, MA: MIT Press 2008. Graham, Beryl, and Sarah Cook Rethinking Curating. Art after New Media. Cambridge, MA: MIT Press 2010. Karp, Ivan, et al., eds. Museum Frictions. Public Cultures/ Global Transformations. Durham, NC, London: Duke UP 2006. Knell, Simon J., et al., eds. Museum Revolutions. How Museums Change and Are Changed. London, New York NY: Routledge 2007.

15. IFLA - Library Services to Multicultural Populations Section (ed.) The IFLA Multicultural Library Manifesto: The Multicultural Library. 2006. Accessed 18 August 2011: http://www.ifla.org/VII/s32/pub/MulticulturalLibraryManifesto.pdf .

16. Gibson, Hannah, Anne Morris and Marigold Cleeve. Links between Libraries and Museums: Investigating Museum-Library Collaboration in England and the USA. Libri (57) 2007: 53-64. Accessed August 16, 2011 www.librijournal.org/pdf/2007-2pp53-64.pdf (www.librijournal.org/pdf/2007-2pp53-64.pdf) ; Walker, Christopher, and Carlos A. Manjarrez. Partnerships for Free Choice Learning: Public Libraries, Museums and Public Broadcasters Working Together. The Urban Institute and Urban Libraries Council. 2004. 22 May 2008 http://www.urban.org/UploadedPDF/410661_partnerships_for_free_choice_learning.pdf .

17. Zorich, Diane M., Waibel Gunter and Ricky Erway. "Beyond the Silos of the LAMs: Collaboration Among Libraries, Archives and Museums". Report produced for OCLC Research, 2008. Accessed August 16, 2011, p. 5, www.oclc.org/research/publications/library/2008/2008-05.pdf (www.oclc.org/research/publications/library/2008/2008-05.pdf) .

18. Bishoff, Liz. "The Collaboration Imperative". Library Journal, 2004. Accessed August 16, 2011 http://www.libraryjournal.com/article/CA371048.html

19. Innocenti, Perla, MacKenzie Smith, Kevin Ashley, Seamus Ross, Antonella De Robbio, Hans Pfeiffenberger, John Faundeen. Towards a Holistic Approach to Policy Interoperabilityin Digital Libraries and Digital Repositories. The International Journal of Digital Curation 6 (2011): 1, accessed August 16, 2011, http://ijdc.net/index.php/ijdc/article/view/167/235

20. IDABC. " European interoperability framework for pan-European eGovernment services". Luxembourg: European Commission 2004.

21. More information on the forthcoming International Conference for MeLa Research Field 03 are available at http://wp3.mela-project.eu/wp/pages/research-field-03-international-conference .

22. For the concept of 'imagined communities' see the seminal book by Anderson, Benedict. Imagined Communities. Reflections on the Origina and Spread of Nationalism, Verso, London and New York, new edition 2006.

23. Research Field 03: Online Survey, http://wp3.mela-project.eu/wp/pages/research-field-03-online-survey . Accessed March 12, 2012.

24. See for example Fuchs, Dieter and Andrea Schlenker. "European Identity and the Legitimacy of the EU". EU FP6 Consent Network of Excellence, 2006. Last accessed September 23, 2011. www.eu-consent.net/click_download.asp?contentid=1258 (www.eu-consent.net/click_download.asp?contentid=1258) .

25. Karp, Ivan, et al., eds. Museum Frictions. Public Cultures/ Global Transformations. Durham NC & London: Duke University Press, 2006.

Ptolemy's Geography and the Birth of GIS

Isaksen, Leif
leifuss@googlemail.com
University of Southampton, UK

The *Geographike hyphegesis* of Claudius Ptolemy (c. 90-c. 168 CE) – more commonly known as the '*Geography*' but hereafter referred to as *GH* – is perhaps the most important text in the history of digital cartography. While its fame derives from the enormous catalogue of ancient places that have made it indispensable for those studying ancient geography, its central innovations of presenting locations as tables of coordinates and offering multiple projections by which to represent them make it the foundational work for all modern Geographic Information Systems (GIS). Yet despite this historic legacy it remains a deeply problematic text (Dilke 1987). The sheer quantity, breadth and apparent precision of the catalogue greatly surpass that of every other known work of ancient geography. How did a lone scholar, working with only published materials, compile such a comprehensive overview of the inhabited world (*oikumene*)? And what provoked his revolutionary shift in methodology?

Several research projects have used digital methods to examine aspects of Ptolemy's Geography in recent years. Most significant is Stückelberger and Graßhoff's database of coordinates (2006) which forms the baseline data for the research presented. Despite creating this excellent resource, their analysis (Stückelberger & Mittenhuber 2009) remains a fairly traditional close reading of the text, makes scant explicit use of the digital data. Tsorlini (2009) used digital methods to show coordinate discrepancies between different manuscripts, although these largely focus on the medieval maps, the provenance of which remains unclear (Mittenhuber 2010). Kleineberg et al. used digital transformations in order to 'georectify' the coordinates in the region of Germania Magna and thus propose the identification of many hitherto unidentified locations with modern sites (2010).

Conventional wisdom is that *GH* is the culmination of a long tradition of Greek geographical and astronomical science, augmented and improved at each stage by contemporary sources such as merchant's reports, itineraries and astronomical observations (Berggren & Jones 2000: 25-30). This paper will present several arguments based on digital methods to demonstrate that this view is almost certainly wrong and that Ptolemy's treatise is a highly original work that fuses two distinct traditions for a specific purpose: the calculation of local celestial and solar phenomena for astrological prognostication. This goal required the combined data of Greek *geographic maps* – which represented the inhabited world schematically but in relation to cosmological constants such as the poles and equator – and Roman *chorographic maps*, which represented the world in a series of densely populated but geospatially unanchored regional images. It was the challenge of bringing together these very different traditions which led Ptolemy to apply the data management techniques he had used for astronomical data in the *Almagest* and thus the lay the groundwork for late medieval geography and ultimately the GIS systems of today.

While *Digital Humanities* is not the forum to lay out a full argument for Ptolemy's motives and the nature of his sources – which will be expressed in a forthcoming paper – the following preliminary points should be made. First, his astrological ambitions are no surprise. Ptolemy's *Tetrabiblos* became the principle work on the topic for almost a millennium and he explicitly refers to the practice of calculating celestial phenomena for terrestrial locations in the *Almagest* (2.3; 8.6), *GH* itself (1.1; 8.2) and the *Manual to the Handy Tables*. No other practical motive is given for *GH*'s construction and thus there is no reason to believe that it was ever intended for an alternative purpose. The second point is that while we have virtually no extant examples of pre-Ptolemaic geographic or chorographic maps, the written sources – including Ptolemy's distinction between the two practices in *GH* (1.1) – offer some important clues. The mathematical precision of *geographia* is explicitly stated by Ptolemy, but evidence as to the paucity of its content can be adduced from fragments of the three authors most likely to have been sources to Ptolemy – Eratosthenes, Hipparchus and Marinos of Tyre. Ignoring those on the Persian Royal Road, the places they mention are overwhelmingly (~90%) located on coastlines and boundaries. This is stark contrast to *GH* in which interior locations outnumber liminal ones by about 3:2. In contrast, both Ptolemy and Strabo (*Geog*. 2.5.17) refer to the richness of chorographic content but a strong indication of their unsuitability for astrology is given by the absence of any geodetic information for the Agrippan 'World Map' and other chorographic sources.

1. Statistical distributions of precision

Ptolemy's coordinates are expressed as fractions of a degree of latitude (measured from the equator) and longitude (measured from an arbitrarily defined prime meridian in the Atlantic ocean). Superficial

examination makes clear that the smallest latitudinal and longitudinal increment is 1/12 of a degree (or 5 minutes) of arc. As real world locations are themselves distributed randomly, we would expect a relatively even distribution of Ptolemaic locations with respect to the fraction of degree upon which they fall. In fact their distribution is highly uneven (Figure 1), a phenomenon also observed by Marx (2011). We know that some locations (which we shall refer to as 'high precision' locations) must have been assigned to the nearest twelfth of a degree, as they fall on divisions that are indivisible by lower denominators.[1] However, the uneven distribution makes it clear that significantly more of the coordinates are assigned at lower levels of precision: sometimes just to the nearest degree. The challenge is to establish which of them have been plotted more or less precisely. While it is not possible to do this by considering each location individually the fact that Ptolemy's principles are so closely aligned with GIS make it an ideal tool for investigating the data's underlying structure.

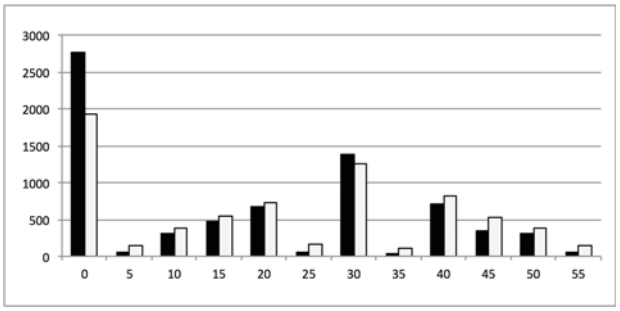


Figure 1: Quantity of coordinates falling on latitudinal (white) and longitudinal (black) degree divisions (in minutes)

2. Spatial distribution of precision

The first clues as to the spatial distribution of 'high precision' locations emerge from plotting all locations which are definitely assigned to the nearest twelfth of a degree in a GIS[2] (Figure 2). This demonstrates beyond all doubt that a core body of data – comprising Hispania, Italia and its major island neighbours, Greece, Asia Minor, the Levant and Egypt – contains virtually all known coordinates assigned to this high level of precision. The phenomenon appears in both manuscript traditions. In contrast, the peripheral regions, which include such well-established roman regions as Gaul and Africa Proconsularis, are almost entirely devoid of 'high precision' locations. Furthermore, although latitude is generally more likely to be of high precision than longitude, this is not caused by the well-known 'Problem of longitude' (Sobel 1995). Latitude can be hopelessly incorrect in some 'high precision' regions (such as Sardinia) so it is evident that these values have not been determined from empirical measurements.

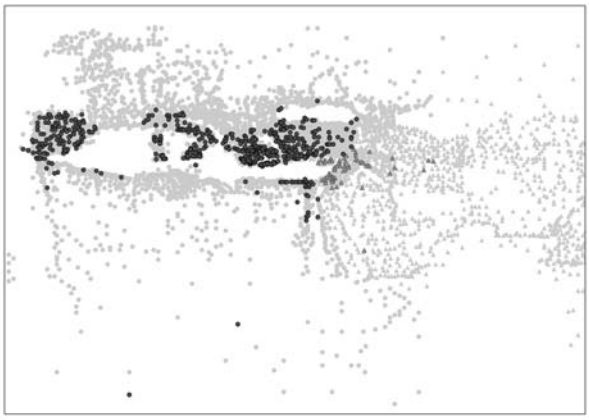


Figure 2: Coordinates know to be assigned at a precision of 1/12 of a degree of latitude or longitude. The effect is observed in both Ξ (black) and Ω (red) recensionsFigure 2: Coordinates know to be assigned at a precision of 1/12 of a degree of latitude or longitude. The effect is observed in both Ξ (black) and Ω (red) recensions

3. Linear interpolation

The order in which Ptolemy's coordinates are listed is not random. The description of each region follows a strict sequence commencing with boundaries, followed by physical geography, continuing with inland cities and concluding with offshore islands. The boundaries and physical features are typically linear in form and so topographically adjacent locations are listed consecutively to facilitate the drawing of lines or iconography between them. Islands are also listed in the sequence in which they follow a coastline wherever possible. Much more interesting for this analysis are the cities of the interior, for there is no natural order for Ptolemy to follow and so they may betray information as to the manner in which they were obtained.

We can visualize this order very easily by simply 'joining the dots' in sequence and colour coding the individual regions. Once again, the results display a striking disparity between the core regions and the periphery. In the core data, coordinates group together in tightly bunched clusters. At the periphery they are generally 'sketched out' in linear zig-zagging rows, usually horizontally (Europe, much of Asia) or vertically (Africa), but occasionally in a NE-SW sweep (India). A comparison with the 'high precision' coordinates over these features clearly demonstrates a correlation between this stylistic variation and the 'core' data we identified previously (Figure 3).

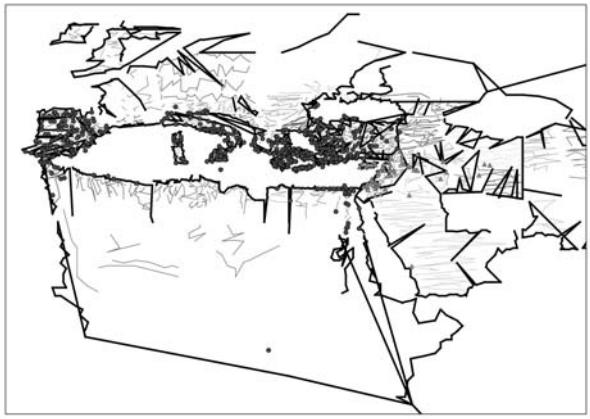


Figure 3: Detail showing lines interpolated between settlements overlaid with 'high precision' coordinates

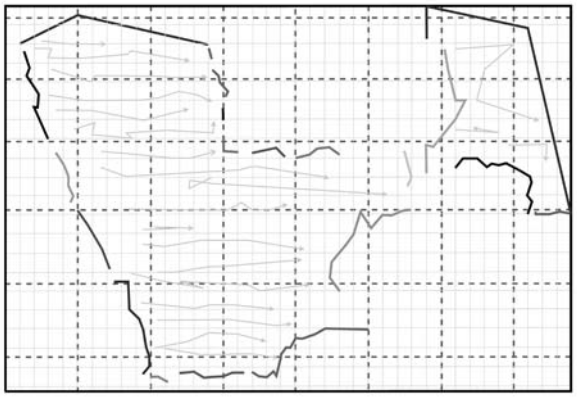


Figure 4: Ptolemy's map of Arabia Felix and Karmania showing how tribal divisions and structure are associated with traditional Greek parallels and divisions of 'hour intervals'

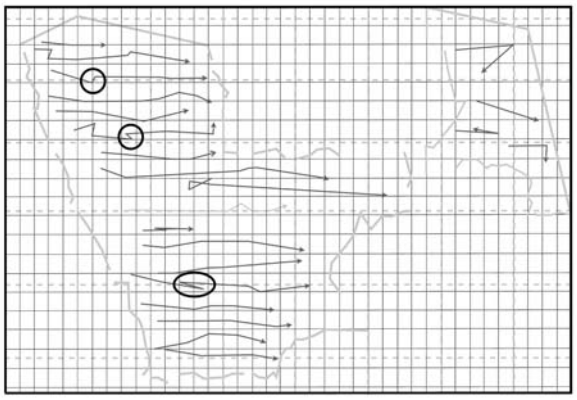


Figure 5: Ptolemy's map of Arabia Felix and Karmania showing how the order of undifferentiated cities of the interior is influenced by the Hipparchan coordinate system

4. Two Coordinate Systems

We can push this further. Despite the apparent regularity of Ptolemy's 180 x 360 coordinate grid (first developed by Hipparchus (Dicks 1960: 32-33)), he also makes reference to a traditional series of irregular Greek parallels (based on the length of the longest day) and tripartite divisions of the 'hour intervals' - the 'time-zones' of antiquity. So which coordinate system is Ptolemy using? The answer is most clearly illustrated by examining Ptolemy's regional map of Arabia Felix and Karmania. Ptolemy lists a series of tribal and geographic divisions along the coast of the Arabian peninsula and it is apparent that their locations are defined with reference to the traditional coordinate system (Figure 4). Yet the interior, which contains no differentiation and is simply a list of towns conforms, to the entirely separate Hipparchan framework (Figure 5).

We must be careful to jump to conclusions too readily for, as we have seen, there is much regional variation in Ptolemy's sources, but a consistent pattern emerges in both core and periphery. Boundaries make reference to the coarse-grained and irregular, traditional Greek framework whereas internal cities generally conform to the fine-grained and regular coordinate system. Returning to our earlier argument, there is an obvious explanation. Ptolemy's goals required him to use Greek *geographic* maps as a framework if he was to compare terrestrial and celestial locations. Yet we know that the data they provide is scanty. His solution was to peg *chorographic* maps to this global framework, thereby eliminating systemic error at the cost of local accuracy.

References

Berggren, J. L., and A. Jones (2000). *Ptolemy's Geography. An Anntoated Translation of the Theoretical Chapters*. Princeton: Princeton UP.

Dicks, D. R. (1960). *The Geographical Fragments of Hipparchus*. London: Athlone Press.

Dilke, O. A. W. (1987). The Culmination of Greek Cartography in Ptolemy. In J. B. Harly and D. Woodward (eds.), *The History of Cartography*. Chicago: U of Chicago P, vol. 1, pp. 177-200.

Kleineberg, A., et al. (2010). *Germania und die Insel Thule*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Marx, C. (2011). On the precision of Ptolemy's geographic coordinates in his Geographike Hyphegesis. *History of Geo- and Space Sciences* 2(1): 29-37.

Mercier, R. (2011). *Ptolemaiou Procheiroi Kanones. Ptolemy's Handy Tables*. Leuven: Peeters.

Sobel, D. (1995). Longitude. The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time. London: Fourth Estate.

Jones, H. L. (1917). The Georgraphy of Strabo. London: Heinemann, vol. 1.

Stückelberger, A., and G. Graßhoff (2006). *Ptolemaios. Handbuch der Geographie*. Basel: Schwabe.

Stückelberger, A., and F. Mittenhuber (2009). *Ptolemaios. Handbuch der Geographie. Ergänzungsband*. Basel: Schwabe.

Toomer, G. J. (1984). *Ptolemy's Almagest*, London: Duckworth.

Tsorlini, A. (2009). Higher Order Systematic Effect in Ptolemy's Geographia Coordinate Description of Iberia. *e-Perimetron* 4(2): 117-130.

Notes

1. At 5 minutes (1/12), 25 minutes (5/12), 35 minutes (7/12) and 55 minutes (11/12)
2. Those that fall on the traditional 2nd, 4th, 11th, 12th, 13th, anti-2nd and anti-4th Greek parallels (which also happen to be 'high precision' for arbitrary reasons) have also been removed here for reasons which will become clear.

Tracing the history of Noh texts by mathematical methods. Validitating the application of phylogenetic methods to Noh texts

Iwata, Yoshimi
sakunoshippo@gmail.com
Doshisha University, Japan

1. Introduction

Noh is a classical Japanese stage art consisting of singing and dancing with musical accompaniment. The song lyrics have been documented and are known as Noh texts. It is generally accepted that there are 240 songs, which are still performed today. However, some academic scholars claim that there are more than 500 Noh texts in existence. The texts take on different forms depending on the Noh school and the era in which they were edited.

Very little analyses have been carried out on Noh texts by applying mathematical methods. Therefore, in previous studies, Noh text was analyzed by identifying the authors of each Noh play using mathematical methods. This study employed mathematical and statistical methods to analyze 21 songs including worrier play category from Noh plays. Specific usage within lexical categories (auxiliary verb, adjective) was identified. As in Noh play, there are manuscripts which differ by the Noh school and the era being edited, the manuscripts under study need to be chosen adequately. Therefore, it is necessary to trace the history of each Noh text using phylogenetic analysis. At the same time, one has to be aware of the differences between engineering and literary approaches.

2. Purpose of research

The purpose of this research is to establish a new methodology in Noh field by employing mathemathecal and stylometric techniques. First, data is collected from various research materials and a database is constructed that facilitates preliminary analyses to verify items in the collected data or to draw a complete map for further studies. Practical analyses directly connected to interpretations are based on the preliminary analyses. Although, in Japan, these processes are common in both engineering and literary approaches, attitudes to analyses are

completely different. In engineering approaches, final conclusions are derived from practical analyses. In contrast, in literary approaches, analyses and interpretations are essentially synonymous and interpretive understanding is given more importance than scientifically based analyses. Therefore the discussions with other fields, such as information science, are required in each research process. Additionally, linguistic and cultural barriers make foreign researchers to spend a lot of time and care on understanding the research materials. Standing on this point, versatile methodology not to depend on researchers' skills of language is required. Normalized Compression Distance (NCD) method, the author focused on in this paper, calculates similarity of texts objects by using compression algorithm without any knowledge about the research materials.

3. Previous Researches

Two previous research applied biological methodologies to cultural phenomena.

A. Split decomposition, Spectronet

Tamaki Yano (2005)

Tamaki Yano applied Phylogenetic methods to collections of traditional Japanese 'waka' poems edited in 1215 [1]. The term 'waka' is a form of poem handed down from ancient times in Japan. The research materials had some versions that departed from the original, and the research aimed to identify an archetype for the differing collections. In this study, the genealogy of each collection is visualized by neighbor-net, a distance-based method for constructing phylogenetic networks that is based on the neighbor-joining algorithm of Saitou and Nei [2].

B. Database and Semantic Analysis for Fine Arts Based on Researchers' Experience –The Case of 'A Hundred Beauties' Drawn by YAMAGUCHI SOKEN

Yuka Iemura, Yu Fujimoto (2010)

In Iemura and Fujimoto's research, NCD was used to calculate the similarity distance because of its simplicity. Suppose, for instance, there are two files named 'x' and 'y,' C(xy) is the compressed size for the concatenation of files 'x' and 'y.' C(x) denotes the compressed size of 'x' and C(y) denotes that of 'y.' Fujimoto and Iemura attempted to apply this method to XML documents and the results were visualized using neighbor-net. They concluded that the results properly reflected differences between each XML document [3].

$$\mathrm{NCD}(x,y)=\frac{C(xy)-\min\{C(x),C(y)\}}{\max\{C(x),C(y)\}}$$

4. Analysis

This study examined *Tadanori* (忠度), one of the 240 existing songs. *Tadanori* is one of the oldest accessible manuscripts, and anyone can browse the full version of the photocopied manuscript. In addition, this masterpiece is popular even now, and thus, there is a considerable relevant material that makes it possible to trace the history of the Noh texts (Table 1). Eleven different *Tadanori* manuscripts were transcribed as data files. The size column in Table 1 shows the byte size of the manuscripts.

Node name	Manuscripts	School	Era	Size (byte)
kp_16c	Komparu jihitsu bon	Komparu	16c	8,020
kp_17c	Kurumaya bon	Komparu	17c	8,110
kz_c1605	Konetsu bon	Kanze	16c	8,118
tx_16c	Komiyama motomasa shikigo bon	Unknown	16c	8,210
kz_1629	Kan'ei uzuki bon	Kanze	1629	8,240
kz_1630	Kurosawa gentarou bon	Kanze	1630	8,002
kz_1925	Kanze Kaitei you hon	Kanze	1925	8,000
kz_1931	Kanze_Rufu bon	Kanze	1931	7,996
kz_c1700	Ryukou bon	Kanze	20c	8,012
kz_1713c	kanze-nobumitubon	Kanze	1713	7,887
hs_1979c	jibyoushibon	hosho	1979	7,957

Table 1: Node names and manuscripts

A. NCD

To create text files for NCD, all kanjis were converted to Japanese hiragana using Visual Basic for Applications (VBA).

NCD was calculated by Hyakka-One, which is a NCD calculating software developed by Fujimoto. The software currently supports zip, gzip, and bzip algorithms. General zip compression was used in this study. Although the original advocates of NCD recommend the quartet method for clustering, neighbor-net was used to visualize the genealogy of Noh-hon. The distance matrix calculated by Hyakka-One was exported to phylogenetic tree software, Splits Tree4, and visualized as a phylogenetic network called a neighbor-net. The neighbor-net derived from the NCD matrix is presented in Figure1.

B. Phylogenetic approach

The phylogenetic approach involves two methods. The first is a method for detecting differences between each group of phrases and calculating distances for a phylogenetic tree. The second, which is used to verify the results of the NCDs, is based on the maximum parsimony method. PAUP*4.0

software was used for this verification in this study. PAUP* is commonly used in cladistics or phylogenetics. To apply the second method, data cleaning is required to conflate each Noh text and extract the different phrases, make groups of phrases, and replace the groups with single letters. The process of separating Noh texts into phrases is not required by the NCD method. The methods provide different results based on differences between phrases. PAUP* returned 841 results. Although the lengths of edges are shorter than the NCD results, they are similar. These result were integrated using super-network and the integrated results were displayed using Splits Tree4.

5. Results

Three results were obtained by the distance method by SplitsTree4 (Figure 2), the Super-Network by the maximum parsimony trees (Figure 3), and the distance method by NCD (Figure 1). In all three results, the kp and kz groups are apparently separated. This means the texts can be categorized according to schools. With regard to hs_1979, all three networks show that hs_1979 is closer to the kz group. Related to the forms of performance, it is generally accepted in the Noh community that the hs group and the kz group are part of the same group, while kp belongs to other groups that were not included in this study. Therefore, these three results support the generally accepted ideas. Although NCD tends to be criticized because of the difficulty involved in verifying the results, in this study, the NCD approach gives enough results on grouping and can be applied to preliminary analysis. The advantage of NCD is the simplicity of its data cleaning process. Because of this simplicity, foreign researchers who are not fully familiar with the Japanese language or researchers of literature and have a limited knowledge of engineering can handle the data. The other two methods require familiarity with both Japanese and engineering. NCD has a disadvantage for the practical analysis stage. Because NCD depends on a complex compression algorithm, it is difficult to perform verification using the original texts.

6. Conclusion

The NCD approach has been successfully applied to Noh research. In this initial study, the results from the NCD method were not significantly different from the other two common methods, and it is reasonable to conclude that the NCD method can support traditional Noh studies equally as well as the other two methods. In future research, the author will attempt to apply NCD to other Noh plays and determine if it is valid to use the method for preliminary analysis. It is acknowledged that it will be necessary to develop database schemas that can potentially generate text plans for calculating NCD.

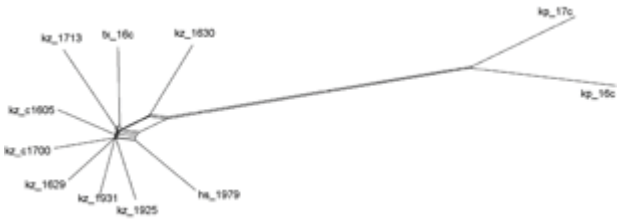


Figure 1: Neighbor-net for texts of Tadanori derived from NCD distance

Figure 2: Neighbor-net for texts of Tadanori derived from standard distance method

Figure 3: Super-network for texts of Tadanori derived from the maximum parsimony method (PAUP*)

References

Yano, T. (2005). Split decomposition, Spectronet. Japan. *The Special Interest Group Notes of IPSJ SIG Computers and the Humanities,* pp. 33-40.

Saitou, N., and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4: 406-425.

Iemura, Y., and Y. Fujimoto (2010). Database and Semantic Analysis for Fine Arts Based on Researchers' Experience – The Case of 'A Hundred Beauties' Drawn by YAMAGUCHI SOKEN. *Proceedings of Third Japan Art Documentation Society,* pp. 12-15.

Omote, A. (1965). *The Study of Kozanbunkobon – the part of Utaibon –*. Tokyo: Wanya Bookstore.

Yokomichi, M., and A. Omote (1960). *Youkyokusyu* 1 & 2 (Tokyo: Iwanami Bookstore).

Computing and Visualizing the 19th-Century Literary Genome

Jockers, Matthew
mjockers@stanford.edu
Stanford University, USA

1. Overview

In literary studies, we have no shortage of anecdotal wisdom regarding the role of influence on creativity. Consider just a few of the most prominent voices:

1. 'Talents imitate, geniuses steal' – Oscar Wilde (1854-1900?).[1]
2. 'All ideas are second hand, consciously and unconsciously drawn from a million outside sources' – Mark Twain (1903).
3. 'The historical sense compels a man to write not merely with his own generation in his bones, but with a feeling that the whole of the literature … has a simultaneous existence' – T. S. Eliot (1920).
4. 'The elements of which the artwork is created are external to the author and independent of him …' – Osip Brik (1929).
5. *Anxiety of Influence* – Harold Bloom (1973).

Whether consciously influenced by a predecessor or not, it might be argued that every book is in some sense a necessary descendant of, or necessarily 'connected to, ' those before it. Influence may be direct, as when a writer models his or her writing on another writer,[2] or influence may be indirect in the form of unconscious borrowing. Influence may even be 'oppositional' as in the case of a writer who wishes to make his or her writing intentionally different from that of a predecessor. The aforementioned thinkers offer informed but anecdotal evidence in support of their claims of influence. My research brings a complementary quantitative and macroanalytic dimension to the discussion of influence. For this, I employ the tools and techniques of stylometry, corpus linguistics, machine learning, and network analysis to measure influence in a corpus of late 18th- and 19th-century novels.

2. Method

The 3,592 books in my corpus span from 1780 to 1900 and were written by authors from Britain, Ireland, and America; the corpus is almost even in terms of gender representation. From each of these books, I extracted stylistic information using techniques similar to those employed in authorship attribution analysis: the relative frequencies of every word and mark of punctuation are calculated and the resulting data winnowed so as to exclude features not meeting a preset relative frequency threshold.[3] From each book I also extracted thematic (or 'topical') information using Latent *Dirichlet* Allocation (Blei, Ng et al. 2003; Blei, Griffiths et al. 2004; Chang, Boyd-Graber et al. 2009). The thematic data includes information about the percentages of each theme/ topic found in each text.[4] I combine these two categories of data – stylistic and thematic – to create 'book signals' composed of 592 unique feature measurements. The 'Euclidian" metric is then used to calculate every book's distance from every other book in the corpus. The result is a distance matrix of dimension 3,592 x 3,592.[5]

While measuring and tracking 'actual' or 'true' influence – conscious or unconscious – is impossible, it is possible to use the stylistic-thematic distance/ similarity measurements as a proxy for influence.[6] Network visualization software can then be used as a way to organize, visualize, and study the presence of influence among of books in my corpus.[7] To prepare the data for use in a network environment, I converted the distance matrix into a long-form table with 12,902,464 rows and three columns in which each row captures a distance relationship between two books. The first cell contains a 'source' book, the second cell a 'target' book, and a third cell the measured distance between the two. After removing all of the records in which the target book was published before, or in the same year as, the source book,[8] the data was reduced from 12,902,464 records to 6,447,640. This data and a separate table of metadata were then imported into the open source network analysis software package Gephi (2009) for analysis and visualization.

3. Analysis

Networks are constructed out of nodes (books) and edges (distances). When plotted, nodes with less similarity (i.e. larger distances between them) will spread out further in the network. Figure 1 offers a simplified example of three imaginary books.

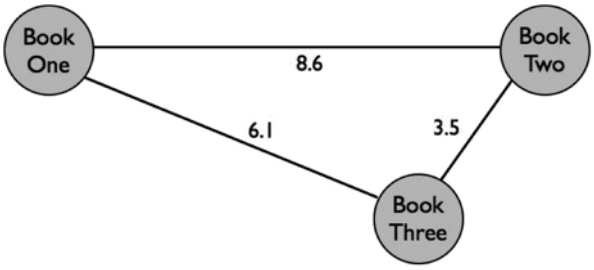


Figure 1: a sample network with edge numbers representing measured distances between nodes

While it is not possible to show the details of the entire network here, it is possible to display several of the most obvious macro-structures. Figure 2, for example, presents a zoomed out view of the network with book nodes colored according to dates of publication.[9]

Figure 2: The 19th-century novel network colored according to publication date

The shading of nodes and edges according to publication date reveals the inherently chronological nature of stylistic and thematic change. The progressive darkening of the nodes from east to west allows us to see, at the macro-scale, how style and theme are changing and evolving over time. [10]Also seen in this image is a 'satellite' of books in the northwest. This satellite represents a 'community' of novels that are highly self-similar but at the same time markedly different from the books in the main network cluster. [11]When the network is recolored according to gender (figure 3), a new axis can be seen splitting the network into northern and southern sectors along gender lines.

Figure 3: The 19th-century novel network colored according to author-gender

This visualization (Figure 3) reveals that works by female authors (colored light gray) and male authors (black) are more stylistically and thematically homogeneous within their respective gender classes. As a result of this similarity in 'signals,' female-authored books cluster together on the south side of the main network, while male-authored books are drawn together in the north.[12] These two 'views' of the network allow us to begin imagining the larger macro-history of thematic-stylistic change and influence in the 19th-century novel. What is not obvious in this macro-view, however, is that a great many of the individual books we have traditionally studied are in fact 'mutations' or outliers from the general trends. Harriet Beecher Stowe's *Uncle Tom's Cabin*, for example, clusters closer to the works of male authors, and Maria Edgeworth's *Belinda* has a signal that does not become dominant for forty years after the date of Belinda's publication. Also absent from the macro-view are the individual thematic-stylistic 'legacies'. Using three measures of network significance (weighted in-degree, weighted out-degree and Page-Rank), [13]I will end my presentation with the argument that Jane Austen and Walter Scott are at once the least influenced (i.e. most original) of the early writers in the network and, at the same time, the most influential in terms of the longevity, or 'fitness,' of their thematic-stylistic signals. The signals introduced by Austen and Scott position them at the beginning of a stylistic-thematic genealogy; they are, in this sense, the literary equivalent of *Homo erectus* or, if you prefer, Adam and Eve.

References

Bastian, M., S. Heymann, et al. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*.

Blei, D. M., T. L. Griffiths, et al. (2004). *Hierarchical topic models and the nested Chinese restaurant process*. Cambridge, MA: MIT Press.

Blei, D. M., A. Y. Ng, et al. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993-1022.

Bloom, H. (1973). *The anxiety of influence; a theory of poetry*. New York: Oxford UP.

Bloom, H. (2011). *The anatomy of influence: literature as a way of life*. New Haven, Conn.: Yale UP.

Brik, O. M. (1929). Teaching Writers.

Chang, J., J. Boyd-Graber, et al. (2009). *Reading Tea Leaves: How Humans Interpret Topic Models*Advances in Neural Information Processing Systems 22.

Eliot, T. S. (1920). *The sacred wood; essays on poetry and criticism*. London: Methuen.

Garcia, M., and C. Martin (2007). Function Words in Authorship Attribution Studies. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* **22**(1): 49-66.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* 22(3): 251-270.

Hoover, D. L. (2001). Statistical Stylistics and Authorship Attribution: An Empirical Investigation. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* 16(4): 421-444.

Hoover, D. L. (2008). *Quantitative Analysis and Literary Studies. A Companion to Digital Literary Studies*. Oxford: Blackwell.

Martindale, C., and D. McKenzie (1995). On the Utility of Content Analysis in Author Attribution: The Federalist. *Computers and the Humanities* 29(4): 259-270.

Team, R. D. C. (2011). *R: A Language and Environment for Statistical Computing*. Vienna: Austria, R Foundation for Statistical Computing.

Twain, M., and A. B. Paine (1975). *Mark Twain's letters*. New York: AMS Press.

Yang, Y., and J. Pedersen (1997). *A comparative study on feature selection in text categorization*. Proceedings of the 14th International Conference on Machine Learning (ICML '97), Nashville, Tennessee.

Zhao, Y., and J. Zobel (2005). Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science*. Berlin: Springer, pp. 174-189.

Notes

1. Routinely attributed to Wilde, but of uncertain origin, Wilde probably 'borrowed' this quip.
2. An extreme case may be Fielding's *Shamela*, which attempts to satirize Richardson's *Pamela*.
3. Features with a corpus mean less than 0.10 were excluded. This resulted in 92 features being retained. For more on feature selection practices see: Martindale and McKenzie (1995), Yang and Pedersen (1997), Hoover (2001), Zhao and Zobel (2005), Garcia and Martin (2007), Grieve (2007), Hoover (2008)
4. For this research, the topic model was set to extract 500 latent topics.
5. 'Distance' here is a measure of stylistic and thematic similarity. Distances were calculated using the default settings of the 'dist()' function in the R statistics package. See: Team (2011)
6. This is not simply a 'proxy' of convenience. It is, in fact, an ideal proxy, especially so when we consider that even plagiarism (the most obvious form of influence) can be accidental.
7. I'm grateful to my colleague Elijah Meeks who sees the world through network analysis goggles and who aided me in visualizing and analyzing this data.
8. Influence only works in one direction!
9. All network layouts employ Gephi's built-in Force Atlas 2 algorithm.
10. Remember that time is not a feature in the similarity calculations. The arrangement of the nodes in a chronological fashion is a byproduct of the way in which the books signals change in a regular, chronological, manner.
11. In the larger presentation of this work, I will offer an explanation for why the 499 books in this isolated cluster are split off from the main network. One unifying element is time; most are books from a similar time period, but the full explanation is more nuanced.
12. Even in the satellite community, sub clustering by gender is apparent.
13. Weighted in-degree is a measure of influence coming into a node whereas weighted out-degree provides a measure of the influence a given node exerts on subsequent nodes. The Page-Rank algorithm can be used to gauge the significance/power/importance of a book in the overall network.

Using the Google Ngram Corpus to Measure Cultural Complexity

Juola, Patrick
juola@mathcs.duq.edu
Duquesne University, USA

It is common to hear assertions that culture is complex, that language is complex, and that while the complexity of language is a universal constant, the complexity of culture is increasing as technology and the increased pace of modern life creates new complexities. These statements are usually based on subjective assessments, often tinged by nostalgia for the 'good old days.' Can questions of cultural complexity be addressed quantitatively?

Previous work (e.g. Juola 1997, 2008) has been able to use information theory to address this question. The basic idea is that a sample of language is 'complex' if it contains a lot of information, defined formally as the size of the computer program that would be necessary to (re)create the sample from scratch, a measure more formally known as Kolmogorov complexity. This can be approximated by compressing the text sample and looking at the size of the resulting file – the larger the resulting file, the more complex the original. Alternatively, one can compute complexity directly using Shannon's (1948) famous formula for information entropy based on a concept of the underlying linguistic 'events.' In any case, linguistic complexity can be measured observing discourse-controlled samples of language, essentially by comparing several (linguistic) versions of the same text, such as translations of the Bible or of a specific novel, and observing whether one language yields systematically larger measurements than another. Previous work suggests that no such systematic pattern exists, and that all languages are indeed roughly equal in complexity.

Key to this approach is the idea of discourse control; we are measuring how difficult it is to express a specific fixed concept in a given language and comparing it to the same concept expressed in an another language. Culture, however, can be treated as the set of concepts that people choose to express. By eliminating the restriction of discourse control and instead investigating language chosen freely by the cultural participants, we may be able to tease apart the interaction between cultural and linguistic complexity. In particular, we can distinguish between linguistic and cultural complexity as follows: a language is complex if there is a lot of information contained in a topic-controlled discourse. A culture is complex if there is a large range of topics for discourse, or alternatively a lot of information contained in topical choice. Therefore, if we compare the complexity (however measured) of two language samples that are not topic-controlled, but instead are in some sense representative of the breadth of discourse present in a culture, we can calculate the differences attributable to discourse variation, and hence to cultural complexity.

As an illustrative example, we follow the approach of Spenser (1900; cited by Denton 2004), in that 'complex' means 'containing many different interdependent parts.' A complex political system has many parties and power groups, many different roles and offices, and many relationships among them. In a political discourse, many if not most of these parties and power groups would need to be explicitly named and distinguished from each other. By contrast, an autocratic monarchy is relatively simple: there is the monarch and then everyone else. A game is complex if it has many rules and strategies. A culture is complex if it contains many heterogeneous aspects such as technological specifications, social stratification, multilevel administrative hierarchies, or a large amount of object or object-types. Continuing this line of reasoning, a complex culture is one with lots of 'stuff' and where people do lots of things to or with 'stuff,' where 'stuff' here refers not only to physical objects but also to people, groups, activities, abstractions, and so forth – anything that can be discussed among the group.

We therefore apply the previous methodology to a different sort of corpus; an uncontrolled corpus that represents the breadth of cultural experience. If the information contained in such a corpus is high, then we can say the culture is complex. Several corpora may be suitable for this purpose; we have chosen to study the Google Books Ngram Corpus (Michel et al. 2010). This contains all of the n-grams from the millions of books in the Google Books database, something like 20 million books, or approximately 4% of all books ever printed. While not strictly speaking representative (for example, 'publishing was a relatively rare event in the 16th and 17th centuries,' and 'many more books are published in modern years'), and of course typically only the literate can write or publish books, this nevertheless gives us a time-stamped window into the scope of culture. Furthermore, by focusing on n-grams (and specifically on 2-grams, word pairs), we can observe not only the distribution of 'stuff,' but also some of the relationships between 'stuff' – for example, the number and range of word pairs beginning with 'expensive' will inform us about changing opinions regarding money and the types of goods considered luxurious and pricey.

We therefore used the Google Books American 2-Gram Corpus to measure changes in the complexity of American culture at ten-year intervals between 1900 and 2000. This corpus simply contains a frequency list of all two word phrases used in American-published books in any given year. For example, the phrase 'hamburgers with' appeared only 8 times in print in 1940, compared to 45 in the year 2000. Focusing strictly on the US during the 20th century avoids many of the problems with mass culture, as publishing was a well-established industry and literacy was widespread. However, the number of books published in this time of course tended to increase. Our first observation, then, is that culture may be increasing simply from the number of different things to talk about. The number of different word pair types per year increased dramatically, nearly doubling from 1900 to 2000, as given in table 1.

Year	# types
1900	17,769,755
1910	22,834,741
1920	22,409,426
1930	19,745,549
1940	20,369,679
1950	23,632,749
1960	27,379,411
1970	34,218,686
1980	34,458,083
1990	37,796,626
2000	41,654,264

Table 1

This alone indicates an increase in the complexity of written culture, although this process is not continuous and some years during the Depression show a loss. To confirm the overall upward trend, we have also calculated the Shannon-entropy of the 2-gram distributions, attached as table 2.

Year	Entropy (bits)
1900	17.942357
1910	18.072880
1920	18.072325
1930	18.133058
1940	18.241048
1950	18.336162
1960	18.391872
1970	18.473447
1980	18.692278
1990	18.729807
2000	18.742085

Table 2

This further analysis illustrates that a more sophisticated measure of complexity shows a continuous process of increasing complexity, even in times when (for example due to economic downturn) the actual volume of words published decreases. Even when people are writing less, they still have more 'stuff' about which to write, showing the cumulative nature of culture (today's current events are tomorrow's history, but still suitable material for discussion and analysis – part of culture).

We acknowledge that this is a preliminary study only. Google Books offers cultural snapshots at much greater frequency than ten-year intervals. Google Books also offers corpora in other languages (including German, French, Spanish, Russian, and Hebrew) as well as another English-speaking culture. Use of a more balanced corpus (such as the Google Books English Million corpus, a corpus balanced at about 1 million words/year to offset increased publication), or the BYU Corpus of Historical American English might help clarify the effects of publication volume. Analysis of n-grams at sizes other than 2 would illustrate other types of complexity -- in particular, 1-grams (words) would show changes in lexical but not syntactic complexity and hence an analysis of 'stuff' but not what people do with 'stuff.' Despite these weaknesses, we still feel this paper illustrates that culture-wide analysis of abstractions like 'increasing complexity' is both practical and fruitful.

References

Denton, T. (2004). Cultural Complexity Revisited. *Cross-Cultural Research* 38(1): 3-26.

Juola, P. (1998). Measuring Linguistic Complexity : The Morphological Tier. *Journal of Quantitative Linguistics* 5(3): 206-213.

Juola, P.. (2008). Assessing Linguistic Complexity. In M. Miestamo, K. Sinnemaki, and F. Karlsson (eds.), *Language Complexity: Typology, Contact, Change*. Amsterdam: John Benjamins.

Michel, Jean-Baptiste, Y. Kui Shen, A. Presser Aiden, A. Veres, M. K. Gray, W. Brockman, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, St. Pinker, M. A. Nowak, and E. Lieberman Aiden (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* (Published online ahead of print: 12/16/2010).

Spencer, H. (1900). *First principles* (6th ed.). Akron, OH: Werner.

‘All Rights Worth Recombination’: Post-Hacker Culture and ASCII Literature (1983-1993)

Katelnikoff, Joel
katelnikoff@ualberta.ca
University of Alberta, Canada

1. ASCII Literature

This presentation will focus on a literary and cultural movement of the computer underground that featured texts written in ASCII code and distributed by means of the Bulletin Board Systems (BBSes) that flourished from the early-1980s until the mid-1990s. Independently-produced ASCII publications (also known as textfiles, t-files, g-files, philes, etc.) offered computer users a means by which to share information and opinions with one another. This movement, at its height, also gave rise to an innovative literature that was shaped by the technological environment in which it was produced and by the digital culture that it was connected to.

This presentation will focus primarily on two of the earliest serial ASCII publications: *PHRACK and Cult of the Dead Cow*. These publications were founded in 1985 (shortly after the film *WarGames* and the book *Hackers: Heroes of the Computer Revolution*). These publications share several things in common: they are collaborative projects, written and edited by pseudonymous computer users in their late teens / early twenties, they contain illicit information about hacking, phreaking, and mischief, and they are uncensored and boundary-pushing. They both refuse to copyright their material, with the hope and expectation that their publications will be freely uploaded and downloaded throughout the computer underground. And most importantly, both of these publications feature literary writings that innovate new forms and styles that are strongly influenced by the techniques and tenets of hacker culture, particularly in its connection to copyright.

2. Rhizomatic transfer

When the earliest ASCII zines emerged, the computer underground consisted of innumerable hubs of digital culture, primarily in the form of independently-operated, predominantly non-profit bulletin board systems. BBSes functioned much like answering services, routing incoming calls into an interface that would allow callers to send messages, upload and download files, and play door games. Most BBSes were very small and had one telephone line, meaning that only one user could access the system at any given time. These small BBSes were not directly linked to one another (unlike hypertext models), but instead were connected through their users, most of whom would frequent several other BBSes, downloading and uploading files from board to board.

Because BBS culture had no hierarchy of organization and because most BBSes were non-profit, there was no major presence of corporate media organization (e.g. newspapers, book publishers) within BBS culture. The literati of the BBS world would instead emerge from a new chorus of voices: namely, those who were willing to publish their works without any remuneration and were willing to allow their works to be copied and distributed freely throughout the computer underground. In exchange for this, the writers would have the potential to proliferate from BBS to BBS, city to city, allowing for the distant (but actual) possibility of becoming major contributors to the shape of computer underground culture. As an acknowledgement of their willingness to allow users to do what they will with the text, most of these files either exclude any statement of copyright (*Phrack*), or compose puns such as ‘all rights worth shit’ (*Cult of the Dead Cow*). The publishers implicitly acknowledge that, in order for their work to proliferate, they must allow the work to be freely downloaded from BBSes and freely uploaded to other BBSes.

3. Open source

If information wants to be free, what does this freedom entail? ASCII publishers allow their works to be shared, transferred, and read without cost to the reader or creator. The files can be limitlessly replicated with minimal concern for material costs. But beyond the ability to distribute and use these files freely, *Cult of the Dead Cow* and *Phrack* themselves take liberties that extend beyond use and distribution. These publications demonstrate a willingness to modify the work of others in order to produce new works of their own, in the way that we might envision software programmers using open-source code. What follows are three successive stages of open source literature: reprinting, modification, and recombination.

4. Reprinting

‘The Conscience of a Hacker’ (also known as ‘The Hacker Manifesto’) is probably the most well-known essay of the ASCII era. It was originally printed in *Phrack* 1:3, Phile 7 (1986). It was later

republished in *Cult of the Dead Cow* 12, without any acknowledgement that the essay had previously been published in *Phrack*, and without the knowledge of the author. Moreover, *cDc* used 'The Conscience of a Hacker' to overwrite the original content of issue 12 (the transcribed lyrics of Metallica's *Master of Puppets* album). For *Cult of the Dead Cow*, it was not only their own copyright that was 'worth shit,' but also the copyright claims of other publishers and writers, both digital and mainstream. In 1988, they republished a rant entitled 'Fuck the World,' which strongly influenced the next ten years of ASCII literature. The article, ironically, was a reprint from a print publication entitled *Forced Exposure*. Although the editor of *cDc* (Swamp Rat) does give credit to the author and original publisher in the issue, the header of the issue clearly reads 'FUCK THE WORLD / by Swamp Rat,' leaving generations of readers confused as to who truly authored the article, and to what extent the text is 'by' Swamp Rat. Here, unauthorized reprinting seems to take its place alongside hacking and explosives manuals as yet another form of illicit information.

5. Modification

In *Cult of the Dead Cow* and *Phrack*, freedom of information also extends to the modification of text. In academia, scholars frequently build upon the work of their colleagues, citing the findings of other scholars as part of the process of developing their own work. In ASCII publications this is also common, but without the same insistence of formality. The most parasitic example of this was the practice of tagging, where BBS sysops would obtain published text files, modify them by adding advertisements for their BBSes to the header or footer of the file, and continue to distribute the file by means of their BBS. The modified versions would proliferate, often under the same file name, which would limit the circulation of (and possibly overwrite) copies of the original version.

Phrack and *Cult of the Dead Cow* also modified texts, but for different reasons and in different ways. Instead of inserting their own text into other peoples' publications, they would bring other people's writing into their publications. This was often done without the knowledge or consent of the writer. In some instances, they would replicate a substantial portion of another text (or an entire text), and then add to or subtract from that text at will. One of *Phrack*'s regular features, 'Phrack World News,' would republish articles from corporate media outlets, occasionally interjecting their own undercutting commentary in square brackets. The inversion is clear: this is what the mainstream is saying about us, and this is what we are saying about what them.

Phrack's most documented exploit in modification is their publication of a scaled-back version of a document that had been illegally downloaded from a 'secure' BellSouth computer, describing some of the workings of their Enhanced 9-11 system. Even though the contents of the file were lightweight and contained no indication of being proprietary or sensitive information, its republication demonstrated that the systems at BellSouth could be hacked, by affiliates of *Phrack*. The editor was indicted on charges of wire fraud, interstate transportation of stolen property, and computer fraud and abuse.

6. Recombination

The central concept of hacking is to generate a flexible and adaptable mastery of codes, and to develop a mastery that transcends a system. This is central to the hack and to the hacker ethos. Like any other codes, legal codes are equally susceptible to disruption and modification, as demonstrated in these four publications, which demonstrate a libertarian attitude in their handling of illicit information (in terms both content and copyright). Like any other codes, literary and linguistic codes are likewise susceptible. Even in these early publications, there is a tendency toward the recombinant (that is, writing that borrows liberally from multiple sources). From the inception of ASCII literature, writers tend to transform, and recontextualize language, style, and form from other ASCII publications and from BBS culture. Writers blur the boundaries between information and narrative, purpose and play, text and intertext.

The most prominent example of recombination that took place in this era is probably 'The *cDc* #200 Higgledy-Piggeledy-Big-Fat-Henacious-Mega-Mackadocious-You-Can't-Even-Come-Close-So-Jump-Back-K-BOOMIDY-BOOMIDY-BOOM File.' This file, designed to simulate the scrolling display of an early-1990s BBS, follows the narrative arc of a BBS experience, with every message board and file area transforming into narrative, each narrative venturing into the most iconic tropes and mythemes of the computer underground, corporate culture, and ASCII literature: Internet Relay Chat, *Teen Beat*, Encyclopedia Brown, K-Rad slang, *Phrack* magazine, 'The Real Pirate's Guide,' George Bush, Malcolm X, and old school 40-column all-caps ASCII. This issue begins to demonstrate what can be done with the structures and contents made familiar by other publications. It demonstrates a writing that takes freely and gives freely, which is never proprietary, and for whom all rights remain shit. Most importantly, it leads into the next generation of ASCII literature, where writers continue to borrow from one another, to build cultural mythologies

collectively, and to follow the hacking tenets set out by Steven Levy: 'All information should be free. Mistrust authority – promote decentralization. You can create art and beauty on a computer.'

Evaluating Unmasking for Cross-Genre Authorship Verification

Kestemont, Mike
mike.kestemont@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Luyckx, Kim
kim.luyckx@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Daelemans, Walter
walter.daelemans@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Crombez, Thomas
thomas.crombez@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

In this paper we will stress-test a recently proposed technique for computational authorship verification, 'unmasking' (Koppel et al. 2004, 2007), which has been well received in the literature (Stein et al. 2010). The technique envisages an experimental set-up commonly referred to as 'authorship verification', a task generally deemed more difficult than so-called 'authorship attribution' (Koppel et al. 2007). We will apply the technique to authorship verification across genres, an extremely complex text categorization problem that so far has remained unexplored (Stamatatos 2009). We focus on five representative contemporary English-language authors. For each of them, the corpus under scrutiny contains several texts in two genres (literary prose and theatre plays).

1. Background: cross-genre authorship verification

In authorship verification, the given text may have been written by one of the candidate authors, but could also be written by none of them. Note that this *open case* scenario is typical of forensic applications: the author of e.g. a bomb letter is not necessarily among the suspect candidate authors. In the case of a suicide letter (potentially faked by a murderer), it is highly likely that this is the only suicide letter the victim ever wrote. In absence of similar material, it is difficult to extract reliable style markers from pre-existing writings to determine authorship of the letter.

Authorship *across genres* is an issue that is being paid all too little attention in present-day research. The few remarks that have been made on this issue agree that authorship attribution is difficult within a single textual genre, even more difficult with several topics involved, and likely to be extremely difficult with several genres involved (Luyckx & Daelemans 2011). Although it is generally assumed that an author will display stable style characteristics throughout his oeuvre, irrespective of genre, this remains speculative in the absence of systematic empirical investigation. Consequently, cross-genre authorship verification deserves much more attention than it has attracted so far.

2. Unmasking

Unmasking is a fairly complex meta-learning approach to authorship verification. Koppel et al. (2007) observed in earlier experiments that a small number of features had a lot of discriminatory power. It is indeed common for authors to use 'a small number of features in a consistently different way between works'. Such features often relate to topic-related, narrative, or thematic differences. As a result, a limited number of features can wrongfully maximize the differences in writing style between two works of identical authorship.

The unmasking approach tests the *robustness* of a stylistic model by deliberately impairing it over a number of iterations, each time removing those features that are most discriminative between the two texts. The resulting 'degradation curves' display many sudden drops in accuracy: when the most telling features are removed during each iteration, it becomes increasingly difficult to differentiate between two texts. In the case of two texts of non-identical authorship, however, a far larger number of features is discriminative, causing less dramatic drops in accuracy during degradation. Using training material in the form of a series of same-author and different-author degradation curves, Koppel et al. (2007) try to verify whether previously unseen degradation curves are of (non-)identical authorship.

The unmasking technique is especially attractive for authorship verification across genres, because of the interference between genre markers and authorial style markers. It might help remedy genre-related artifacts in that superficial genre-related differences between same-author texts in different genres will be filtered out easily and removed from the model early in the degradation process. After the removal of these non-essential stylistic features, one could hypothesize that only features more relevant for authorial identity will be preserved.

3. Methodology and evaluation

Our unmasking implementation closely adheres to the original description of the procedure. In the experiments, we have used the same generic parameter settings as tentatively adopted by Koppel et al. (2007): a *chunk size* of 500 tokens, n=250, m=10 and k=3. The main difference is, that a 'leave-one-text- out validation' is carried out on these curves for evaluation purposes, whereas k-fold cross-validation was applied in the original paper. We train an SVM classifier on the training curves and have it classify each of the test curves as a *same-author* or *different-author* curve. When all predictions have been collected, one can report on the overall classification accuracy and macro-averaged F1-score.

4. Corpus and selection of texts

The corpus we collected for the experiments in cross-genre authorship verification consists of published texts by five contemporary authors: Edward Bond, David Mamet, Harold Pinter, Sam Shepard, and Arnold Wesker. The main criterion for selecting an author was the availability of texts in more than one literary genre. Theatre and prose were the genres these five authors were most productive in, so these were chosen for the experiments. In our corpus, applying a text length threshold of 10,000 words (cf. Sanderson & Guenter 2006) resulted in 11 prose works and 23 theatre plays. We experimented with a complete matrix of authors and genres, allowing both intra-genre and cross-genre experiments for all authors. Digitization of the material involved three steps: scanning, OCR'ing, and manual post-correction.

5. Intra-genre experiments

Figure 1 shows degradation curves for an experiment on the eleven prose works in the corpus. Solid lines represent *same-author* curves, whereas dotted lines represent *different-author* curves. All curves display downward slopes, with decreasing cross-validation accuracies, as more predictive features get eliminated in each iteration. For *same-author* curves, however, it is clearly visible that the effect of degradation generally sets off sooner and more dramatically. *Different-author* curves are more robust and yield higher cross-validation accuracies, even when a large number of strongly discriminative features is deleted. Intersections between both curve types are minimal. A leave-one-text-out validation test on this set of curves confirms the success of the approach: the overall accuracy amounts to 96%, which is only just over the F1 score of 95%. This result confirms the

potential of unmasking for authorship verification in prose work collections.

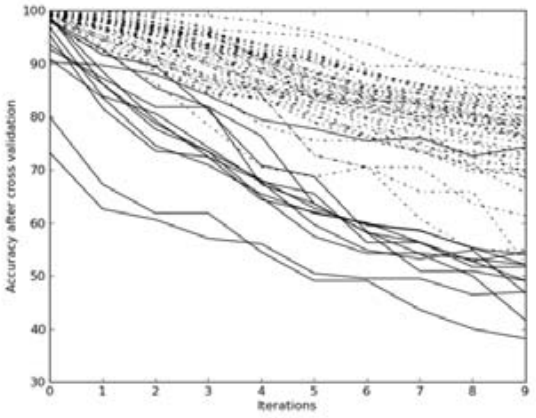


Figure 1: Unmasking on prose texts by five authors

A second experiment has been carried out on the 23 theatrical works in the corpus. Figure 2 displays a much less clear-cut differentiation of the *same-author* curves and their *different-author* counterparts, suggesting that the unmasking approach (with its default settings) is less effective for the theatrical section of the corpus. The leave-one-text out validation confirms this, yielding an overall accuracy of 84% and an F1 score of 62%.

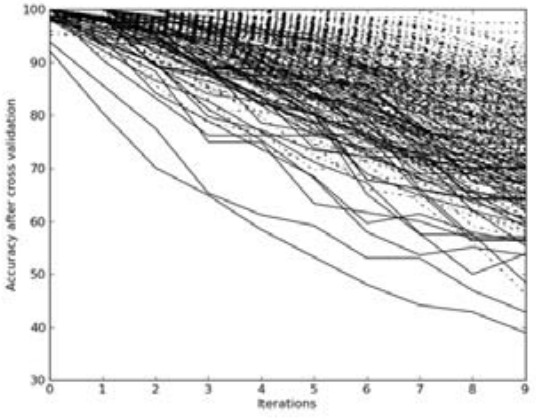


Figure 2: Unmasking on theatre plays by five authors

6. Cross-genre experiments

So far, the unmasking procedure has been mainly investigated for text pairs within the same text variety, although Koppel *et al.* (2007) report on a successful application of the technique to Hebrew-Aramaic texts across different topics. It is an interesting question whether the degradation differences between *same- author* and *different-author* curves would also hold for pairs of texts that do not belong to the same genre. A leave-one-text-out validation, however, shows poor performance of unmasking in this experiment, with an overall accuracy of 77% and a macro-averaged F1 of 56%.

7. Interpretation

After unmasking has been applied, the individual degradation curves allow for interpretation of results. Figure 3 visualizes the elimination process for Pinter's play *The Caretaker* and Mamet's prose text *The Old Religion*, who were personal friends. Mamet even acknowledged Pinter as a key influence on his work. The limited degradation in accuracy demonstrates that these Mamet and Pinter texts appear to adopt well-distinguishable styles. Figure 3 shows early elimination of names of principal characters (*davies*, *mick* and *aston* vs. *mark* and *pete*), personal pronouns that relate to a text's narrative perspective (*i, you*), and colloquial language (*aint*). Moreover, typical genre-features (e.g. the director's indication *pause*) are deleted as anticipated.

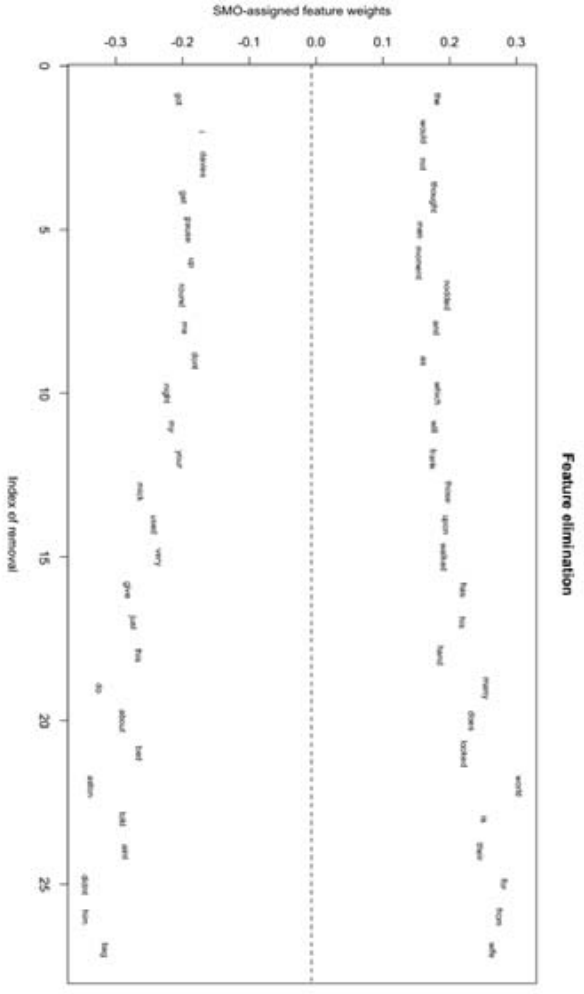


Figure 3: Visualization of the feature elimination process for Pinter's play The Caretaker and Mamet's prose text The Old Religion

8. Conclusion

The experiments reported on in this paper confirm that unmasking is an interesting technique for computational authorship verification, especially yielding reliable results within the genre of (larger) prose works in our corpus. Authorship verification, however, proves much more difficult in the theatrical part of our corpus. The original settings for the various parameters often appear to be genre-specific or even author-specific, so that further research on optimization is desirable. Finally, we have shown that interpretability is an important asset of the unmasking technique.

Funding and Acknowledgements

Kestemont is a Ph.D. fellow of the Research Foundation – Flanders (FWO). The research of Luyckx and Daelemans is partially funded by the FWO project 'Computational Techniques for Stylometry for Dutch'. The research by Crombez is partially funded by the FWO project 'Mass Spectacle in Flanders'. The authors would like to acknowledge Sarah Bekaert's work on the digitization of the corpus.

References

Koppel, M., J. Schler, and E. Bonchek-Dokow (2007). Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research* 8: 1261-1276.

Luyckx, K., and W. Daelemans (2011). The Effect of Author Set Size and Data Size in Authorship Attribution. *Literary and Linguistic Computing* 26(1): 35-55.

Sanderson, C., and S. Guenter (2006). Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia, pp. 482-491.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60(3): 538-556.

Stein, B., N. Lipka, and P. Prettenhofer (2011). Intrinsic Plagiarism Analysis. *Language Resources and Evaluation* 45(1): 63-82.

Koppel, M., and J. Schler (2004). Authorship Verification as a One-Class Classification Problem. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada,

Literary Wikis: Crowd-sourcing the Analysis and Annotation of Pynchon, Eco and Others

Ketzan, Erik
eketzan@lavabit.com
Institut für Deutsche Sprache, Germany

1. Introduction

Annotation of complicated texts – the *Bible*, the works of Shakespeare, experimental fiction – is a familiar concept. In 2005 I had an idea: what if I used a wiki to create such a guide? Would anyone contribute to it? Would anyone read it? Would the results be any good? Since then, hundreds of users have annotations thousands of pages of experimental fiction by the authors Thomas Pynchon and Umberto Eco, and the projects *PynchonWiki* and *Umberto Eco Wiki*.

What have we learned from these projects? What can other digital humanities and crowd-sourcing projects learn from their successes and failures? The Eco wiki project is still ongoing, but I plan to present my findings and insights at DH 2012.

2. The Queen Loana Wiki

When Umberto Eco's novel, *The Mysterious Flame of Queen Loana,* was published a few months later, I launched what I called the *Queen Loana Annotation Project*, a wiki organized by chapter and page.[1] Eco's novel was a perfect test case for the experiment I envisioned, a literary annotation wiki. First, Eco frequently quotes texts without attribution, which led to wiki entries like:

P. 15, 'you always said you could resist anything but temptation'

quotation from *Lady Windermere's Fan* by Oscar Wilde.

Second, many references in the novel were confusing to readers, making my wiki a *useful* resource. According to a *Village Voice* review at the time, 'Early reviews have dismissed *Mysterious Flame* as nostalgic and at times so personal as to be impenetrable. Eco concedes he wrote it with his own generation in mind. "It's a book for Italian people of my age".' Thanks to the wiki, though, readers could easily read up on all those references, with entries like this:

P. 382, 'Rina Fort'

On November 30, 1946, a woman and her three children were found beaten to death in their apartment in Milan. In January, 1950, Caterina Fort, the lover of the woman's husband, was condemned to life imprisonment for the murders.

Third, Eco has written dozens of books and he frequently references the same incidents, themes and ideas across them. For example:

P. 297, A rose by another other name

from *Romeo & Juliet*.

Eco remarked in *Postscript to The Name of the Rose* that everyone assumed that the title, *Name of the Rose*, was a reference to this same line of Juliet. Eco emphasizes, however, that he meant his title to mean the exact opposite: names *are* important.

This kind of entry requires more than just Googling. Only someone truly familiar with Eco's previous works could have written it, thus adding another *useful* element to the wiki on par with traditional published literary analysis and commentary.

In short, the *Queen Loana Annotation Project* was a modest success. A couple dozen contributors from around the globe added hundreds of high-quality entries. This group of contributors, which briefly felt like a 'community', annotated seemingly every last historical, literary, and artistic reference in Eco's novel. The quality of the annotations was also surprisingly good, on par with a publishable guide book.

3. Pynchonwiki.com

The next year, 2006, a musician and writer named Tim Ware converted his extensive online notes on the novels of Thomas Pynchon into a wiki, which he aptly named Pynchonwiki.com. The works of Pynchon were also perfect for a literary wiki, as Pynchon's novels teem with countless obscure references which, when analyzed, often illuminate his bigger themes and messages. I joined Pynchonwiki as one the earliest contributors (under the handle 'Bleakhaus') and created a section for page-by-page annotation in the manner of the *Queen Loana Annotation Project*. This one took off beyond the wildest expectations of a literary annotation wiki geek.

According to Ralph Schroeder and Matthijs den Besten, scholars at Oxford's Internet Research Institute and e-Research Centre who published a paper[2] on Pynchonwiki, some successes of Pynchonwiki include:

- Over 235 contributors
- Over 455,000 words (over twice as long as *Moby Dick*)
- 1350+ alphabetical entries
- 4000+ page-by-page entries
- >10,000+ edits

Schroeder and den Besten also concluded that:

- 'the Wiki had covered all pages within three months of publication'
- 'the Pynchon Wiki can be put into the context of growing interest in computer supported collaboration and e-Research'
- 'A common theme emerging from the facts and figures presented is the tremendous attraction of the page-by-page style of annotation in the "Against the Day" wiki'
- 'Wikis thus seem to be a good tool [for tasks] where endless detective work is called for, and this may apply to other areas of e-Research or online collaboration'
- 'the Wiki is bound to encourage learning among contributors'

Section for Pynchon's other novels followed shortly at Pynchonwiki.com, and in short order there existed annotations for almost all of the thousands of pages in Pynchon's oeuvre.

4. What next? Umberto Eco Wiki and linguistic exploration

Over the next few years, I thought vaguely about how the lessons and successes of the *Queen Loana Annotation Project* and *Pynchonwiki* could be repeated. I created a site, Literarywiki.com, for my own experiments in annotating different kinds of texts. But I quickly discovered that most novels do not benefit from extensive commentary or annotation. Only certain kinds of novels, those that teem with information, references, and/or enigmas, benefited from such supplementary information. The novels of Pynchon and Umberto Eco, I came to believe, were exceptions rather than the rule.

To explore these issues in greater depth, I will launch, this coming week, the first of November, 2011, a new website named Umberto Eco Wiki, which I am creating through the support of my current employer, the Institut für Deutsche Sprache (Mannheim, Germany).[3]To coincide with the German- and English-language editions of Umberto Eco's new novel, *The Prague Cemetery*, I will launch this wiki in the spirit of the above-named projects.

Umberto Eco Wiki has secions for German and English annotations, and I plan to explore how

successful and useful such a resource in multiple languages can be.

My proposed short paper for Digital Humanities 2012 will document the history of the literary wiki as summarized above, relate the progress on the new Umberto Eco Wiki, explore issues relating to the literary wiki concept, including how its lessons can be applied to other digital humanities areas.

Notes

1. The Mysterious Flame of Queen Loana Annotation Project was initially online at http://queenloana.wikispaces.com/ . Later it was moved to my own Literarywiki.org, and its contents will soon be migrated to the Umberto Eco Wiki (a project of the Institut für Deutsche Sprache), at http://eco.ids-mannheim.de
2. Schroeder, R., and M. L. den Besten (2008). Literary Sleuths Online: e-Research Collaboration on the Pynchon Wiki. Information, Communication & Society 11(2): 25-45. Available at SSRN: http://ssrn.com/abstract=1086671
3. Umberto Eco Wiki, available at http://eco.ids-mannheim.de

Social Network Analysis and Visualization in 'The Papers of Thomas Jefferson'

Klein, Lauren Frederica
lauren.klein@lcc.gatech.edu
Georgia Institute of Technology, USA

The silences endemic to the archive of slavery have long presented a range of challenges to the literary scholar: How does one account for absences in the archival record, both those inscribed in the archive's contents, and those introduced at the time of the archive's construction? How does one account for the power dynamics at work in the relationships between the enslaved men and women who committed their narratives to paper, and the group of (mostly white) reformers who edited and published their works? How does one identify and extract meaning from the unique set of documents that do remain – letters, inventories, ledger books, and personal narratives – documents that, in the words of Susan Scott Parrish, the literary critic, we must struggle to make 'mean something more?' (2010). And how does one do so without reinforcing the damaging notion that African American voices from before emancipation – not just in the archival record, but the voices themselves, are silent, and irretrievably lost?

This final, critical challenge is what has prompted scholars from across the humanities, such as the literary critics Stephen Best (2009, 2011) and Saidiya Hartman (2008), the sociologist Avery Gordon (2008), the archivist Jeanette Bastian (2003), and the historian Jill Lepore (1998), to call for a shift away from what Best has identified as 'a logic and ethic of recovery,' to a new focus, instead, on animating the mysteries of the past (2011). However, each of these scholars proposes traditional methods of analysis and criticism for animating such mysteries.

This short paper will instead draw upon a set of tools and techniques associated with social network analysis and visualization in order to propose a new method for animating the mysteries of the past. While social network analysis and visualization tools have been applied to literary texts (Drouin 2006-), historical datasets (Wilder 2010-), and correspondence networks (Findlen et al. 2008-), these tools have not yet been applied to historical archives with the aim of illuminating the relationships among people mentioned in the archive's content.

For this task, I employed a named entity recognizer (NER) developed at Stanford, and included in its suite of CoreNLP (Natural Language Processing) tools, in order to identify the names of the people mentioned in the archive's content (Manning et al. 2010). I then developed my own co-appearance script, in Python, in order to determine which people were mentioned in the same document as each other, and how many times those people appeared together. I then formatted the data to be displayed using Protovis, a javascript-based visualization toolkit also developed at Stanford (Bostock & Heer 2010).

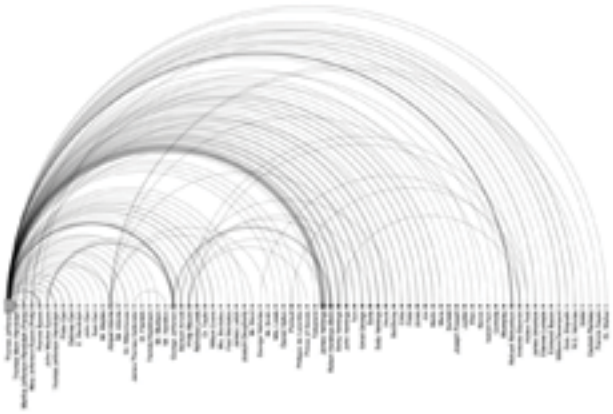


Figure 1

My paper will focus on a subset of documents included in *The Papers of Thomas Jefferson*, those that make reference to members of the enslaved Hemings family (Figure 1). The descendants of Elizabeth Hemings, including Sally Hemings, the woman with whom, whether consensually or not, Jefferson maintained a lifelong relationship and bore six of his children; and James Hemings, who trained as a *chef* de cuisine in the high French style, and then worked as the head cook at Jefferson's Monticello plantation until his emancipation, in 1796, function as an ideal test case both for assessing the limits of the computing methodologies employed, and for determining how the analysis and visualization of social networks can, in fact, animate the mysteries of the past – those that Best, Hartman, and others, have sought to identify and amplify through more traditional scholarly techniques.

A significant limit of employing NER software on this particular dataset is that enslaved men and women were most often referred to by first name alone. For instance, not only are there many Jameses mentioned in *The Papers of Thomas Jefferson*, but James Hemings was also called Jamie, Jimmy, and even *Gimmé* while in France. After running my co-appearance script, I was required to go through the results by hand in order to determine which names referred to the same individual, and in that event, combine the associated data. Because of the eighteenth-century style of the Jefferson Papers, I was also required to remain attentive to any NER errors, as well as to individuals clearly alluded to in the Papers, but not referred to by name at all.

The test case of the Hemings family also points to the limits of the 'arc diagram' model for visualizing complex relationships. While preferable to a force-directed layout, which obscures the distinct relationships among individuals, its layout implies a linearity that is not consistent with the web of relationships contained within the dataset. My paper will thus also propose how hive plots and chord diagrams might function as better models for visualizing such network data (Krzywinski 2009, 2011). In this way, I will suggest how contemporary data visualization research can have a direct impact on a range of literary and cultural investigations.

Furthermore, by demonstrating how a visualization of this particular social network allows us to see the historical traces of the Hemings family as presences, not absences, as scholars such as Best, Hartman, Gordon, Bastian, and Lepore, have each argued for, I begin to address an issue of mounting concern within the community of digital humanities scholars, as voiced by Johanna Drucker (2009) and Alan Liu (2011), among others, about the need to reinscribe humanistic inquiry as the central focus of digital humanities work. The presence of the Hemings family in *The Papers of Thomas Jefferson*, as unearthed by the methods described above, suggests not only how digital tools might be applied to a diverse set of cultural and literary questions, but also how these methods might expose – and animate – the diversity of experience embedded in significant records of early American cultural life.

References

Bastian, J. (2010). *Owning Memory, How A Caribbean Community Lost Its Archives and Found Its History*. Westport, CT: Libraries Unlimited.

Best, St. (2011). Neither Lost nor Found: Slavery and the Visual Archive. *Representations* 113(1): 150-163.

Best, St., and M. Sharon (2009). Surface Reading: An Introduction. *Representations* 108(1): 1-21.

Bostock, M., and J. Heer (2010). *Protovis*. `http://mbostock.github.com/ protovis/`

Drouin, J. (2006-). *Ecclesiastical Proust Archive*. http://www.proustarchive.org/

Drucker, J. (2009). *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. Chicago: U of Chicago P.

Findlen, P., et al. (2008-). *Mapping the Republic of Letters*. https:// republicofletters.stanford.edu/

Gordon, A. (2008). *Ghostly Matters: Haunting and the Sociological Imagination*. Minneapolis: U of Minnesota P.

VariaLog: how to locate words in a French Renaissance Virtual Library

Lay, Marie Hélène
marie-helene.lay@univ-poitiers.fr
University of Poitiers, France

1. Introduction

The efficiency of search engines is based on the principle that the information sought can be retrieved by 'looking for words' conveying the information and that these words can be identified thanks to the string of characters they are comprised of. This view takes for granted that the words are always spelt in the same way and that they comply with orthographic rules.

Such is not the situation which prevails for the texts produced during the French Renaissance period. Therefore the availability of older texts for purposes of archiving and disseminating the cultural heritage tradition raises a particular problem. In texts edited in French before the 18th century, spellings are not consistent, as proper spelling has not been 'invented' yet. One and the same word may therefore be spelt in a variety of forms. This is not only a time-related variation, as would be expected from the evolution of the language between the 15th and the 17th century. In one and the same book many different spellings may be identified for the one and the same word: for the word *côté*, either *coté, cotté, cote, costé,* or *couste* could be used, the verb *savoir* may be spelt either *scavoir* or *sçavoir*, *'je sais'* may be spelt *'ie sçay'*, and its past participle *'su'* may appear as *'sceu'*.

It is therefore necessary to adapt search engines based on word form identification if they are to render the service expected. Several strategies can be envisaged and the purpose of this paper is to focus on those which resort to linguistic expertise, either included in the documents themselves (by annotation) or into the search engine (by query extension). The solutions considered are produced in the context of the Virtual Humanistic Library Project and its evolution (`http://www.bvh.univ-tours.fr/`). This part of the project called **VARIALOG**, is financed by a Google Digital Humanities Research Award.

2. Methodological alternatives

The BVH/VHL context, considered here, is that of a highly expert environment of a relatively moderate size aiming at a complete editorial treatment and the dissemination of annotated and validated resources. Within this context, two solutions have been designed:

1. Texts annotation with linguistic information gained from lemmatization. The forms retrieved, whatever their spelling, are lemmatized under a canonic form which then becomes the pivot of further requests: for example the lemma for *nuit* groups together forms like *nuits* (which is 'regular french') or *nuyctz* (old written form). A first solution, **HUMANISTICA** (Lay 2000) was based on the adaptation of a probabilistic tagger/lemmatizer. The results achieved were satisfactory but the adaptation of the analyzer had to be started all over again to take into account specific features of this high heterogeneous corpus. Another solution, **ANALOG** (Lay 2010a, 2010b) was therefore developed. It provides an annotation computer assisted environment, and is currently being used in the BVH project. But the enrichment of text through linguistic annotation is a slow and costly process. Though this solution is very useful to go on producing a reference environment, it is nonetheless desirable to provide efficient query tools on texts already available but not yet annotated.
2. Query extension, without requiring the lemmatization process. The aim is not to produce exactly the right forms (like in EEBO -VosPos-, Impact, ToTrTaLe, LGeRM, or for old czech, or old German projects). We will do so, in order to help in an editorial process (**DISSIMILOG**), but here, we just want to spot all the written forms which could correspond to a query, being insensitive to variation.

3. VariaLog : Principles

To solve the problem of spelling variation, one has to go back to observational evidence. Two directions may be taken in this respect: either observe the texts or observe the variants attested for a given form.

1. Concerning text observation, the aim is to evaluate the number of forms for a given text which do not correspond to the norm. Moreover, one must take into consideration the extent to which the texts can be compared. We intend to illustrate this with two short extracts from Montaigne and Rabelais, two authors of paramount significance.
2. Concerning the observation of variants attested for one word, the idea is to formulate the rules which govern the production of abnormal forms. We will then build rules to extend queries, turning the search of a word into the search of all the forms assumed by this word, and match the results with forms in texts.
   Comparing the searched forms and their spelling in text, a typology of the situations occurring may be offered. The form being searched is the same one (*raisons/raisons*) or the link can be very weak (*impératrice/empériere*). Between these two types, a whole gradation of situations can be organised on a linguistic basis: relations between sounds and different ways of spelling in modern french (*c=ss; n=nn; r=rr; s=z; t=th; ai=ei,ai,ey,ay,oi,oy; [uv]=u,v; u=eu)*, flexionnal history (*serais/seray/serois)* and morphological history (*hôpital/hospitalier*; *forêt/forestier; advis/avis)*. Due to the structural instability of this linguistic data, equivalences between character strings are difficult to track statistically and no model-based approach can be developed. But linguistic knowledge helps recognize regular replacement patterns, which can be turned into rules.
3. The next point which needs to be taken into account is the relevance of the rules: they have to help find all the forms concerned (low silence, good recall), and to avoid generating too much noise (good precision).

To test the first results, a small corpus of 7 words (*vices/une/face/fesse/lu/vu/souverain*) has been transformed by the substitution rules mentionned above. The results do contain all the relevant forms, but the 7 words have been extended to 118445 forms. There is obviously some correlation between the length of the word and the number of generated words due to the combinatory process.

The solution chosen to fix that problem is to describe, for each rule, the context in which the substitution is allowed. This aims at constraining their application strongly, and limits their productivity. This contextualisation is based on a good knowledge of the linguistic process involved. In the example given below, 8 simple rules are transformed into 9 more complex rules. Most of the time, one simple rule will be derived into 5 to 15 contextualised rules.

(?<\=[bdflmnprstv])u=eu	ain=ein,ain,eyn,ayn
(?<\=[aeiouy])c(?\=[eiy])=ss	^s(?\=[eiy]) = c
(?<\=[aeiouy])ss(?\=[eiy])=c	(?!^.+)v = u
(?!^.+)n(?<!.+$)=nn	^u = v
(?!^.+)r(?<!.+$)=rr	s$ = z

The results achieved are satisfactory: the rules produce all the linguistically permissible variants, and the number of variants is much lower. The 7 words generate 37 forms.

4. VariaLog : Tool description

The tool itself is thought to be really user-friendly especially for the tuning of rules and the evaluation of their consequences (efficiency and non regression tests). It is a free available java program which first transforms a list of words into an extended list of forms, using that for a rules set. Having done this, the need is to localise the different forms attested in the old spelling in a text, according to the requested form. The output file of this last part of the process is an html file with a graphical highlighting (or bold character) of the identified variant. Moreover, each form is connected to a bubble showing the rules used to derive the variant. A table containing the summary of the used rules for the text is also available: the human validation process is quite friendly. This tool is being put forward to be integrated to an XTF platform.

5. Conclusion

Using a rule-based approach, VariaLog is designed to identify all the written forms that are likely to correspond to a query, since it is insensitive to variations in spelling. The recall rate (tested on 5000 forms) provides evidence that all the linguistically permissible variants in French are produced by the rules, so long as the problem is simply one of spelling (*nuit/nuyct*)and not a morphological one (e.g. *impératrice/empérière*). As far as precision is concerned, the rules may sometimes generate more ambiguity than anticipated. If 'o' becomes 'ou', then, *école* becomes *écoule*, which is not an acceptable variation, but *volant* becomes *voulant*, which is an acceptable variation; as a result, *volant* will correspond to *vouloir* ('want'), thus increasing the ambiguity of this form which means already 'flying, robbing, wheel, flounce, shuttlecock'. The generated ambiguity is no different from standard ambiguities, even in an orthographic environment.

Already used to search several French dialects, VariaLog can be used to process any form of spelling variation, in any language. One just needs to adjust one's own specific spelling rules or dictionary. Our aim is to help locate spelling variation efficiently. User feedback is most welcome.

References

Baron, A., and Rayson, P. (2009). http://ucrel.lancs.ac.uk/publications/cl2009/314_FullPaper.pdf In M. Mahlberg, V. González-Díaz, and C. Smith (eds.), *Proceedings of the Corpus Linguistics Conference, CL2009*. University of Liverpool, UK, pp. 20-23.

Burnard, L. (1995). http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_en.xml *Proceedings of the Second Language Engineering Conference*, 1995.

Craig, H., and R. Whipp (2010). http://llc.oxfordjournals.org/content/25/1/37.short *Literary and Linguistic Computing* 25(1): 37-52.

Demonet, M.-L., and M. H. Lay (2011). Digitizing European Renaissance prints: a 3-year experiment on image-and-text retrieval. *International Workshop on Digital Preservation of Heritage* (IWDPH07). Kolkata, 2007.

Erjavec, T. (2011). http://www.aclweb.org/anthology/W/W11/W11-1505.pdf *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2011, Portland, pp. 33-38.

Hana, J., A. Feldman, and K. Aharodnik (2011). http://www.aclweb.org/anthology-new/W/W11/W11-1502.pdf *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 2011, Portland, pp. 10-18.

Lay-Antoni, M.-H. et al. (2000), http://lexicometrica.univ-paris3.fr/jadt/jadt2000/pdf/89/89.pdf *jadt 2000,* Lausanne.

Lay, M.-H., et al. (2010). http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-1045-1056_106-Lay.pdf *jadt 2010,* Rome.

Sánchez Marco, C., G. Boleda, and L. Padró (2011). http://ilk.uvt.nl/LaTeCH2011/slides/01_Sanchez-Marco_etal.pdf *ACL-HLT Workshop*, 2011, Portland, pp. 1-9.

Scheible, S., R. J. Whitt,M. Durrell, and B. Bennett (2011). Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. *ACL-HLT Workshop*, 2011, Portland, pp. 10-18.

Souvay, G. and J.M. Pierrel (2009). http://www.doaj.org/doaj?func=abstract&id=812845 *TAL* 50(2): 149-172.

Thaisen, J. (2011). http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-346.xml;query=;brand=default *Digital Humanities Conference Abstracts, 2011,* Stanford.

http://www.bvh.univ-tours.fr/

http://www.c-tei.org/

http://www.bvh.univ-tours.fr/XML-TEI/index.asp

http://www.bvh.univ-tours.fr/Epistemon/philologic.asp

http://xtf.cdlib.org/documentation/programming

http://www.monkproject.org/

http://eebo.chadwyck.com/help/whatis_wh.htm

http://impactocr.wordpress.com/

http://panini.northwestern.edu/mmueller/vospos.pdf

DeRiK: A German Reference Corpus of Computer-Mediated Communication

Lemnitzer, Lothar
lemnitzer@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Beißwenger, Michael
michael.beisswenger@uni-dortmund.de
Technische Universität Dortmund, Germany

Ermakova, Maria
ermakovamd@googlemail.com
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Geyken, Alexander
geyken@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Storrer, Angelika
angelika.storrer@uni-dortmund.de
Technische Universität Dortmund, Germany

1. Project background and focus of the paper

In view of the increasing amount of reading and writing that people do on the Internet, corpus designers who set out to provide balanced corpora that include all relevant text types of contemporary language should also include samples of genres of computer-mediated communication (CMC) such as e-mail, weblogs, microblogging on Twitter, discussion boards and wiki discussions, chats and instant messaging conversations, and communication in social network sites. In our paper we present selected aspects of an ongoing project that aims at building a reference corpus of German CMC, called *DeRiK* ('Deutsches Referenzkorpus zur internetbasierten Kommunikation').[1] DeRiK is a joint initiative of TU Dortmund University and the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW). The corpus will be integrated into the lexical information system provided by the BBAW project *Digitales Wörterbuch der deutschen Sprache* (*DWDS*, www.dwds.de (www.dwds.de)).[2]

In our paper we will focus on the role of the DeRiK component in the DWDS framework (section 2) and

on CMC-specific issues of corpus annotation (section 3).

2. Integrating CMC discourse into a corpus of contemporary German: motivation and application fields

DWDS (www.dwds.de (`www.dwds.de`)) is a lexical information system developed by and hosted at the BBAW. The system offers one-click-access to three different types of resources (Geyken 2007):

a) *lexical resources:* a common language dictionary,[3] an etymological dictionary, and a thesaurus;

b) *corpus resources:* a balanced reference corpus (called 'DWDS core corpus') of German ranging from 1900 up to now, a set of additional newspaper corpora, and specialized corpora;

c) *statistical resources* for words and word combinations.

These resources are displayed alongside one another in separate panels (cf. Fig. 1). The system offers the choice among several views, i.e. between several profiles with predefined panel combinations.

Figure 1: Web frontend of the DWDS system (`http://www.dwds.de`)

The CMC component *DeRiK* ('Deutsches Referenzkorpus zur internetbasierten Kommunikation') will be integrated into this framework both as an independent panel and as a subcorpus of the DWDS core corpus. The sampling of the CMC genres in DeRiK is guided by the findings of the 'ARD/ZDF-Onlinestudie', a German online usage survey conducted once per year (www.ard-zdf-onlinestudie.de). The findings of this survey allow us to derive a key for the composition of our corpus. However, for practical reasons the project will set out to acquire data of only those CMC applications on the Internet for which the users have explicitly granted permission for (re-)distributing and (re-)using their written utterances for non-commercial purposes/ academic research.[4] The first partial corpus of DeRiK will include discourse from *Wikipedia* talk pages, a selection of forum and weblog discussions, chat conversations, and postings of selected *Twitter* users who have licensed their tweets under Creative Commons. We hope that in the long term changes in IPR restrictions within the domain of discourse on the Internet will enable us to apply more principled methods of data sampling.

The integration of the CMC reference corpus into the DWDS system may be valuable for various research and application fields, for example:

a) Language variation, language change and stylistics: A general-language corpus that includes a CMC component will provide a broad empirical basis (a) for further, corpus-based investigations of the usage and dissemination of CMC-specific phenomena across linguistic varieties and digital genres, and (b) for comparative analyses of the features of CMC discourse and of "traditional" written genres (e.g. newspaper, fiction, scientific writing, nonliterary prose); it will thus facilitate to track and describe how new linguistic patterns and communicative genres emerge.[5]

b) Lexicology and lexicography: Besides genre-specific discourse markers and netspeak jargon (like 'lol' *laughing out loud* or 'imho' *in my humble opinion*), new vocabulary is characteristic for CMC discourse, e.g. 'funzen' (an abbreviated variant of 'funktionieren' *to function*) or 'gruscheln' (a function of a German social network platform, most likely a blending of 'grüßen' *greet* and 'kuscheln' *cuddle*). There are also CMC-specific processes of lexical-semantic changes, e.g. the broadening of the concept of 'Freund' (*friend*). Up-to-date lexical resources should document and describe these tendencies by integrating CMC data into their data basis. Once the first partial corpora of the DeRiK corpus are made available in the DWDS system, it is intended to extend the DWDS dictionary component with entries describing new lexemes that have evolved from CMC discourse. In addition, the DWDS corpus system will then allow one to track how new vocabulary from CMC discourse (such as the examples mentioned above) spreads into 'traditional' genres (e.g. newspaper, fiction, nonliterary prose).

c) Language teaching: CMC has become an important part of everyday communication. Language- and culture-specific properties of CMC should thus also be regarded in communicative approaches to Second Language Teaching. In this context, the DeRiK corpus and the documentation of CMC vocabulary in the DWDS dictionary may be useful resources. In school teaching, students with

German as a native language may use the DWDS system to compare written language with CMC and to explore how style varies across different genres.

3. Annotation of CMC-specific phenomena

One advantage of integrating DeRiK into the DWDS system is that users can profit from the DWDS corpus annotation and querying facilities: The corpus resources which are currently available in the DWDS system are lemmatized with the *TAGH* morphology (cf. Geyken & Hanneforth 2006) and tagged with the part-of-speech tagger *moot* (cf. Jurish 2003). The corpus search engine *DDC* (*Dialing DWDS Concordancer*) supports linguistic queries on several annotation levels (word forms, lemmas, STTS part-of-speech categories), filtering (e.g. by text type) and sorting options.

Since all corpus resources in the DWDS system are encoded according the guidelines of the Text Encoding Initiative (TEI-P5), the project aims to use and customize TEI for the appropriate base-level annotation of the CMC sub-corpus. For this purpose, we have developed a TEI-compliant annotation schema that

- provides a macro-structure of CMC discourse which should cover as many genres as possible (see section 3.1);
- provides a suggestion for the description of CMC-specific phenomena which is oriented mainly on surface features (for instance, the annotation will cover interaction signs of various types; see section 3.2 for details).

The TEI-related details of this schema are described in Beißwenger et al. (2012).[6] The discussion in this paper will focus on two annotation issues: The representation of CMC-specific micro- and macrostructures (section 3.1) and the annotation of typical 'netspeak' elements (section 3.2).

3.1. Annotation of CMC-specific micro- and macrostructures

We introduced the category *posting* as a basic element to capture CMC micro- and macrostructures. A posting is defined as a content unit that is being sent to the server 'en bloc'. Postings can usually be recognized by their formal structure, even if they have different forms and structures across CMC genres. This facilitates the automatic segmentation and annotation of CMC micro- and macrostructures.

We use the term *microstructure* to refer to the internal structure of postings. There are cases in which a posting consists of exactly one portion of text. In other CMC genres, e.g. in discussion groups, postings may contain divisions and markup used by the author to structure their content.

We use the term *macrostructure* to describe how the postings are sequenced. While microstructures are generated by an individual author, macrostructures do not emerge from the actions of just *one* user but from all posting activities of *all* users involved in a CMC conversation plus server routines for ordering the incoming postings.

In our TEI schema, we represent structures on the microstructure level (which result from the planning and composition decisions of *one* author) using the <*p*> element ('paragraph') from the TEI standard. Structures on the macrostructure level, in contrast, are described using the <*div*> element ('division'). In addition, we differentiate between two major types of CMC macrostructures:

- *logfile* structures, which arrange the postings in a linear chronological order based on when they reached the server.
- *thread* structures, which arrange the postings in a sequence and use two dimensions with specific semantics: the *above/below* dimension representing a temporal 'before/after' relation; the *left/right* dimension (by indentation), which usually symbolizes the topical affiliation of one posting to a previous posting.

3.2. Annotation of interaction signs

The corpus-based investigation of 'netspeak' jargon is interesting in many research contexts (style variation and language change, discourse management, language teaching etc.). Our annotation schema copes for a set of 'netspeak' phenomena which we term 'interaction signs'. The term builds on the category 'interaktive Einheiten' which has been introduced in the *Grammatik der deutschen Sprache*, a three-volume scientific grammar of the German Language (cf. Zifonun et al. 1997), to classify *interjections* (such as 'hm' or 'oh my god') and *responsives* (such as 'yes' and 'no') in spoken discourse[7]. In contrast to part-of-speech-categories, interaction signs are not syntactically integrated and do not contribute to the compositional structure of sentences. In spoken discourse, they serve as devices for conversation management, i.e. they can be used to express reactions to the partners' utterances or to display the speaker's emotions. Our category 'interaction sign' includes interjections and responsives as well as four CMC-specific classes (cf. Fig. 2):

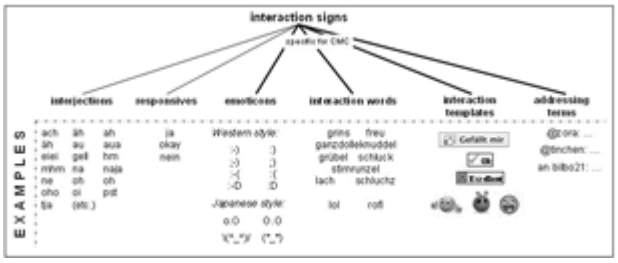


Figure 2: Typology of interaction signs (with examples)

1. *Emoticons* are iconic units that have been created with the keyboard and which typically serve as emotion or irony markers or as responsives. Being of iconic origin, the use of emoticons is not restricted to a specific language. However, different styles of emoticons exist – e.g, Western style emoticons such as *:-)*, *:-(*, *;-)* or the *:)*, or Japanese style emoticons such as *(^_^)*, *\(^_^)/*, *(*_*)*.
2. *Interaction words* are symbolic linguistic units whose morphologic construction is based on a word or a phrase. They may describe gestures or facial expressions, e.g. **g** (< "grins" *grin*), **fg** (< *fat grin*), *s* (< *smile*), or they are used for the simulation of actions and events.
3. *Interaction templates* are units that the user does not generate with the keyboard but which are generated automatically from a file with a previously prepared text or graphical element after the user has activated a template. Due to their generation from predefined templates, we do not classify them as emoticons, even though some of them may have similar functions. Amongst others, this category includes *.*gif* files. Many of them portray not just facial expressions but can depict almost everything; in the case of animated graphics they can even portray entire scenes as moving pictures.
4. *Addressing terms* are units which are used to address an utterance to a particular interlocutor. The most widely used form here is the one made out of the <@>-character together with a specification of the addressee's name.

4. Conclusion and outlook

Up to now, many assumptions about the Internet's impact on language change have been based upon small datasets. As a new component within the DWDS system, the DeRiK corpus is meant to be a resource for the investigation of language usage in CMC genres on a broader empirical basis. The annotation schema sketched in section 3 is used and evaluated in the ongoing work of the DeRiK project. The categories proposed in this schema will have to be further discussed within the CMC community. We consider the development of this schema as a first step towards the development of an annotation standard that will facilitate cross-language, cross-genre, and micro-diachronic investigations of CMC phenomena on the basis of corpora. The schema focuses on linguistic aspects, but it is open for extensions motivated through other fields of research, i.e. cultural studies or sentiment analysis.

References

Beißwenger, M. (2000). *Kommunikation in virtuellen Welten: Sprache, Text und Wirklichkeit*. Stuttgart: Ibidem.

Beißwenger, M., and A. Storrer (2008). Corpora of Computer-Mediated Communication. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook,* vol. 1. Berlin: de Gruyter, pp. 292-308.

Beißwenger, M., M. Ermakova, A. Geyken, L. Lemnitzer, and A. Storrer (2012). A TEI schema for the Representation of the Computer-mediated Communication. *TEI (Text Encoding Initiative) Journal*. (submitted 2012).

Biber, D., et al. (1999). *Longman Grammar of Spoken and Written English*. Edinburgh: Pearson.

Biber, D., et al. (2002). *Longman Student Grammar of Spoken and Written English*. Edinburgh: Pearson.

Blake, B. J. (2008). *All About Language*. New York: Oxford UP.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge UP.

Crystal, D. (2011). *Internet Linguistics. A Student Guide*. New York: Routledge.

[DUDEN-4[5]] **DUDEN** (1995). *Die Grammatik*. 5th ed. Mannheim: Bibliographisches Institut.

[DUDEN-4[7]] **DUDEN** (2005). *Die Grammatik*. 7th ed. Mannheim: Bibliographisches Institut.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In Ch. Fellbaum (ed.), *Collocations and Idioms*. London: Continuum, pp. 23-40.

Geyken, A., and Th. Hanneforth (2006). TAGH – A Complete Morphology for German based on Weighted Finite State Automata. In: A. Yli-Jyrä, L. Karttunen, and J. Karhumäki (eds), *Finite State Methods and Natural Language Processing – Proceedings of FSMNLP, 5th international workshop,* Helsinki 2005, Heidelberg: Springer (= Lecture Notes in Artificial Ingelligence 4002), pp. 55-66.

Greenbaum, S. (1996). *The Oxford English Grammar*. New York: Oxford UP.

Herring, S. C. (1996). Introduction. In: S. C. Herring (ed.), *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins, pp. 1-10.

Herring, S. C., ed. (1996). *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins.

Herring, S. C., ed. (2010). Computer-Mediated Conversation, Part I. *Special Issue of Language@Internet* 7. http://www.languageatinternet.org/articles/2010 (accessed 11 March 2012)

Jurish, B. (2003). A Hybrid Approach to Part-of-Speech Tagging, Final report. *Project 'Kollokationen im Wörterbuch'*. Berlin-Brandenburgische Akademie der Wissenschaften, Berlin. http://www.ling.uni-potsdam.de/~moocow/pubs/dwdst-report.pdf (accessed 11 March 2012)

McArthur, T., ed. (1998). *Concise Oxf ord Companion to the English Language*. Oxford: Oxford UP.

Reynaert, N., O. Oostdijk, H. de Clercq, et al. (2010). Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Paris, 2693-2698. Online: http://eprints.eemcs.utwente.nl/18001/01/LREC2010_549_Paper_SoNaR.pdf

Runkehl, K., T. Siever, and P. Schlobinski (1998). *Sprache und Kommunikation im Internet. Überblick und Analysen*. Opladen: Westdeutscher Verlag.

Schiffrin, D. (1986). *Discourse markers*. Cambridge: Cambridge UP.

Storrer, A. (2012, in press). Sprachstil und Sprachvariation in sozialen Netzwerken. In B. Frank-Job, A. Mehler, and T. Sutter (eds.), *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*. Wiesbaden: VS Verlag für Sozialwissenschaften.

[TEI-P5] **TEI Consortium** (eds., 2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/ (accessed 11 March 2012).

[WDG]: **Klappenbach, R., and W. Steinitz, eds.** (1964-1977). *Wörterbuch der deutschen Gegenwartssprache (WDG)*. 6 Bände. Berlin: Akademie-Verlag.

Werry, C. C. (1996). Linguistic and interactional features of Internet Relay Chat. In S. C. Herring (ed.), *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins, pp. 47-63.

Zifonun, G., L. Hoffmann, B. Strecker, et al., eds. (1997). *Grammatik der deutschen Sprache*. 3 Bände. Berlin, New York: de Gruyter.

Notes

1. Cf. http://www.empirikom.net/bin/view/Themen/DeRiK . The project is embedded in the scientific network 'Empirische Erforschung internetbasierter Kommunikation' (http://www.empirikom.net/), funded by the Deutsche Forschungsgemeinschaft (DFG).
2. Another corpus of contemporary language which aims to include a CMC subcorpus is the Dutch SoNaR project (Reynaert et al. 2010).
3. This dictionary is based on a six-volume paper dictionary, the *Wörterbuch der deutschen Gegenwartssprache* (WDG, en.: 'Dictionary of Contemporay German') published between 1962 and 1977 and compiled at the Deutsche Akademie der Wissenschaften (cf. [WDG]).
4. This is typically communicated by assigning certain subtypes of the 'Creative Commons' License to CMC documents or to web applications which specify the terms of (re-)use and (re-)distribution of their content.
5. Overviews of the features of CMC discourse from a linguistic perspective can be found, e.g., in Herring (ed., 1996, 2010), Werry (1996), Runkehl et al. (1998), Beißwenger (2000), Crystal (2001, 2011), Beißwenger & Storrer (2008), and Storrer (2012). The DeRiK corpus will include annotations of selected phenomena which are often described as being typical for language use on the Internet.
6. The RNG schema file, a TEI-compliant ODD documentation as well as encoding examples are available at http://www.empirikom.net/bin/view/Themen/CmcTEI .
7. In other grammars these units are described as *interjections* (e.g., Greenbaum 1996; McArthur et al. 1998; Blake 2008) or *Interjektionen* (DUDEN-4[7]), *inserts* (Biber et al. 1999, 2002), *discourse markers* (Schiffrin 1986), *discourse particles* or *Gesprächspartikeln* (DUDEN-4[5]).

Estimating the Distinctiveness of Graphemes and Allographs in Palaeographic Classification

Levy, Noga
nogaor@gmail.com
Tel-Aviv University, Israel

Wolf, Lior
wolf@cs.tau.ac.il
Tel-Aviv University, Israel

Dershowitz, Nachum
nachumd@post.tau.ac.il
Tel-Aviv University, Israel

Stokes, Peter
peter.stokes@kcl.ac.uk
King's College London, UK

1. Introduction

Within the discipline of palaeography, the 'morphological' approach tries to describe the letter-shape as a whole, so a letter may be described as a 'Caroline *a*' or as an 'insular *r*'. Aspects of this approach are visible in almost all palaeographical handbooks, particularly those that provide alphabets or selections of letter-forms. An example is Albert Derolez's, *Palaeography of Gothic Manuscript Books,* which also provides a useful discussion of morphology as a palaeographical method.

Commonly, a morphological system of descriptors contains two main categories: One category is the grapheme or perhaps, more correctly, character: the letter as an abstract entity but with physical form, such as *a, æ*, or a single punctuation mark. The second category is the allograph, namely, a particular way of writing the letter; typical examples include 'Caroline' or 'insular'.

A key question, given any system of descriptors, is to evaluate the relative importance of each of the components of this system. For example, it is well known among palaeographers that the grapheme *a* is very distinctive for late Anglo-Saxon minuscule (Ker 1957; Dumville 1988; Stokes 2005); however, subjective evaluation of distinctiveness could potentially be misleading. It is therefore quite useful to conduct a statistical analysis of significance, and potentially contribute thereby to the practice of palaeography, provided the results can be presented in a meaningful form.

In this work, we employ methods that are commonly used for mining insights from biological experiments regarding underlying genetic mechanisms. We show that in the context of palaeography such an approach also provides insightful observations.

2. Methods

A dataset consisting of 456 scribal handwritings in English Vernacular minuscule, ca. 990 – ca. 1035, is used (Stokes 2005). 'Scribal handwriting' here refers to a single stint or block of writing by one person; these samples are spread across some 198 manuscripts and range from the main text of the book to later additions and notes or glosses between the lines or in the margins.

The handwritings were described using 289 descriptors (Stokes 2007-2008), where each descriptor indicates whether a certain grapheme (or group of similar graphemes) written as specific allograph(s) appear in the manuscript, as well as forms of certain parts of letters such as ascenders, descenders, and pen-angle. Every sample of handwriting is described by its known or predicted place of writing (where possible) and the estimated range of dates of writing. For classification purposes, the dataset was divided into classes that are homogeneous in time and place.

First, we measure for each descriptor how informative it is. This is done using the *information gain* method (Mitchell 1997), which is often used for feature selection in text categorization tasks. The information gain score measures the decrease in entropy when a descriptor is given vs. the baseline in which it is absent. That is, the information gain measures the discriminative power added by each single descriptor.

Measuring the power of each descriptor by itself is of limited power. That a certain descriptor is ranked high tells us very little about how informative other similar descriptors are. Using an analogy to biology, it is helpful to know which genes express differently in a specific experiment testing different classes of biological conditions. However, if we want to learn about biological functions and processes, we must go beyond the level of the specific gene by aggregating the information from multiple genes of the same family. In palaeography, the importance of a specific character, for example, cannot be reliably detected just by looking at the ranking of one of the associated descriptors, no matter how highly it is ranked.

The Gene Set Enrichment Analysis (GSEA) is a statistically valid tool to evaluate how prominent a set (that is, a family) of descriptors is (Subramanian

2005). This computational method determines whether an a priori defined set of features presents itself in a statistically significant and coherent manner among a ranked list of descriptors.

The input of the GSEA method is a ranked list of features and a list of families of features. Context is important, and so we conduct one experiment using families derived from graphemes and another experiment for families of allographs. The GSEA method is based on elaborate order statistic mechanisms that can compare, on a leveled ground, between large and small families of features.

3. Results

The information gain method was used to automatically rank all 289 descriptors. The first, most significant, 30 features are listed in Table 1.

Rank	Descriptor	Rank	Descriptor
1	æ__horned	16	7__high_top_right
2	a__horned	17	d__bilinear
3	s__round	18	g__convex_top
4	æ__angled tongue	19	g__tail_in_middle
5	d__vert_tipped	20	æ__rotund_minim
6	f__deep_split	21	y__hooked_tail
7	æ__high_e	22	æ__low_lig
8	æ__flat_topped	23	7__convex_top
9	g__tail_oblong	24	æ__tail_non_bulging_lig
10	a__flat_topped	25	e__angled_SW
11	æ__Caroline	26	a__angled_top
12	ð__vertical_tip	27	c__flat_hook
13	c__angled_SW	28	Asc.__forked_trailing_to_left
14	a__teardrop	29	Dcl__predominant_d
15	7__vertical_desc.	30	Aspct__rounded

Table 1: The 30 single descriptors that were found to be the most informative using the information gain score

GSEA was then applied to these results, finding 16 families of graphemes (and stylistic issues, like aspect) that are 'enriched' in a statistically significant way, and 20 families of 'enriched' allographs. Tables 2 and 4 show the enriched families. The GSEA tool also provides another output: a list of descriptor families that are statistically speaking irrelevant. These lists (Tables 3 and 5) contain 5 graphemes and 9 allographs whose descriptors are ranked so consistently low that it is unlikely to be by chance. As can be seen, the GSEA results contain both expected results and surprising ones; see (Ker 1957; Dumville1988, 1993, 1994) for general background for script of this period.

Rank	Grapheme+	Normalized ES
1	æ	1.73
2	7 (Tironean nota)	1.31
3	a	1.31
4	c	1.15
5	f	1.10
6	d	0.86
7	g	0.82
8	Aspect	0.79
9	e	0.76
10	y	0.64
11	Ascender	0.60
12	h	0.58
13	Þð	0.58
14	k	0.57
15	Descender	0.56
16	s	0.50

Table 2: The 16 graphemes (and stylistic issues) that were found to be discriminative in a statistically significant way by the GSEA method. The enrichment score (ES) reflects the degree to which a feature-set is overrepresented at the top. The normalized ES accounts for differences in set size and the correlations between the sets

It is unsurprising that the ash (æ) is most significant, since it is a combination of *a* and *e*, both of which are in themselves significant (Dumville 1988). It is quite surprising that the Tironean nota for *and* (which looks like the numeral 7) is discriminating, as this is not noted by any of the palaeographers cited. It also came as a surprise that *c* and *y* turned out to be highly significant, because they are not generally recognised as such, though some of their specific forms certainly are (Dumville 1993). It was also surprising, from a palaeographer's viewpoint, to see *h* in the list, as it is not included in those published by palaeographers (Ker1957; Dumville 1993; 1994; Stokes 2005).

Rank	Graphemes+	Normalized ES
1	Minim	-1.20
2	Pen	-0.90
3	hmn	-0.76
4	r	-0.70
5	ð	-0.62

Table 3: The five graphemes and stylistic issues that were found to be irrelevant to the discrimination task to a statistically significant degree

Here, it is very interesting to see that ð is insignificant, since its form varies very widely between scribes. Perhaps some of its features are better discriminants than others.

Rank	Allograph	NormalizedES
1	HORNED	1.83
2	CONVEX_TOP	1.51
3	FLAT_TOPPED	1.48
4	ANGLED_SW	1.32
5	ANGLED_TONGUE	1.30
6	HORIZ._TONGUE	1.15
7	TEARDROP	1.08
8	BILINEAR	1.08
9	SEMI_CAROLINE	1.07
10	ROUND	1.03
11	ANGLED_TOP	0.92
12	ANGLED_SHOULDER	0.91
13	LOW_LIG	0.90
14	ROTUND_MINIM	0.89
15	FLAT_HOOK	0.82
16	SHORT	0.82
17	LONG_TONGUE	0.76
18	CAROLINE	0.75
19	BULGING_LIG	0.74
20	SQUINTING	0.73

Table 4: The 20 allographs that were found to be discriminative in a statistically significant way by the GSEA method

Table 4 is also quite insightful. While HORNED is widely accepted as significant (Ker 1957), ANGLED_SW (angled southwest quadrant) is not, although it has been suggested as discriminative (Stokes, 2005). LONG_TONGUE is not attested in the literature as significant, as are the related ANGLED_TONGUE and HORIZ._TONGUE: it will be interesting to study if they are strongly correlated, and if they are even more distinctive in combination. Since TEARDROP, ROUND, and SEMI_CAROLINE are all forms of the letter *a*, their significance is expected and is supported by related literature on the minuscule of the period (Ker, 1957; Dumville, 1994), and is strongly argued as relevant for this period in the thesis from which this dataset is taken (Stokes 2005). ROTUND_MINIM is unrecognized as being discriminative. LOW_LIG (low ligature) is less recognized in the literature than TALL_LIG (Ker, 1957) which is correlated with it. BULGING_LIG is well recognized (Ker 1957; Stokes 2005). SQUINTING is recognised as distinctive in Latin (Dumville 1993) but not in the script normally used for the vernacular.

Rank	Allograph	Normalized ES
1	MINIM_LENGTH	-1.35
2	TALL	-1.05
3	MINIMS_ROUNDED	-1.04
4	LONG	-0.98
5	SHORT_HOOK	-0.89
6	TURNED_DOWN_TONGUE	-0.81
7	WEDGED	-0.78
8	ATTACK_STROKE	-0.66
9	STRAIGHT_BACKED	-0.56

Table 5: The nine allographs that were found to be irrelevant in a statistically significant manner to the discrimination task. Most of these results are expected since these allographs are either very common in the corpus or present the 'default' values for the script of the period

4. Discussion

Computerized systems that perform digital palaeography have been criticized in the past for reducing script entirely to statistical processes that are themselves difficult or impossible to evaluate (Stokes 2009). In fact, the struggle to elicit meaning out of statistical inference tools is common to many scientific domains. Here we begin to show that, by using the appropriate statistical tools, computers can be used to mine meaningful insights in palaeography.

The proposed method is not without its limitations. First, it relies on a specific definition of 'distinctive descriptors' that is derived from the choice of the feature ranking algorithm used. The IG method used focuses on the ability to discriminate between the classes; by choosing another ranking method one could focus, for example, on scribal variation, which is also a question of interest to palaeographers. A second limitation is that the method cannot go beyond the assumptions made in the initial coding system. For example, there is a normalization quality to GSEA, in the sense that, if the palaeographer recorded more varieties of *a,* say, than of other letters, precisely because he expects that letter to be more significant, the GSEA method will counterbalance this by looking at the overall distribution of all varieties. However, presumably there could have been other letters and features that are in fact more distinctive but that were not recorded in the database because they were mistakenly deemed to be relatively insignificant. Future work should therefore aim to augment the database with automatically extracted features, with the potential benefit of adding a new (robotic) perspective to morphological analysis. Another limitation of the method is its dependence on verbal descriptors for features which are visual, or, indeed, which are a function of the physical movements of the scribe's arm, hand and pen, particularly given the lack of standard palaeographical terminology for such detailed features (Stokes 2011-2012).

These difficulties could potentially be overcome by providing greater rigour in nomenclature and by connecting these labels to particular images. Both approaches are already being tested in the DigiPal project (http://digipal.eu/), and discussions are already underway to combine the work done there with that described here.

Despite these limitations, the method described is still very promising. One of the difficulties in palaeographical study is the vast quantity of detail that must be processed, and so helping the human expert to identify distinctive features would be enormously beneficial in managing that data. For instance, identifying patterns in variation such as those by region, date or group of scribes would be invaluable in identifying manuscripts by their writing. Applying the method to other corpora of scribal handwriting could also lead to valuable insights. For example, if particular features prove to be discriminative across many different script-systems then this could be an important clue into identifying scribes independently of the script that they wrote, something very necessary given that most scribes routinely wrote in a number of very different scripts. Finally, with further testing and refinement, this approach promises an important and large step towards a method for distinguishing handwriting that can be described explicitly, that can be communicated effectively in ways that palaeographers and other humanities scholars can understand, and that can also enjoy the support of quantitative data. This goal is one that has been sought for a very long time (Stokes 2009).

References

Bishop, T. A. M. (1971). *English Caroline Minuscule*. Oxford: Clarendon Press.

Dumville, D. N. (1988). Beowulf come lately: some notes on the palaeography of the Nowell Codex. *Archiv für das Studium der neueren Sprachen und Literaturen* 225: 49–63.

Dumville, D. N. (1993). *English Caroline Script and Monastic History*. Boydell: Woodbridge.

Dumville, D. N. (1994). English Square Minuscule script: the mid-century phases. *Anglo-Saxon England* 23: 133–64.

Ker, N. R. (1957). *Catalogue of Manuscripts Containing Anglo-Saxon*. Oxford: Clarendon Press

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Stokes, P. A. (2005). *English Vernacular Script ca 990–ca 1035*. Cambridge [unpublished Ph.D. dissertation].

Stokes, P. A. (2007-2008). Palaeography and image-processing: some solutions and problems. *Digital Medievalist* 3. http://digitalmedievalist.org/journal/3/stokes/ (accessed 15 March 2012).

Stokes, P. A. (2009). Computer-aided palaeography: present and future. In Rehbein, M., *et al*. (eds), *Kodikologie und Paläographie im Digitalen Zeitalter*. Norderstedt: Books on Demand. 309–338. http://kups.ub.uni-koeln.de/2978/ (accessed 15 March 2012).

Stokes, P. A. (2011-2012). Describing script [Parts i–]. *DigiPal Project Blog*. King's College London. http://digipal.eu/blogs/blog/describing-handwriting-part-i/ (accessed 15 March 2012).

Subramanian, A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102(43): 15545-50.

Academic Research in the Blogosphere: Adapting to New Opportunities and Risks on the Internet

Littauer, Richard
richard.littauer@gmail.com
Universität des Saarlandes, Germany

Winters, James
wintzis@gmail.com
Katholieke Universiteit Leuven, Belgium

Roberts, Sean
S.G.Roberts@sms.ed.ac.uk
University of Edinburgh, UK

Little, Hannah
H.R.Little@sms.ed.ac.uk
University of Edinburgh, UK

Pleyer, Michael
Pleyer@stud.uni-heidelberg.de
Ruprecht-Karls-Universität Heidelberg, Germany

Benzon, Bill
BBenzon@mindspring.com
QuestionCopyright.org, USA

Interdisciplinary research with the intent of publication to a wide audience is increasingly desired. *Blogging* offers new opportunities for academics to collaborate with researchers from other fields and integrate data. The power to publish results and theories freely and get rapid feedback has positive and negative potential implications. On the positive side, new ideas can be presented and discussed easily with progress potentially much faster than a traditional journal peer-review. The internet also provides a forum to engage the public about ongoing research, an increasing concern for funding bodies. On the negative side, ideas can appear in public and affect research without being properly assessed. This could dilute the impact of relevant research. We review whether blogging can become part of research, by examining the case study of our blog on language evolution.

ReplicatedTypo.org has received 120,000~ hits in 3 years (counts gathered using Wordpress Analytics) and been awarded 8 editor's selections from *ResearchBlogging.com*. As well as reporting on recent publications and conferences, we have written basic introductions to Linguistics, Evolution, mathematical modeling and animal signaling. As a blog with multiple authors, our interests are varied, but our central research theme – evolutionary approaches to language and culture – remains the same. Our post topics include what makes humans unique, top-down versus bottom-up approaches to language evolution, the evolution of colour terms, Specific Language Impairment and Autism, the effect of second language learners on linguistic structure, cultural evolution and the singularity and genetic correlates of social sensitivity. We've written about the current trend for large-scale statistical analyses of linguistic features and social features, and contributed some of our own including phoneme inventory size and demography, alcohol consumption and morphological complexity and whether linguistic diversity is correlated with traffic accidents. As shown by the examples provided, Language Evolution is an inherently interdisciplinary field with many new ideas and which relies upon new techniques and analyses.

Blogs are a useful source for discovering current research and a forum for open peer review, whether open (from the public) or closed (from co-authors on drafts). Linguistics blogs have been around for many years (e.g. Language Log), but blogs dedicated to language evolution have emerged, too (e.g. Babel's dawn, Shared Symbolic Storage, Culture Evolves!, Biolinguistics Blog, Replicated Typo). One of these blog-authors has even published a book directed at a general audience about the theories he developed on his blog (E. B. Bolles 2011). However, there is no universal consensus on the method or acceptability of citing ideas from blogs. We argue that the devaluing of research and criticism appearing in open forums risks obstructing research. This is not merely a debate in Linguistics; much ink has been spilled on a similar grounds in evolutionary biology, in the so called *#arsenicgate* scandal (see Zimmer 2011). Scholarship will not take advantage of the collaborative potential of the Internet if academic standards are not applied to Internet resources. Concerns about standards and plagiarism, from work that might considered to be in the public domain, are particularly important for blogs that are used to disseminating original work in progress such as small-scale experiments and theoretical essays. A particularly sensitive issue, in that it may stop one blogging about personal research, is public access to experiment data and model code.

On a note unrelated to research dissemination and publication models, writing for blogs can benefit students. It encourages wider reading, engagement with cutting-edge topics and helps integrate students from diverse backgrounds into the language evolution community. In an increasingly competitive academic environment, blogs are a vital tool for career development. We would like to see universities teach new media skills such as blogging.

Our aims as science bloggers on Replicated Typo are: to highlight and discuss new research on language evolution; to engage with the general public by presenting language evolution research in an accessible way; to be a platform for open science research into language evolution. We hope that presentation of research and discussion on the internet can, in conjunction with journal peer-review, lead to more productive, accurate and progressive research. Discussions of posts on our own blog have lead to revisions of our research and new avenues of research as well as collaborations and clarifications of research by the authors of the studies reviewed. Releasing code on our blog has lead to interactions that benefited both the readers and the researcher. We hope that model transparency and the sharing of code can help foster links between language evolution and other fields who use similar techniques and technologies (biology, informatics, etc.).

The aim of this paper is to provide a forum at Digital Humanities 2012 for discussing these issues, and the questions that arise from them, namely:

- Can blogged research be taken as seriously as peer-reviewed research?
- What are the risks of publicly accessible research?
- Is research blogging adaptable to other fields, in particular fields involving minorities or low resource groups?
- Are there particular concerns with running experiments or soliciting feedback online?
- Is the field of academia doing enough towards public engagement on the internet?

We hope that providing a focus for these issues will ensure a productive and balanced response.

References

Bolles, E., ed. (n.d.). *Babel's dawn*. http://wwww.babelsdawn.com.

Bolles, E. B. (2011). *Babel's dawn: A natural history of the origins of speech*. Berkeley, CA: Counterpoint.

Jordan, F. (n.d.). *Culture evolves!* http://evolutionaryanthropology.wordpress.com.

Liberman, M., and G. Pullum, eds. (n.d.). *Language Log*. http://languagelog.ldc.upenn.edu.

Martín, T., ed. (n.d.). *Biolinguistics blog*. http://biolingblog.blogspot.com.

Munger, D., ed. (n.d.). *Research blogging*. http://www.researchblogging.com.

Pleyer, M. (n.d.). *Shared symbolic storage*. http://sharedsymbolicstorage.blogspot.com.

Winters, J., ed. (n.d.). *Replicated typo*. http://www.replicatedtypo.com

Zimmer, C. (2011) Happy Birthday, #arseniclife. *The Loom*, December 2, 2011. http://blogs.discovermagazine.com/loom/2011/12/02/happy-birthday-arseniclife/ (accessed March 14 2012).

Feeling the View: Reading Affective Orientation of Tagged Images

Liu, Jyi-Shane
jyishane.liu@gmail.com
Natioanl Chengchi University, Taiwan

Peng, Sheng-Yang
young.orange@gmail.com
Natioanl Chengchi University, Taiwan

1. Problems

Images are visual representations of human experiences and the world (Molyneaux 1997). An image captures a moment of life and provides a glimpse of human history. With the overwhelming popularity of social media and the convenience of personal digital devices, digital images on web platforms have exploded in astonishing numbers. According to a recent article by an independent online news site (Kessler 2011), the number of photos on Facebook was 60 billion, compared to Photobucket's 8 billion, Picasa's 7 billion and Flickr's 5 billion in early 2011. These digital images embody certain aspects of human nature in a historical scale. Challenges and opportunities abound for new findings of humanities research on digital visual media.

Many images (or photos) on social media are associated with tags that describe the content of the images. As opposed to controlled vocabularies in domain-specific taxonomy, tags are keywords generated freely without hierarchical structure and are characterized as folksonomy (Trant 2009). An image, with its visual representation of the world, can be seen as an anchor that links a set of tags. Mitchell (1994) discussed the relations of images and words and suggested the views that images and words are interactive and constitutive of representation and that all representations are heterogeneous. Therefore, images with tags are integral representations that encode fuller and richer content. This observation seems to support the assumption that, through an image anchor, meaningful linkage among tags exists.

This research focuses on the affective aspect of human life in digital images and attempts to identify a subset of tags that may indicate strong affective orientation. This provides a possibility to read into viewers' affective response to images with certain affect indicative tags that are strongly associated with known affect types. An affective magnitude computation method is also proposed to retrieve images with strong affective orientation.

2. Methodology

Among the major photo web sites, Flickr provides several unique features that seem to be better fit to the purposes of this research. First and foremost, Flickr emphasizes the use of tags for image description and community activity. Flickr photos tend to be associated with more tags than other web site photos. Second, user activities on Flickr involve more direct interaction of images and words with comments and feedbacks. Third, Flickr provides APIs that are conducive to data acquisition.

This research adopts an affective framework based on Russel's circumplex model of affect (Posner 2005), in which 28 emotion words are evenly distributed around a circle centered on the origin of a two dimensional space of affective experiences. The horizontal dimension of valence ranges from highly negative (left) to highly positive (right), whereas the vertical dimension of arousal ranges from sleepy (bottom) to agitated (top). The set of emotion words were reduced from 28 to 12 based on significant occurrence in Flickr tags and equal representation of four affective quadrants. For each affective tag, a total of 1,000 Flickr photos are retrieved based on a popularity index called interesting. The retrieving process also includes a check-and-replace step to make sure that the final sample of 12,000 photos were all unique and from unique users. This is to remove bias and achieve objective representation in data collection.

Point-wise mutual information (PMI), developed in the fields of probability and information theory, is a measure of mutual dependenceof two random variables and has been successfully applied to provide effective word association measurement in text mining research (Church & Hanks 1989; Recchia & Jones 2009). For any two words X and Y, their PMI value is calculated by taking the base-2 logarithm of the ratio of the joint probability distribution p(X, Y) over the independent probability distributions p(X) and p(Y). Maximal positive PMI values indicate strong tendency of co-occurrence, whereas negative values suggest less chance of occurring together and zero means independence. In this research, two tags co-occur if they appear in the tag set of a photo.

For the purpose of retrieving images with strong affective orientation, an affective magnitude computation method is developed. This method combines association strength, affect distribution in tag set, and community feedbacks. First, community feedback parameters such as views, favorites, and comments, are used to retrieve candidate photos with

above average parameter values. Second, for each candidate photo, an affective weight is calculated by the ratio of the number of affect indicative tags in the photo's tag set to the total number of tags in the photo's tag set. Third, for each candidate photo, the affective magnitude of certain affective type or certain affective quadrant is computed by the weighted sum of PMI values of the pairs of the affect types and the affect indicative tags of the photo.

3. Results

The total number of tags in the sample set of 12,000 photos is 258,293, in which 70,066 tags are unique. The frequency distribution of these unique tags shows a long tail. A stemming process is applied to these unique tags to reduce the inflected words to their root form, e.g., girls to girl, beautiful to beauti. A frequency threshold of 1% is also set to filter out less meaningful tags. The stemming and filtering processes result in 386 word forms of tags. These 386 tags are called affect indicative tags because of their high co-occurrence with the 12 affective tags. By calculating PMI values of a pair of an affective tag and an affect indicative tag, a matrix of association strength among 12 affective tags and 386 affect indicative tags are obtained. Table 1 shows a partial matrix of PMI values between affective tags and affect indicative tags.

Table 1

Affect Indicative Tags	Affective Tags in Four Affect Quadrants											
	Q1: positive & more aroused			Q2: negative & more aroused			Q3: negative & less aroused			Q4: positive & less aroused		
	happy	love	excited	distressed	angry	fear	sad	bored	tired	sleepy	relaxed	calm
smile	1.856	-0.460	1.245	-3.381	-1.617	-2.677	1.036	-2.954	-1.816	-1.866	-1.267	-2.954
joy	1.505	1.042	2.283	-3.329	-4.439	-2.077	2.459	-4.474	0	0	-1.543	-4.474
play	1.228	-1.030	1.921	-3.723	-0.401	-0.802	1.597	-1.554	-3.853	-1.886	-1.261	0
jump	1.201	1.024	2.285	-3.092	-1.066	-1.004	1.970	-1.079	0	0	-3.519	-4.249
Christmas	1.121	0.203	1.291	-0.615	-3.555	-1.91	1.876	-0.568	-0.960	-1.377	-1.638	-1.568
children	1.103	-0.139	1.521	-1.775	-1.234	-0.268	-0.748	-1.442	-1.419	-1.843	-2.734	-4.249
colorful	1.092	0.236	0.322	-1.087	-1.705	-1.348	2.763	-0.720	-1.212	-1.656	0.335	0.3421
new	1.059	-0.395	0.658	-1.764	-0.320	-0.4673	1.088	-0.4357	-0.897	-1.377	-1.312	-0.920
kid	1.013	0.224	1.565	-2.268	-1.663	-0.604	-0.647	-0.830	-2.392	-3.010	-1.970	-4.000
fun	0.913	-0.823	2.104	-3.321	-1.374	-1.701	-2.082	-0.522	-1.7534	-3.071	-0.7394	-2.477

Table 2

QI positive & more aroused	QII negative & more aroused	QIII negative & less aroused	QIV positive & less aroused
joy	scary	lonely	serene
jump	scream	tear	quiet
play	pain	cry	peace
smile	horror	wait	lake
fun	decay	candid	landscape
kid	death	rest	sea
children	vintage	yawn	sand
Christmas	dark	nap	boat
kiss	wall	bed	sunset
heart	teeth	alone	ocean

Table 2

The PMI values of an affect indicative tag are further grouped by each affect quadrant to provide distinct affective orientation. The association strength of an affect indicative tag with an affect quadrant is calculated by summing up positive PMI values of the three affective tags in the affect quadrant. Negative PMI values indicate unlikely co-occurrence and are taken as zero during quadrant grouping to avoid incorrect underestimation of co-occurrence union. Table 2 shows the lists of the top ten affect indicative tags that have the most association strength with one of the four affect quadrants.

These affect indicative tags can be used to retrieve images with predicted affective orientation. Given a set of images that include an affect indicative tag in their tag sets, the affective magnitude computation (AffMC) method is then used to determine an image's affective magnitude in the affect quadrant, and thus provides a ranked list of images. The performance of the affective magnitude computation method is compared to that of Google and Flickr, which also include a ranking mechanism on retrieved images. Images retrieved by AffMC, Flickr, Google with top affect indicative tags of each affect quadrant are shown in Tables 3, 4, 5, and 6.

Q1: positive & more aroused affect indicative tag: *joy*	
AffMC	
Flickr	
Google	

Table 3

*Copyrights of all images shown are reserved by original contributors.

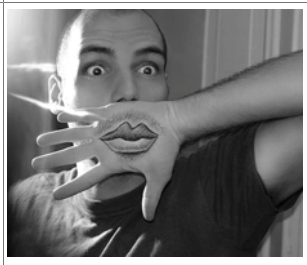
Q2: negative & more aroused affect indicative tag: *scary*	
AffMC	
Flickr	
Google	

Table 4

Q3: negative & less aroused affect indicative tag: *lonely*	
AffMC	
Flickr	
Google	

Table 5

Q4: positive & less aroused affect indicative tag: *serene*	
AffMC	
Flickr	
Google	

Table 6

To evaluate the affective magnitude computation (AffMC) method, top three affect indicative tags for each affect quadrant are used to retrieve images. The number of images retrieved by each ranking method is twelve, with three for each affect quadrant. A total of thirty-six images are given to a group of 136 subjects for testing their emotional response. Survey results indicate that AffMC is better than Flickr and comparable to Google in retrieving images of strong affective orientation. Note that image sources of Google and Flick (and AffMC) are different and Google enjoys a much significant user feedback for effective ranking.

4. Conclusions

This research shows that image tags can be exploited to characterize affect indicative tags. Accordingly, images with strong affective orientation can be identified and retrieved, from which viewers' emotional response can be mostly anticipated. This method can be further developed into an engine for sorting and grouping images into affective types based on the given tags. Potential applications include intelligent visual services for user emotion enhancement or compensation. Another research implication is that significant pairs can be revealed by substantial repetition among loosely associated items and the triangular cluster of a significant pair and the central item provides the basis for functional exploitation.

Acknowledgement

This work was partially supported by the National Science Council, Taiwan, [NSC 100-2221-E-004-013]; and the National Chengchi University's Top University Project.

References

Church, K. W., and P. Hanks (1989). *Word Association Norms, Mutual Information and Lexicography,Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, June 1989, Vancouver, British Columbia, Canada, pp. 76-83.

Kessler, S. (2011). Facebook photos by the numbers, *Mashable*. http://mashable.com/2011/02/14/facebook-photo-infographic/ (accessed 28 October 2011).

Mitchell, W. J. T. (1994). *Picture Theory: Essays on Verbal and Visual Representation*. Chicago: U of Chicago P.

Molyneaux, B. L., ed. (1997). *The Cultural Life of Images: Visual Representation in Archaeology*. London: Routledge.

Posner, J., J. A. Russell, and B. S. Peterson (2005). The Circumplex Model of Affect: An Integrative Approach to Affective Neuroscience, Cognitive Development, and Psychopathology. *Development and Psychopathology* 17(3): 715-734.

Recchia, G. L., and M. N. Jones (2009). More Data Trumps Smarter Algorithms: Comparing Pointwise Mutual Information to Latent Semantic Analysis *Behavior Research Methods* 41: 657–663.

Trant, J. (2009). Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information* 10(1): 1-42.

Characterizing Authorship Style Using Linguistic Features

Lucic, Ana
alucic2@illinois.edu
Graduate School of Library and Information Science, USA

Blake, Catherine
clblake@illinois.edu
Graduate School of Library and Information Science, USA

1. Introduction

For centuries, scholars have been searching for reliable methods to establish the authorship of a work. The underlying premise of this scholarship is that each author has a detectable individual style of writing that can be explored with automated methods.

A range of features have been explored that include function words and character n-grams either individually or in combination with other features (Burrows 1987; Bayeen et al. 2002; Keselj et al. 2003; Argamon & Levitan 2005; Houvardas & Stamatatos 2006), part-of-speech information (Stamatatos et al. 2001), probabilistic context-free grammars (Raghavan et al. 2010), and linguistic features of the text (Koppel et al. 2006). Although many of those experiments consider only lexical features, syntactic features are considered more reliable (Stamatatos 2009: 542), but parsing technologies have only recently achieved the accuracy necessary to consider syntactic features in detail. Once the features are in place, the problem of authorship attribution can be cast as a classification problem and machine-learning methods can be used, such as Naïve Bayes or Support Vector Machine (SVM) that both work well when there are a large number of weak features. For very large collections and for open authorship attribution sets, similarity-based methods are more appropriate than machine-learning methods (Koppel et al. 2011).

Our goal is to analyze the potential of syntactic dependencies to characterize an author's writing style with respect to how an author refers to people. The features that we use are drawn from both semantic and syntactic features. Specifically we first identify personal names – a semantic feature – and then identify local dependencies that surround those personal names – a syntactic feature. Our results demonstrate that these features vary between

authors and that those variations can be exploited to accurately determine authorship attribution.

2. Method

The proposed authorship attribution model emphasizes how an author refers to people. This approach leverages linguistically motivated markers and takes advantage of recent automated methods to identify syntactic dependencies. Consider the following sentence shown in Figure 1: 'Robert Ryan is the prey he captures, along with the girl Janet Leigh' where author 819382 uses a noun appositive to refer to the actress Janet Leigh. Note that there are two structures in this case – one for Janet Leigh and another for Robert Ryan where the author uses the actor as the subject of the sentence. Our method considers one dependency before and one dependency after the person reference. Author 819382's profile suggests a preference for appositives, such as in the sentence 'The claims manager, Lee J. Cobb here, unravels the plot' where the author again uses the appositive when referring to the character Lee J.Cobb.

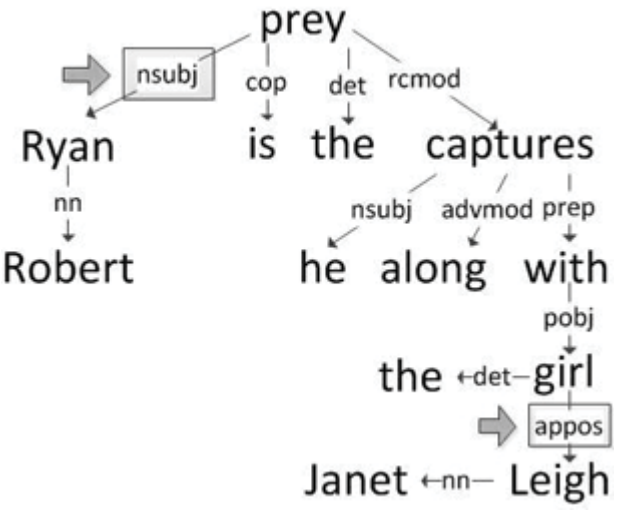


Figure 1: Dependency Grammar showing Robert Ryan as the subjects and Janet Leigh as a noun appositive

In contrast to author 819382, author 463200 prefers to describe several people together such as in the following examples that occur in different reviews: 'Lois Lane and Clark Kent are sent to cover a circus (for some reason)'and 'And there's Gloria DeHaven and June Allyson in bit parts!' (see Figure 2 for the dependency grammar representation). The coordination (cc) and preposition (prep) would also be used as candidate markers for this sentence.

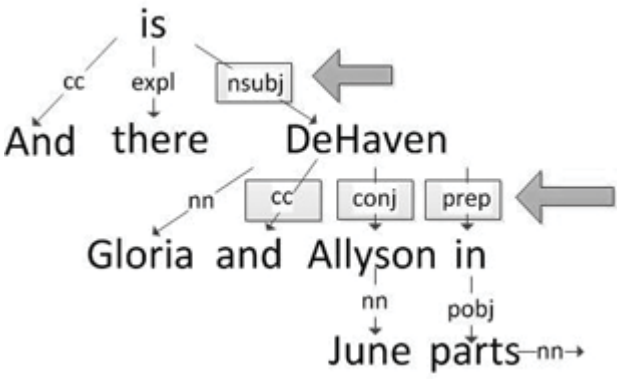


Figure 2: Dependency Grammar showing Gloria DeHaven as the subject and June Allyson in a conjunctive clause

Author 1609079, however, is more likely to use a nominal subject passive grammatical relation when introducing personal names than other reviewers. Figure 3 shows this pattern when referring to Evelyne in the context of the sentence 'Arnold's wife Evelyne is also expected to sacrifice for Edward.' As with the earlier authors, author 1609079 uses other patterns in addition to nominal subject passive before the name within the same sentence. The generated author's profile reflects all of these preferences and is thus well suited to use in discerning one author from another.

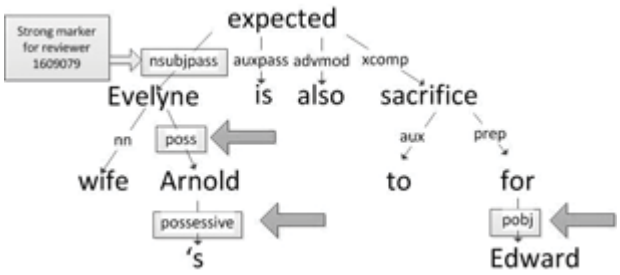


Figure 3: Dependency Grammar of a sentence by author 1609079 who prefers to refer to people as the subject of a passive clause

These examples provide the motivation for our strategy to use local syntactic structures where authors refer to people. The general approach to create an author profile has the following steps. The information shown in parenthesis provides details for the experiments reported in this paper.

1. Collect and identify sentences from the source collection (Natural Language Toolkit recognizer, nltk.org
2. Identify people within each sentence (Named Entity Recognizer, nlp.stanford.edu/software/CRF-NER.shtml)
3. Generate syntactic dependency trees from each sentence (Stanford Lexical Parser version 1.6.9, nlp.stanford.edu/software/lex-parser.shtml)

4. Align the person names with the dependencies and count the local grammatical structures before and after each person reference
5. Normalize the local grammatical structures by dividing by the number of people used by an author

Our method only considered those sentences and only those reviews which included personal names and that had 80 or fewer words. Of the 62,000 reviews, 92% or 57,286 included at least one sentence that satisfied this criterion and of the 919,203total sentences 354,389 referred to people.

3. Results

The evaluations presented in this paper are based on a closed set of 62 candidate authors. The publicly available IMDb62 dataset (Seroussi et al. 2011) was harvested from the IMDB (imdb.com) in 2009 and satisfies Grieve's (2007) requirement that the texts should be written in the same genre and should have originated around the same period. The IMDb62 collection, however, does not satisfy Grieve's requirement that the texts should preferably be on the same topic as the movie review topics depend on the topic and genre of the movie. Only those reviewers who made prolific contributions to the imdb.com web site were selected for inclusion in the dataset so that each of 62 reviewers has 1,000 reviews (the total of 62,000 reviews). Movie reviewers typically refer to actors, directors and characters from a movie, so the IMDb62 dataset provides a rich collection in which to explore authorship style when referring to people.

Results show that authors differ both with respect to the number of person references made and how they refer to people. The majority of authors in the collection (43/62) referred to between 5,000 and 20,000 personal names in all of the reviews, but 15 authors use far fewer names (between 1,086 and 5,000), and 4 reviewers referred to more than 20,000 personal names (up to 33,115). Twenty-nine candidate syntactic features (13 before and 16 after personal names) were selected based on the frequency of use within the entire set of reviews. The two most frequent syntactic patterns before a personal name were nominal subject and object of a preposition while the most frequent syntactic patterns after a personal name were a punctuation mark and possessive modifier.

To evaluate how well the syntactic features could predict the correct author for a review, a test set was created by drawing 5% of reviews for each author at random from the original collection. The author profile described in the methods section was then created using the remaining 95% of the reviews (the training set) and the cosine similarity between the features in the test set and the profile in the training set was used to identify the correct author, i.e. if syntactic features are well suited to authorship disambiguation then the cosine similarly would place the correct author first in the ranked list of candidate authors. Note that all reviews in the test set (approximately 50) were considered together rather than as 50 different reviews and that this process was repeated 10 times with 10 different samples. On average the correct author appeared at rank 1.84 (in the range from 1 to 6.8). Twenty-one of the 62 authors are ranked first and author accuracy increases from 33.8, 67.7, 83.9 and 95.2% as the rank increases from 1 through 4.

With respect to related work, Seroussi et al. (2011) used Latent Dirichlet Allocation (LDA) to model topics within the IMBDb62 dataset and then authorship attribution is calculated by finding the smallest Hellinger distance between the test and training documents. Author attribution accuracy varied between 19 to 68% for the multi-document approach and 25 to 81% for a single-document approach depending on how many topics were used in the model. Seroussi et al.'s experiment (2011) outperformed the KOP method (Koppel et al. 2011) on the IMDb62 dataset by about 15% at 150 topics and the results were statistically significant.

4. Closing Comments

We have presented a new method that leverages local syntactic dependencies to reveal the ways in which an author refers to people. One limitation of this approach is that it is restricted to only those texts and those reviews which contain named entities (in our case personal names); however, authors frequently refer to people and places in history books, historical novels, and fiction works in general. Our preliminary results on the IMDb62 collection are consistent with Stamatatos's speculation (2009) and indicate that grammatical structures , in our case that surround personal names, are indeed useful to differentiate between authorship styles. We posit that further experimentation is required to better understand how individual features, and combinations of features can best be used to determine authorship attribution accurately.

References

Argamon, S., and S. Levitan (2005). Measuring the Usefulness of Function Words for Authorship Attribution. *Proceedings of the 2005 ACH/ALLC conference*. Victoria, BC, Canada, June 2005.

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *LLC* 22(3): 251-270.

Houvardas, J., and E. Stamatatos (2006). N-Gram Feature Selection for Authorship Identification. *AIMSA* 2006: 77-86.

Keselj, V., F. Peng, N. Cercone, and C. Thomas (2003). N-gram-based Author Profiles for Authorship Attribution. *Proceedings of the Conference Pacific Association for Computational Linguistics*. Dalhousie University, Halifax, Nova Scotia, Canada, August 2003.

Koppel, M., N. Akiva, and I. Dagan (2006). Feature instability as a criterion for selecting potential style markers: Special Topic Section on Computational Analysis of Style. *Journal of the American Society for Information Science and Technology* 57(11): 1519-1525.

Koppel, M., J. Schler, and S. Argamon (2011). Authorship Attribution in the Wild. *Language Resources* & Evaluation 45(1): 83-94.

Raghavan, S., A. Kovashka, and R. Mooney (2010). Authorship Attribution Using Probabilistic Context-Free Grammars. *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden, July 2010.

Seroussi , Y., I. Zukerman, and F. Bohnert (2011). Authorship Attribution with Latent Dirichlet Allocation. *CoNLL 2011: Proceedings of the 15th International Conference on Computational Natural Language Learning*. Portland, OR, USA, June 2011.

Stamatatos, E. (2009). `http://onlinelibrary.wiley.com/doi/10.1002/asi.21001/full` A Survey of Modern Authorship Attribution Methods. *J ournal of the American Society for Information Science and Technology* 60(3): 538-556.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2001). Computer-Based Authorship Attribution without Lexical Measures. *Computers and the Humanities*: 193-214.

Investigating the genealogical relatedness of the endagered Dagon languages

Moran, Steven
steve.moran@lmu.de
Research Unit: Quantitative Language Comparison, LMU, Munich, Germany

Prokic, Jelena
j.prokic@lmu.de
Research Unit: Quantitative Language Comparison, LMU, Munich, Germany

Dogon languages are spoken predominately in eastern Mali in West Africa. The Dogon were made famous by Marcel Griaule, a French anthropologist who pioneered Ethnography in France, and worked with the Dogon between 1931-1956. He reported that the Dogon had advanced astronomical knowledge of the Sirius binary star system, knowledge that is not possible without telescope. Since then, the Dogon have been shrouded in controversy and mystery.

As late as 1989, Dogon appeared in reference books as if it were a single language, e.g. Bendor-Samuel 1989 (as found in Blench 2005). The standard encyclopedic reference on the world's languages, the Ethnologue,[1] now lists 14 Dogon languages (Lewis 2009), but this figure is too low. In 2004, an extensive sociolinguistic survey by Hochstetler et al. (2004) estimated no less than 17 distinct languages and described the language family as highly internally divided. Since 2004, much initial survey work on Dogon has been undertaken by Roger Blench, Denis Douyon and Jeffrey Heath's Dogon languages project,[2] which has lead to the 'discovery' of a web of divergent dialects, some of which have been raised to the status of distinct languages based on standard linguistic criteria. Thus, the current Dogon linguistic situation is not at all transparent. Dogon languages are very under-described, many are highly endangered, and all are genealogically not well established (Blench 2005; Heath 2008).

The Dogon languages project provides a tentative detailed inventory of known Dogon languages. There are currently 20 distinct languages grouped (crudely) into eight geographical regions, with no implications for genealogical subgrouping. The internal structure of the Dogon language family is unknown, as is the number of mutually unintelligible languages it contains. In fact, the Ethnologue gives a flat genealogical tree. The position of the Dogon

languages relative to other African language families is also unclear because of Dogon's unique typological characteristics. Its lineage has long been disputed, as summarized in Table 1.[3]

Year	Classification (language family)	Author
1924	Nigéro-Sénégalais	Delafosse
1941	Voltaic (Eng. Gur)	Homburger
1948	Voltaic; Gurunsi	Baumann & Westermann
1951	Mandé	Holas
1952	Mandé	Delafosse
1952	Gur (Fr. Voltaic)	Westermann & Bryan
1953	Voltaic	Bertho
1953	Non-classified	de Tressan
1950/60	Gur	Calame-Griaule
1963	Gur	Greenberg
1971	Gur	Bender-Samuel
1981	Voltaic	Manessy
1981	Volta-Congo	Bendor-Samuel
1993	Unresolved; non-classified	Galtier
1994	Unresolved; non-classified	Plungian and Tembiné
2000	Ijo-Congo	Williamson and Blench
2009	Volta-Congo	Lewis

Table 1: Historical classification of Dogon

In this paper, we use a marriage of digital methods successfully applied in bioinformatics to decode DNA and determine the genetic relatedness of humans, and we apply these methods to language data in an attempt to shed light on the prehistory of the Dogon languages and determine their genealogical subgroupings. The comparative method employed in historical and comparative linguistics (the study of language change to reconstruct the genealogical relatedness of languages) is a very laborious and time-consuming task that involves identifying cognates (words that share a common etymological origin) through their shared meanings and common sound change correspondences (e.g. English 'is', German 'ist', Latin 'est', Indo-European 'esti').

We show how recent advances in the use of quantitative methods in the study of language comparison can be applied to digital data of (endangered) languages to automate the discovery of regular sound correspondences and cognate forms. Next, powerful statistical techniques allow for new insights into the origin and evolution of human languages. We use distance-based methods like Levenshtein (Levenshtein 1965) and character-based methods like Bayesian Markov Chain Monte Carlo methods (Page & Holmes 2006) to induce language family trees from lexical data. We present the first subgrouping hypothesis for the Dogon language family and we will discuss the limits of current quantitative approaches, where the state-of-the-art in computational historical linguistics is heading, and what we can hope to glean from their application to linguistic diversity.

References

Bendor-Samuel, J., E. J. Olsen, and A. R. White (1989). *The Niger-congo languages – a classification and description of Africa's largest language family*, chapter Dogon. Langham: UP of America, pp. 169-177.

Blench, R. (2005). A survey of Dogon languages in Mali: overview. *Ogmios* 26: 14-15.

Heath, J. (2008). *A grammar of Jamsay*. Berlin: Mouton de Gruyter.

Hochstetler, J. L., J. Durieux, and E. Durieux-Boon (2004). *Sociolinguistic survey of the Dogon language area*. SIL International.

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR* 163: 845-848.

Lewis, M. P. (2009). *Ethnologue: languages of the world*, Sixteenth edition. SIL International.

Page, R. D. M., and E. C. Holmes (2006). *Molecular evolution: a phylogenetic approach*. Malden: Blackwell.

Notes

1. http://www.ethnologue.com/
2. http://dogonlanguages.org/
3. See Hochstetler et al. (2004) and Hantgan's Dogon bibliography for references at: http://dogonlanguages.org/bibliography.cfm.

Landscapes, languages and data structures: Issues in building the Placenames Database of Ireland

Měchura, Michal Boleslav
mechrm@dcu.ie
Dublin City University, Ireland

1. Introduction

The Placenames Database of Ireland is a bilingual database which records the official names in Irish and English of geographical objects on the island of Ireland. The database contains over 100,000 entries, is accessible to the public for free (www.logainm.ie (`www.logainm.ie`)) and serves around 150,000 searches per month.

The public-facing website was launched in 2008 in the context of recently enacted legislation which gave more prominence to Irish-language placenames in the Republic of Ireland; for details on the legal and sociolinguistic situation of placenames in Ireland see Mac Giolla Easpaig (2008, 2009). For more insights into the history and role of placenames in Ireland, see Mac Mathúna (1990).

We will not attempt to describe the structure of the Placenames Database of Ireland holisticly here. Instead, we will concentrate on three sub-areas where interesting challenges have arisen. In many cases, the issues presented here are issues that have not been solved satisfactorily in our database yet, or a solution has not been implemented yet. In this sense, the paper is merely a discussion of issues rather than a catalogue of solutions.

2. Linguistic properties of placenames

Placenames databases differ in how much attention they pay to the linguistic properties of the names themselves, such as their gender, inflection and so on. Multilingual databases such as Geonames usually ignore such aspects completely. Smaller databases, however, often focus on only one or two languages and are conceived mainly as a *lexical* database rather than a *geographic* one. Linguistic properties relevant to placenames can be broadly subdivided into the following three areas.

Internal structure. Practically all placenames that consist of more than one word have an implicit internal structure, just like other linguistic expressions. Some aspects of this structure may need to be made explicit in a database. These include definite articles and other textual phenomena that get in the way of alphabetical sorting including initial mutations in Celtic languages; composite names such as *Ballaghgowla and Froghan*; and names with disambiguators such as *Black Lough (South)*.

Combinatorial properties. These are linguistic properties relevant to how the placename interacts with the text in which it is being used and include language-specific features such as gender, grammatical number and inflection paradigms. In addition, the following combinatorial properties are relevant to placenames in particular: how the name can be combined with categorizers such as 'town' or 'county' (*Donegal Town* but *County Donegal*); which locative prepositions are combined with the placename (*i nGaillimh* 'in Galway' but *ar an gCeathrú Rua* 'in Carrarow', literally 'on Carrarow').

Lexical relations to other words in the language and to other placenames. The former include demonyms (terms for 'people from') and derived adjectives which are sometimes irregular. The latter include cases where one geographical object has been named after another, such as *Ballybeg Road* or *Lismore Terrace*.

3. Cross-linguistic issues

A separate cluster of issues stems from the fact that the Placenames Database of Ireland is bilingual. This does not seem to pose a challenge at first; the general principle is that every place has two names, one in each language, and this seems to call for a simple data structure: all we need is two text fields. However, that approach would fail to account for the following phenomena.

Borrowing and gaps. Sometimes, a place has the same name in both languages. An example is an area of Dublin called *Dún Laoghaire*. This is an Irish name which is also used in English with unchanged spelling (but with anglicised pronunciation). An obvious solution would be to simply record *Dún Laoghaire* twice, once as an Irish name and once as an English name. However, this is unsatisfactory as it fails to account for the fact that this name is not really *in* English, it is merely *used* in English.

On the other hand, in some strongly Irish-speaking areas, minor features such as crossroads, fields and wells only have Irish names and there are no known English names. One can only assume that if somebody needed to refer to such a place while speaking English, one would briefly code-switch into Irish to utter the name.

A fairly granular data structure is called for here, one that allows us to capture facts as to whether a name in a given language exists or not, whether the name *used* in a given language is also a name *in* that language, and whether the name has been borrowed from another language.

Anglicisation, translation and re-interpretation. Many English placenames in Ireland have been obtained from the original Irish names by a process of writing down an approximated pronunciation (*Gaoth Dobhair -> Gweedore*). In other cases, when the English name was created first, the method used to coin the Irish name has often been translation, such as *Butler's Bridge -> Droichead an Bhuitléaraigh*. In other cases still, the Irish and English names are independent coinages (example: *Loch Garman/Wexford*).

Translation is sometimes accompanied by re-interpretation. An example of this is an area of Dublin called *Barra an Teampaill/Temple Bar*.The placename originated in English from the personal name *Temple* but was later re-interpreted as the common noun *temple* and hence the Irish name (literally 'the bar of the temple'). Although based on a misunderstanding, it was decided to keep the Irish name as official because it is commonly used.

An ideal data structure would allow us to record these and other etymological relations between names of the same place in different languages. If such relations are explicated and annotated in the database, then we can not only provide better information to users but also extract interesting statistical observations about the relative proportion of these phenomena in the country's body of placenames.

4. Dealing with overlapping hierarchies

Most countries are subdivided into administrative units such as districts, provinces or counties. These units form a hierarchy and such hierarchies are deliberately designed as a *nested hierarchy* to disallow overlapping and facilitate reasoning.

In Ireland, the situation is far from this computational ideal. The basic units (*counties -> baronies -> civil parishes -> townlands*) may have originally been designed as a nested hierarchy but are no longer so because of boundary changes that have not been propagated up and down the hierarchy and because of the introduction of competing hierarchies which overlap with the traditional system.

The consequence when building a placenames database is that we must work with an *overlapping hierarchy* where a child may have more than one parent. This complicates things; for example, reasoning is no longer always possible. If we know that A is in B and that B is in C1 and C2 simultaneously, we can no longer infer whether A is in C1 or C2 or both. In fact, there is no way to know this other than by deriving it from a dataset of geographical boundaries, or by recording it explicitly.

5. Place as an abstract concept

Consider the placename *Dún na nGall/Donegal* which can be found in the north-west of Ireland. The question to ask is: how many places called *Dún na nGall/Donegal* are there in this corner of Ireland? The answer will differ depending on one's perspective. An outsider will see only one, a county of that name. A local inhabitant will probably see two, the county and its capital town of the same name. A local politician will probably see an electoral division of that name and also a town of that name with its town council. In total, there are five units called *Dún na nGall/Donegal* in that part of Ireland.

In our database, these are treated as separate objects which, as far as the database knows, only *happen* to have the same name. That, however, is unsatisfactory as it fails to distinguish a case like this from cases such as the 19 places called *An Baile Mór* (literally 'the large town') which can be found all over Ireland and which genuinely *happen* to have the same name. Another reason why this is unsatisfactory is that it introduces a potential for inconsistency because data such as the Irish name's grammatical information need to be recorded five times instead of once.

An ideal data structure would provide a way to connect several places to a single 'abstract place'. The abstract place would contain all information common to the concrete places, such as names and historical citations, and these would then be inherited by the concrete places.

6. Conclusion

Many of the issues illustrated here are inherently linguistic and stem from the fact that we conceive of our database as primarily a lexical database and only secondarily as a geographical database. A second cluster of issues is caused by the heritage of conflicting and overlapping administrative hierarchies and from the differences in perspective these hierarchies impose.

References

A. Placenames databases

Geonames: `http://www.geonames.org`

Placenames Database of Ireland: http://www.logainm.ie/

B. Literature

Mac Giolla Easpaig, D. (2008). Placenames Policy and its Implementation. In Caoilfhionn Nic Pháidín, Seán Ó Cearnaigh (eds.), *A New View of the Irish Language*. Dublin: Cois Life, pp. 164-177.

Mac Giolla Easpaig, D. (2009). Ireland's heritage of geographical names. *Wiener Schriften zur Geographie und Kartographie* 18: 79-85.

Mac Mathúna, L. (1990). *Ár dTimpeallacht Logainmneacha: Inniu agus Amárach* [our placenames environment: today and tomorrow]. Dublin: Coiscéim.

Interoperability of Language Documentation Tools and Materials for Local Communities

Nakhimovsky, Alexander
adnakhimovsky@colgate.edu
Colgate University, USA

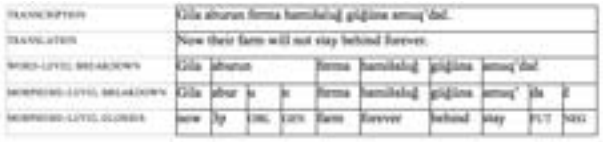
Good, Jeff
jcgood@buffalo.edu
Department of Linguistics, University at Buffalo, USA

Myers, Tom
tommyers@dreamscape.com
N-Topus Software, USA

1. Tools for language documentation

The two main outputs of traditional language documentation are lexicons and corpora of texts. In the last two decades, it has been proposed to augment these in a number of ways (Himmelmann 1998, 2006):

- The primary data for language documentation should be made available in the form of (digital) audio or video recordings made in the field.
- Corpora should consist of time-aligned and annotated transcripts of those recordings. Time alignment makes explicit how a given set of annotations relates to a media segment. Since there is a one-to-one correspondence between text and media segments, the latter can be searched by their text and annotations.
- Text annotations should generally take the form of Interlinear Glossed Text (IGT) (see Palmer & Erk 2007). An example of IGT, from Haspelmath (1993), is shown below. The format has a tree-like structure, which can be represented via a 'nested' table. Its basic components include a transcriptional representation of data from the language being described which is further broken down into words and their component morphemes. Each sentence is associated with a free translation, and each word is associated with (possibly specialized) glosses.



Existing tools supporting the production of lexicons and time-aligned corpora are not, at present, well integrated. One testimony to this is the continuing use of Toolbox (formerly known as Shoebox), a database utility optimized for the creation of IGT and lexicons initially developed in the 1980s by SIL International (SIL). Toolbox lacks native support for time-aligned annotation and, more strikingly, proper data validation. However, one of its key features, the integration of a text database with a lexical database to facilitate automated glossing, has yet to be effectively replicated in any other widely used tool. A recent major revision of Field Language Explorer (FLEx), also from SIL, positions this tool to fill this gap, however (see Rogers 2010 for a recent review). FLEx is now cross-platform and internally uses a native XML format, which can form the basis for its integration into a set of interoperable tools.

In Europe, there is a major center for the development of language documention and language archiving technology (LAT) software at the Max Planck Institute for Psycholinguistics. One of the tools they have produced, ELAN, has become widely adopted for the creation of time-aligned annotations (see Berez 2007 for a review). Toolbox, ELAN and FLEx are the most commonly used specialized tools for language documentation. While all three do some things well, none covers the entire range of language documentation tasks. For example, ELAN has no support for lexicon building, and Toolbox and FLEx have no support for time alignment or audio playback that is of great help in transcribing a media file. In addition, none provides direct support for the creation of publishable outputs, a gap that is most frequently filled by Microsoft Word, even though it offers no ready means of interoperating with custom linguistic software.

While ELAN has import/export modules for Toolbox, there was, until recently, no way to share data between ELAN and FLEx. In 2009, one of the authors (TJM) led the development of a FLEx-import module that was integrated into a release ELAN. This effort was severely handicapped by the limitations of FLEx's native data storage format at that time, which has been partly addressed in its present XML format. Since then, we have been working with both ELAN and FLEx teams to establish a means for lossless two-way interchange of files between ELAN and FLEx. This will make it possible for a linguist to, for example, exploit the time-alignment functionality of ELAN and the lexicon and parsing functionalites of FLEx without losing information produced by either tool when exporting/importing data between them.

The main obstacle to achieving this goal is that the two programs have very different internal data models: ELAN has the capability to create structures that are not replicable in FLEx (for instance, to represent overlapping utterances from more than one speaker), while FLEx allows words to be associated with a wealth of grammatical information that cannot be encoded using ELAN. At the same time, both ELAN and FLEx contain some overlapping information in their representations of IGT. Our proposed solution to achieve full interoperation is to provide both ELAN and FLEx with a unified underlying representation that will represent data from both programs. Some parts of that representation will be ignored by ELAN, and other parts will be ignored by FLEx, but new tools can be developed to ensure that a lossless round trip between the two programs will be possible (see Cochran et al. 2007 for an earlier discussion). This approach can also broaden the interoperability between ELAN and Toolbox, and indeed facilitate interoperability between both Toolbox and FLEx as well.

An obvious choice for a unifying representation are Resource Description Framework (RDF) graphs of the sort associated with the Semantic Web (Allemang & Hendler 2010). One of the main purposes of RDF is specifically to merge heterogeneous representations of overlapping data. It also has a standard query language and a rapidly growing arsenal of tools for development.

We would like to emphasize the generality of this solution: given two programs with overlapping but partially incompatible data models, RDF, perhaps augmented by Web Ontology Language (OWL), can be used as an interlingua that can represent both. (See Farrar & Langendoen 2009 for discussion of the use of OWL in descriptive linguistics.) When this representation is accessed by one of the programs (e.g., by using SPARQL, the query language for Semantic Web data – see DuCharme 2011), parts of the representation will be ignored, but the editing done on the overlapping part can be preserved for eventual access by the other program.

Since May 2011, we have maintained a discussion group in which both ELAN and FLEx developers participate, allowing the development of a concrete implementation of this proposed solution. A key area of progress has been devising an appropriate system for assigning globally unique identifiers to be used in both ELAN and FLEx to allow the overlapping data available to both tools to be effectively tracked.

2. Serving diverse communities through interoperation

An endangered language documentation project typically results in two collections of materials, one addressed to scholars, the other addressed to the local community. The requirements for the scholarly collection are quite precisely defined in terms of allowable data formats, required metadata, and specifications of access restrictions (see, e.g., Johnson 2004; Austin 2006; Nathan 2011). By contrast, there has been little discussion of a general systematic approach to creating materials for local communities, nor even an attempt to create principled specifications for them (though see Nathan 2006; Nakhimovsky & Myers 2003). As a result, materials for local communities are frequently developed as a separate effort, unrelated to the scholarly archive.

The approach presented in the first section of the paper creates an opportunity for a more general solution. We suggest that a unified RDF representation is well-suited for the creation of materials for local communities due to rapidly developing trends in data dissemination technology, including the increasing adoption of Semantic Web technologies as a means of data exchange, the widespread of deployment of Cloud services, and the increasing use of social media in even relatively marginalized communities especially via mobile devices like smartphones (see, e.g., the BOLD project for an example as well as Nathan 2010).

The overall strategy for developing community materials could be as follows. Using SPARQL queries, extract suitable materials from the overall RDF collection (e.g., streamable media, community-selected texts) and assemble them, possibly dynamically, into a community website to be accessed via the internet or a local server. Such a model would allow community members to reassemble the data to suit their needs while also ensuring they have access to the most up-to-date analyses of the linguists. Moreover, in principle, the same basic techniques that permit tool interoperation via a unified RDF representation could also allow communities to enrich the RDF themselves, thereby adding information (e.g., metadata or cultural annotations) to the dataset and allowing them to take a more active role in the construction of the documentary record of their language.

Funding

This work has been supported by the National Science Foundation [0553546 A.N.; 1065619 A.N., T.M; 0715246, J.G.]

References

Allemang, D., and J. Hendler (2010). *Semantic Web for the working ontologist: Effective modeling in RDFS and OWL* (second edition). Amsterdam: Elsevier.

Austin, P. K. (2006). Data and language documentation. In J. Gippert, N. P. Himmelmann & U. Mosel (eds.), *Essentials of language documentation*. Berlin: Mouton de Gruyter, pp. 87-112.

Berez, A. (2007). Technology review: EUDICO Linguistic Annotator (ELAN). *Language Documentation and Conservation* 1: 283-289. http://hdl.handle.net/10125/1718 .

BOLD. http://bold.xpdev-hosted.com/ (accessed March 4, 2012).

Cochran, M., J. Good, D. Loehr, S. A. Miller, S. Stephens, B. Williams, and I. Udoh (2007). Report from TILR Working Group 1 : Tools interoperability and input/output formats. Toward the Interoperability of Language Resources Workshop, Stanford, California, July 2007. http://linguistlist.org/tilr/working-group-reports/Working%20Group%201.pdf .

DuCharme, B. (2011). *Learning SPARQL*. Sebastopol, CA: O'Reilly.

ELAN. http://www.lat-mpi.eu/tools/elan/ (accessed March 4, 2012).

Farrar, S., and D. T. Langendoen (2009). An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In Andreas Witt & Dieter Metzing (eds.), *Linguistic modeling of information and markup languages: Contributions to language technology*. Berlin: Springer, pp- 45-66.

FLEx. http://fieldworks.sil.org/flex/ (accessed March 4, 2012).

Haspelmath, M. (1993). *A Grammar of Lezgian*. Berlin: Mouton.

Himmelmann, N. (1998). Documentary and descriptive linguistics. *Linguistics* 36: 161-195.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In J. Gippert, N. P. Himmelmann & U. Mosel (eds.), *Essentials of language documentation*. Berlin: Mouton de Gruyter, pp. 1-30.

Johnson, H. (2004). Language documentation and archiving, or how to build a better corpus. In P. K. Austin (ed.), *Language documentation and description, volume 2*, 140-153. London: SOAS.

LAT. http://www.lat-mpi.eu/ (accessed March 4, 2012).

Nakhimovsky, A. D., and T. Myers (2003). Digital video annotations for education. *Paper presented at the International Conference on Engineering Education*, Valencia, Spain.

Nathan, D. (2006). Thick interfaces: Mobilizing language documentation with multimedia. In J. Gippert, N. P. Himmelmann & U. Mosel (eds.), *Essentials of language documentation*. Berlin: Mouton de Gruyter, pp. 363-379.

Nathan, D. (2010). Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing* 4: 111-124.

Nathan, D. (2011). Digital Archiving. In P. K. Austin & J. Sallabank (eds.), *The Cambridge handbook of endangered languages*. Cambridge: Cambridge UP, pp. 255-274.

Palmer, A., and K. Erk (2007). IGT-XML: An XML format for interlinearized glossed texts. *Proceedings of the Linguistic Annotation Workshop*. Stroudsburg, Pennsylvania: Association for Computational Linguistics. 176-183. http://www.aclweb.org/anthology/W/W07/W07-1528.

Rogers, Ch. (2010). Technology review: Fieldworks Language Explorer (FLEx) 3.0. *Language Documentation and Conservation* 4: 74-84. http://hdl.handle.net/10125/4471.

Content Creation by Domain Experts in a Semantic GIS System

Nakhimovsky, Alexander
adnakhimovsky@colgate.edu
Colgate University, USA

Myers, Tom
tommyers@dreamscape.com
N-Topus Software, USA

1. Introduction

The focus of this paper is on how an ontology-based GIS system can be populated with class instances and further maintained by domain experts with no support from ontological engineers. This is an old goal of knowledge engineering. We present our approach in the context of a specific Semantic GIS system called EventMap (Nakhimovsky 2010), but the techniques we propose should be of general interest. At the core of our approach are two simple observations. First, instances of a given class can be described by a table in which each row corresponds to an instance and each column to a property. Such tables can be created in a context that allows data validation for data properties and, for object properties, provides access to already created instances of the object class. If the tables are kept in the cloud (e.g., Google spreadsheets) then we have a ready-made environment for joint distributed authoring, essential for creating community-based resources. The second observation is that map layers can be described by KML documents which can be created using a variety of tools ranging from ArcGIS to the free KML editor built into Google Maps / My Places. This is a flexible arrangement that makes simple things easy and difficult things possible.

Specifically, in this paper we present two mechanisms for data entry: via a blog with links to KML data, in which each blog entry corresponds to an event; and via a Web application that reads in the relevant subset of an ontology (expressed as metadata tables) and provides a form that validates new class instances against that ontology.

2. The EventMap Framework

An EventMap is a sequence of annotated maps produced by a GIS system and controlled by a Timeline: each map corresponds to a time interval, and together a map and a time interval represent an Event. The Timeline makes it possible to navigate

event sequences and observe changing event patterns over time; the patterns depend on the query posed to the GIS system. Every event is linked to an annotation that can contain arbitrary web content, including images, multimedia, and output from Web applications. Internally, the content is represented by RDF graphs; specific event sequences, their maps, and their annotations are produced by a SPARQL query (DuCharme 2011).

Instead of thinking of an EventMap as a sequence of maps each corresponding to a time interval and annotated by a web page, one can think of it as a sequence of pages each annotated by a map. A good example of this kind of book is McEvedy (2003). EventMap is a framework for creating such books online.

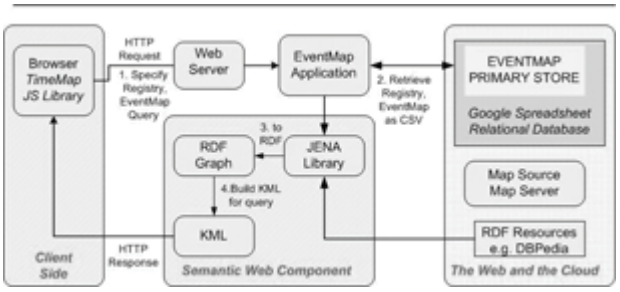


Figure 1: An Event Representation

Figure 1 shows a Google map overlaid with a scanned and edited map of Afghanistan's borders in 1879. It corresponds to the historical event of the treaty of Gandamak, formalizing the British occupation of the mountain passes into Afghanistan in response to rapid Russian advance into Central Asia. Pink strips on the Timeline correspond to events and link to event descriptions. The controls help navigate the events.

2.1. The Architecture of the Framework

The overall structure of the EventMap framework is shown in Figure 2. EventMap data are kept in a Primary Store, with links to other materials. The Store can be in any data structure that can be serialized in the Comma-Separated Values (CSV) format; this includes spreadsheet and relational database. The Store can be located in the cloud, which creates a framework for collaborative authoring that does not require advanced computer skills or server maintenance. We have been using Google spreadsheets, partly because Google provides HTML forms for editing them, and we can write code to validate user input and pre-populate the forms with data that the author must verify but rarely needs to type.

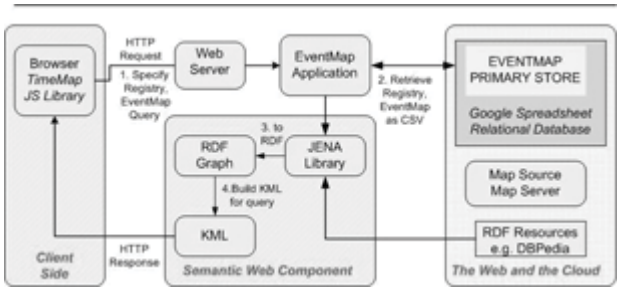


Figure 2: EventMap Architecture

In more detail, the Primary Store contains a number of Repositories, each a container of material on a specific topic, such as the history of Afghanistan borders, or the history of Arctic research during the Second International Polar Year. A repository contains the following kinds of materials:

- data tables, each corresponding to a class of objects or events
- metadata tables, one for each data table; they describe the table's schema.
- an OWL ontology that defines the classes of objects and events in data tables, and provides axioms about them.

When a repository is loaded, all the data tables are linearized as CSV (Comma-Separated Values) and passed to a Java servlet, together with the ontology. The servlet invokes the Jena library to build one big RDF graph out of the tables and the ontology. Individual EventMaps (e.g., the story of a particular Arctic expedition, or the changing patterns of networks of international cooperation) are produced by SPARQL queries from the repository graph, and sent to the timemap library (Timemap 2012) that controls the display of the map and the timeline.

3. Creating a Repository

To create a new repository one needs to

- add to the ontology as needed;
- populate data tables, using an HTML form with ontology-based validation;
- create map layers, as KML files;
- create annotations, which are Web content.

Only the first of these tasks requires a practicing ontologist; the rest is done by domain experts. The ontology and the data tables are aligned, in the sense that if a class has a certain property then the data table has a column whose header is the name of that property. In addition, if the class name has a Wikipedia entry with an infobox, we use the same property names as the infobox, gaining additional alignment with DBPedia.

4. Eventmap Authoring

We can usefully distinguish two levels of authoring: working with events within an existing EventMap, and creating a new EventMap. To add an event, one first needs to create a KML file with a placemark or several Placemarks for the event, and an HTML file with the event's description. This is the creative part; the rest is just filling out a form with time data and links to resources. To edit an existing event the user works with the same tools, but requires a different level of access to the repository. Creating a new EventMap involves creating a metadata table for it, which may be too complex for some domain experts, although most of them will probably be able to copy an existing metadata table and modify it for their needs.

To recapitulate, we are trying to provide a range of options for data entry by domain experts. At one end are users who have access to ArcGIS, Adobe Dreamweaver, and competent designers who transform their knowledge and data into compelling maps and Web pages. At the opposite end are users who are content with simple vector-graphics layers, and a simple arrangement of text and images, to tell a story or to visualize a process. For this latter case, we provide a Blogger-based interface that makes it possible to create an EventMap as a sequence of blog entries. The only constraint is that each entry needs to specify the start and end times of the event, and a link to the KML file that describes the placemark(s).

5. Comparison to similar efforts

As a story-telling and history-writing framework, EventMap resembles two other efforts: (Visual Eyes 2012) (formerly History Browser) and (Pleiades 2012). Compared to Visual Eyes, EventMap requires less of a learning curve, and is more cleanly based on standard technologies. Compared to Pleiades, EventMap is more flexible because based on dates rather than a closed list of named 'epochs.'

EventMaps should also be considered in the context of the larger Historical GIS effort (Gregory & Ell 2007). While EventMaps cannot match the scope of large national databases of geospacial data, Google spreadsheets' storage capacity and computational power provide an adequate platform for many projects in which GIS databases are used. EventMaps can use a database as well, but databases are expensive to create, maintain, and make globally accessible, while Google servers offer an unparalleled global network for free.

Acknowledgements

This work was supported by the National Science Foundation [7017695 to A.N.].

References

DuCharme, B. (2011). *Learning SPARQL*. Sebastopol, CA: O'Reilly.

Gregory, I., and P. S. Ell (2007). *Historical GIS: Technologies, Methodologies and Scholarship*. Cambridge: Cambridge UP.

McEvedy, C. (2003). *The New Penguin Atlas of Ancient History: Revised Edition*. London: Penguin.

Nakhimovsky, A. (2010). Timelines, Google Maps, and Visualization of History. Paper presented at the Biannual European Social Science History Conference (ESSHC) Ghent, April 13-16.

Pleiades (2012). `http://pleiades.stoa.org/` . Accessed March 1, 2012.

Timemap (2012). http://code.google.com/p/timemap/ Accessed February 28, 2012.

Visual Eyes (2012). `http://www.viseyes.org/` . Accessed March 1, 2012.

From Preserving Language Resources to Serving Language Speakers: New Prospects for Endangered Languages Archives

Nathan, David John
djn@soas.ac.uk
SOAS, UK

This paper will describe seismic shifts that are currently taking place in the field of endangered languages archiving, drawing on research and implementation at the Endangered Languages Archive, together with developments among other innovative archives reported at a recent workshop *Language Documentation and Archiving*. Endangered languages archives, a new type of facility of which several have sprung up over the last 15 years, have brought new considerations into account:

- endangered languages materials are predominantly in the language documentation genre, newly created through fieldwork conducted in areas where languages are threatened
- such language documentation material consists initially, fundamentally, and crucially of media recordings (audio and video) of spontaneous, naturalistic language use such as conversation (Himmelmann 1998)
- due to the nature and contexts of communities whose languages are endangered, the identities of speakers and the content of their recordings can be sensitive, so that nuanced access protocols must be designed and implemented.

These considerations have brought new types of participants into engagement with each other, some for the first time. Linguists are now found to be working at one moment with members of remote communities and soon after with audio-video specialists, archive technicians, and even journalists and filmmakers. Over the last decade, some researchers have advocated for more direct community participation (e.g., Grinevald 2003), and some archives have steadily built up trust relationships with communities. However, the field has generally remained in a 'steady state', emphasising language as data and building infrastructure for research interoperability, in a 'language resources' approach.

Now, the situation is suddenly changing, largely as a result of shifting expectations amongst language community members. Following projects funded by DoBeS[1], ELDP,[2] and DEL,[3] more members of communities have been exposed to language documentation and become aware not only of its potential but also of its methods. Digital technologies central to language documentation – media devices (including mobile phones), computers, and the World Wide Web – are rapidly becoming available to people in language endangerment 'hotspots', such as in much of Africa, Asia and South America. The shift has been precipitated by growing expectations of participation, control, and personal and local relevance, fostered by social networking platforms such as Facebook, Orkut, YouTube and Twitter. In fact, many communities are first experiencing the internet through these platforms and see the web as a more social, participatory and personally relevant space than those who were 'first generation' netizens.

While many communities are keen to participate in the new media landscape (deliberately or incidentally as new providers of language documentation), the challenge is to reconcile their contributions with the structures and policies of archives. Mechanisms need to be created and monitored to determine if increased access to a variety of naturalistic language recordings creates more value for their audiences than is lost through perceived methodological limitations in data collection (Trilsbeek & König 2011). Despite the potential problems of crowdsourcing, it provides a way for language speakers to establish real links with resources, rather than being merely 'participant metadata'. Here is an example: the web is currently trending towards mash-up pages, mobile 'apps', and aggregating portals. These gather resources based on a particular user's preferences, and display them according to topic, geographic location, or language. But what happens when that user wants to view information connected to a specific person, say language speaker X? Unless speaker X is a true participant in the digital platform as a member, owner, or even curator, rather than a mere meta-data-point, then such a speaker-centred page will be an incomplete and insipid representation, with weak implementation of access control because nobody except speaker X can properly decide who he/she wants to share with. We are fortunate, therefore, to see the maturation and continual rise of online and mobile social networking and innovative 'apps' which personalise individuals' interactions, provide further exemplars for implementing participatory models, and organically solve protocol problems through individually-managed access control.

Mary Linn (2011), for example, has proposed a framework called Community Based Language Archiving, in which the language community is involved in every step, from documentation planning to curating to dissemination. It turns traditional

archiving on its head, because the primary curatorial task, namely contextualisation, becomes about the context of *the users themselves*, because it is they (especially as community members newly welcomed into the archive ecology) who ultimately determine the success of archives in meeting their goals.

Other proposals are for decentralised, web-based functions that allow language speakers to interact with archives' existing resources. They can add further materials, comments, or contextualisation. They can identify themselves or their relatives in order to claim their moral rights in recordings and other materials, and make corrections to erroneous data, interpretations, and attributions (Garrett 2011).

Existing archivists also need to regain participatory roles. The era of the engineer as the pace-setter is giving way to usage-driven design and evolution, catering for the shifting interests of users. Archivists, whose job entails understanding their audiences, can contribute not only to contextualisation of materials and curation of exhibitions etc but also to software design issues.

Participants at a workshop on language documentation and archiving (Nathan ed. 2011) came up with a number of emerging trends that can be expected to be increasingly influential dynamics for our archives:

Form of documentation: Despite extensive theorisation of documentation in previous work, there has been little discussion of the *form* of documentation: its granularity, structure, organisation, links, and how it is to be navigated. Archives which attempt to provide attractive and usable interfaces that encourage user engagement will find themselves frustrated with current models for language documentation, which still see documentation as consisting of unstructured collections of data. New genres of expression will be needed, and these will result from collaborative efforts among documenters, archives, and contributors and users.

Community curation and contextualisation: 'Community curation' represents a paradigm-changing challenge, changing roles within archives. The original contributors and their language-learning community members become the principal curators, presenting a radical but rational inversion: the archive concept of 'context' is no longer that of the materials or their (supposed) provenance, but of the *users,* where users are often the language speakers. Contextualisation of materials is at the heart of archiving, but in many of our current archives, the art of contextualisation has been replaced by the science of software development. However, communities may wish to play a role in framing the interpretation of *their* materials to *others* (cf. Christen 2011: 197). Similarly, community access to materials is not reducible to file transfer, but entails access to *meaning* (Christen 2011: 194). Issues of quality control (Trilsbeek & König 2011) become just one of many elements to be renegotiated between technical 'best practices' and community aspirations and activities.

Promotion: Archives need to do more than acquire, curate, preserve and disseminate materials. To reach target audiences for language revitalisation, archives also need to actively *promote* the materials they hold (Wilbur 2011; Woodbury 2011). Archives need to develop relationships with their audiences that are not based purely on access to language materials, for the success of archive outreach may depend on first developing contact, relationships and trust in order to encourage usage or other participation with the materials. Ultimately, social networking provides promotion through social endorsement.

Publishing: Paradoxically, even as archives seem to become fragmented and fluid as they respond to audience participation, they will come to be seen less as archives and more as publishers, due to their increased attention to genre, exhibitions, promotion, and other outreach.

Endangered languages archives are an essential component of ethical and effective responses to language endangerment. However, the 'language resources' approach is giving way to a participatory one, because neither the quality of documentary materials, nor the effectiveness of technologies are meaningfully measurable without considering audiences and usages. Full engagement with language speakers through social networking will provide new sources of data for researchers as well as new forms of cultural repatriation for communities, and new ways of supporting threatened languages.

References

Christen, K. (2011). Opening archives: Respectful Repatriation. *American Archivist* 74(1): 185-210.

Garrett, E. (2011). Web software for participant-driven language archiving. In D. Nathan, David (ed.), *Proceedings of Workshop on Language Documentation and Archiving*. London: SOAS, pp. 71-72.

Grinevald, C. (2003). Speakers and documentation of endangered languages. In P. K. Austin (ed.), *Language Documentation and Description* 1: 52-71.

Himmelmann, N. (1998). Documentary and Descriptive Linguistics. *Linguistics* 36: 161-195.

Linn, M. S. (2011). Living archives: A community-based language archive model. In D. Nathan (ed.), *Proceedings of Workshop on Language*

Documentation and Archiving. London: SOAS, pp. 59-70.

Nathan, D., ed. (2011). *Proceedings of Workshop on Language Documentation and Archiving*. London: SOAS.

Trilsbeek, P., and A. König (2011). Increasing the usage of endangered languages archives in the years to come. In D. Nathan (ed.), *Proceedings of Workshop on Language Documentation and Archiving*. London: SOAS, pp. 45-49.

Wilbur, J. (2011). Think globally, archive locally: Opportunities and challenges in working with local archiving institutions. In D Nathan (ed.), *Proceedings of Workshop on Language Documentation and Archiving*. London: SOAS, pp. 51-58.

Woodbury, A. (2011). Archives and audiences: toward making endangered language documentations people can read, use, understand, and admire. In D. Nathan (ed.), *Proceedings of Workshop on Language Documentation and Archiving*. London: SOAS, pp. 11-20.

Notes

1. Dokumentation bedrohter Sprachen, funded by the Volkswagen Foundation. http://www.mpi.nl/DOBES .
2. Endangered Languages Documentation Programme, SOAS, funded by the Arcadia Fund. http://www.hrelp.org/ .
3. Documenting Endangered Languages, National Endowment for the Humanities. http://www.neh.gov/grants/guidelines/del.html

Retrieving Writing Patterns From Historical Manuscripts Using Local Descriptors

Neumann, Bernd
neumann@informatik.uni-hamburg.de
University of Hamburg, Germany

Herzog, Rainer
herzog@informatik.uni-hamburg.de
University of Hamburg, Germany

Solth, Arved
solth@informatik.uni-hamburg.de
University of Hamburg, Germany

Bestmann, Oliver
7bestman@informatik.uni-hamburg.de
University of Hamburg, Germany

Scheel, Julian
julian@jusst.de
University of Hamburg, Germany

1. Introduction

Computer-supported retrieval of manuscripts based on the visual features of a query image is a highly desirable, but rarely available service for manuscript research. The service could be used, for example, to check whether a manuscript, specified by a copy, is contained in a museum collection. This kind of retrieval is often approximated by making use of an index based on textual annotations, and thus requires extensive manual preparation. Retrieval based on a query image without annotations, on the other hand, promises to be mainly automatic and also support interesting applications beyond document retrieval. Most importantly, this service can allow retrieval of manuscripts containing the query image as a detail. For example, one could find manuscripts where characters are written in a specific way, exemplified in the query image. Moreover, one could search for the occurrence of writing patterns consisting of arbitrary graphical features, in short graphs. Similar graphs, retrieved from different manuscripts, may contribute valuable information about a possible scribe identity or a common origin of manuscripts.

In this contribution we describe a novel approach for graph retrieval based on local descriptors at 'interest points'. Interest points (IPs) specify locations of strong 'cornerness' of the image intensities and thus provide reasonably stable reference points for

local descriptions. They have proved their worth in many image analysis applications, in particular in image retrieval solutions based on SIFT features (Lowe 2004). Different from SIFT features, our descriptors are not scale and rotation invariant, although tolerant to small variations, giving rise to a distinctly superior performance and efficiency while preserving its usefulness for many concerns of manuscript research. Basically, each descriptor consists of the structure tensors (Koethe 2003) in the neighborhood of an IP, thus giving a precise account of the local gradient distribution. Depending on the resolution of the manuscripts and the chosen size of the neighborhood, the descriptor of a single IP may comprise several hundred feature values, called *IP features* in the sequel. For highly detailed query images, for example depicting a Chinese character, there may be a large number of IPs (in our experiments up to 40), each of which is recorded with its location and its feature values in a local descriptor.

For retrieval, a target image is processed essentially in the same way as the query image, i.e. IPs and IP features are determined. This is done only once and off-line, comparable to establishing an annotation. The main retrieval task is then to find a subset of IPs in a target image whose relative locations and feature vectors best match the query descriptors. We will present an approach which profits from prior segmentation of the target image into possible matching candidates, but can also be applied to large datasets without segmentations. It is controlled by a simple probabilistic model for the kind of differences between query and data which should be tolerated for a retrieval, with parameters which can be adjusted by a manuscript researcher.

First experimental results with Chinese characters in historical manuscripts indicate that our descriptor is tolerant with respect to a certain amount of variations of the same character, yet quite discriminative with respect to structurally different characters, promising high precision and recall.

In the remainder of this abstract, we describe related work in Section 2, give details about the technical implementation in Section 3, and finally report about experimental results in Section 4.

2. Related Work

Retrieving graphs from manuscripts is a special case of Content-Based Image Retrieval (CBIR), a well-established research field with a rich set of methods (Datta et al. 2008). But CBIR applied to manuscripts has found very little attention in this community. Most relevant work builds on handwriting recognition which is increasingly applied to historical manuscripts (Fischer et al. 2009; Yosef et al. 2007; Lavrenko et al. 2004; Shrihari et al. 2005). Most approaches so far use retrieval based on words or characters (Lavrenko et al. 2004; Adamek et al. 2007; Srihari et al. 2005; Zhang et al. 2004) which limits the applicability to handwritings with clear word or character separation. A more general approach, as followed in this contribution, relies on retrieving a spatial configuration of graphical features, extracted from the query (Su et al. 2009). For all approaches, the key question is which features to use for the comparison of query and data. It has been shown that the spatial distribution of gradients gives the best results (Zhang et al. 2004; Ding et al. 2007). In most approaches, descriptors based on simple gradient computations (Srihari et al. 2005; Ding et al. 2007) are assigned to a fixed mesh covering a segment. In our work, descriptors are based on structure tensors (Koethe 2003) at IPs determined for each query independent of a segmentation. Furthermore, we use a novel probabilistic model applicable to IPs.

Since our focus is on the retrieval of arbitrary writing patterns, we do not discuss work on OCR here, although OCR could also be an application area worth exploring with our approach.

3. Technical Implementation

Due to the restricted length of this abstract, we cannot provide many technical details here. The IPs in our approach are computed using the Harris Corner Detector (Harris and Stephens, 1988) with the improvements introduced in (Koethe, 2003). Fig. 1 shows typical results achieved for manuscripts with Chinese and Arabic handwritings.

Local descriptors are computed for each IP by combining the structure tensors of all points in the neighbourhood (in our examples of size 11 x 11) into a large feature vector characterizing strength and directionality of intensity gradients.

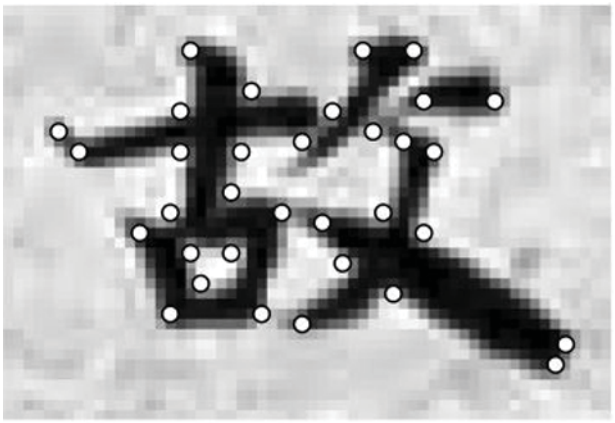


Figure 1a: Interest points (IPs) determined (a) for a Chinese character in a 60 x 41 image and (b) for a 86 x 107 segment of Arabic handwriting

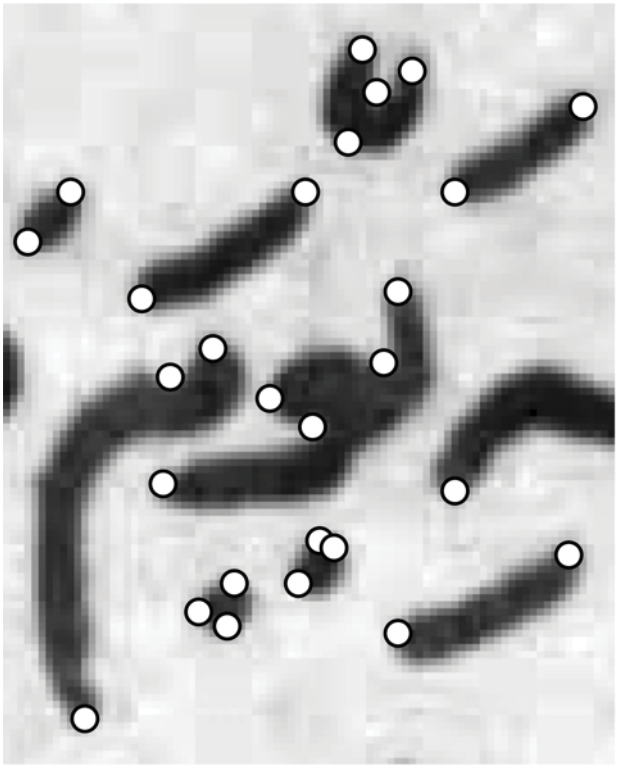


Figure 1b: for a 86 x 107 segment of Arabic handwriting

To retrieve matching patterns from target data, IPs of the query are incrementally compared with IPs of the data using a probabilistic model comprising the following boolean probabilities, both for the hypothesis H_1 that the target is a match, and the hypothesis H_0 that it is not a match:

P_{IP} target descriptor missing in window at query location

P_{loc} target descriptor dislocated from query location

P_{feat} target descriptor features differing from query descriptor features

Let A be pairs of descriptors of query and target, $P(H_0)$ and $P(H_1)$ the prior probabilities of the respective hypotheses, then for a hypothesis test we have to evaluate:

$$\frac{P(A|H_1)}{P(A|H_0)} = \frac{P_{ip}(A|H_1)\,P_{loc}(A|H_1)\,P_{feat}(A|H_1)}{P_{ip}(A|H_0)\,P_{loc}(A|H_0)\,P_{feat}(A|H_0)} > \frac{P(H_0)}{P(H_1)}$$

The comparison is formulated as two hypothesis tests, the first whether the query pattern is contained in the target, and the second whether the target pattern, constrained by a successful first test, is contained in the query. The target is considered a match of the query, if both tests succeed.

4. Experimental Results

We have carried out first retrieval experiments with Chinese and Arabic manuscripts. Fig. 2 left shows a section of the Fo shuo Tiwei jing (British Library Or.8210/S. 2051). The left-most white box marks a character used as a query, the other boxes show matching characters found by the retrieval system. There have been no false negatives, but the second hit in Column 6 is a false positive, although quite similar to the query. Similar results have been achieved for other queries, applied to a manuscript section with 364 characters.

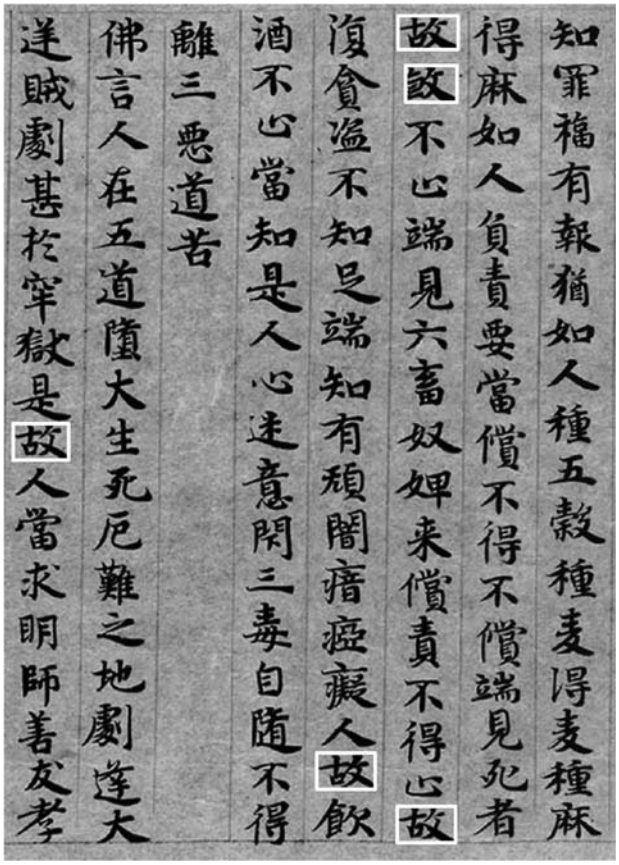


Figure 2a: Retrieval from a Chinese (left) and an Arabic manuscript (right), see text

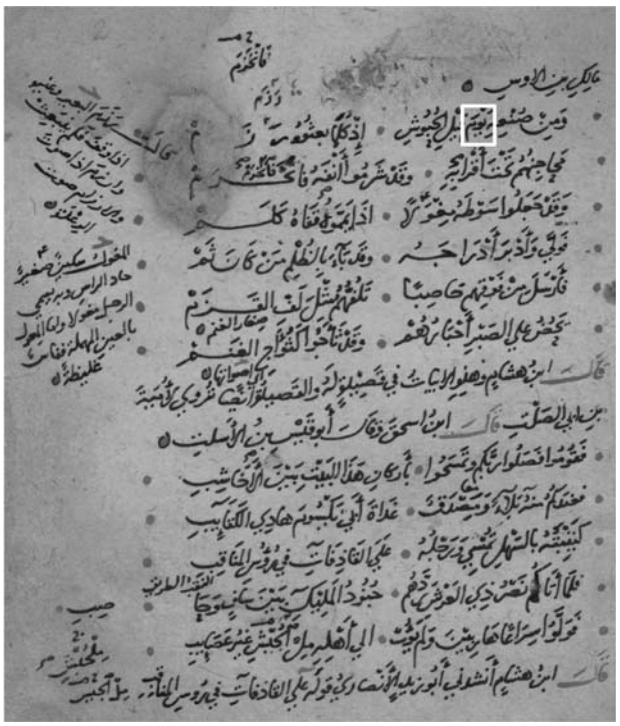


Figure 2b right shows part of Vollers0015,S.2 from the Refaiya Library in Leipzig. The section marked with a box (shown also in Fig. 1b) was used as a query for the same manuscript and retrieved as the only hit as to be expected. Further experiments are on the way

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the Research Group 'Manuscript Cultures in Asia and Africa' and the Collaborative Research Center for the Study of Manuscript Cultures SFB 950.

References

Adamek, T., N. E. O'Connor, N. Murphy, and A. F. Smeaton (2007). Word Matching Using Single Closed Contours for Indexing Handwritten Historical Documents. *International Journal on Document Analysis and Recognition* 9(2-4): 153-165.

Datta, R., D. Joshi, J. Li, and J. Z. Wang (2008). Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys* 40(2): 1-60.

Ding, K., Z. Liu, L. Jin, and X. Zhu (2007). A Comparative Study of Gabor Feature and Gradient Feature for Handwritten Chinese Character Recognition. *Proceedings International Conference on Wavelet Analysis and Pattern Recognition,* pp. 1182-1186.

Fischer, A., M. Wüthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz (2009). Automatic Transcription of Handwritten Medieval Documents. *Proceedings 15th International Conference on Virtual Systems and Multimedia,* pp. 137-142.

Harris, C., and M. Stephens (1988). A Combined Corner and Edge Detector. *Proceedings 4th Alvey Vision Conference,* pp. 147–151.

Koethe, U. (2003). Edge and Junction Detection with an Improved Structure Tensor. *Proceedings 25th DAGM Symposium,* pp. 25-32.

Lavrenko, V., T. Rath, and R. Manmatha (2004). Holistic Word Recognition for Handwritten Historical Documents. *Proceedings Document Image Analysis for Libraries (DIAL),* pp. 278-287.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision* 60(2): 91-110.

Srihari, S. N., C. Huang, and H. Srinivasan (2005). A Search Engine for Handwritten Documents. *Proceedings SPIE-IS&T Electronic Imaging,* pp. 66-75.

Su, T.-S., T.-W. Zhang, D. J. Guan, and H. J. Huang (2009). Off-line Recognition of Realistic Chinese Handwriting Using Segmentation-free Strategy. *Pattern Recognition* 42(1): 167-182.

Yosef, I. B., I. Beckman, K. Kedem, and I. Dinstein (2007). Binarization, Character Extraction, and Writer Identification of Historical Hebrew Calligraphy Documents. *International Journal on Document Analysis and Recognition (IJDAR)* 9: 89-99.

Zhang, B., S. N. Srihari, and C. Huang (2004). Word Image Retrieval Using Binary Features. *Proceedings Document Recognition and Retrieval XI,* pp. 45-53.

Distractorless Authorship Verification

Noecker Jr., John
jnoecker@jgaap.com
Evaluating Variations in Language Laboratory,
Duquesne University, USA

Ryan, Michael
mryan@jgaap.com
Evaluating Variations in Language Laboratory,
Duquesne University, USA

1. Introduction

In traditional authorship attribution, our task is to assign an author label (or a similar categorical label such as genre, publication date, etc.) to a work of unlabeled authorship. In the closed-set problem, we assign a label from a set of potential authors for whom we have some labeled training data. In the open-set problem, we also allow for the answer 'none of the above'. We build upon this here with the authorship verification task, which is essentially an open-class authorship attribution problem with only one author in the candidate pool. Thus for a given document D and candidate author A, we attempt to answer the question 'Was D written by A?'.

2. Background

Previous approaches to this problem [1-2] have involved the creation of a distractor set, which is normally controlled for genre, tone, length, etc. and performing an a traditional authorship attribution-style analysis to see whether the unlabeled document is more likely to be by the candidate author or one of the authors in the distractor set. This approach is not ideal because it relies heavily on the creation of an appropriate distractor set. That is, enough information was available about the candidate author and the training documents to choose a set of distractor authors that were appropriate for the task. Thus, although these methods performed well at the verification task, they do not lend themselves well to automation. Indeed, the entire result of this type of authorship verification hinges upon the documents chosen for the distractor set. This creates a sort of chicken-and-egg problem wherein it is necessary to know the answer in order to evaluate the suitability of the distractor set, yet it is necessary to know whether the distractor set is appropriate in order to evaluate the results.

We will attempt to remedy the errors introduced by the distractor set by eliminating the set entirely. Instead, we will consider only the document in question as well as a sample of writing known to belong to the candidate author. Thus, the validity of the verification task hinges only on obtaining a representative model of the candidate author's work, a requirement shared by traditional verification tasks, and does not involve any guesswork.

3. Materials and Methods

3.1. Distractorless Authorship Verification

Goal: Given a document *D*, and a candidate author *A*, determine the likelihood that *D* is written by *A*.

Method:

1. Compile a set of *training data*, which is known to be written by *A*.
2. Compile a *model* from the training data. This is normally accomplished by extracting linguistic or token-level features from the text and compiling a feature vector using any of various standard techniques from the authorship attribution field. We will label this feature vector $M = \langle m1, m2, \ldots, mn \rangle$.
3. Extract a *feature set*, *F*, from *D* in the form of $F = \langle f1, f2, \ldots, fn \rangle$, where *fi* corresponds to *mi* for all *i*.
4. Choose a 'distance like' function, δ, such that if $\delta(x, y) > \delta(x, z)$, we can say that *x* is 'closer to' or 'more similar to' *y* than to *z* (in some meaningful way).
5. Choose a threshold, *t*, such that if $\delta(M, F) > t$, we accept the premise that *M* and *F* are written by the same author, *A*. This threshold is found empirically by analyzing the average δ between documents of the same author.

3.2. The Corpora

To evaluate the performance of our authorship verification algorithms, we used two publically-available corpora. We made use of the Ad-hoc Authorship Attribution Competition corpus (AAAC) [3] and the PAN 2011 Authorship Identification Training Corpus [4].

Ad-hoc Authorship Attribution Competition Corpus

The AAAC was an experiment in authorship attribution held as part of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities.

The AAAC provides texts from a wide variety of different genres, languages and document lengths, assuring that the results will be useful for a wide variety of applications. The AAAC corpus consists of 264 training documents, 98 test documents and 63 authors, distributed across 13 different problems (labeled A-M). A description of each problem is found in Table 1.

Problem	Language	Description	Number of Authors	Training Docs.	Test Docs.
A	English	Student Essays	13	38	13
B	English	Student Essays	13	38	13
C	English	Novels	4	17	9
D	English	Plays	3	12	4
E	English	Plays	3	12	4
F	Middle English	Letters	3	60	10
G	English	Novels	2	6	4
H	English	Speech Transcripts	3	3	3
I	French	Novels	2	5	4
J	French	Cross-Genre	2	5	2
K	Serbian-Slavonic	Cross-Genre	3	14	4
L	Latin	Poetry	4	6	4
M	Dutch	Student Essays	8	48	24

Table 1: AAAC Breakdown

PAN 2011 Authorship Identification Training Corpus

The PAN 2011 Authorship Identification Training Corpus consists of 'real-world texts' (described by the originators as 'often short and messy' [4]). These texts appear to come primarily from the Enron Email Dataset [5]. Each text contained authorship information within the text itself, which was removed during the preprocessing stage. They are included to ensure that the results obtained herein are applicable in real-world situations, and to avoid overtraining on the AAAC data. The PAN corpus contained 5,064 training documents and 1,251 test documents across 10 authors.

3.3. Preprocessing

Preprocessing was performed on the text based on current best practices from traditional authorship attribution. Both corpora were preprocessed to standardize whitespace and character case. Any sequence of whitespace characters in the documents converted to a single space, and all characters were converted to lower case. As previously mentioned, for the PAN corpus, we also removed the author tags from each document.

3.4. Features

We used a variety of features, also chosen for their known performance in traditional authorship attribution, in our approach. Current research [6] shows that character n-grams are strong performers in traditional authorship attribution tasks. As such, we have focused on these features, examining results for character n-grams for n from 1 to 20. For completeness, we also provide results for word n-grams for n from 1 to 10. Here, a word is defined as any series of non-whitespace characters separated by whitespace. The n-grams are generated using a sliding window of size *n* and slide 1. We have limited ourselves to these simple features as they can be calculated very rapidly and without risk of error (such as that introduced by imperfect part-of-speech taggers), and thus lend themselves well to rapid, confident analysis.

3.5. Author Model

In order to perform the distractorless authorship verification, it is necessary to accurately model the writing style of the candidate author. We accomplished this using the centroid of the feature vectors for each training document. The centroid was calculated by using the average *relative* frequency of each event across the training documents, to adjust for variations in training document length.

3.6. Analysis Method

For this study, we have limited ourselves to the use of a normalized dot-product (*Cosine Distance*) analysis method. This method was shown in [7] to be among the best performing and simplest methods for authorship attribution. The advantage to using the Cosine Distance, particularly in conjunction with the simple features described above, is that it is possible to perform this verification extremely quickly, even on very large data sets. In order to answer the verification task, we need only consider the dot product of the unknown document with the candidate author model. So, let:

$$\delta(M,F) = \frac{M \cdot F}{\|M\|\|F\|} = \frac{\sum_{i=1}^{n} m_i f_i}{\sum_{i=1}^{n} m_i^2 \sum_{i=1}^{n} f_i^2}$$

3.7. Evaluation Metrics

For each set of results below we report the 'accuracy' of a particular experiment. For the purposes of this paper, we define the accuracy as the number of correctly classified texts divided by the total number of attempts.

4. Results

4.1. AAAC Corpus Results

Characters

For the AAAC Corpus with character n-grams, we achieved our best results using Character 12-grams. Our highest accuracy was 87.44% with a threshold, t, of t=0.099387. For character n-grams of various values of n, our best accuracies varied from about 86% to about 88% as seen in Figure 1.

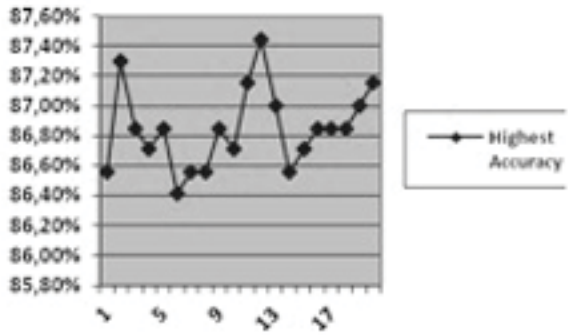


Figure 1: AAAC Character n-grams by Highest Accuracy

Words

For the AAAC Corpus with word n-grams, we achieved our best results using Word 4-grams. Our highest accuracy was 88.04. These results were achieved by setting a threshold t=0. For word n-grams of various values of n, our best accuracies varied from about 86% to about 88% as shown in Figure 2.

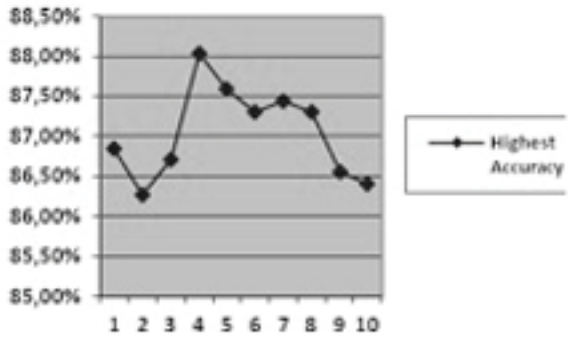


Figure 2: AAAC Word n-grams by Highest Accuracy

4.2. PAN Corpus Results

Characters

For the PAN Corpus with character n-grams, we achieved our best results using Character 7-grams. Our highest accuracy was 92.23% with *t = 0.1643* and *t = 0.1707* respectively. For character n-grams of various values of n, our best accuracies varied from about 90% to about 92% as shown in Figure 3.

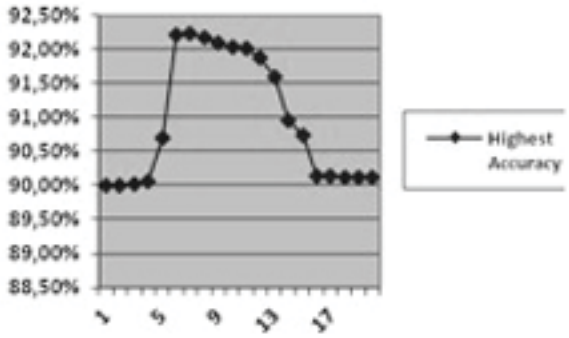


Figure 3: PAN Character n-grams by Highest Accuracy

Words

For the PAN Corpus with word n-grams, we achieved our best results using Word 2-grams. Our highest accuracy was 91.53%. These results were achieved by setting a threshold, *t*, of *t = 0.1518*. For word n-grams of various values of n, our best accuracies varied from about 90% to about 92% as shown in Figure 4.

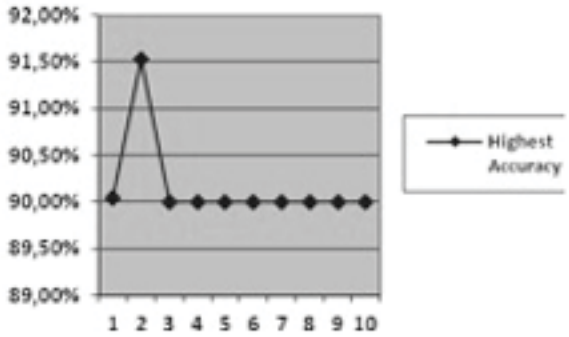


Figure 4: PAN Word n-grams by Highest Accuracy

5. Discussion and Conclusion

These results are as expected, and show that distractorless authorship verification is a viable technique with high accuracy even on extremely difficult problems. The lower accuracy rates on the AAAC Corpus are expected, as this corpus contains many times more authors than the PAN corpus with fewer training documents per author. Despite this, we were able to achieve accuracy rates of up to 88% on this more difficult corpus, and 92% on the 'easier' PAN corpus.

It should be noted that the results of this technique are extremely tunable. For instance, we can tune for any combination of accuracy, precision and recall as desired. That is, this distractorless authorship verification technique allows us to tune the Type I (false positive) vs. Type II (false negative) error rates

depending on the application of the technology. For instance, in a forensic context we may want to err on the side of saying 'no', and thus prefer to reduce false positives, while in a less stringent application we may wish to improve the overall accuracy at a cost of possibly having more false positives.

Given the performance on both the PAN corpus, which consisted mainly of real-life e-mail messages, and the AAAC corpus, which contained a wide variety of documents, the distractorless authorship verification technique shows promise for a wide range of genres and document lengths, and appears to work across a variety of languages, all without much tuning. Indeed, the only difficulty appears to be in finding an appropriate threshold, *t*, for given candidate author. This is possible by analyzing the average δ between documents by the candidate author. It can also be approximated from a large corpus, as was done here, although these results make it clear that there will be some of the same problems in controlling the corpus for genre, document length, etc. that were present in forming a distractor set for traditional authorship verification.

Future work will focus on improving these results, mainly through the addition of confidence ratings for the verification answers. That is, by allowing the system to decline to answer some percentage of the verification questions asked (or, effectively, to answer 'I don't know'), we have seen some improvement in these experimental results. Although full results for this study are not yet available, we have seen that by dropping the 20% 'most difficult' verification problems, we see an increase in accuracy to approximately 96% (from 92% on the strict binary problem). Whether or not this is truly an increase in accuracy depends upon the intended application of the technology, but we believe this future work will provide interesting results in application-specific tradeoffs, as described above.

Overall, we believe we have shown this distractorless authorship verification to be a useful tool on the stylometrist's workbench. Although no tool is itself a panacea, we are also planning efforts to combine this technique with a mixture-of-experts style voting system, effectively using multiple distance functions and feature sets on the same problem to increase confidence in our answer. Finally, we hope to both more fully explore the process of determining the appropriate threshold without the need for extraneous texts, and to find an acceptable 'default' threshold for cases where there is little training data.

References

[1] Koppel, M., and J. Schler (2004). Authorship Verification as a One-Class Classification Problem. *ICML '04 Proceedings of the Twenty-First International Conference on Machine Learning*. New York, NY

[2] Koppel, M., and J. Schler (2004). Text Categorization for Authorship Verification. *8th Symposium on Artificial Intelligence*.

[3] Juola, P. (2004). Ad-Hoc Authorship Attribution Competition. *ALLC/ACH 2004 Conference Abstracts*. Gothenberg: University of Gothenberg.

[4] PAN 2011 Lab (2011). Uncovering Plagiarism, Authorship, and Social Software Misuse. *Authorship Identification Training Corpus*. Amsterdam.

[5] Enron Email Dataset. `http://www.cs.cmu.edu/~enron/` .

[6] Juola, P., and M. Ryan (2008). Authorship Attribution, Similarity, and Noncommutative Divergence Measures. In *Selected Papers from the Chicago DHCS Colloquium*. 2008. Chicago, IL,: Chicago Colloquium on Digital Humanities and Computer Science.

[7] Noecker Jr, J., and P. Juola (2009). Cosine Distance Nearest-Neighbor Classification for Authorship Attribution. In *Proceedings from Digital Humanities 2009*. College Park, Md.: Digital Humanities.

Cataloguing linguistic diversity: Glottolog/ Langdoc

Nordhoff, Sebastian
sebastian_nordhoff@eva.mpg.de
Max Planck Institute for Evolutionary Anthropology, Germany

Hammarström, Harald
harald_hammarstroem@eva.mpg.de
Max Planck Institute for Evolutionary Anthropology, Germany

1. Overview

Glottolog/Langdoc is a comprehensive database linking 180k bibliographical references to 21k languoids (language families, languages, dialects). It provides extensive query possibilities for human users and subscribes to the principles of Linked Open Data (Heath & Bizer 2011) as far as machine users are concerned.

The aim of Glottolog/Langdoc to provide near-total bibliographical coverage of descriptive resources to the world's languages. Every reference is treated as a resource, as is every 'languoid' (Good & Hendryx-Parker 2006). References are linked to the languoids which they describe, and languoids are linked to the references described by them.

Computational treatment and modeling of language resources has so far mainly concentrated on major languages with a research tradition in NLP and some commercial viability. When we leave the industrialized countries, languages resources become very scarce. Treebanks or annotated corpora seem like fanciful ideas when the sum total of resources treating a language amounts to a description of its verbs and a treatise of its phonology from a local university, which is for instance the case of the Niger-Congo language Aduge. Before one can start thinking about developing a WordNet or similar larger resources for these languages, one must take stock of the resources which exist, however arcane they might be. This is one of the aims of the Glottolog/Langdoc project. The resources are tagged for resource type (grammar, word list, text collection etc), macroarea (roughly, continents), and language.

2. Use cases

Four different user groups can be distinguished: language diversity researchers, statisticians, Semantic Web engineers, and linguistic empiricists.

2.1. Linguistic diversity researchers

The first groups covers linguists interested in the world-wide distribution of linguistic diversity (cf. Evans & Levinson 2009), for the largest part typologist. These researchers are for instance interested in the distribution of subject, verb, and object in the languages of the world (SVO, SOV, VSO, OVS, OSV, VOS), or in the size of the phonemic inventory. The emerging patterns can be related to human cognition on the one hand hand (SOV and SVO have distinct processing advantages Hawkins 2004) and known migration patterns of humans as they settled the global land mass (Atkinson 2011). In order to acquire the necessary data points, description of various languages have to be perused, respecting genetic and geographical. This means that substantial bibliographical information has to be collected. Glottolog/Langdoc aims at providing near-total coverage of literature of the world's lesser known languages, including grey literature. Note that Glottolog/Langdoc only provides the bibliographical records, not the references themselves. All bibliographical information can be downloaded as txt, html or bibtex. Zotero integration is also provided. The provision of references is complemented by links to sites where a copy could be obtained (WorldCat, GoogleBooks, Open Library).

2.2. Statistical analysis

The links established between 180k references and 21k languoids allow for statistical analyses of the following kinds:

- What is the descriptive coverage of a particular language?
- What is the descriptive coverage of a particular language family or area (Hammarström & Nordhoff, in press)?

	Austronesian	non-Austronesian	Total
grammar	93 (17.82)%	114 (13.82%)	207 (15.37%)
grammar sketch	104 (19.92%)	148 (17.94%)	252 (18.71%)
phonology or sim.	55 (10.54%)	54 (6.55%)	109 (8.09%)
wordlist or less	270 (51.72%)	509 (61.70%)	779 (57.83%)

- How many languages have so far been described (Hammarström & Nordhoff 2011)?

- Status of the least described language families in the world (Hammarström 2010)?
- In which geographic area is the research focus on phonology, in which area is syntax deemed more interesting?

2.3. Semantic Web engineers

Next to XHTML, Glottolog/Langdoc data are also available as RDF, making use of a number of established ontologies such as rdfs, skos,[1] gold,[2] lexvo,[3], wgs84,[4] bibo,[5] and Dublin core.[6] This means that Glottolog/Langdoc data can be integrated into other projects making use of the aforementioned ontologies.

2.4. Empiricists

There are a number of researchers who are at unease with the current way how languages are defined and language codes assigned by the current registrar, SIL international. Some spurious languages do get codes, while some existing languages do not get a code. Some language families see a multiplication of their members (e.g. there are over 40 Quechuan languages) while this 'splitter' approach is not observed in other areas of the world (Nordhoff & Hammarström 2011). SIL draws on the Ethnologue[7] for the set of living languages it provides codes for. The Ethnologue lists about 7000 languages, but does not always disclose the information the inclusion is based on. As a result, there are a number of 'languages' which do not seem to relate to anything in the real world, dangling pointers so to speak. They have an Ethnologue entry, a name, a code assigned by SIL, but no referent. Glottolog has the stated goal to give a reference for every language it includes. The project distinguishes 'established languoids', for which scientific documentation is available, for 'provisional languoids', for which we do not have dedicated descriptions. The long term goal is to get rid of 'provisional languoids' by either finding a description, making them 'established', or by discarding them. This will allow linguists to always be able to ascertain on what grounds a given language is argued to exist.

3. Technologies

Glottolog/Langdoc draws on over 20 input bibliographies, which are enriched with information about document type and languages covered using machine-learning techniques (Hammarström, 2011). The project uses the pylons web framework. Glottolog/Langdoc supports content-negotiation (XHTML and RDF) and is part of the Linguistic Linked Open Data Cloud (Chiarcos et al. 2012).

4. Comparison

There are a number of related projects with slightly different foci. The **WEBALL** project for instance[8] lists genetic, bibliographic, and geographic information about the languages of Africa, but not beyond. We incorporate a recent version of the WEBALL database in Glottolog/Langdoc. **OLAC** [9] aggregates references to linguistic resources, but has a slightly different data model. Furthermore, OLAC uses federated data aggregated via OAI-MHP while Glottolog/Langdoc uses a static repository. OLAC does include genetic information, but this seems to be ad hoc. For instance, OLAC includes *Austronesian* and *Eastern Malayo-Polynesian*, but not the lower grouping *Oceanic* (still over 1000 languages). Furthermore, the coverage of OLAC and Glottolog/Langdoc is different. OLAC has more information on major languages, an area Glottolog/Langdoc disregards. Many of the higher numbers from OLAC come in fact from a single documentation project hosted at the MPI for Psycholinguistics in Nijmegen. These high numbers have technical reasons (every recording of the MPI archives is counted as one resource) and do not translate to an equally high number of publications. The following table gives the top languages as far as number of references are concerned for OLAC and Glottolog/Langdoc. A dagger † signals languages which profit from overcounting of resources from corpus1.mpi.nl.

OLAC	**Glottolog/Langdoc**
English (9044)	Swahili (1826)
German (5770)	Hausa (1542)
Dutch (5239)	Nama (Namibia) (1270)
Japanese (4317)	Afrikaans (1155)
Spanish (3331)	Central Yupik (1129)
Turkish (3091)	Standard French (1059)
French (1964)	Zulu (1048)
Yuracare (1576) †	South Levantine Arabic (990)
Oxchuc Tzeltal (1390) †	Tlingit (987)
Central Yupik (1312)	Gwich'in (905)
Turkish Sign Language (1157) †	Yoruba (900)
Yele (1152) †	Kabyle (879)
Aleut (1135)	Aleut (740)
Beaver (1080) †	Thai (730)
Dutch Sign Language (1063) †	Koyukon (728)
Tlingit (1028)	Xhosa (728)
North Alaskan Inupiatun (1024)	Akan (709)
Gwich'in (955)	Pulaar (703)
Czech (943)	Ewe (693)
Polish (892)	Tswana (690)

Multitree's[10] focus is on gathering all genetic classifications of the world's languages, not so much on references. Glottolog uses Multitree information

for lower levels, but an own, more conservative, classification for higher levels. Multitree has 141 families while the Glottolog main tree has 429 (including isolates).

The Ethnologue[11] finally lists languages and references, but does not subscribe to the document-centric approach Glottolog/Langdoc employs. These projects cover similar ground, but with different specializations. As an example, Ethnologue lists 45 references for Hausa, while Glottolog lists 1826; The figures for Thai are 37 and 730, respectively. Ethnologue has a similar number of languages families as Multitree (132), which is less 'splitting' than Glottolog. The Ethnologue lists population figures and language development; this information is not found in Glottolog.

The current technical limitations mean that there is a substantial duplication of work. It is hoped that the use of RDF and related technologies will lead to a decrease of this duplication of work. There is for instance no need that all these projects keep their own database mapping geographical coordinates and countries to languages, neither is there a need to have several databases of language names. Publication of these resources according to the principles of Linked Data (Chiarcos et al. 2012) will mean that the data can easily be repurposed and integrated into other applications.

Glottolog/Langdoc provides URIs for references and languoids so that other resources can easily link to or retrieve from Glottolog/Langdoc. While Glottolog/Langdoc takes a critical stance towards ISO 639-3, all relevant information can nevertheless also be accessed via the ISO 639-3 code.

References

Atkinson, Q. D. (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science* 332: 346.

Chiarcos, C., S. Nordhoff, and S. Hellmann, eds. (2012). *Linked Data in Linguistics. Representing Language Data and Metadata*. Springer. Companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

Evans, N., and S. Levinson (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Cognitive and Brain Sciences* 32: 429-492.

Fabre, A. (2005). Diccionario Etnolingüístico y guía Bibliográfica de los Pueblos Indigenas Sudamericanos. Book in Progress at http://butler.cc.tut.fi/fabre/BookInternetVersio/Alkusivu.html accessed May 2005.

Good, J., and C. Hendryx-Parker (2006). Modeling Contested Categorization in Linguistic Databases. *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. Lansing, Michigan. June 20-22, 2006 http://www.linguistlist.org/emeld/workshop/2006/papers/GoodHendryxParker-Modelling.pdf.

Hammarström, H. (2010). The Status of the Least Documented Language Families in the World. *Language Documentation & Conservation* 4: 177-212.

Hammarström, H. (2011). Automatic Annotation of Bibliographical References for Descriptive Language Materials. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke (eds.), Proceedings of the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, *LNCS*, vol. 6941. Berlin: Springer, pp. 62-73.

Hammarström, H., and S. Nordhoff (2011). How many languages have so far been described? Paper presented at NWO Endangered Languages Programme Conference, Leiden, April 2011.

Hammarström, H., and S. Nordhoff (in press). Achievements and Challenges in the Description of the Languages of Melanesia. In M. Klamer and N. Evans (eds.), Melanesian languages on the Edge of Asia, Special Issue of *Language Documentation & Conservation*.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford UP.

Heath , T., and C. Bizer (2011). *Linked Data - Evolving the Web into a Global Data Space*. San Rafael: Morgan & Claypool.

Maho, J. (2001). *African Languages Country by Country: A Reference Guide*, *Göteborg Africana Informal Series*, vol. 1. Department of Oriental and African Languages, Göteborg University, 5th ed.

Nordhoff, S., and H. Hammarström (2011). Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In Proceedings of the First International Workshop on Linked Science 2011, *CEUR Workshop Proceedings*, vol. 783*CEUR Workshop Proceedings*, vol. 783. URL http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/LISC/nordhoff.pdf.

Notes

1. http://www.w3.org/2004/02/skos/
2. http://linguistics-ontology.org
3. http://lexvo.org
4. http://www.w3.org/2003/01/geo/wgs84_pos
5. http://bibliontology.com/
6. http://dublincore.org/
7. http://www.ethnologue.com
8. http://sumale.vjf.cnrs.fr/Biblio/
9. http://www.language-archives.org/
10. http://multitree.org/
11. http://www.ethnologue.com

Geo-Temporal Interpretation of Archival Collections Using Neatline

Nowviskie, Bethany
bethany@virginia.edu
University of Virginia, USA

Graham, Wayne
waynegraham@virginia.edu
University of Virginia, USA

McClure, David
dm4fn@virginia.edu
University of Virginia, USA

Boggs, Jeremy
jkb2b@virginia.edu
University of Virginia, USA

Rochester, Eric
err8n@virginia.edu
University of Virginia, USA

In late 2009, with funding from the National Endowment for the Humanities, the Scholars' Lab at the University of Virginia Library began development of Neatline, a tool with which students, scholars, and archivists can express the geo-temporal dimensions of literary or historical collections: http://neatline.org/ . These expressions take the form of customizable, highly interpretive, interlinked, and interactive web-based timelines and maps. They are built using open-source software and standards-based approaches to geospatial data, and they permit scholars to draw heavily on digitized archival content and standardized metadata created by cultural heritage institutions. That said, each use of Neatline is imagined as a carefully-designed narrative or exhibit – a subjective story told in time and space through small-scale interpretive decision-making, rather than (as is more commonly pursued in our era of 'big data') an algorithmically-derived or data-driven geotemporal information visualization. In the broadest terms, Neatline is conceived as a contribution – in the visual vernacular – to multidisciplinary, place-based scholarship using primary sources.

This paper describes: 1) the theoretical goals of Neatline, including a role for iterative sketching (or graphesis) and hand-craftedness in digital interpretive toolbuilding; 2) the decision-making process which led us to architect Neatline not as a stand-alone tool, but as a set of mix-and-match plugins for Omeka, an open-source

platform for online collections and exhibits; and 3) preliminary assessment of ongoing work on the project, undertaken since 2010 as a partnership between the Scholars' Lab at UVa Library and the Roy Rosenzweig Center for History and New Media at George Mason University.

As an NEH-funded prototyping project, Neatline made three claims to innovation or to a shifting of the landscape of geo-temporal visualization in the humanities. First, by building on primary resources expressed in EAD (Encoded Archival Description) metadata – among other standards, such as VRA Core – Neatline creates a pragmatic path for collaboration among metadata providers and scholars. Although EAD has long been employed by academics working in concert with archivists, as in the Walt Whitman Archive, its use has typically been straightforwardly bibliographical, as a finding aid or for the production of catalogs of manuscripts and letters. EAD has rarely been used as a stepping-stone to rich, interpretive or theory-based expression (much less visualization) of the *content* of those primary resources. The Neatline project aims to demonstrate the value of archival metadata to interpretive scholarship, and thereby strengthen connections among scholars and collections stewards.

Next, Neatline aims for ease of use by scholars new to the digital humanities. This was a major argument in our bid for NEH funding, and our rationale for shifting from development of a stand-alone, downloadable tool (installable as a single, server-side application), to a functionally atomized array of interchangeable plugins for Omeka (`http://omeka.org`) is arguably the most important contribution of the project to the current scene of humanities computing software development. Neatline was imagined as a self-contained, self-service, single-function tool that nonetheless allowed expert users easy access 'under the hood' to customize and contribute to its open source code. For a stand-alone tool, this would have been the right approach; but we quickly realized that we had the opportunity to model a more productive and collaborative set of open source software practices. The Scholars' Lab has now shared source code for 9 completed or in-progress Neatline-related plugins with the Omeka developers' community, and Omeka forms the backbone for basic content management functionality in our project. We feel strongly that our shift to Omeka plugin production retains or enhances all of the desired qualities of our originally-proposed system, while adding two great benefits. First, Neatline has become mix-and match. In other words, no longer must users assent to the entire, ideal Neatline workflow in order to make use of our work – and in fact we are seeing great interest in and use of individual plugins in Omeka user communities and scholarly and archival contexts far removed from those interested in geo-temporal interpretation. Secondly, our close collaboration with the Omeka team and open source developers' community is leading to advancements in the core code of Omeka itself, again benefiting a far wider audience than we anticipated. Not only were we able to leverage Omeka as a technical and social framework for Neatline, but our work has made the system a more attractive option for research and special collections libraries – even those with sophisticated technical and repository infrastructure of their own.

Finally, Neatline makes a theoretical contribution to the digital humanities by emphasizing hand-crafted visualization as a mode of praxis and scholarly inquiry. (This is a practice we have, after the experimentation of the UVa's former 'SpecLab' thinktank and the scholarship of Johanna Drucker, called *graphesis*.) Low-tech sketching and storyboarding is regularly taught as part of the earliest design processes for digital projects in the Scholars' Lab. A side effect of our work is to demonstrate the value of iterative interpretation and knowledge-production manifested in visual form – particularly within fields like history and literary studies, in which interpretation of visual artifacts is rarely taught and drawing is infrequently modeled as (in William J. Turkel's formulation) 'a way of knowing.' We have emphasized this by designing drawing and editing interfaces that are relatively simple to use and nearly identical to a finished, end-user's view. In Neatline, scholars are always sketching, erasing, and sketching again their arguments on the screen.

Subjectivity in space and time is a matter of interest in Neatline. Our map and timeline-related plugins (in progress) are designed to offer users the ability to model multiple backgrounds or alternatives independently from the foreground of their critical attention, and express all of these fields and their interrelation visually. The spatial and temporal foreground stands as a place for scholarly commentary, including textual annotation as well as intervention in the visual field by means of the freehand drawing of lines and shapes. Meanwhile user-specified Neatline backgrounds can be empirical or unabashedly subjective – precisely geo-referenced or wholly speculative. For example, one local test case, which we will demonstrate, uses the maps and letters of U.S. Civil War cartographer Jedediah Hotchkiss. GIS (Geographic Information Systems) layers are brought into Neatline from modern satellite imagery, from scans of Hotchkiss's own surveyed-and-drawn battlefield maps, and from the historical atlases that served as the surveyor's mental 'base layer.' Against these, we have

georectified an informal, child's-eye map Hotchkiss drew in a letter, to allay his young daughter's concerns about his welfare. Neatline allows all of these documents to be plotted and annotated in space and time.

We will describe not only the theoretical basis of the project, but its concrete products, including improvements to the Omeka codebase and the creation of a suite of open-source plugins. Most of these plugins are usable independently but gain value in combination, comprising the Neatline system.

1. *EAD Importer*: allows for the easy and easily-adjusted import of Encoded Archival Description metadata into an Omeka exhibit.

2. *Neatline Maps*: provides map display through one or more GeoServer instances or through any specification-compliant WMS service.

3. *Neatline Features*: allows users to draw on map layers, encoding and edit geospatial shape information.

4. *Neatline Time*: incorporates the well-known Simile Timeline Javascript framework into Omeka to provide chronological visualization of Omeka items.

5. *Neatline Theme*: provides facilities for combining Neatline and Omeka information into unified interactive presentations.

In addition, our work on Neatline led us to develop five further plugins which users may employ as part of their Omeka exhibits. These plugins greatly extend the capacity of Omeka and make it a more attractive option not only for end-users, but for better-resourced libraries, digital humanities centers, and cultural heritage institutions – the constituency best positioned to contribute further development time to the open source code of Neatline and Omeka alike.

1. *FedoraConnector*: makes it possible to display, comment on, annotate, and otherwise employ objects inheriting behaviors from a Fedora Commons repository.

2. *GenericXmlImporter*: permits users to import any arbitrary, flat XML data into Omeka.

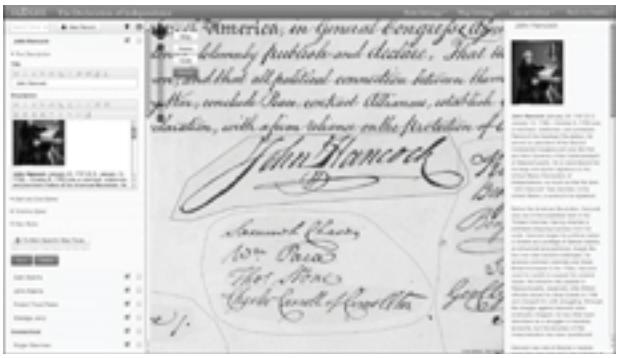
3. *SolrSearch*: allows use of the Solr search engine with Omeka, facilitating improved search and implementing faceted browsing.

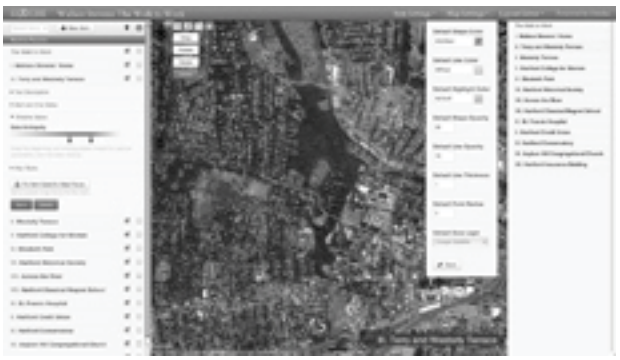
4. *TeiDisplay*: allows users to render TEI files in HTML format and attach them to Omeka items. This plugin integrates with SolrSearch for indexing.

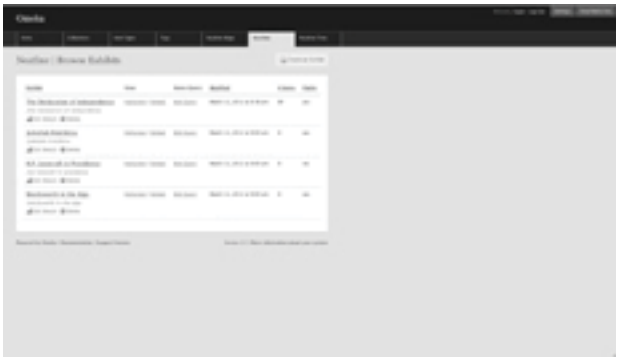
5. *VRACoreElementSet*: allows users to bring the VRA Core Element Set (designed for description of visual resources) into Omeka.

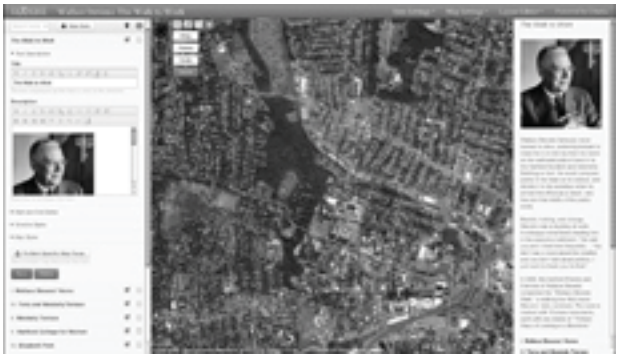
The project continues under the rubric of an 'Omeka + Neatline' partnership with the Center for History and New Media, funded through 2013 by the United States Library of Congress, with a next major code release planned for January 2012. DH 2012 in Hamburg will coincide with the release of Neatline demonstration projects and end-user documentation at `http://neatline.org/`.

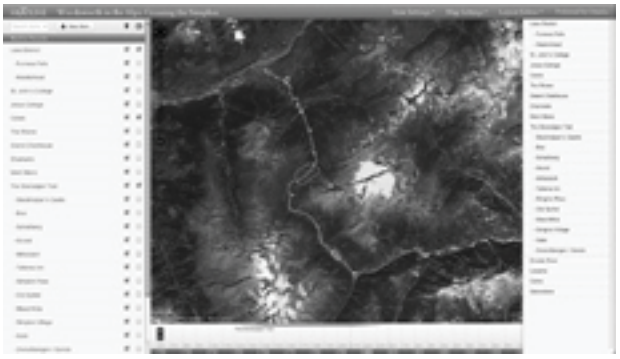
Neatline operates on archival metadata – itself already an interpretation of a literary or historical collection – to allow scholars to illustrate connections among documents and the spatial and temporal dimensions that arise through their reading. In this, it embodies a theme of much work in the Scholars' Lab: that method is a path to argument, and that interpretive digital humanities scholarship may be best enacted in iterative, visual modes.











References

Catapano, T., K. M. Price, and K. Walter (2005). The Walt Whitman Archive: Archivist-Scholar Collaboration in Description and Representation. *Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, Victoria, British Columbia, 2005.

Drucker, J., and B. Nowviskie (2004). Speculative Computing: Temporal Modelling. *A Companion to Digital Humanities*. Oxford: Blackwell.

Drucker, J. (2009). *SpecLab: digital aesthetics and projects in speculative computing*. U of Chicago P.

Kramer-Smyth, J., M. Nishigaki, and T. Anglade. *ArchiveZ: Visualizing Archival Collections*. Available: `http://www.archivesz.com/`

Neatline: Plot Your Archive in Space and Time. University of Virginia Library Scholars' Lab. Available: http://neatline.scholarslab.org/

Papers of Jedediah Hotchkiss, Accession #2822 and #2907, Special *Collections*, *University of Virginia* Library, Charlottesville, Va.

Scholars' Lab Omeka Plugins. University of Virginia Library Scholars' Lab. Available: `http://www.scholarslab.org/projects/omeka-plugins`

Turkel, W., and D. Elliott (2010). Rapid Prototyping to Support Experimental History. *Playing with Technology in History* Symposium. 30 April 2010. Available: `http://www.playingwithhistory.com/`

Enriching Digital Libraries Contents with SemLib Semantic Annotation System

Nucci, Michele
mik.nucci@gmail.com
Università Politecnica delle Marche, Italy

Grassi, Marco
margra75@gmail.com
Università Politecnica delle Marche, Italy

Morbidoni, Christian
christian.morbidoni@gmail.com
Università Politecnica delle Marche, Italy

Piazza, Francesco
f.piazza@univpm.it
Università Politecnica delle Marche, Italy

1. Introduction

The advent of revolutionary communication tools, like social media and image/video sharing, have transformed the Web into a giant collection of resources created by millions of people around the planet, representative of their different cultures and ideas. In this context, the Digital Humanities community have understood that Digital Libraries (DL) should no longer be simple 'expositions' of digital objects but rather provide interaction with users, enabling them to contribute with new knowledge, e.g. by annotating and tagging digital artefacts (Arko et al. 2006). This provides a more engaging user experience, enabling scholars to benefit from the digital world in their everyday work and researchers to work collaboratively (David et al. 2008). The crowdsourcing paradigm has been experimented, as in (Holley 2010), by leveraging users annotations to enrich the DL, helping in curating contents (correcting or signalling typos, etc.) or upload new digital material[1].

Existing annotations tools, however, are often limited to context and tags, which provide poor semantics, and, when they offer more expressive annotations, use proprietary or non-interoperable formats to represent semantics. In other words, the knowledge created by advanced users carefully annotating contents with valuable information is often lost, since it is not directly reusable by applications.

In this paper we present the prototypal annotation system developed within the SemLib EU project. The SemLib Annotation System (AS) can be easily integrated into existing DL or used through a bookmarklet to annotate generic Web pages, addressing annotations at different level of complexity and expressivity. In the prototype, users can write simple comments, augment them with links to Linked Data[2] (LOD) and semantic tags. But if they are scholars and need more expressivity they can create complex structured relations among digital objects, connecting text within different documents and relating to specific vocabulary terms.

Annotations are structured as RDF data, which is stored to a remote annotation server and can be consumed by third party applications as 'slices' of a single collaborative RDF graph.

2. The Role of Semantic Annotations

Data interoperability and reuse are the main achievements that the DL research community is trying to address by looking at Semantic Web (SW) and LOD technologies. As argued in previous researches (Kahan et al. 2001; Haslhofer 2010; Gerber et al. 2010) and witnessed by the Europeana initiative, which is adopting a RDF based data model (EDM), great expectations come for the semantic technologies. Semantically structured data is expected to provide the ability to mash-up heterogeneous information and establish connections among digital objects independently provided by different institutions.

Data interoperability and flexibility of aggregation foster reuse and enable serendipity: unexpected reuse of data by different persons and in different contexts from the one data was produced in. Users annotations, when properly represented as semantic data, can play an important role in this context, as they link documents to data.

On the one hand, semantic annotations represent precise relations and metadata and they are reusable with relatively low effort to augment the DL's data. Here, it is very important to expose easily accessible APIs to 'slice' the RDF data produced. On the other hand, LOD annotations contain links to open datasets or ontologies that can be contextualized. This allows applications to merge annotations and augment their metadata with other semantic. As an example, if documents are connected to a world wide linked data graph, composed by big datasets, vocabularies and user annotations, software applications can let users explore such a graph and possibly discover unexpected and interesting connections among them.

3. Simple Use-Case

Let us shortly describe a simple use case that illustrates the kind of scenario that AS should address.

Alice is a researcher in politics and annotates a transcription of an election rally. She writes a comment to a passage and then clicks the 'extract tags' button. AS uses DBPedia Spotlight[3] (or possibly other multi language entity extraction tools) to propose a set of entity tags, using remote metadata to build a preview with pictures, descriptions, etc. Alice chooses a number of tags including the politician 'X' who is mentioned in the transcription and then she loads an 'emotions vocabulary' that allows her to express a precise relation (e.g. 'is a manifestation of') between the annotated sentence and the term 'rage' from the vocabulary. Alices annotations are collected, among others by a virtual community of users, into a Web site exposing a RDF based faceted browser, where they are merged with further metadata about the DBPedia entities cited in annotations and ontologies.

A second user, Bob, made a similar annotation, semantically linking to the politician 'Y' from DBPedia. Since in DBpedia the two politicians are connected by their metadata (e.g. they are from the same party or they are born in the same city, etc.), the two entities can have one or more paths that connect them in the graph. Bob can explore such paths and discover a relation among his annotation and the one from Alice. In addition, when Bob search for the keyword 'anger', to find objects related to that emotion, he will be able to find the transcription annotated by Alice, since an emotion ontology is available and it defines 'rage' as a narrower term of 'anger'. At this point, Bob could also decide to create a new semantic annotation expressing a similarity between the two transcriptions, for example, using a simple 'is similar to' relation.

Using the AS, DL administrators or DL owners/ maintainers can make their own annotations but they can also select relevant end-user contributions, aggregate them and then publish back as trusted/ official annotations, importing them to enrich the DL.

4. Annotations Data Model

Annotations represent a peculiar type of resources, specifically conceived to add information to other existing resources. Annotations acquire therefore full significance in relation with the target resource and other contextual information, such as its author, its creation date and the vocabulary terms used. Properly structuring an annotation using SW technologies is therefore necessary at twofold level to clearly separate contextual metadata from its content. The AS data model reuses and extends OAC ontology (Sanderson et al. 2011) to encode contextual metadata and to attach annotations to involved Web resources. Since OAC specification makes no assumption on the kind of body an annotation can have, one of the first issues that have been tackled in SemLib project is how to represent annotations that have an RDF graph as body. In AS data model, annotation body is a named graph containing a set of RDF statements. This allows to represent semantically structured annotation content basing on the standard support for named graphs provided by RDF triplestores and SPARQL in addition to query and access only little slices of the entire collaborative knowledge base (KB).

The annotation storage is agnostic with respect to the ontologies used to represent the informative content of annotations. Pluggable ontologies and SKOS taxonomies can be used to encode annotations content according to standardized domain vocabularies to foster data interoperability between different communities and DL.

5. Annotations Sharing Model

In AS users collect their annotations in notebooks to organize their work, aggregating annotations by topic or task, and to easily make available to others subsets of their annotations enabling collaborative scenarios. Notebooks are identified by dereferenciable URLs that applications can use to retrieve RDF-encoded annotations and relative metadata in different formats.

Sharing a notebook is as easy as sharing its URL on the Web, similarly to what happens for popular file sharing platforms and common mechanisms like e-mails, wikis or social networks can be used. For each notebook, the owner can easily set fine-grained rights access, also making it public or sharing it only with selected users.

Before creating and sharing annotations each user must signing into the annotation system. To make as easy as possible the integration of the annotation system in existing DL, single sign-on systems like OpenID and OAuth have been used. Different authentication systems can be supported developing dedicated plugins.

6. Prototype

At the time of writing, a fully working AS prototype[4] has been developed and will be further enhanced in the continuation of the project. The prototype can be used in any existing Websites through a dedicated

bookmarklet, which will injects the needed Javascript code into the Web pages.

System architecture, shown in Fig. 1, is typically client-server. The client-side component, which comprises a set of modules implementing the graphical user interface to create and browse annotations, is developed in Javascript. The server-side component is a RESTful Web service which implements the main application logic and the storage for the annotations based on a native Sesame triplestore with the support for semantic inference.

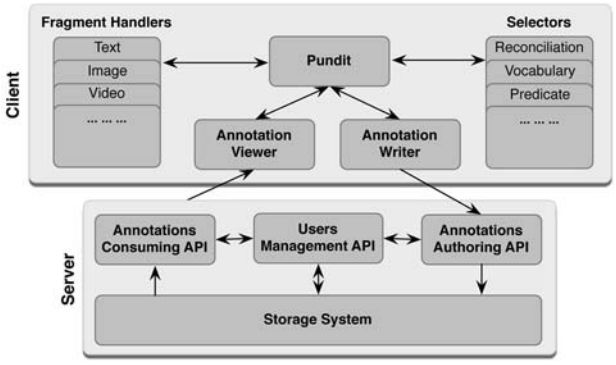


Figure 1: Simplified architecture of the Annotation System

The AS prototype allows users to compose RDF statements using an on purpose GUI (Fig. 2), connecting selected items with properties coming from pluggable ontologies.

Figure 2: An example of the annotations editor

Selected items can be DL resources, named entities coming from custom SKOS vocabularies or Web reconciliation services like Freebase[5] Web pages or fragments of these. For example, users can create structured annotations in Dante Alighieri's DL page to reference other pages of the DL about Italian writers that have been influential for him, to provide additional information about his place of birth referring to Florence entity in Freebase or to link text-excerpts of the Divine Comedy coming from the Gutenberg Project[6].

The AS prototype includes a module for comments and tags. At the moment, it provides basic functionalities including entity extraction from user comments (based on DBPedia Spotlight) and semantic tagging.

To demonstrate how annotation can be consumed basing on simple REST API, the AS prototype provides also an example faceted browsing facility, implemented using Simile Exhibit[7].

7. Conclusions

In this work, SemLib annotation system have been introduced, discussing its data and annotation model and presenting a working annotation system prototype, discussing how this is expected not only to foster annotation sharing between DL and user engagement but also to allow the application of crowdsourcing paradigm in the creation of added value for the DL.

Fundings

The research leading to these results has received funding from the European Union's Seventh Framework Programme managed by REA-Research Executive Agency [SEMLIB - 262301 - FP7/2007-2013 - FP7/2007-2011 - SME-2010-1].

References

Arko, R. A., K. M. Ginger, K. A. Kastens, and J. Weatherley (2006). *Using annotations to add value to a digital library for education* Online: `http://www.dlib.org/dlib/may06/arko/05arko.html` (accessed on 7th March 2012).

David, S., M. Nucci, and F. Piazza (2008). Talia: a Research and Publishing Environment for Philosophy Scholars. *Proceedings of the Digital Humanities 2008 Conference*, Oulu, Finland, 25th-29th June, 2008.

Gerber, A., and J. Hunter (2010). Authoring, Editing and Visualizing Compound Objects for Literary Scholarship. *Journal of Digital Information* 11.

Haslhofer, B., E. Momeni, M. Gay, and R. Simon (2010). Augmenting Europeana Content with Linked Data Resources. *Proceedings of the 6th International Conference on Semantic Systems (I-Semantics)*, September 2010.

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine, The Magazine of Digital Library Research*, March/April 2010.

Kahan, J., and M. R. Koivunen (2001). Annotea: An Open RDF Infrastructure for Shared Web Annotations. *Proceedings of the 10th international conference on World Wide Web*, pp. 623-632.

Morbidoni, C., M. Grassi, and M. Nucci (2011). Introducing SemLib Project: Semantic Web Tools for Digital Libraries. *Proceedings of the International Workshop on Semantic Digital Archives – sustainable long-term curation perspectives of Cultural Heritage held as part of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL)*, Berlin, 29th September 2011.

Sanderson, R., and H. van de Sompel (2011). *Open Annotation: Beta Data Model Guide*, Online: http://www.openannotation.org/spec/beta/ (accessed on 7th March 2012).

Notes

1. BBC WW2 People's War: http://www.bbc.co.uk/ww2peopleswar/
2. Linked Data: http://linkeddata.org/
3. DBPedia Spotlight: http://dbpedia.org/spotlight
4. For for a working demo see http://metasound.dibet.univpm.it/release_bot/release/semlib-client_0.61-demo_page/examples/demo.html
5. Freebase: http://www.freebase.com/
6. Gutenberg Project: http://www.gutenberg.org/
7. Exhibit: http://www.simile-widgets.org/exhibit/

The VL3: A Project at the Crossroads between Linguistics and Computer Science

Nuñez, Camelia Gianina
cnunez3@uwo.ca
The CulturePlex Laboratory, The University of Western Ontario, Canada

Mavillard, Antonio Jiménez
ajimene6@uwo.ca
The CulturePlex Laboratory, The University of Western Ontario, Canada

The current study will introduce the VL3 (Virtual Language Learning Lab) a multidisciplinary project that brings together a variety of disciplines: from Theoretical Linguistics, Second Language Acquisition and Second Language Pedagogy to Computer Science and Natural Language Processing (NLP). The VL3 stands as a great example of how humanistic knowledge, when combined with the power of the technology, can lead to the creation of innovative tools, beneficial to society at large.

Specifically, the VL3 aims to provide a virtual environment where (Spanish) second language learners can work towards improving their communicative skills in the target language, by participating in a set of predefined conversation scenarios that closely mimic real life situations.

As mentioned earlier, the VL3 is heavily based on research within the fields of Linguistics, Second Language Acquisition and Second Language Pedagogy; as such an important part of this presentation will focus on introducing and explaining the theoretical context for the development of the current project. We will then proceed to show how this research has been applied in order to further develop and refine the existing spoken dialogue system technology used by our partner, Natural Language Inc.

Although the importance of linguistic theory (phonology, morphology, syntax) to the advances in the field of NLP has been pointed out by many (Matthews1993; Schultze & Gupta 2010), it continues to be a relationship that needs to be explored in more detail. Without a doubt, greater attention paid to linguistic research would lead to even more significant developments for natural language processing technologies. Skeptics of such technologies claim that computer programs will

never be able to account for the complexity of the human language (Salaberry 1996) and although it may well be the case that we will not be able to fully describe something as unpredictable as human language, it is only by maintaining a close relationship with linguistics, its discipline of study, that any real attempt can be made.

As such, in order to meet our objective of creating a highly developed technology capable of processing human language as a whole with the highest degree of accuracy possible, we will recur to recent theories of linguistics and second language acquisition.

Language technologies such as ones described above, are used in telephone conversations where a person needing information is speaking to a machine trained to understand human language (call centers, for example). Although these technologies have enjoyed great success, they are most often limiting, in sense that they are only trained to deal with highly specialized language. For example a NL software employed by an internet service provider will only 'understand' language that is specific to this context. In fact, it is often the case that these technologies are trained to only pick up on certain key words (home internet, high speed, technical problems, billing, etc.) uttered by the person on the phone and from that on, infer the purpose of the call. Should the caller ask a question 'outside' this scenario, the machine will most likely not understand. As such, it can be said that most of these technologies have divided language into specific contexts and have focused on teaching computers to 'understand' scenario-specific language. That is, they are trained to deduce the meaning of an utterance by picking up on certain key words.

The VL3 thus proposes to adopt and further adapt these technologies to the field of foreign language learning and offer students much needed opportunities to practice their target language in one-on-one conversations that imitate real life situations. This of course implies the need of less restrictive NL technology that is able to understand human language as a whole rather than only context-specific language. Consequently, such a project demands a NL technology that is based on grammatical analysis strategy rather than keyword spotting.

The technology used by Natural Language Inc., our partner in this project, is already superior to other existing ones mainly because it employs grammatical strategies for understanding human language. Nonetheless, our research experience within the fields of Linguistics, Second Language Acquisition and Second Language Teaching, will allow us to further develop the existing technology of Natural Language Inc. in order to successfully use it as a second language-learning tool. Any speaker of a foreign language is well aware of the challenge that lays in the process of learning a language other that one's own as well as the many different factors that come at play in the learning process. The field of Second Language Acquisition has shown us that one such factor is input in the foreign language. In fact, input has been described as 'the single, most important concept of second language acquisition' (Gass 1997). Input specifically refers to exposure to the target language (oral, written, formal, informal etc.) and no one would argue that a second language can be learned in isolation and with no exposure to it.

Furthermore, it has been suggested that aside from input, interaction with a native speaker also plays an important role in the process of second language development. (Long 1981, 1983, 1996). As such, producing the target language as part of the learning process (Swain 1985, 1993, 1995, 1998, 2005) has been suggested as beneficial to the learner. One of the benefits of interaction as part of the process of second language learning is what is known as 'negative evidence' That is, when the learner says something that their interlocutor does not understand, after some 'negotiations for meaning' take place, the interlocutor may model the correct language form. Consequently, this feedback received by the learner on their production will turn into more positive input. Interaction strategies between native and non-native speakers such as modifications, simplifications, paraphrasing etc. have been vastly studied as well and have clearly shown their benefits to the language learner.

Evidently, linguistic research can provide us with valuable information for the development of a project such as the VL3. Throughout this presentation you will have the opportunity to see how such research has been applied to the creation of an effective online tool that will significantly impact the field of second language pedagogy.

References

Gass, S. (1997). *Input, Interaction and the second language learner*. Mahwah, NJ: Erlbaum.

Gupta P., and M. Schultze (2010). Human Language Technologies (HLT). Module 3.5. In G. Davies (ed.), *Information and Communications Technology for Language Teachers (ICT4LT)*, Slough, Thames Valley University [Online]. Available from: `http://www.ict4lt.org/en/en_mod3-5.htm` [Accessed 10 December, 2010].

Long, M. (1981). Input, interaction and second language acquisition. In H. Winitz (ed.), *Native Language and Foreign Language Acquisition*.

Annals of the New York Academy of Sciences 379: 259-278.

Long, M. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics* 4(2): 126-141.

Long, M. (1996). The Role of the Linguistic Envronment in Second Language Acquisition. In W. R. Ritchie and T. J. Bhatia (eds.), *Handbook of Second Language Acquisition*. San Diego: Academic Press, pp. 413-468.

Matthews C. (1993), Grammar frameworks in Intelligent CALL. *CALICO Journal* 11(1): 5-27.

Salaberry R. (1996). A Theoretical foundation for development of pedagogical tasks in computer mediated communication. *CALICO Journal* 14(1): 5-34.

Swain, M. (1985). Communicative Competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass and C. Madden (eds.), *Input in second language acquisition*. Rowley, MA: Newbury House, pp. 235-253.

Swain, M. (1993). The Output Hypothesis: just speaking and writing aren't enough. *The Canadian Modern Language Review* 50: 158-164.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook and B. Seidlhofer (eds.), *Principles and Practice in the Study of Language*. Oxford: Oxford UP.

Swain, M. (1998). Focus on form through conscious reflection. In C. Daughty and J. Williams (eds.), *Focus on Form in Classroom Second Language Acquisition*. New York: Cambridge UP.

Swain, M. (2005). The Output Hypothesis: Theory and Research. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*. New Jersey: Lawrence Erlbaum, pp. 471-483.

'Eric, you do not humble well': The Image of the Modern Vampire in Text and on Screen

Opas-Hänninen, Lisa Lena
lisa.lena.opas-hanninen@oulu.fi
Department of English, University of Oulu, Finland

Hettel, Jacqueline
jacqueline.hettel@gmail.com
University of Georgia, USA

Toljamo, Tuomo
toljatuo@mail.student.oulu.fi
Department of English, University of Oulu, Finland

Seppänen, Tapio
tapio.seppanen@oulu.fi
Department of English, University of Oulu, Finland

This paper discusses the image of the vampire in the television series *True Blood* and the books by Charlaine Harris on which the series is based. Our analysis is based on cultural imagology as developed by Johnson (2005, 2006) and Lotman (1990). To this end we have used CATMA, the concordance program developed at the University of Hamburg, and added multimodal functionality to it. By time stamping and aligning the image, the soundtrack and the subtitling, we can mark up and investigate both the linguistic and the paralinguistic markers that describe the characters and their emotions. This means that we are able to move between a concordance of the subtitling and the film itself; in addition, we have linked the scenes of the film to the novels of Charlaine Harris, on which the series is based. Thus, this paper not only discusses the images created in the novels and the television series, but also presents the tools we have built to enable us to carry out the research.

Beliefs and legends about vampires have existed probably as long as mankind. The earliest known records are found among the Assyrians (Wright 2006). Lores and legends of some form of the vampire can be found all around the world. The vampire has been described as an evil, bloodsucking creature, a force of darkness. He is seductive, smart, and a powerful predator. Sometimes the vampire is seen as a spirit; some legends say they can change into animals, even control them (Nobleman 2007). Vampire literature began with the Gothic tradition, a fiction that was intended to evoke a sense of dread in the reader. During Victorian times the vampire became more human and a powerful archetype. The

common image, which began with Bram Stoker's *Dracula*, was one of a gentleman dressed in evening wear and a long cloak; he was smart, polite, wealthy and well-dressed. Of all classic horror creatures, he has had the most sex appeal and has thus easily been able to attract humans.

The image of the vampire has once again changed. He (or she) is now beautiful, erotic, powerful, young and quite human-like. He has human emotions, is dangerous when called for, but loving and good to those humans he is attached to – a modern-day Heathcliff, a dark and brooding sexual fantasy. This is largely due to Anne Rice, who changed the image of the vampire, and vampire literature more generally, with her character Lestat. She achieved this by ignoring the old conventions of what a vampire can or cannot do; she changed the traditional Gothic settings by moving them from isolated houses and areas to central locations and beautiful homes; she made the vampire the protagonist, thus moving away from the more traditional narration by the victim, and focused on their feelings and their anxieties, making them quite sympathetic (Overstreet 2006).

We investigate the *True Blood* television series and the books it is based on. These were written by Charlaine Harris between 2002 and 2010, in other words well after L.J. Smith's *Vampire Diaries* series and after the television series *Buffy the Vampire Slayer*, but before Stephanie Meyer's *Twilight* series. True Blood features Sookie Stackhouse, a human in her late twenties, a waitress in a local bar, and her relationship to Bill Compton, quite the Southern gentleman, and Eric Northman, the Viking, two vampires who are almost opposites of each other. We set out to investigate what kind of an image the viewer gets of Eric; this we did through investigating the adjectives used to describe him and those used by him. Thus our research questions were: How do others perceive Eric? How does Eric behave? How does Eric perceive others? How does he describe others? To carry out the investigation, we needed a concordance program, but the problem of course was that there simply wasn't one that could also handle the video. Thus we turned to the CATMA team at Hamburg University and asked whether we could use their program and build on it. We have now built multimodal functionality on top of CATMA and carried out our investigation on the speech of the series, calling up the video of each instance in the KWIC concordance (see Fig. 1). We then marked up the text of the original book series and linked the instances describing Eric in the book to those in the TV series.

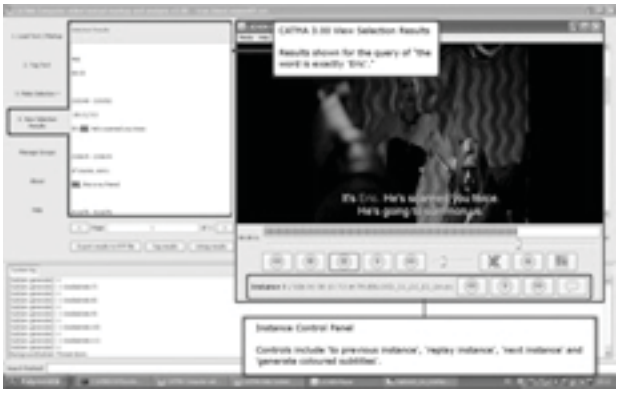


Figure 1

Our results show that in the television series *True Blood*, there is little description of Eric; the viewer's image of Eric is shaped largely by comments made about his behavior, such as 'Eric, you do not humble well' or '*I'm* not afraid of Eric' and 'Eric has had a thousand years of practicing deceit.' However, the image of Eric is not only constructed through this. In fact, the multimodality of the television series plays a crucial role in the construction of the characters more generally. Eric, for example, is created through the things he wears, the way he speaks, the things he does and the expressions on his face. Sometimes, for example, Sookie criticizes Eric harshly and his facial expression indicates that he is quite hurt, although he does not say so. Thus it is crucial that in addition to analyzing the actual speech of the characters, one can access the visual images as well – much like in modern conversation analysis, which also relies on sound and images, in addition to the actual words spoken.

Another interesting aspect of the investigation is the comparison of how the image of Eric differs between the original novels and the television series. It seems that much of what is conveyed to the viewer through the images in the television series forms what in literary studies one would think of as that part of the creation of a character in a novel which is left to the reader's imagination. It is like an interpretation of the character, perhaps even to be seen as the actor's interpretation, which then gets reinterpreted by the viewer. The image of Eric is also created through the juxtaposition of him with Sookie's other lover, Bill Compton, who was turned into a vampire after the Civil War and who is quite the Southern gentleman, particularly as opposed to Eric, who is the rogue; he is the ruthless Viking, an absolute ruler, at times rather cruel, and a sharp businessman. He is attractive to women, whom he uses ruthlessly, but he is also very protective of those he really cares for, which often is not reflected in the words he speaks, but rather in his facial expressions and gestures.

Thus, it becomes evident that in order to carry out an analysis of data such as this, it is crucial that

one has access to tools which are able to handle the multimodality of the data. Our decision to not reinvent the wheel, but to build upon an already available tool, was the right choice, but there is still work to be done to make it even more applicable to other data and other types of investigations.

References

Johnson, A. W. (2005). Notes Towards a New Imagology. *European English Messenger* 14(): 50-^58.

Johnson, A. W. (2006). New Methodologies: Imagology, Language and English Philology. In H. Anttila, J. Gear, A. Heikkinen and R. Sallinen (eds.), *Linguistic Topics and Language Teaching*. Oulu: Oulu UP, pp. 7-27.

Lotman, Yu. M. (1990). *Universe of the Mind: A Semiotic Theory of Culture*, tr. Ann Shukman, intro Umberto Eco. London and New York: Tauris.

Nobleman, M. T. (2007). *Vampires*. Oxford: Raintree Publishers.

Overstreet, D. W. (2006). *Not Your Mother's Vampire. Vampires in Young Adult Fiction*. Lanham, MD: The Scarecrow Press.

Wright, D. (2006). *The Book of Vampires*. 2nd revised and enlarged edition. New York: Dover Publications.

Electronic Deconstruction of an argument using corpus linguistic analysis of its on-line discussion forum supplement

O'Halloran, Kieran Anthony
kieran.o'halloran@kcl.ac.uk
King's College, University of London, UK

1. Introduction

In the last few years, one technological innovation has been the appending of discussion forum facilities to online arguments, such as in online versions of newspapers. The facility allows readers to post responses to an argument and to debate issues raised in it. Such online discussion forums can be regarded as electronic supplements to these arguments.

The aim of this presentation is to report on research which highlights the utility value of this electronic supplementarity for critical reading of arguments (see O'Halloran, 2010, 2011 2012). I show how a content analysis of an online discussion forum appended to an argument can illuminate repression or marginalisation in the latter of concepts which are used in normal discussion of the argument's topic. In turn, I demonstrate how this can reveal where the rhetorical structure of the argument is unstable, where it deconstructs. Since, to conduct such content analysis, I use corpus linguistic method – the analysis of collections of electronic texts – I refer to this approach to critical reading as *Electronic Deconstruction*. The theoretical stimulus for this approach comes from an appropriation of 'the supplement', an idea of the French philosopher, Jacques Derrida.

2. Keyword Analysis

2.1. Concepts normally used for discussion of a topic

If an online discussion forum is sufficiently large, and contains a substantial number of critical posts, it can provide illumination of *what* concepts are normally used in discussion of a particular topic regardless of *how* these concepts are used, e.g. whether these concepts are assented to or not. Should certain concepts be salient in the forum as a whole but absent from, or at best marginal

in, the original argument, this can offer insights into what the argument might be said to repress or marginalise from normal conceptual discussion of its topic. Since the discussion forum is directly related to the original argument, then arbitrariness will have been considerably reduced in producing these insights.

2.2. Keywords

Salient concepts in an online discussion forum, or any collection of electronic texts, can be revealed through keyword analysis using appropriate corpus linguistic software. A keyword is 'a word which occurs with unusual frequency in a given text…by comparison with a reference corpus' (Scott 1997: 236). Keywords are established through statistical measures such as log likelihood (see Dunning, 1993). A log likelihood value of ≥ 7 ($p < 0.01$) confers keyness on a word. The larger the log likelihood value ≥ 7, the greater the salience of the keyword. Importantly, the log likelihood value, as a statistical measure, reduces arbitrariness in what is selected as salient.

Comparison of the highest value keywords in a discussion forum with words in the original argument can be illuminating. For this presentation, the following types of keyword are relevant:

- high value keywords in the forum which are absent from the original argument. These are candidates for the status of repressed concepts from the argument;
- high value keywords in the forum which are used infrequently in the original argument. These are candidates for the status of marginalised concepts in the argument.

Keyword analysis should not only be quantitative. They also need to be qualitatively explored to understand how they are being used. Qualitative exploration of keywords in the discussion forum can, in turn, strengthen judgements as to marginalised / repressed candidacy in the original argument.

In order to mobilise how I use keywords in a discussion forum supplement for purposes of Electronic Deconstruction, I take as stimulus how Derrida conceives of 'the supplement'.

3. Discussion forums as supplements

3.1. Derrida's supplement

We normally think of the word 'supplement' as meaning something extra, an add-on. For Derrida, the supplement is more subtle. This is because he illuminates how a supplement 'adds only to replace' a lack of something. So, while the supplement may seem like an add-on and thus *outside* that which is supplemented, in fact it becomes simultaneously *inside* that which it is added to. Derrida writes that every supplement:

…harbors within itself two significations whose cohabitation is as strange as it is necessary […] [The supplement] adds only to replace. It intervenes or insinuates itself *in-the-place-of* (Derrida 1976 [1967]: 144-145).

Take, for example, a shop sign outside a bicycle shop. It is outside the shop not part of the inside. It is an add-on, an extra, signaling the nature of the shop. However, in being outside the shop it 'adds to replace' a 'lack' inside the shop – the shop cannot function unless it can attract custom. The shop sign is thus simultaneously an add-on and an essential part of the shop – it is both outside and oddly 'inside' the shop as well.

3.2. Discussion forums as Derridean supplements

If we apply the logic of the supplement to online discussion forums appended to arguments, then a discussion forum is not just outside the original argument. It is not just an add-on, an extra. On the logic of the supplement, concepts which are normally used to discuss a topic as reflected in keywords in an online discussion forum - but which are absent in the original argument – can be seen as 'lacking' in the argument. In turn, this can offer insights into repression or marginalisation in the argument (*relative* to the particular supplement). Furthermore, since keywords are generated non-arbitrarily, we have in turn a non-arbitrary basis for intervening in the argument. We can use these keywords to 'add to replace' what can be perspectivised as deficient in normal discussion of the argument's topic, intervening to replace an absence *inside* the argument with keywords *outside* the argument. Via the logic of the supplement, the border between an argument and its discussion forum supplement is porous.

To be clear, my use of Derrida's notion of the supplement is an appropriation of his work. Though Derrida is synonymous with an approach to critical reading called 'Deconstruction', I am not doing Derridean Deconstruction. While I appropriate the idea of supplements as being simultaneously inside and outside the things that are supplemented, this does not mean that I concur with Derrida's vision where meaning is ultimately undecideable. Indeed, a corpus linguistic approach illuminates Derrida's focus on undecideability in language and meaning to be misguided (see O'Halloran 2012).

3.3. Using the supplement to investigate deconstruction in an argument

The next stage is to trace the extent to which an intervention in the argument 'to add to replace' leads to instability in its cohesion. An argument's rhetorical structure is dependent on effective cohesion. If cohesion is disturbed by this intervention, then the credibility of the argument diminishes.

4. The argument topic: 'New Atheism'

4.1. Orientation

I demonstrate the method of Electronic Deconstruction on an argument in an online version of a British newspaper.[1] Written by the journalist Brendan O'Neill, the argument criticises 'new atheism', the atheism associated with thinkers such as Richard Dawkins. The entire argument rests on a rhetorical structural opposition: 'old atheism' (e.g. the atheism of Karl Marx) is good while 'new atheism' is bad. Using the corpus linguistic software WMatrix (Rayson 2008), O'Halloran (2010, 2011) generates keywords in the discussion forum appended to the argument.[2] 'Faith' is a keyword with high statistical significance, its meaning mostly equivalent to 'religious belief'; 'belief' in the sense of 'religious belief' also has a high keyness value. Qualitatively examining how the keywords are used in discussion forum posts, I found that discussion of 'new atheism' is in relation to criticism of belief/ faith in a supernatural power. That 'religious belief' or 'faith' are key concerns of 'new atheism' can be corroborated by examining relevant websites / books by 'new atheists'[3]

Turning my attention to the argument, 'faith' is absent from it and 'religious belief' is marginal, only occurring twice; in contrast, the argument contains 21 instances of the general category 'religion'. O'Neill describes 'new atheists' as being critical of 'religion' generally rather than of 'faith' / 'religious belief' in a supernatural power. In sum, we can say that relative to normal discourse for 'new atheism' (as reflected in the discussion forum), there is a deficiency in the argument of 'faith' / 'religious belief'.

4.2. Exploring possible deconstruction in rhetorical structure

Below is an example of deconstruction in O'Neill's argument. Sentences 8-10 set up the rhetorical structure of 'new atheism' (bad) versus 'old atheism (good)':

8. History's greatest atheists, or the 'old atheists' as we are now forced to call them, were humanistic and progressive, critical of **religion** because it expressed man's sense of *higher* ethical purpose in a deeply flawed fashion [emphasis added].

9. The new atheists are screechy and intolerant; they see **religion** merely as an expression of mass ignorance and delusion.

10. Their aim seems to be, not only to bring belief [in] God crashing back down to earth, but also to *downgrade* mankind itself [emphasis added].

Notice 'new atheists' are described as LOW and 'old atheists' as HIGH (see italics).

What happens to the cohesion of this rhetorical structure if I 'add to replace' the deficiency of 'faith' and 'religious belief' in normal discussion of 'new atheism' (as reflected in the discussion forum) in sentences 9 and 10? (I do not intervene in sentence 8 since this relates to 'old atheism'):

8. History's greatest atheists, or the 'old atheists' as we are now forced to call them, were humanistic and progressive, critical of **religion** because it expressed man's sense of *higher* ethical purpose in a deeply flawed fashion [emphasis added].

9. The new atheists are screechy and intolerant; they see **{religion}** *religious belief* merely as an expression of mass ignorance and delusion [emphasis added].

10. Their aim seems to be, not only to bring *belief [in]* God crashing back down to earth, but also to *downgrade* mankind itself [emphasis added].

As a result of this intervention, the general category of 'religion' in sentence 8 is no longer in cohesion with this general category in sentence 9. Since the rhetorical structure of 'old atheists' = good / high and 'new atheists' = bad / low is dependent on this cohesion, the rhetorical structure can be said to deconstruct. Thus, once normal discourse for 'new atheism' is instituted, the argument can be evaluated as falling apart structurally.

References

Derrida, J. (1976 [1967]). *Of Grammatology* [trans G.C. Spivak] Baltimore: Johns Hopkins UP.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-74.

O'Halloran, K. A. (2010). Critical reading of a text through its electronic supplement. *Digital Culture and Education* 2(2): 210-229.

http://www.digitalcultureandeducation.com/cms/wp-content/uploads/2011/06/DCE1022_ohalloran_2010.pdf

O'Halloran, K. A. (2011). Limitations of the logico-rhetorical module: Inconsistency in argument, online discussion forums and Electronic Deconstruction. *Discourse Studies* 13(6): 797-806.

O'Halloran, K. A. (2012). Electronic deconstruction: revealing tensions in the cohesive structure of persuasion texts. *International Journal of Corpus Linguistics* 17(1): 91-124.

Rayson, P. (2008). Wmatrix: a Web-based Corpus Processing Environment. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix.

Scott, M. (1997). PC analysis of key words – and key key words. *System* 25(2): 233-45.

Notes

1. The argument consists of 926 words (42 sentences). The whole argument and the discussion forum appended to it can be found at: http://www.guardian.co.uk/commentisfree/2007/dec/30/thenewatheism.
2. The discussion forum appended to the argument consists of 365 individual posts, which in total come to 69, 252 words. In order for keywords to be revealed, WMatrix compares the discussion forum with a *reference corpus*of around 1 million words of written English (which WMatrix accesses online).
3. In a three minute clip, Richard Dawkins summarises a central argument of 'The God Delusion': 'The great majority of people on this planet do **believe** that there is some kind of a supreme being…historically it's always been very important to people their **belief** in some sort of a supreme being…I give in the book the argument…that there is no supernatural supreme being and that **belief** in such a being can under some circumstances be rather a bad thing' [my bold]. The clip can be viewed at: http://www.amazon.com/The-God-Delusion-Richard-Dawkins/dp/0618918248/ref=sr_1_1?ie=UTF8&qid=1330788377&sr=8-1 [Accessed March 2012].

Citygram One: Visualizing Urban Acoustic Ecology

Park, Tae Hong
park@gsu.edu
Georgia State University, USA

Miller, Ben
miller@gsu.edu
Georgia State University, USA

Shrestha, Ayush
ashrestha2@student.gsu.edu
Georgia State University, USA

Lee, Sangmi
esangmi@gmail.com
Georgia State University, USA

Turner, Jonathan
johnturner@me.com
Georgia State University, USA

Marse, Alex
aemarse@gmail.com
Georgia State University, USA

1. Introduction

The surfaces and spaces of our planet, and particularly metropolitan areas, are not only manifestations of visible physical shapes, but also include dimensions that go beyond the ocular. Furthermore, these invisible energies are often in a state of flux and could potentially yield insights into the living and breathing dimensions of our surroundings that are underexplored in contemporary mapping research. The Citygram project explores and develops ideas and concepts for novel mapping paradigms that reflect aspects of dynamicity and 'the invisible' pertinent to cityscapes. The project's main goal is to contribute to existing geospatial research by embracing the idea of time-variant, poly-sensory cartography. Our multi-layered and multi-format custom maps are *data-driven* and based on continuous data streams captured by remote sensors. These dynamic maps can inform us of the flux in our built environments, such as parks, museums, alleys, bus stations, freeways, and university campuses.

Citygram is a large-scale project divided into a number of manageable iterations. The current iteration focuses on visualizing acoustic ecology and computationally inferring meaningful information such as emotion and mood. We are currently focusing our attention on small public spaces at the following

university campuses: Georgia State University and Georgia Institute of Technology in Atlanta; California Institute of the Arts in Los Angeles; and Massey University in Wellington, New Zealand. Our aim is to render dynamic, non-intrusive acoustic maps that are scale accurate and topologically oriented.

2. Current Mapping Paradigms and Technologies

The most commonly used mapping paradigms are represented by Google Maps and and Bing Maps (PC World 2010). Many supplemental features are provided by these platforms: street-views, webcams (1 frame/hour), and pre-recorded videos taken from locations. Other mapping systems, such as UCLA's *Hypercities* ('Hypercities' 2010), BBC's *Save Our Sounds* ('Save Our Sounds' 2011) and the *Locus Stream Project* ('Locus Sonus > Locustream' 2011) are more creative. *Hypercities* is an interactive mapping system that allows temporal navigation via layers of historic maps, which are overlaid on modern maps via the *Google Earth* interface. *Save Our Sounds* gathers crowd-sourced audio snapshots, archives them, and makes 'endangered sounds' available on a web interface. The *Locustream SoundMap* project is based on the notion of networked 'open mic' audio streaming where site-specific, unmodified audio is broadcast on the Internet through a mapping interface. It seems that the project has its roots in artistic endeavors where 'streamers,' persons who install the *Locustream boxes* in their apartments, share 'non-spectacular or non-event based quality of the streams' thereby revisiting a mode of listening exemplified by the composer John Cage.

3. Citygram One: Goals, Structure, and Focus

The project's main goals are to (1) investigate potential avenues for capturing the flux of crowds, machines, ambient noise, and noise pollution; (2) map, visualize, and sonify sensory data; (3) automatically infer and measure emotion/mood in public spaces through sound analysis, pattern recognition, and text analytics; (4) provide clues to waves of contemplation and response in public spaces; (5) provide hints into the invisible dynamics between Citygram maps and aspects of urban ecology such as crime and census statistics, correlations to municipal boundaries, population density, school performance, real-estate trends, stock-market correlations, local, regional, national, and global news headlines, voting trends, and traffic patterns; (6) seek additional regional, national, and international collaborators.

Figure 1 shows the system overview. The data collecting devices (d) are small computers that wirelessly transmit feature vectors ($v[n]$) to a central server. The minicomputers autonomously run the signal processing modules for the optimal efficiency of the entire system. Low-level acoustic feature vectors streamed to the server are *unintelligible* and cannot be exploited to reconstruct the original audio waveforms or 'eavesdrop' on private conversations. This protects the privacy of individuals in public spaces. However, to enable users to hear the 'texture' of a space without compromising privacy we employ a modified granular synthesis techniques (Gabor 1946) which results in shuffling the temporal microstructure of audio signals.

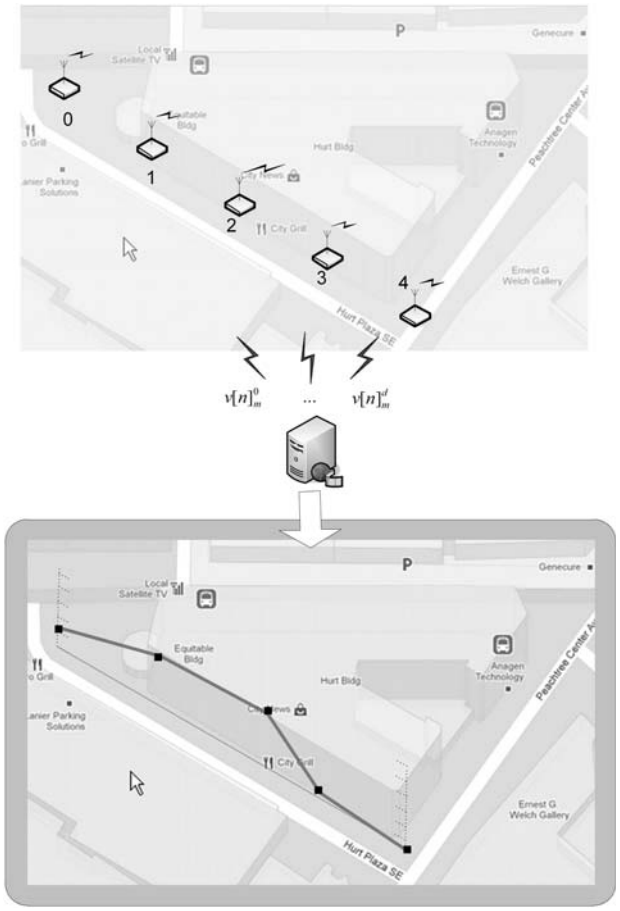


Figure 1: System architecture Top: Autonomous remote sensing devices transmitting feature vectors Bottom: Measurement of acoustic energy (dB)

Citygram focuses on inferring useful information from invisible energies. One type of avisual information is environmental emotion or mood. Although interdisciplinary sentiment research is still in its nascent stages, automatic mood detection has found increasing interest in the field of music (Liu et al. 2003; Wieczorkowska et al. 2006; Meyer 2007), speech analysis (Chuang & Wu 2004; Vidrascu 2005; Schuller 2008; Shafran 2011), face-recognition (Cohn & Katz 1998; Yoshitomi et al. 2000; Nagpal

et al. 2010), and natural language processing (Xia et al. 2008; Tao 2004). As far as emotion detection in music is concerned, much of the research can be traced to Kate Hevner's (Hevner 1936) seminal work in musical expression (see Figure 2) and other models developed by Thayer and Tellegen-Waston-Clark (Thayer 1989; Tellegen et al. 1999) as shown in Figure 3 and 4.

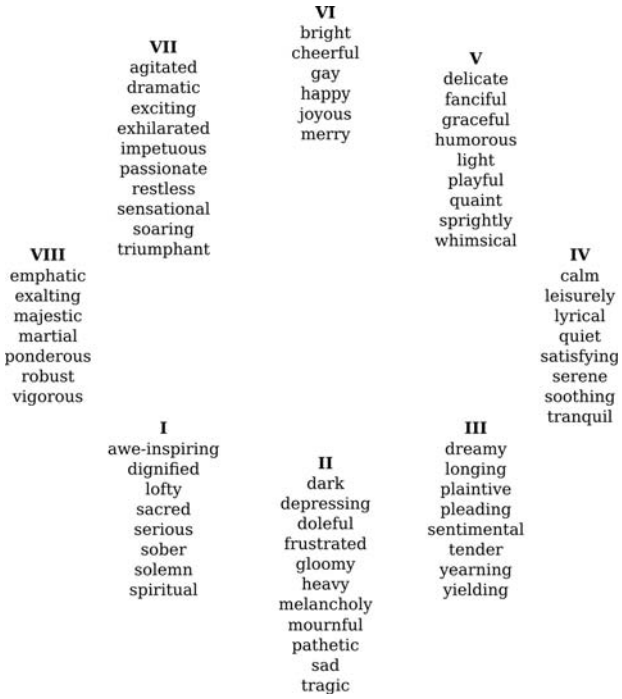


Figure 2: Hevner's adjective circle

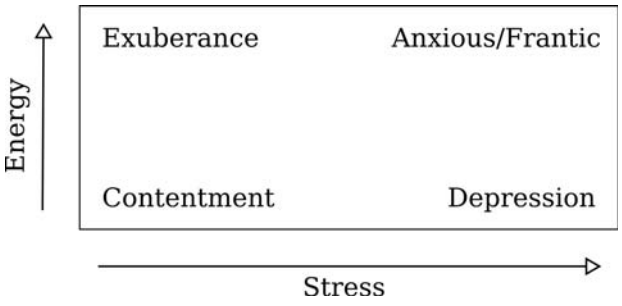


Figure 3: Thayer's model of mood

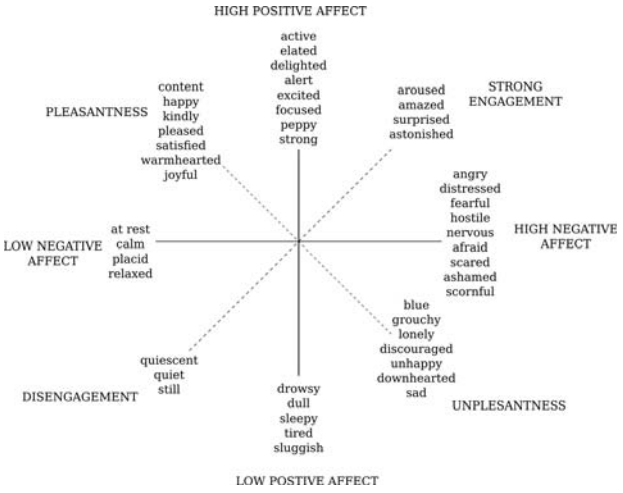


Figure 4: Tellegen-Waston-Clark mood model

4. Automatic Emotion Detection

The fundamental architecture of automatic emotion detection is shown in Figure 5. Features are extracted from audio signals, followed by training of the machine learning (ML) algorithms via a set of training samples; at the final stage, the machine classifies emotion ratings for unknown signals. Although much of the emotion/mood detection research for sound has been in the realm of music, there is strong potential that algorithms and methodologies similar to those used in music will translate to non-musical signals – many of the low-level feature vectors are *timbral* rather than *musical* and reflect acoustic dimensions of sound.

We are exploring application of radial and elliptical basis functions (RBF/EBF) artificial neural networks (ANN) for automatic emotion classification as it has shown promise in previous research (Park 2004; Park and Cook 2005). We plan to compare Support Vector Machines (SVM) with RBF/EBF ANNs as each is based on contrasting approaches – RBF/EBF is hyperspherical/hyperellipsoid based and SVM is hyperplane based. We are also investigating 'knowledge-based' approaches for automatic verbal aggression detection in acoustically complex social environments already in use in public spaces (van Hengel et al. 2007). Feature vectors will be archived on our server to provide users a hub for future research on public spaces.

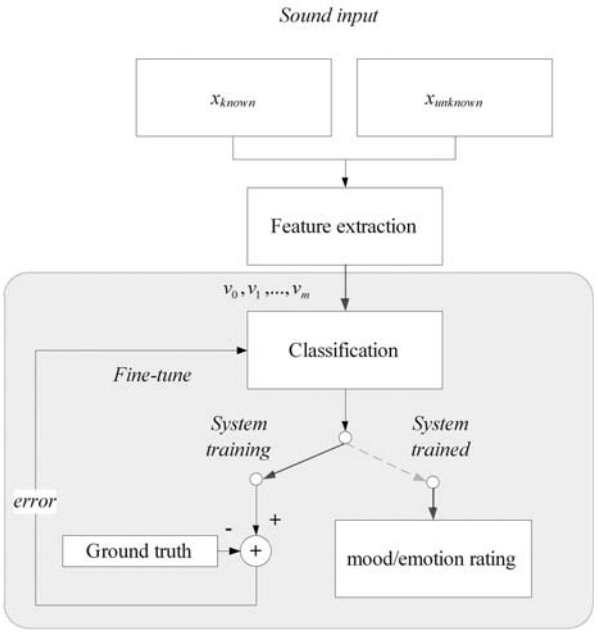


Figure 5: Mood/Emotion Detection Architecture

Natural language processing techniques will further enrich the project's assessment of environmental sentiments via site-specific emotional readings from on-line sources. Recent examples of localized text sentiment visualization from blogs and micro-blogs can be found in 'Pulse of the Nation' (Mislove et al. 2010) and 'We Feel Fine' (Harris & Kamvar 2011). In conjunction with spatial acoustic emotion detection, this analysis will help us render comprehensive, dynamically changing emotion indices through our on-line visualization mapping formats. The general structure for sentiment analysis in the context of ML is similar to general practices in ML for sound; both apply algorithms aimed at revealing patterns through features. We are currently concentrating on keyword spotting (Ortony et al. 1987) supplemented by 'common sense reasoning.' Keyword spotting relies on a pre-established lexicon of scored, significant language. Common sense reasoning utilizes a database like Open Mind Common Sense (OMCS) to infer meaning from text. A toolkit like ConceptNet is used in conjunction with OMCS to facilitate sentiment analysis. This combination will allow us to recognize both pre-coded and emergent sentiments. For instance, a keyword approach utilizes lists of significant language and modifiers to recognize that 'unhappy,' preceded by 'very,' is more significant than just 'unhappy' by itself. Recognizing that a capitalized or exclamatory modifier, i.e.'VERY!!,' accentuates the keyword and intensifies its positive or negative valence (intrinsic attractiveness vs. aversiveness) scale, requires OMCS.

This research, to better understand the dynamic life and mood of urban spaces via acoustic ecology, mapping, pattern recognition, and data visualization, is the first iteration in the larger Citygram project. Subsequent iterations of this project will include analyzing additional types of avisual energies, such as electromagnetic fields, humidity, temperature, color, and brightness and extending this dynamic, poly-sensory cartographic data into a resource for inquiry on topics ranging from urban ontology and climate change to sonification, and environmental kinetics.

References

Baum, D. (2006). Emomusic – Classifying Music According to Emotion. *Proceedings of the 7th Workshop on Data Analysis*, Kosice, Slovakia, July 2009.

British Broadcasting Corporation (2011). Save Our Sounds. `http://www.bbc.co.uk/worldservice/specialreports/saveoursounds.shtml` (accessed 10 October 2011).

Chuang, Z., and H. Wu (2004). Multi-Modal Emotion Recognition from Speech and Text. *International Journal of Computational Linguistics and Chinese Language Processing* 9(2): 45-62.

Cohn, J., and G. Katz (1998). Bimodal Expression of Emotion by Face and Voice. *Proceedings of the Sixth ACM International Multimedia Conference,* New York, NY, September 1998.

Cohn J., and G. Katz (1998). Workshop on Face / Gesture Recognition and Their Applications. *Proceedings of the Sixth ACM International Multimedia Conference*, New York, NY, September 1998.

Davies, A. (2011). *Acoustic Trauma : Bioeffects of Sound*. Honors Thesis, University of South Wales.

Devillers, L., V. Laurence, and L. Lori (2005). Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection. *Neural Networks* 18(4): 407-422.

Fragopanagos, N., and J. Taylor (2005). Emotion Recognition in Human-Computer Interaction. *Neural Networks* 18(4): 389-405.

Gabor, D. (1946). Theory of Communication. *Journal of the Institute of Electrical Engineers* 93: 429-457.

Gupta, P., and N. Rajput (2007). Two-Stream Emotion Recognition for Call Center Monitoring. *Proceedings of Interspeech,* Antwerp, Belgium, August 2007.

Harris, J., and S. Kamvar (2011). We Feel Fine and Searching the Emotional Web. *Proceedings of the Web Search And Data Mining Conference,* Hong Kong, China, November 2011.

Hevner, K. (1936). Experimental Studies of the Elements of Expression in Music. *American Journal of Psychology* 48: 246-268.

Kessous, L., G. Castellano, and G. Caridakis (2004). Multimodal Emotion Recognition in Speech-Based Interaction Using Facial Expression, Body Gesture and Acoustic Analysis. *J Multimodal User Interfaces* 3: 33-48.

Li, T., and M. Ogihara (2003). Detecting Emotion in Music. *Proceedings of the International Society for Music Information Retrieval*, Baltimore, Maryland, October 2003.

Li, T., and M. Ogihara (2004). Content-Based Music Similarity Search and Emotion Detection. *Proceedings of the International Society for Music Information Retrieval*, Quebec, Canada, May 2004.

Liu, D., L. Lu, and H. Zhang (2003). Automatic Mood Detection from Acoustic Music Data. *Proceedings of the International Society for Music Information Retrieval,* Baltimore, Maryland, October 2003.

Liu, H. (2002). Automatic Affective Feedback in an Email Browser. In *MIT Media Lab Software Agents Group*.

Liu, H., H. Lieberman, and T. Selker (2003). A Model of Textual Affect Sensing Using Real-World Knowledge. *International Conference on Intelligent User Interfaces,* Miami, FL, January 2003.

Liu, H., and P. Singh (2004). Conceptnet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal* 22: 211-26.

Locus Sonus (2008). LocusStream. `http://locusonus.org/w/?page=Locustream` . (accessed 10 October 2011).

Meyers, O. C. (2007). *A Mood-Based Music Classification and Exploration System*. Master's thesis, Massachusetts Institute of Technology.

Mislove, A., S. Lehmann, and Y. Ahn (2010). Pulse of the Nation: U.S. Mood Throughout the Day Inferred from Twitter. `http://ccs.neu.edu/home/amislove/twittermood/` . (accessed 1 November 2011).

Nagpal, R., P. Nagpal, and S. Kaur (2010). Hybrid Technique for Human Face Emotion Detection. *International Journal of Advanced Computer Science and Applications* 1(6).

Null, C. (2010). Which Online Mapping Service Is Best? *PC World*. `http://www.pcworld.com/article/206702-2/which_online_mapping_service_is_best.html` (accessed 30 October 2011).

Ortony, A., G. L. Clore, and M. A. Foss (1987). The Referential Structure of the Affective Lexicon. *Cognitive Science* 11: 341-364.

Park, T. H. (2004). *Towards Automatic Musical Instrument Timbre Recognition*. Ph.D. thesis, Princeton University.

Park, T. H., and P. Cook (2005). Radial/ Elliptic Basis Functions Neural Networks for Musical Instrument Classification. *Journées d'Informatique Musicale*.

Presner, T. (2010). Hypercities: Building a Web 2.0 Learning Platform. In A. Natsina and T. Kyalis (eds.), *Teaching Literature at a Distance*.London: Continuum Books, pp. 171-182.

Schnaps, J. (2009). Age and Clarity of Imagery. `http://sites.google.com/site/earthhowdoi/Home/ageandclarityofimagery` (accessed 10 October 2011).

Schuller, B. W. (2008). Speaker, Noise, and Acoustic Space Adaptation for Emotion Recognition in the Automotive Environment. *Proceedings of ITG Conference on Voice Communication (SprachKommunikation),* pp. 1-4.

Shafran, I., and M. Mohri (2005). A Comparison of Classifiers for Detecting Emotion from Speech. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* May 2005.

Singh, P., T. Lin, E. Mueller, G. Lim, T. Perkins, and W. Zhu (2002). Open Mind Common Sense: Knowledge Acquisition from the General Public. *Lecture Notes in Computer Science*, pp. 1223-1237.

Tao, J. (2004). Context based emotion detection from text input. *Proceedings ofInternational Conference on Spoken Language Processing,* Jeju Island, Korea, October 2004.

Tellegen, A., D. Watson, and L. A. Clark (1999). On the dimensional and hierarchical structure of affect. *Psychological Science* 10: 297-303.

Thayer, R. (1989). *The Biospychology of Mood and Arousal*. Cambridge: Oxford UP.

van Hengel, P., and T. Andringa (2007). Verbal aggression detection in complex social environments. *Fourth IEEE International Conference on Advanced Video and Signal Based Surveillance*, London, United Kingdom, September, 2007.

Vidrascu, L., and L. Devillers (2005). Annotation and detection of blended emotions in real human-human dialogs recorded in a call center. *Proceedings of the IEEE International Conference*

on Multimedia and Expo, Amsterdam, Netherlands, July 2006.

Wieczorkowska, A., P. Synak, and Z. W. Ras (2006). Multi-Label Classification of Emotions in Music. *Intelligent Information Processing and Web Mining* 35: 307-315.

Xia, Y., L. Wang, K. Wong, and M. Xu (2008). Sentiment vector space model for lyric-based song sentiment classification. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, Association for Computational Linguistics, Stroudsburg, PA, USA.

Yoshitomi, Y., S. I. Kim, T. Kawano, and T. Kitazoe (2000). Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face. *IEEE International Workshop on Robot and Human Interactive Communication*, Osaka, Japan, September 2000.

Towards Wittgenstein on the Semantic Web

Pichler, Alois
Alois.Pichler@uni.no
University of Bergen, Norway

Zöllner-Weber, Amélie
amelie.zoellnerweber@gmail.com
University of Bergen, Norway

1. Introduction

Ludwig Wittgenstein's 20,000 pages of manuscripts and typescripts (his 'Nachlass') display his continuous philosophical development and contain revisions, rearrangements and 'multiple versioning.' Their publication poses a number of challenges, for book as well as for digital editions (Huitfeldt 1994). Since its creation in 1990, the Wittgenstein Archives at the University of Bergen (WAB) has tried to meet these challenges through digital editorial philology. In 2000, WAB's 'Bergen Electronic Edition' (BEE) of Wittgenstein's Nachlass was published by Oxford University Press, and in 2009, 5000 pages from the Nachlass were made freely available (Open Access) on the Web (`http://www.wittgensteinsource.org/` , cf. Fig. 1). Since 2001, XML-based text encoding (TEI P5) has been one of several central ingredients through which WAB has worked continuously to improve access to Wittgenstein's manuscripts.

Figure 1: Screenshot of the Wittgenstein Source Bergen Text Editon

One key aspect of making Wittgenstein better and widely available is to prepare his works for the Semantic Web (Krüger 2007, Pichler 2010, Erbacher 2011). This, in turn, creates new possibilities for achieving deeper insights into his philosophy and intertextual relations within and between his work(s). A case study for developing an ontology

for Wittgenstein's *Tractatus* (1922) was published in 2007 (Zöllner-Weber & Pichler 2007). Work with the ontology received a significant boost from the EU-supported DISCOVERY project (2006-09, Smith 2007) and established that the ontology should permit interlinked browsing of philosophical concepts and primary sources. Therefore the Wittgenstein ontology includes both classes for sources and for philosophical concepts as well as properties which relate these classes. The ontology was implemented using OWL. A visualization tool (cf. Fig. 2) which allowed for coordinated use of both classes was created: Philospace/SwickyNotes which is today being further developed towards a collaborative Web based ontology browser and annotation tool (Morbidoni & Nucci 2010).

Figure 2: Extract from the Wittgenstein ontology in Philospace/SwickyNotes

From its start, the project was faced with a major challenge: Wittgenstein's work and philosophy are highly dynamic, multi-facetted and in continuous development. One example is Wittgenstein's concept of 'meaning' which in his *Tractatus* is primarily linked to reference, in the middle period to rules, in the *Philosophical Investigations* (1953) to usage and practice, and in his last years includes yet other aspects such as 'the atmosphere of a word.' Even within the same period and work a term can be used in different and not always consistent ways. Many of Wittgenstein's works are characterized by presenting a range of different views and approaches (cf. Fig. 3). Wittgenstein commentators often disagree and propose conflicting interpretations. How can such challenges be dealt with in ontologies?

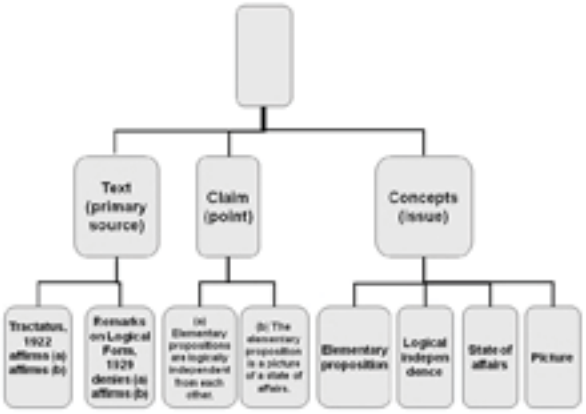


Figure 3: Elements of the Wittgenstein ontology

In our view, providing a frame for documenting and explicating different models of the world and its domains belongs to the assets of an ontology – irrespective of these models agreeing with one other or not. Disagreement and alternative proposals can be made visible, sharable and debatable exactly through such an ontology. We regard the implementation of an ontology in the field of philosophical interpretation as an opportunity rather than as a hindrance.

When conceiving the ontology, WAB analysed different dictionaries and indices of Wittgenstein's work, including Hanjo Glock's valuable *Wittgenstein Dictionary* (Glock 1996), and philosophical dictionaries and works of reference more generally. Obviously, WAB needed a bottom-up rather than a top-down approach since its ontology should support the use of specific texts, such as the 5000 pages of Wittgenstein Source. Another important aspect was that the ontology should match and support specific tasks that WAB and its partners, including guest researchers, were involved in, such as teaching a class on Wittgenstein's *Tractatus*, or joint research on Wittgenstein's philosophy of psychology. In this way, the ontology could be tested in research as well as learning environments and developed further in cooperation with Wittgenstein experts.

2. Description of the ontology

On the top level, the Wittgenstein ontology is divided into two branches: *Source* and *Subject* (see Fig. 4). The *Source* class houses primary and secondary sources relevant for Wittgenstein research; the subclass *PrimarySource* divides further into Wittgenstein sources and external sources (e.g. Augustine's *Confessiones*). The lowest subclass of Wittgenstein primary sources is *Bemerkung* and denotes the Wittgensteinian *remark*: typically not longer than half a page and separated from other remarks by one or more blank lines.

The *Subject* class contains four subclasses: *Issue, Perspective, Point* and *Field*. *Issue* refers to subjects dealt with by Wittgenstein himself or in Wittgenstein research, e.g. 'elementary proposition', 'logical independence', 'picture' and 'state of affairs'. They denote the subjects which one would typically expect to be included in the 'subject index' of a comprehensive study of Wittgenstein's philosophy. *Point* is a specific category developed at WAB: it refers to the point made or confirmed by an individual Wittgensteinian remark and typically contains an entire statement or a question like 'The elementary proposition is a picture of a state of affairs'. Via the *Source* properties *discussesIssue* and *hasPoint* instances of *Point* can be interlinked with instances of *Issue*.

Figure 4: Main classes of the Wittgenstein ontology

In order to provide for different modelling of the relations a Wittgensteinian subject can have, as well as to serve different user needs, WAB introduced the class *Perspective*. This class contains subclasses which each represent a specific grouping or filtering of the instances of *Issue* and *Point*. As an example the following may serve: Logical analysis (which is an instance of *Issue*) is central to the method of the philosophy from the perspective of the *Tractatus*, while it is not from the perspective of the *Investigations*. A course on the *Tractatus* will draw positive attention to that term, while a course on the *Investigations* may not even have it in its list of important subjects or start from the following instance of *Point*: 'Logical analysis is not central to philosophy'. One and the same instance of *Issue* can enter different configurations, and even configurations which contrast with each other. For example, the subjects 'elementary proposition' and 'logical independence' are in the *Tractatus* (1922) and "Some Remarks on Logical Form" (SRLF, 1929) combined to configurations contrasting with each other. The first work claims: 'Elementary propositions are logically independent from each other', while the later denies this. The *Tractatus* perspective will include the claim 'Elementary propositions are logically independent from each other', while the later work's perspective will include a negation (property *deniesPoint*) of this claim or the explicit claim: 'Elementary propositions are not logically independent from each other' (see Fig. 3).

In addition, the class *Field* was created as suggested by Grenon & Smith 2011. This field connects the Wittgenstein domain with other domains in philosophy and research more generally, e.g. epistemology or aesthetics. Users not familiar with Wittgenstein at all can enter Wittgenstein research with fields and debates from this class as their starting points. Each instance of *Issue* can be affiliated with one or more such field.

The name of an *Issue* instance consists of a German-English compound; this enables both German and English speakers to work more smoothly on the ontology, and in addition provides welcome means for precisely disambiguating and identifying the domain's instances. Users browse the ontology with a specific language set as preference: each instance can, via *rdfs:label*, in principle be identified and rendered in all languages. Currently, the ontology contains about 700 *Issue* instances which are rendered in German, English and Danish. Renderings in Norwegian, French and Finnish are being added. WAB's vision is that users of the ontology shall as much as possible be enabled to browse and access the Wittgenstein ontology in a language of their preference. The ontology is intended for use in the research and learning environments of different languages.

Properties are key in WAB's ontology: they provide the relations which allow for interlinked browsing of the ontology and Wittgenstein primary sources. Properties of *Bemerkung* which is the most important instance of *Source*, include: isPartOfSource, hasAuthor, hasDifferentVersion, refersToPerson, refersToText, discussesIssue, hasPoint, deniesPoint a.o. Properties of an *Issue* instance can be: assertsPerspective, isRelatedToIssue, isRelevantForField, a.o. Each *Bemerkung* of Wittgenstein's *Nachlass* shall eventually be associated by *discussesIssue* to at least one instance of *Issue*, but also take on *Point* properties such as *hasPoint*, directing to the class *Point*. Thereby, a scholar searching for discussions of certain issues will be steered to certain remarks, and vice versa looking at a specific remark one's attention will be drawn also to specific issues or points.

3. Summary

WAB's Wittgenstein ontology is still in its infancy, especially with regard to its *Subject* branch. WAB naturally cannot achieve its goal of a comprehensive Wittgenstein ontology alone: while the first version of the ontology and its overall structure were developed in the frame of the DISCOVERY project by WAB staff, it has since 2010 been strengthened by international cooperation (see `http://wab.uib.no/wab_philospace.page`). A key challenge in developing this

ontology was to reflect Wittgenstein's dynamic and multi-facetted work, but also Wittgenstein research's different interpretations of this work. This was dealt with by introducing among others the class *Perspective*, allowing for concurring models in and about Wittgenstein's work while maintaining a technically valid ontology. The instances of *Issue* are the only stable entities of the Wittgenstein subject domain; through the subclasses of *Perspective* they can be combined and filtered in different ways which can even contradict each other. The ontology is related to Wittgenstein primary sources via properties, enabling fast and easy studying of Wittgenstein within a common platform. In this way, research and learning about Wittgenstein can be significantly enhanced.

References

Erbacher, C. (2011). Unser Denken bleibt gefragt: Web 3.0 und Wittgensteins Nachlass. In S. Windholz and W. Feigl (eds.), *Wissenschaftstheorie, Sprachkritik und Wittgenstein*. Heusenstamm: ontos, pp. 135-146.

Glock, H. (1996). *A Wittgenstein dictionary*. Oxford: Blackwell.

Grenon, P., and B. (2011). Foundations of an ontology of philosophy.*Synthese* 182(2): 185-204.

Huitfeldt, C. (1994). Toward a Machine-Readable Version of Wittgenstein's *Nachlaß*. Beihefte zu *editio* 6: 37-43.

Krüger, H. W. (2007). Wittgenstein registrieren. In *Papers of the 30th International Ludwig Wittgenstein-Symposium*. Kirchberg a.W.: ALWS, pp. 119-121.

Morbidoni, C., and M. Nucci (2010). *SwickyNotes user guide*. `http://www.swickynotes.org/docs/SWickyNotesStartingGuide.pdf`.

Pichler, A. (2010). Towards the New Bergen Electronic Edition. In N. Venturinha (ed.), *Wittgenstein After His Nachlass*. Houndmills: Palgrave Macmillan pp.157-172.

Smith, D. C. P. (2007). Re-Discovering Wittgenstein. In: *Papers of the 30th International Ludwig Wittgenstein-Symposium*. Kirchberg a.W.: ALWS, pp. 208-210.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. London: Kegan Paul.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.

Zöllner-Weber, A., and A. Pichler (2007). Utilizing OWL for Wittgenstein's Tractatus. In: *Papers of the 30th International Ludwig Wittgenstein-Symposium*. Kirchberg a.W.: ALWS, pp.248-250.

Uncovering lost histories through GeoStoryteller: A digital GeoHumanities project

Rabina, Debbie L.
drabina@pratt.edu
Pratt Institute, USA

Cocciolo, Anthony
acocciol@pratt.edu
Pratt Institute, USA

1. Brief description

In this paper, we first describe the design and implementation of the GeoStoryteller project. GeoStoryteller is a project where learners engage with archival photos and multimedia narratives in historically relevant places using a combination of augmented reality technology and web-based delivery via mobile devices. The initial application of GeoStoryteller is in partnership with the Goethe-Institut New York, the worldwide cultural organization of Germany. In this partnership, learners – particularly German language students in the United States – engage with content that details the historical events and makes use of real places related to German immigration to New York City (1840-1945). Second, based on the design and implementation of this project, we offer a framework for creating location-based mobile learning projects that could be used by others interested in implementing similar projects.

2. Designing GeoStoryteller

GeoStoryteller builds upon recent research and development from the emerging subfield of the Digital Humanities known as GeoHumanities, which is a term that highlights the growing interconnections between geography and the humanities. GeoStoryteller strengthens this connection specifically by layering historic narratives on a specific location and by making it available to learners' Internet-enabled mobile device. In the case of the application of GeoStoryteller with the Goethe-Institut – named German Traces NYC – learners can view location-sensitive maps and lists of historic sites related to German immigration to New York. Once a learner arrives at a physical site, he or she can see historic photos layered against the imagery visible through the camera's phone (also known as augmented reality), as well as watch videos about that particular site that include historic narratives and archival photos. Users can also play short trivia games that answers can only be found from being on physical location and post their accomplishments to Facebook or Twitter.

For developing the videos – which we also refer to as GeoStories – we turned to published, non-published, digital and non-digital sources available from cultural and memory institutions (libraries, archives, museums and historical societies). Using these sources we developed historical narratives as text, and set those to audio recordings that were augmented with archival photos.

We then used geotechnologies to deliver the narratives to users' mobile devices (smart phones, such as iPhone, Android, Blackberry, and tablets such as iPad) at the places where these events occurred. Geotechnologies used include global positioning systems (GPS), digital mapping services (available through Google) and Layar, an augmented-reality browser for mobile devices.

For example, users can find such sites as the Ottendorfer Library, the oldest public library in Manhattan, opened originally to support the German immigrant community in Kleindeutschland – or Little Germany (today known as the East Village; see Figure 1). While on site, learners can learn about the people who crated the physical site, why they created the site (e.g., philosophies that motivated their action) and how that site has changed over time to reflect cultural and world events (e.g., anti-German sentiment during both world wars).

3. Framework for creating location-based mobile learning projects

From our experience developing GeoStoryteller, we offer a framework that can be used in the planning and implementation of similar projects.

At the core (the inner section) of this framework is traditional humanities research and development (see Figure 2). In the case of the GeoStoryteller project, this included print, non-print, digital, digitized, non-digital, archival, and published sources, all used to construct historical narratives of German immigration in New York. Placing humanities research and development at the core of the framework highlights what distinguishes a digital humanities project from other digital projects. Additionally, this type of research and development could be done with other fields of the humanities, including philosophy, literature, and the arts.

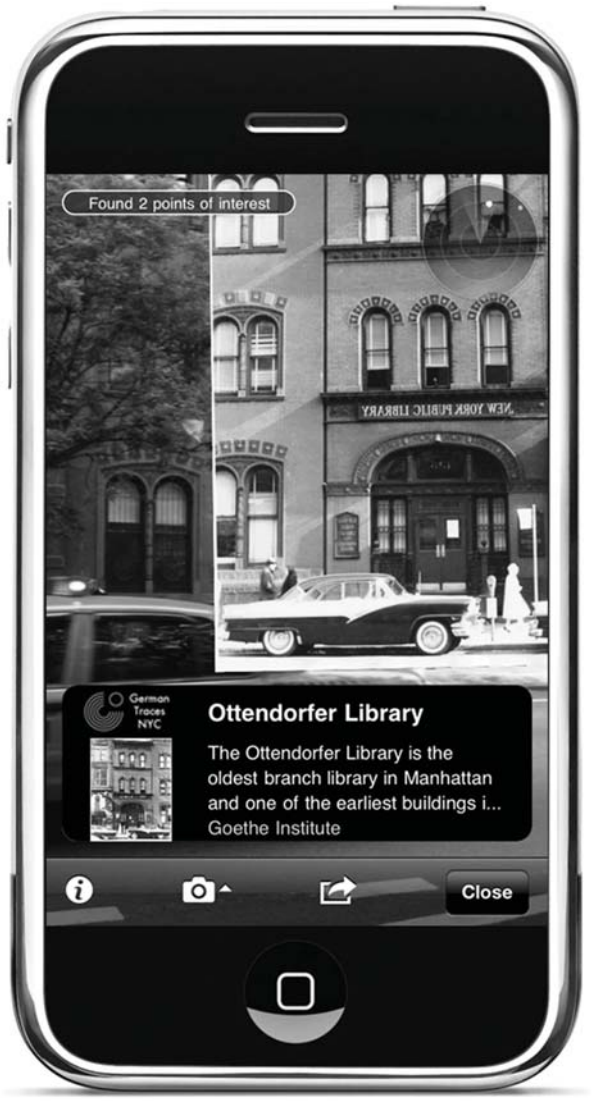


Figure 1. Augmented-reality interface for retrieving multimedia stories

Next, we proceed to the 'Theory and Interface Development' ring. At this stage, researchers must consider the theory that will inform their project, and how this theory will be reflected in the user interface available to the learner. For example, some of the theoretical questions we asked were, how does situating historical content in physically relevant locations affect learner engagement? And does making augmented-reality content available to learners further enhance engagement? Such questions necessarily lead to decisions about how to present the content within a digital interface.

In the third ring, learners engage with the socio-technical environment created by the researchers. In the case of GeoStoryteller, this includes not simply a user interface, but also the physically relevant locations that the interface prompts the user to explore. Additionally, social interactions may occur during this stage among multiple others in the environment (e.g. a librarian in the Ottendorfer library). It cannot be assumed that the best learning experience comes from the digital device, but could result from the serendipitous interaction in the real environment.

From this stage, we proceed to the fourth ring, which is the formal user research. In this stage, we use traditional (e.g., surveys and interviews) and digital (e.g., tags, hits, time stamps) social science research methods to uncover the working of the interface, address the theoretical questions, and evaluate the learning outcomes and user engagement. These studies can then influence any earlier stages (e.g., specific areas of content that did not engage the interest of learners could be revised, confirming or refuting the learning theory, or requiring changes to the user interface).

Figure 2: A Process for Digital Humanities Research and Development

4. Relevance to conference themes

The GeoStoryteller project addresses several of the Digital Humanities 2012 conference themes and aspects of digital humanities described in the Call for Papers. By focusing on vertical histories – or presenting narratives that describe events at a specific locale throughout a period of over 100 years – we introduce students to how immigrant communities, their language and cultures, contributed to the fiber of New York City as we know it today. We draw parallels between the prejudices against, hardships, discrimination and finally successful integration of the predominant immigrant community of the 19th and early 20th century immigrant and those of today. GeoStoryteller is a completely bilingual project that serves both to

uncover lost histories of New York and demonstrate their relevance to contemporary geohistory.

Workflows as Structured Surfaces

Radzikowska, Milena
mradzikowska@gmail.com
Mount Royal University, Canada

Ruecker, Stan
sruecker@id.iit.edu
IIT Institute of Design, USA

Rockwell, Geoffrey
grockwel@ualberta.ca
University of Alberta, Canada

Brown, Susan
susanirenebrown@gmail.com
University of Guelph, Canada

Frizzera, Luciano
lucaju@me.com
University of Alberta, Canada

INKE Research Group
sruecker@id.iit.edu
University of Alberta, Canada

At DH 2011 in Stanford, we introduced the concept of structured surfaces, where we extended the metaphor of interactive digital pins on a digital map into interactive digital pins on any data visualization (Radzikowska et al. 2011). We proposed that such structured surfaces would provide a means for people working with digital collections to gain insight into their material while also providing them a means of formulating an argument about the data.

More recently, in considering scholarly editions, we became interested in how the concept of the structured surface could be applied to the editorial process – in effect, producing an interface derived from the flowchart of activities that an editor can use to manage the movement of a submitted article or other item of text (such as a book review) through the stages from its initial appearance in the editor's inbox to the final step where it is ready to send to the printer.

Accordingly, we have designed and are in the process of producing experimental prototypes of a series of these workflow interfaces for different kinds of editorial jobs, including the publication of journal articles, the writing of XML-encoded original material based on Orlando biocritical entries on women writers, and the editing of a scholarly edition (e.g. Figure 1). In these designs, the structured surface consists of the flowchart used by the editor,

and the pins represent the articles, with size of pin indicating length of article in words.

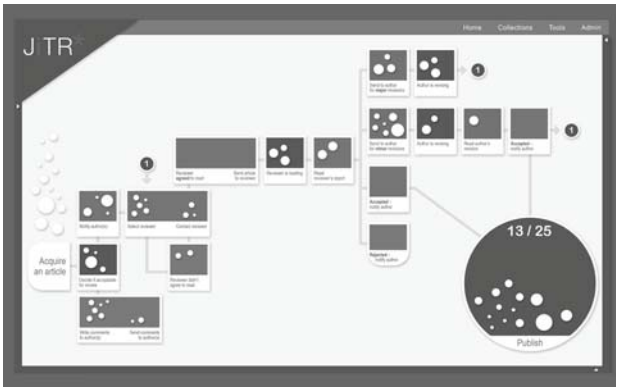


Figure 1: This workflow diagram shows the movement of a journal article through the publication process from the perspective of the journal editor who is managing that process

Numerous other systems exist for managing workflow in the context of digital text production, such as that devised by the Public Knowledge Project for its Open Journal Systems and its Open Monograph Systems. The interfaces here seek to retain the modularity and customizability associated with such systems but to combine them with a rich prospect visualization (Ruecker, Radzikowska & Sinclair 2011) of the collection being managed than has hitherto been implemented. Recent work in computing science has focused both on workflow analysis (e.g. Schroeder 2009) and on algorithms for automatic layout (e.g. Albrecht et al. 2010).

However, a workflow management tool, as useful it might be to the editor who has not had the advantage of using one yet, is in itself not a particularly interesting object. It is essentially a structured checklist, composed of a collection of stages with transitions, including occasional decision points between them. But once such a concept is available for interrogation, it can begin to introduce additional considerations.

For example, why is the structure of the stages limited to a simple bucket? Would it not be useful to be able to distinguish different kinds of material moving through the system? It might possible, for instance, to automatically identify file types so that digital text is treated one way by the tool, while digital film is treated another. Or the editor may wish to produce a variant of the workflow for book reviews as opposed to peer-reviewed articles. So we have begun to experiment with ways of visually structuring the boxes that represent stages (Figure 2).

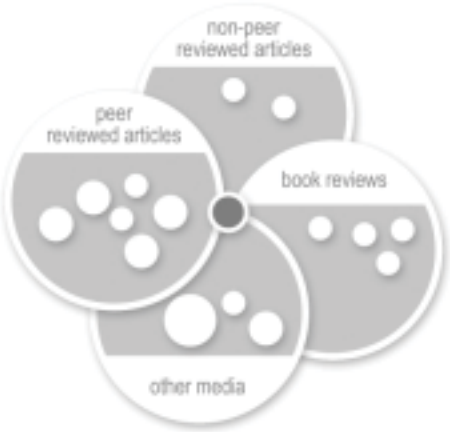


Figure 2: An initial attempt to turn the workflow stage itself into a kind of structured surface

Once we have the concept of the system automatically doing something to the documents, it is a short step to recalling that the original goal of the structured surface project was to provide the user with a set of analytics in the form of a data visualization and an argument made by the superimposition of pins. In that context, it seems reasonable to consider what kinds of analytical processes might be applicable to various stages within the workflow, and also what kind of information might be useful to the editor at each stage. For example, when an article is with a reviewer, the editor is likely to be primarily interested in the following information:

- the title of the article
- the name and contact information of the reviewer
- the dates when the article went to the reviewer and when it is expected back

A rollover of the pin within the 'at the reviewer' stage should therefore by default produce this list of information, with access to further information relegated to the 'more...' link. At the next stage, however, when the article is back from the reviewer(s), what the editor would like to know is the results of the review. A rollover of the pin while in that bucket should therefore produce the title of the article and the name of the reviewer(s), but also the contents of the review.

On the analytics side, there may be processes (like reviewing) where the article is passed to another person, but there may also be processes where the article is passed to some algorithmic treatment, such as an automated step of metadata enhancement where the system makes the first pass at named

entity recognition for people and places. Since one of the affordances of the JiTR environment is that any collection or subset of a collection can be passed to a tool, perhaps the contents of the stages within the workflow could serve as subsets for this kind of treatment. An editor might then, for instance, decide to take a look at a KWIC concordance for a term in every article in the 'accepted' bucket for the current edition, or run a word cloud on all the articles that have been accepted or rejected. The wordcloud of acceptances might provide some guidance in writing an introductory editorial, for instance, while the latter might help the editor see whether there is a gap in reviewer expertise at work. Support for these sorts of curiosity-driven activities within the editorial process itself has the potential to change the nature of the task in a variety of ways.

Finally, the concept of an interface that combines workflow management with analytics is of considerable interested to the Canadian Writing Research Collaboratory (CWRC) project, which seeks to lower the bar to participation in online resource creation and enhancement by mainstream scholars without significant expertise in digital humanities, and in so doing lower the overhead of producing scholarly materials on the web. In order to ensure quality control within a system such as CWRC's, a workflow process needs to be highly intuitive as well as modular, robust, and well integrated with other interface components. For these reasons, CWRC is experimenting with the structured surfaces interface and the JiTR environment as a major component of its interface, seeing the potential of it for managing everything from query results to worflow management to membership coordination. The CWRC project has presented the interface to Collaboratory members at two workshops this year and is extending the prototype to handle Orlando-like workflows and interact with an external workflow engine.

This presentation will outline the transfer of the structured surface concept to workflows, show the sketches and demonstrate the prototypes, and report on user feedback to date. It will provide an analysis of the implications of this approach to supporting the editorial process, suggesting that there are a number of potential new affordances that make the workflow a useful addition to the editor's toolbox. An interface which integrates a tightly defined workflow alongside the ability to invoke other tools, we argue, can also encourage new kinds of thinking about possible improvements or additions to the workflow.

References

Albrecht, B., P. Effinger, M. Held, and M. Kaufmann (2010). An automatic layout algorithm for BPEL processes. *Proceedings of the 5th international symposium on Software visualization* (SOFTVIS '10). ACM, New York, NY, USA, 173-182. DOI=10.1145/1879211.1879237 http://doi.acm.org.login.ezproxy.library.ualberta.ca/10.1145/1879211.1879237

Canadian Writing Research Collaboratory. http://www.cwrc.ca

Open Journal Systems Documentation. http://pkp.sfu.ca/files/OJSinanHour.pdf

Open Monograph Project. http://pkp.sfu.ca/omp

Orlando Project. http://orlando.cambridge.org

Radzikowska, M., S. Ruecker, S. Brown, P. Organisciak, and the INKE Research Group (2011). Structured Surfaces for JiTR. Paper presented at the *Digital Humanities 2011 conference. Stanford, June 19-21, 2011.*

Ruecker, S., M. Radzikowska, and S. Sinclair (2011). *Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-Prospect Browsing*. Farnham, Surrey: Ashgate Publishing.

Schroeder, W. (2009). New tools for task workflow analysis. *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems* (CHI EA '09). ACM, New York, NY, USA, 3877-3882. DOI=10.1145/1520340.1520587 http://doi.acm.org.login.ezproxy.library.ualberta.ca/10.1145/1520340.1520587.

Code-Generation Techniques for XML Collections Interoperability

Ramsay, Stephen
sramsay.unl@gmail.com
University of Nebraska-Lincoln, USA

Pytlik-Zillig, Brian
bpytlikz@unlnotes.unl.edu
University of Nebraska-Lincoln, USA

XSLT has often been criticized for its verbosity.[1] Control structures, in particular, can require many more lines of code (and many more indications of start and end state) than nearly any modern programming language. Yet for all this, XSLT has features that make it an ideal language both for code generation (automatic generation of XSLT using XSLT itself) and as a target output format for other languages. In this paper, we discuss our attempts to tackle the problem of collections interoperability by exploiting both features of XSLT.

The rules governing the validity of particular XML instances are usually set forth in one of several standard forms of specification (such as RELAX NG,[2] XML Schema,[3] or the older Document Type Definition[4]). Such schemas typically specify the nature, number, and sequence of permissible elements, ensure their correspondence to particular data types, and enforce the overall referential integrity of the instance. We have developed a system (called Abbot) that uses schema definitions located in a target schema to generate a stylesheet that will effect the transformation of one or more document collections into instances that validate against that same schema.[5]

In the trivial case, where the target schema describes a proper subset of the collections in question, Abbot operates more-or-less automatically, but more complex transformations are also possible. One can, for example, give the system two collections and have it generate a stylesheet that makes one collection conform to the schema of the other. One can also make several collections target an entirely different schema. In these latter cases, it becomes necessary to describe particular mappings in a configuration file, but that configuration uses a simple syntax unrelated either to that of a schema language or XSLT.

The key step here is the automatic generation of an XSLT stylesheet. Our choice of XSLT as the language that generates that stylesheet might at first seem slightly perverse, but because XSLT is a homoiconic language – a language in which the primary representation of the language is itself a data structure in that same language – code generation can be undertaken through the use of metapgramming (in which code is passed into another, more abstract layer and evaluated).[6]

The Abbot system begins by running a 'meta-stylesheet' (analogous to a higher-order function in a traditional functional language) on both a target schema and a configuration file. The configuration file, while not written in XSLT, is nonetheless converted into XSLT by the surrounding runtime (using a translation method we discuss below). By default, that target schema describes TEI Analytics – a TEI subset that provides an encoding scheme optimized for text analysis. This meta-stylesheet generates, as its only output, a conversion stylesheet used for the actual transformation of the documents. This latter transformation yields files that will, in the majority of cases, validate against the target schema.

When Abbot reads the target schema, it accounts for all elements and associated attributes and generates a default XSLT template for each element. These default templates reflect the general assumption that elements and attributes in the input files resemble their counterparts in the target schema. If, for example, <foo n="001"/> exists in the input file and is specifically allowed in the target schema, then Abbot will pass the element through unaltered under the assumption that the element is fully valid. Anything beyond that needs to be articulated in the configuration file.

Ultimately, the custom transformations set forth in the configuration file need to be instantiated as XSLT templates and included in the conversion stylesheet at runtime. For example, to replace the <temphead> element with <teiHeader> (its TEI P5 counterpart), the system would need to generate the following:

```
<transformation type="xslt" activate="yes">
  <desc>substitute 'temphead' with 'teiHeader'</desc>
    <xsl:template match="*[lower-case(name())='temphead']" priority="1">
   <teiHeader>
    <xsl:apply-templates/>
   </teiHeader>
 </xsl:template>
</transformation>
```

To replace spaces with underscores in the extent attribute of the <gap> element (a considerably more

complex operation), requires substantially more code:

```
<transformation type="xslt" activate="yes">
  <desc>add underscore to 'gap' @extents containing a space</desc>
                <xsl:template match="*[lower-case(name())='gap']" priority="1">
    <xsl:choose>
      <xsl:when test="contains(@extent,' ')">
        <xsl:element name="gap">
          <xsl:for-each select="@extent">
            <xsl:choose>
              <xsl:when test="contains(.,' ')">
                <xsl:attribute name="extent">
                  <xsl:value-of select="replace(.,' ','_')"/>
                </xsl:attribute>
              </xsl:when>
              <xsl:otherwise>
                <xsl:copy-of select="."/>
              </xsl:otherwise>
            </xsl:choose>
          </xsl:for-each>
          <xsl:apply-templates/>
        </xsl:element>
      </xsl:when>
      <xsl:otherwise>
        <xsl:copy-of select="."/>
      </xsl:otherwise>
    </xsl:choose>
  </xsl:template>
</transformation>
```

Depending on the particular situation, the configuration file might have to contain dozens of hand-built templates for performing subtle transformations that cannot be deduced from the schema. But here, we undertake a second code-generation step using Clojure – a dialect of Lisp that runs on the Java Virtual Machine.

Because Lisp is also a homoiconic language, it too is well suited to code that reads and writes code. Moreover, XML is itself a first-class datastructure in Clojure, which can be easily (and lazily) transformed into a map object in which descendant nodes are represented as nested vectors. The problem of parsing a configuration file (in which complicated XSLT transformations are rendered in the form of a radically simplified DSL), becomes a matter of parsing the file into a map structure. Clojure can then trivially transform that map directly into XML (XSLT), which can be inserted at runtime into the conversion stylesheet. The first XSLT example above becomes something like:

```
temphead -> teiHeader
```

The second, more complicated example might be expressed as:

```
gap[@extent='/ /'] -> gap[@extent='/_/']
```

In this way, Abbot becomes not merely a framework for effecting interoperability of XML document collections, but a general purpose XML transformation framework that avoids the need for XSLT itself.[7]

Thinking of XSLT as an intermediate form – a language that is targeted much as a compiler might target assembly – allows us to imagine radically simplified document transformation languages that can (potentially) exploit the full range of XSLT itself. In the case of Abbot, radical simplification is possible, in part, because the problem domain is itself highly constrained. But such constraints constitute precisely the rationale for domain-specific languages that try to map a user's domain knowledge to a simplified syntax. Such languages, while smaller and simpler than more general-purpose languages, often still require the full range of language design tools (lexers, parser generators, the specification of a grammar, and so forth). Exploiting the homoiconicity of languages that possess this feature – including XSLT itself – makes the process of designing a 'mini-language' considerably easier.

References

Adler, S. (1997). A Proposal for XSL. *World Wide Web Consortium* (W3C) `http://www.w3.org/TR/NOTE-XSL.html` (accessed 31 October 2011).

McIlroy, D. (1960). Macro Instruction Extensions of Compiler Languages. *Communications of the ACM* 3(4): 214-220.

Pytlik-Zillig, B. (2009). TEI Analytics: Converting Documents into a TEI Format for Cross-Collection Text analysis. *Literary and Linguistic Computing* 24(2): 187-192.

Pytlik-Zillig, B. (2011). TEI Texts that Play Nicely: Lessons from the MONK Project. *Journal of the*

Chicago Colloquium on Digital Humanities and Computer Science 1(3): 1-5.

Unsworth, J. (2011). Computational Work with Very Large Text Collections: Interoperability, Sustainability, and the TEI. *Journal of the Text Encoding Initiative* 1 http://jtei.revues.org/215 . (accessed 21 October 2011).

Notes

1. A Web search for 'XML verbosity' will bear this out amply, though we note that this 'loquaciousness' is itself consistent with the design goals of XML, which was intended to be a human readable format. In the first formal XSL proposal, the authors explicitly state that 'Terseness in XSL markup is of minimal importance' (see Adler 1997).
2. http://relaxng.org/
3. http://www.w3.org/XML/Schema
4. The Document Type Definition is currently defined by the XML specification itself, though it descends from earlier specifications associated with Standard Generalized Markup Language (SGML). The Extensible Markup Language (XML) specification is at http://www.w3.org/TR/REC-xml
5. We have called this method 'schema harvesting' (see Pytlik-Zillig 2011).
6. Homoiconicity is a property of several programming languages, including REBOL, SNOBOL, PostScript, XQuery, Prolog, and all dialects of Lisp. The concept of homoiconicity is first set forth in Douglas McIlroy's 1960 article, 'Macro Instruction Extensions of Compiler Languages' (see McIlroy 1960).
7. Abbot cannot, of course, replace XSLT in all circumstances, since the DSL we are describing is not intended to capture the entire semantics of XSLT. Still, we imagine that Abbot can be usefully employed in many situations in which large bodies of texts are being transformed, even if interoperability is not a primary concern.

Uncertain Date, Uncertain Place: Interpreting the History of Jewish Communities in the Byzantine Empire using GIS

Rees, Gethin Powell
gpr26@cam.ac.uk
University of Cambridge, UK

1. Introduction

The presentation will address the problem of how to display uncertain historical data in a web-based Geographical Information System (GIS). In recent years the use of internet GIS has allowed the general public to access large volumes of spatial data. Although GIS has been applied in specific areas of academic research, the technology has largely remained the preserve of specialists, as it can be difficult to use, requiring training and experience (Boonstra 2009). Yet the application of GIS to humanities research in general, and to the interpretation of historical data in particular, has the potential to develop and answer innovative research questions (Bodenhamer, Corrigan & Harris 2010).

The 'Mapping the Jewish Communities of the Byzantine Empire' project will use Internet GIS technology to disseminate data that outline the history of these communities. The project is based at the Centre for Advanced Religious and Theological Studies, Faculty of Divinity, University of Cambridge and is funded by the European Research Council. Textual, epigraphic and archaeological data allow the project to establish the locations, dates and other attributes of Jewish communities. Prior to our work these data were scattered, fragmentary and difficult to access. The project incorporates varied data in an Internet GIS which provides an efficient method of dissemination through dynamic, interactive maps. The GIS will feature an easy-to-use interface making data accessible through a standard web browser. The project team hopes that this accessibility will promote the inclusion of the Byzantine Jewish communities in the study of Byzantine, Jewish and minority history. A further aim is that medieval historians with little experience of GIS will become familiar with the technology and its potential through use of the website.

The spatial perspective that the project offers will allow research into Jewish communities to develop in new directions. For example, the project's GIS could be used to aid interpretation of the involvement of Jews in trade, the effect of political change on their distribution and lives, and the factors that influenced the development of separate Jewish residential quarters. The focus of this presentation will be the role of Jews in the Byzantine economy. A robust understanding of geography is important for interpreting trade and therefore the perspective that GIS offers makes it a valuable aid to understanding historical economies. Characteristics of Jewish communities such as their location and date are crucial to evaluating their interaction with the wider economy and are fundamental to the structure of the project's database. Yet these data are beset by uncertainty, causing significant problems with their representation and interpretation in GIS.

2. Uncertain Data in GIS

The presentation of historical data using GIS is not straightforward as the technology has developed in the context of empirical research. Symbols used in any form of mapping can convey an unwarranted air of reliability by masking differences between features represented in a single category. The original sources on which maps are based are rarely revisited by scholars in the same way that references to texts are. Maps have a veneer of objectivity making them easily manipulated to suit particular interpretations (Monmonier 1996). Similar problems affect GIS, and the technology has, since the late-twentieth century, been criticised for being over-reliant on a positivist interpretative framework. Historical or other cultural data are often uncertain and therefore the application of GIS in these areas requires a different approach.

Uncertainty can be divided into three categories. First are factors that are inherent in defining the object of study; second are problems with sources of information and their interpretation; and third are issues with the representation and reduction of these data to feature in a database (Plewe 2002). Uncertainty can arise in each of these areas where characteristics are, for example, contested, ambiguous or unreliable. These problems are particularly acute in interpreting historical data from medieval and earlier periods (Bartley & Campbell 1997). To be a useful resource for historical research, our website must incorporate and communicate the complexity of these data in a format that is straightforward enough to be understood quickly and by non-specialists. This presentation will examine uncertainty in the location and date of Jewish communities as these attributes are fundamental to the structure of the project's database.

3. Representing Uncertainty Through Transparency

The project has developed a system of symbols that communicates uncertainty in dating and locating Jewish communities. Degree of uncertainty is represented by variation in the transparency of symbols. To give an example, the date of a person's birth can be known with varying degrees of certainty. Where one source may provide the date to the exact day, another might only give the year. Dates may also be reached by comparison or interpretation. If the symbol that represents a piece of data in the project GIS is more transparent, the user is made aware that there is less certainty regarding its date. Uncertain dates will be displayed as periods with reference to a time series. The system of symbols will communicate the certainty with which the existence of a Jewish community at a particular date is known. This can allows users of the website to examine the Jewish presence within the wider framework of Byzantine economic history with an appropriate level of confidence.

The system will also use transparency to communicate uncertainty regarding where Jewish communities were located. The location of a Jewish community can be questioned due to, for example, ambiguous or imprecise references in sources as well as differences in their interpretation. The degree to which a symbol is transparent will be determined by the number of possible locations. The siting of Jewish communities was influenced by and was an influence on their role in the economy. Preliminary analysis supports the view that some Jewish communities resided in particular locations for economic reasons. Internet GIS can help historians gain a better understanding of these relationships providing that uncertainties in the data are communicated.

To explicate the representation of uncertainty, the website must communicate the sources and types of uncertainty to the user. Information on uncertainty and the transparency of symbols will be made accessible through clicking on symbols. The sources used to produce the GIS will be communicated through metadata. The symbols will also allow access to other attribute data provided in the sources including religious sect, occupation and name. These attributes provide further structure to the database and form the basis of searches.

4. Summary and Implications

The historical study of medieval economies would benefit from greater integration with geography, particularly to develop a detailed understanding of the spatial character of trade. Our project's GIS can help to develop this understanding by disseminating

historical data whilst being sensitive to the needs of medieval historians. It is hoped that the system used for communicating uncertainty in our project's Internet GIS will be taken up by future historical projects that deal with uncertain data.

References

Bartley, K., and B. M. S. Campbell (1997). Inquisitiones Post Mortem, GIS, and the Creation of a Land-use Map of Medieval England. Transactions in GIS 2: 333-46.

Bodenhamer, D., J. Corrigan, and T. Harris, eds. (2010). The Spatial Humanities: GIS and the Future of Humanities Scholarship. Indiana University Press, Indiana.

Boonstra, O. (2009). Barriers Between Historical GIS and Historical Scholarship. In D. Bodenhamer and P. Ell (eds.), International Journal of Humanities and Arts Computing 3. Edinburg: Edinburgh UP, pp. 3-7.

Monmonier, M. S. (1991). How to Lie with Maps. London: U of Chicago P.

Plewe, G. (2002). The Nature of Uncertainty in Historical Geographic Information. Transactions in GIS 6: 431-56.

Code sprints and Infrastructure

Reside, Doug
dougreside@gmail.com
New York Public Library for the Performing Arts, USA

Fraistat, Neil
nfraistat@gmail.com
Maryland Institute for Technology in the Humanities (MITH), University of Maryland, USA

Vershbow, Ben
benjaminvershbow@nypl.org
New York Public Library, USA

van Zundert, Joris Job
joris.van.zundert@gmail.com
Huygens Institute for the History of The Netherlands, The Netherlands

Over the last five years, several national and supranational funding bodies have invested in large digital humanities infrastructure projects designed, in part, to reduce the proliferation of projects inventing and reinventing small, commonly needed tools. During the same period, several very modestly funded projects have attempted to achieve the same results by holding working meetings – called 'code sprints' – to develop commonly needed tools collaboratively. Interedition, The Center for History and New Media's One Week | One Tool, MITH's XML Barn-Raising, and New York Public Library's Tilden Papers Project have each hosted a series of code camps (also known as boot camps or code sprints) that bring together groups of humanities scholars with serious programming expertise to spend a week working on a tool of value to each of them. The process has proven to be so productive that it has recently been adopted by one of the large infrastructure projects, Project Bamboo, in 'CorporaCamp.' In this panel, we will examine the advantages and limitations of the small, agile methods of code sprints and how they may be supported by the large, sustainable infrastructures currently under construction by larger projects.

The advantages of codesprints are numerous. By spending time in rapid prototyping; by gathering scholars who are also coders (rather than those with only one skill or the other); and by establishing a policy of, to quote Dave Lester, 'More Hack, less yak!', worksprints quickly determine the real challenges facing academic software development and often make significant headway towards solving them.

'One week, one tool' produced Anthologize which continues under active development and use to this day; Interediton has produced, among other things, CollateX (a modular automated collation workflow); the MITH barn raising produced a prototype of a web-based XML editor: ANGLES; and NYPL's codesprint significantly refactored the code in the Internet Archive's BookReader tool and prepared it for future extensions by the participants. These sprints not only build tools, they lay the groundwork and point the way for the sort of infrastructure that is truly needed.

Nonetheless, these sprints occasionally encounter problems. Setting up code-sharing mechanisms, installing needed dependencies, and teaching coding dialects such as jQuery or Node.js to participants unfamiliar with them often absorbs a full day of work. Differing expectations and desires by participants can sometimes (though surprisingly rarely) threaten to derail progress. Documenting work so that it can be taken up later or by others is sometimes not prioritized to the extent it should be. Further, although there is often much to do at the end of the sprints, participants frequently find it difficult to continue working when they return home as competing and more immediate priorities take precedence. It is possible that large infrastructure projects such as Bamboo, Dariah, and TextGrid could provide the organizational and administrative infrastructure needed to make codesprints more effective and their work more sustainable.

This paper will discuss four code sprints (outlined below), recount 'lessons learned', and discuss how big infrastructure projects and light-weight, rapid development efforts such as these may support each other.

Sample case studies:

MITH 'Barn Raising': In 2010, **Doug Reside** (Digital curator for the Performing Arts, NYPL) organized a 'barn raising' to produce a web-based XML editor in his last weeks at the Maryland Institute for Technology in the Humanities. The event brought several participants from Canada and elsewhere in the United States to College Park, Maryland, but also organized a group of coders to participate remotely via Skype and IRC. After the first two days of the sprint, the group divided into two groups, one concerned with building a WYSIWIG editor and one wanting to replicate the core functionality of popular XML editors in a JavaScript-based web application. The second group consisted almost entirely of remote participants, but arguably built more code in the course of the sprint than those working together in the same room.

BookReader Sprint: In 2011, **Benjamin Vershbow** organized New York Public Library's worksprint to extend the Internet Archive's JavaScript widget, BookReader. The goal of the project was to refactor the Internet Archive's existing code to make it more modular and extensible. Around a dozen participants were brought to New York from libraries around North America for four days of hacking on the code base. In an attempt to be open to the many desired use-cases of the BookReader, NYPL invited participants with a diverse set of skills and desires which, in retrospect, probably dispersed some of the productive energy of the sprint into efforts that could not be realistically achieved in the course of a few days. However, the sprint is notably in that it began to extend an existing, heavily used, code base supported by a major organization.

Interedition: Over the past three years, **Joris van Zundert** organized ten boot camps as part of European Cost Action IS0704 'Interedition'. The boot camps varied in participation from 5 to 15 scholars, developers, and scholarly developers from the wider European region and the US. The boot camps focused transcription, annotation, and collation as primary scholarly tasks in producing (digital) scholarly editions that could effectively be supported by common models and tools. The Interedition boot camps have resulted in various new tools in the form of web services – of which CollateX probably is best known – and considerable progress of development of existing tools (Juxta, eLaborate etc.). However, the production of tools is paradoxically 'just' a side effect of the Interedition endeavor. Interedition's main objective is furthering the interoperability of tools used in the production of scholarly editions as a means of enhancing the sustainability of both tools and digital editions. One of Interedition's findings was that it is pivotal to such sustainability that there must be an academic platform supporting researcher-developers' interaction and collaboration in a most concrete way: interoperability and integration of tool development is best *done together*.

Corpora Camp: Neil Fraistat and **Seth Denbo** organized Corpora Camp as part of project Bamboo. CorporaCamp was a key step in the design process for Project Bamboo's Corpora Space, which will enable the curation and exploration of data across the boundaries of large structured collections. The primary goal of Corpora Camp was to see if over the space of three days participants could make a prototype tool for visualization and analysis function across three different collections. While the 'work' of the workshop involved building this tool – which we called WoodChipper – the tool itself was only one of several important outcomes. CorporaCamp not only tested our assumptions about the larger design process for Bamboo Corpora Space, but the rapid development process of the workshop required us constantly to balance our long-term

goals – experimenting with a distributed, extensible architecture – against our desire to have a working prototype implemented at the end of the three days. In many cases the team had two development threads running in parallel, with one group working on a more general solution and another on a simpler fallback. This process provide a better sense of the problems and decisions – and the range of consequences of those decisions – that would be faced with in developing Bamboo architecture and applications.

Digital Genetic Criticism of RENT

Reside, Doug
dougreside@gmail.com
New York Public Library, USA

While textual critics of medieval and classic texts require expertise in paleography, those studying contemporary works, and in particular contemporary multimodal works such as musical theater, require a similar sort of facility with digital forensics. In this paper I will detail the digital forensic techniques (and a few dumb hacks) that I have used to create a detailed biography of the musical, *RENT,* from it's earliest its earliest incarnation in 1989 to the final version edited by Larson and saved just two weeks before his death in 1996. I will also suggest a set of skills those attempting similar work may wish to acquire.

Jonathan Larson, the Broadway composer and lyricist, used his personal computer as a creative tool throughout his brief career from the mid-1980s to his premature death at 36 in January of 1996. After his death, around 180 of the floppy disks he used to save his work were donated to the Library of Congress. In 2007, I was permitted to create complete, bit-for-bit, data images of these disks and study any file related to the creation of Larson's most famous musical, *RENT*. These disks contain several dozen drafts of *RENT*.

In some cases, multiple drafts are recoverable from a single file. In 1992, when most personal computers ran only about 5% as fast as an iPhone 4, saving an entire file to a floppy disk was a potentially time consuming activity. While writing RENT, Larson mostly used Microsoft Word 5.1, which had a feature called 'fast save.' The fast save feature sped the saving process somewhat by replacing the entire file only once every 14 saves or so. In most cases, it would simply append changes to the end of a file with information about where they belonged in the original document. When the file was opened in Word 5.1, the software would integrate these changes back into main text; however, by opening this file with a simple text editor, it is possible to see the text of the last full save along with all the emendations made during fast saves in an apparently uninterrupted text stream stored at the bottom of the file.

Unfortunately, the Word 5.1 standard has no openly published documentation. Thus, while additional text stored by a 'fast save' are visible, the way in

which this text maps onto the 'base text' is not immediately obvious. Fortunately, with the help of Microsoft Research, I have obtained documentation that describes the standard (unfortunately provided under a non-disclosure agreement), which describes how this mapping was accomplished. Therefore, although this paper will not include a complete description of the specification for legal reasons, it will demonstrate how several versions of a Larson's text might be reconstructed from a single Word file.

For example, throughout 1992 there are several files and folders on Larson's disks that contain versions specially formatted for the laser printer. In 1992 Larson generally used a 'dot-matrix' printer that worked by pushing ink from a small typewriter-like ribbon onto paper in a series of small dots that composed the desired letters or the image. The printer would advance the paper through a spool in increments as tall as the printer ribbon as each line of dots was printed. Although this method of printing was relatively inexpensive and was capable of representing image that could be composed out of the dots (rather than, say, a fixed set of letters and numbers), it produced copies that, even in the 1990s, appeared amateurish. There were a finite (and relatively small) number of dots available on any line which made text appear jagged, and ink would transfer from the ribbon inconsistently, creating a faded, uneven look. Larson was content with this technology for rough drafts, but when he needed a copy to give to a potential producer, he would print the text on a laser printer (perhaps at a copy service store).

There was a complication, however. Text formatting conveys a good deal of meaning in musical theater libretti. Stage directions are often represented alongside spoken text as typed marginalia; overlapping lines sung in unison or counterpoint are represented in different columns. At this period, the relatively low resolution of the screen Changes in typeface or margins can change the line lengths and create wrap-around text where a hard break was intended creating an amateurish appearance that the use of the laser-printer was supposed to prevent. Larson's usual font 'New York,' like many of those designed for early versions of the Macintosh operating system, was a 'bitmap' typeface, meaning that each character was stored as a small image. Just as a small 'thumbnail' image downloaded from a website becomes fuzzy or 'pixelated' when it's size is increased much beyond the original size, the size of a bitmap font cannot be attractively increased beyond the original size of the character images. Moreover, the quality of the characters themselves – how smooth or jagged they appear – will not change even if the monitor or the printer quality improves. 'Vector' fonts, like the 'Courier' font used in Larson's 'laser' drafts, on the other hand, are generated not from an image file but from an equation that describes the lines and curves of the characters. The result is that the characters can be smooth as the monitor or printer can make them and can be scaled up to any size without loss of quality. Moreover, 'Courier' was designed as a 'monospace' font – that is, each character occupies the same amount of space (an 'M' is the same width as an 'I'). 'Monospacing' makes the process of aligning text much easier than it would be with fonts of variable width as one can be confident how many characters will fit on a line (regardless of what the characters might be). Thus, for those drafts which needed to look especially good on the page as well as on the screen, Larson would reformat his text to meet the requirements of the printer.

Thus, there are folders on Larson's disks with titles like 'Rent Laser' that contain versions of the script formatted for higher quality printing. By consequence revising the script during the rehearsal process, when all copies should be printed on a laser printer, was difficult. In some cases, probably during the rehearsal or preview period when it was necessary to distribute frequent changes to a cast, Larson found it preferable to print only revised passages on a laser printer and then to literally cut and paste these revisions onto an older laser-printed version and photocopy the result. These pasted-together drafts, preserved in the Library's paper collection, exemplify the awkward hybrid of centuries old print traditions with the advantages and quirks of new media and serve as a strikingly appropriate metaphor the history of the text to which they bear witness.

Of course, *RENT,* is hardly the only text for which both digital forensic and bibliographic expertise will be needed to conduct a contemporary *critique génétique*. Indeed, editors and textual scholars of most work composed after (at the latest) 1995, are likely to encounter digital drafts among their witnesses. This paper is intended to serve as a kind of early exemplar for the kind of techniques a contemporary digital textual scholar must learn and practice.

On the Internet, nobody knows you're a historian: exploring resistance to crowdsourced resources among historians

Ridge, Mia
m.ridge@open.ac.uk
Open University, UK

Crowdsourcing, the act of taking work once performed within an organisation and outsourcing it to the general public in an open call (Howe 2006), is increasingly popular in memory institutions as a tool for digitising or computing vast amounts of data, as projects such as Galaxy Zoo and Old Weather (Romeo & Blaser 2011), Transcribe Bentham (Terras 2010) and the Australian Newspapers Digitisation Program (Holley 2010) have shown. However, the very openness that allows large numbers of experts and amateurs to participate in the process of building crowdsourced resources also raises issues of authority, reliability and trust in those resources. Can we rely on content created by pseudonymous peers or members of the public? And why do academics often feel that they can't? This paper explores some of the causes and forms of resistance to creating and using crowdsourced resources among historians.

'Participant digitisation' is a specialised form of crowdsourcing in which the digital records and knowledge generated when researchers access primary materials are captured at the point of creation and potentially made available for future re-use. Through interviews with academic and family/local historians, this paper examines the following: the commonalities and differences in how these two groups assess the provenance, reliability and probable accuracy of digital resources; how crowdsourcing tools might support their working practices with historical materials; the motivations of historians for sharing their transcriptions and images in a public repository; the barriers that would prevent them from participating in a project that required them to share their personally-digitised archives; and the circumstances under which they would selectively restrict content sharing. From this preliminary investigation, the paper will go on to consider implications for the creation of digital humanities resources for academic and amateur users.

References

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine* 16(3/4).

Howe, J. (2006). The Rise of Crowdsourcing. *Wired* 14.06. [Online] Available from: `http://www.wired.com/wired/archive/14.06/crowds.html`

Romeo, F., and L. Blaser (2011). Bringing Citizen Scientists and Historians Together. In J. Trant and D. Bearman (eds), *Museums and the Web 2011: Proceedings*. Toronto: Archives & Museum Informatics.

Terras, M. (2010). Digital curiosities: resource creation via amateur digitization. *Literary and Linguistic Computing* 25(4), 425-438.

Formal Semantic Modeling for Human and Machine-based Decoding of Medieval Manuscripts

Ritsema van Eck, Marianne Petra
nanne.rve@gmail.com
Rijksuniversiteit Groningen, The Netherlands

Schomaker, Lambert
schomaker@ai.rug.nl
Rijksuniversiteit Groningen, The Netherlands

In recent years, archival institutions storing historic handwritten document collections have been somewhat reluctant to invest in digitization (scanning). The reasons for this are threefold. Firstly, early, rigorous digitization campaigns of the past (1985-2000) were costly and sometimes disappointing. Secondly, the quality of Optical Character Recognition (OCR) techniques is very low for *handwritten* manuscripts. Thirdly, the digitization of historical documents and the permanent storage of the resulting digital images is costly. Such investments are generally only considered worthwhile if it is possible to access and browse the digitized material quickly and easily, in a manner which satisfies a large user base.

The *Monk* system was developed in answer to the afore-mentioned problem: the aim of this project is build an interactive search engine for browsing handwritten historical documents, using machine-learning and pattern classification. Ultimately, it should become possible to 'google' for a search-term in large collections of digitized handwritten document images. Traditional OCR fails in handwriting, because patterns are largely connected, noisy and difficult to read, even for humans. In addition, historical script types vary tremendously according spatio-temporal origins. Furthermore, traditional Automatic Handwriting Recognition (AHM) techniques require many examples of a single word or letter class from a single writer to be considered feasible in this case, and training occurs off line, in a lab or company. Therefore, an innovative new bootstrapping method based on user-based word labeling was developed to train a computer system, *Monk,* to read historical scripts (Zant, van der et al. 2009). This system has been used successfully in the first massive processing of early twentieth century handwriting in the Dutch 'Cabinet of the Queen.' The success of this endeavor stimulated the research in the direction of older, more difficult material.

2. A Difficult Case: the Leuven Alderman's Rolls

The Leuven city archive is currently in the process of digitizing its collection of the Leuven Alderman's rolls, early fifteenth century. This text is the urban record of voluntary jurisdiction and disputes settled by the Aldermen; individual cases are described in short legal acts. As a pilot study, the book of the year 1421 (MS SAL 7316, Leuven City Archive, Belgium) was ingested by the *Monk* system, with the aim of achieving handwriting recognition on the current Gothic minuscule these hundreds of pages are written in. Indeed, this is no easy feat, considering that contemporary humans need special training to be able to read this script. However, even if one – as a human – can decipher series of individual characters (letters, abbreviations, glyphs) in the text, this does not directly lead to understanding. Firstly, there is a linguistic barrier: the rolls are written mostly in Latin and occasionally in Middle Dutch. Secondly, in order to understand what the text means one needs much background knowledge about the type of text and the administration of justice in medieval Leuven. As the *Monk* system still makes many mistakes in its attempts to read the script of the Leuven alderman's rolls, it seems that additional modeling is needed to improve results of computer-based reading.

3. Contextual Modeling: Language and Semantics

The idea of top-down linguistic support for bottom-up word hypotheses generated by a script classifier such as *Monk* – is not entirely new. The use of a statistical language model to improve offline handwriting recognition has been proposed (Zimmerman & Bunke 2004). However, the use of a general language model or syntax will not do in case of the Leuven Alderman's rolls. Firstly, the text employs a register of highly artificial legal language. Secondly, no formal language models are existent for the Middle Dutch and Medieval Latin it employs and no encoded text corpus exists for the automatic generation of a 'stochastic grammar' (Manning & Schütze 1999).

Therefore, a semantic model was developed to support recognition by *Monk* in the rolls. While syntactic modeling may fail due to input variation and variability, semantics are invariably present in any utterance of language. A semantic framework can support both the human- and machine-based decoding. In the case of the Leuven Alderman's rolls, meaning, i.e. higher knowledge levels, supports the recognition of words and letter forms at the lexical and orthographical level. Human readers, as opposed to digital document processors, sample text

opportunistically and take into account contextual information eclectically (Schomaker 2007). To distinguish the /v/ from the /b/ in current gothic minuscule the human reader may rely on contextual information where an uninformed computer program cannot. More specifically, the interpretation of the ink patterns on a page is determined by an expectancy on what to find in the text. In case of an 'IOU', for instance, the reader expects to find, somewhere, the names of two persons, a sum of money, and possibly a date of payment. The more difficult the deciphering of patterns at the small scale, the more important is the role of such additional high-level information and expectation at a larger scale of observation: the semantics expressed by a text block of about a paragraph or administrative entry ('item') in the case of an administrative archive.

4. Semantic Support

During the present project a framework for disclosing the semantics contained in the digital image of the manuscript page was attempted. First, the accomplishments of historical philological method and manuscript studies were put into practice to create a 'world model' as introduced in the field of artificial intelligence (Schank 1975), specifying objects, their attributes and their mutual (abstract) relationships. Such a model is the necessary 'historical backdrop' for understanding the text. It includes vital information elements, such as the specific type of legal text and its place in the process of administering voluntary justice (Synghel G. v. 2007; Smulders 1967).

Secondly, a typology of types of acts was formulated, with the help of scholarship on the comparable alderman's rolls of Den Bosch (van Synghel 1993; Spierings 1984). This typology of acts functions as a fundamental part of the interpretative apparatus. Exhaustive knowledge about which key phrases encode for a certain type of act, immensely help in identifying the real-life legal transaction an act refers to. However, the gap between high level meaning and the 2-dimensional format of the document still needs to be bridged.

5. Towards a Formal Semantic Modeling of the Act

In the research field of document structure and lay-out analysis a distinction is commonly made between 'geometrical layout' and 'logical content'; referring to specific visual regions on the page, and their content and/or functional label, and reading order. The logical structure can then be mapped to the relevant geometrical regions on the page, to achieve optimal analysis of the document image (Haralick 1994; Namboodiri & Jain 2007).

During the second phase of the project 'The Legal transactions' a formal semantic model was constructed based on the basic distinction between geometrical and logical representation, viz. digital image and semantic content respectively. The geometrical branch is subdivided in polygonal *regions of interest* (ROIs), which again break down into smaller nodes such a 'ink' and 'background'. The logical branch describes the structure of semantic content in legal acts. By interlinking the relevant nodes the two branches using unique identifiers (idROIs), it becomes possible to cross-refer between manuscript page and content.

Several XML schemes for layout and content modeling exist; however, to date none of these standards is designed for handwritten text (Stehno, Egger, & Retti 2003; Bulacu, van Koert & Schomaker 2007). These formalisms are suitable for regular, printed text and rectangular elements. Due to the erratic nature of the page layout and line formation in medieval handwritten manuscripts a system of polygonal visual regions connected to the semantic model by unique identifier codes needs to be used for *Monk*, comparable to layout modeling tools such as TrueViz and GEDI.

None of these formalisms provide an easy solution for linkage between layout and the semantic level: such a connective layer needs to be developed for each particular document type. If it is elaborated in Web Ontology Language (OWL) or Resource Description Framework (RDF), the 'aldermen' semantic model can be used to train *Monk* to recognize the structural elements of acts and formulate hypotheses about their meaning. Conversely, the same model may assist in interactive teaching of humans as to where to locate the desired information in an act. The most exciting finding of this study was that superficial low-level ink patterns and glyphs can be used to infer clues about the semantic content, which in turn constrains the large number of possible text interpretations.

As an example, a selection of the most basic logical components in the legal acts was made:

- the start of a new act is signaled by '*Item*' (X_0,Y_0)
- the *semantic anchor* reveals the type of legal transaction (X_1,Y_1)
- names of presiding aldermen and (X_2, Y_2)
- the date give the act legal force (X_3,Y_3)

- the 'solvit' glyph signals that a charter was issued and seal-money paid (X_4,Y_4)

From a number of pages, 22 acts were manually analyzed in terms of the geometric distribution of element in the horizontal plane. Based on the relative geographical location of these semantic elements and their probability of occurrence in the plane, Monk may weigh his bottom-up hypotheses using a Bayesian model (see fig. 1). The small sample includes only the most basic semantic elements on the page; many more may be added to improve recognition accuracy.

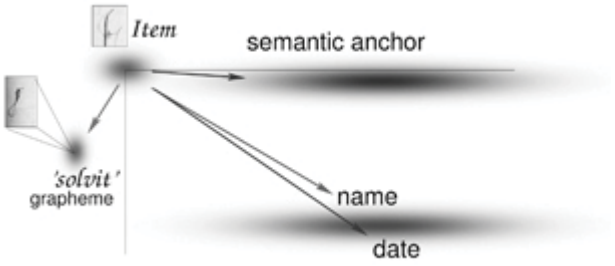


Figure 1

References

Balacu, M., R. van Koert, and L. Schomaker (2007). *Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen: Proceedings of the 9th International Conference on Document Analysis and Recognition*. Curitiba, Brazil, September 2007.

Haralick, R. M. (1994). *Document Image Understanding: Geometric and Logical Layout: Proceedings of the Computer Vision and Pattern Recognition Conference*. Seattle, June 1994.

Manning, C. D., and H. Schütze (1999). *Foundations of Statistical Language Processing*. Cambridge, Mass.: MIT Press.

Namboodiri, A., and A. Jain (2007). Document Structure and Layout Analysis. B. B. Chaudhuri (ed), *Digital Document Processing*. London: Springer, pp. 29-48.

Schank, R. C. (1975). *Conceptual Information Processing*. Amsterdam: North-Holland Publishing Company.

Schomaker, L. (2007). Reading Systems: An Introduction to Digital Document Processing. In B. B. Chaudhuri (ed), *Digital Document Processing*. London: Springer, pp 1-26.

Smulders, F. W. (1967). Over het Schepenprotocol. *Brabants Heem* 19: 159-65.

Spierings, M. (1984). *Het Schepenprotocol van's-Hertogenbosch, 1368-1400*. Tilburg: Stichting Zuidelijk Historisch Contact.

Stehno, B., A. Egger, and G. Retti (2003) METAe – Automated Encoding of Digitized Texts. *Literary and Linguistic Computing*, 18: 77-88.

Synghel, G. van (1993). *Het Bosc" Protocol: een Praktische Handleiding*. 's-Hertogenbosch: Werken met Brabantse bronnen 2.

Synghel, G. van (2007). *"Actum in camera scriptorum oppidi de Buscoducis": De stedelijke secretarie van's-Hertogenbosch to ca. 1450*. Hilversum: Verloren.

Zant, T. van der, et al. (2009). Where are the Search Engines for Handwritten Documents. *Interdisciplinary Science Reviews* 34: 228-39.

Zimmerman, M., and H. Bunke (2004). *Optimizing the Integration of a Statistical Language Model in HMM based Offline Handwriting Text Recognition: Proceedings of the 17th International Conference on Pattern Recognition*. Cambridge, UK, August 2004.

The Swallow Flies Swiftly Through: An Analysis of Humanist

Rockwell, Geoffrey
geoffrey.rockwell@ualberta.ca
Department of Philosophy and Humanities Computing, University of Alberta, Canada

Sinclair, Stéfan
stefan.sinclair@mcgill.ca
Department of Languages, Literatures & Cultures, McGill University, Canada

We're 14 years old now, a venerable age in this medium, like everything else somewhere between coming into being and going out of it, 'like the swift flight of a single sparrow through the banqueting-hall where you are sitting at dinner on a winter's day with your thegns and counsellors. In the midst there is a comforting fire to warm the hall; outside, the storms of winter rain or snow are raging. This sparrow flies swiftly in through one door of the hall and out through another. While he is inside, he is safe from the winter storms; but after a few moments of comfort, he vanishes from sight into the wintry world from which he came.'[1]

HUMANIST has been one of the few stable discussion spaces where people in the digital humanities can discuss the field. The metaphor of the swallow flying though the banqueting-hall captures what we would like our commons to be, warm and comfortable compared to the winter outside, and a feast of methods, ideas, and conversation. This paper will fly though the hall of HUMANIST on the wings of analytical methods. Specifically this paper will take a retrospective look at the evolution of the digital humanities through an analysis of HUMANIST using Voyant, a suite of web text tools we have developed for handling large corpora like this. In the paper we will therefore,

- Ask why we might look at our history now?
- Discuss the data analyzed and the methods used
- Discuss the shift from humanities computing to the digital humanities
- Look at the effect of the web on the field

1. Why look back now?

In the space of a very few years the digital humanities seems to have gone from a marginal field trying to gain respect to a favorite of many university administrators. Digital humanists now need to define (and justify) what the digital humanities is to people who *ask us* – instead of trying to explain it to anyone willing to listen. It is difficult to pin down exactly when this transition happened, but one important moment was when William Pannapacker wrote in the Chronicle that 'Amid all the doom and gloom of the 2009 MLA Convention, one field seems to be alive and well: the digital humanities.' He continued with 'the digital humanities seem like the first "next big thing" in a long time'.[2] As John Unsworth noted in his 'State of the Digital Humanties, 2010' address to the Digital Humanities Summer Institute (DHSI), when a field is perceived to have jobs (while other humanities fields are seeing a dramatic decline) 'it's bound to attract some notice, especially among bright, goal-oriented graduate students who are approaching the job market.'[3] As universities try to get into the game by posting not jobs but clusters of jobs, and graduate students try to adapt to those jobs, issues of definition and self-definition became important.

Alternatively this self-conscious turn could be due to funding agencies noticing and creating programs for the digital humanities like the *Image, Text, Sound and Technology* program of SSHRC that first made awards in 2003-4 and the Office of Digital Humanities programs of the US National Endowment for the Humanities. We shouldn't underestimate the degree to which funding programs can get attention.

The self-conscious turn can also be seen in papers looking at the field and asking about it. Some examples include a paper by Wang and Inaba presented in Taipei and then published under the title, 'Analyzing Structures and Evolution of Digital Humanities Based on Correspondence Analysis and Co-word Analysis.'[4] Another is by Patrik Svensson titled 'The Landscape of Digital Humanities.'[5] Both essays provide a 'birds-eye' view of the field, though in very different ways. Wang and Inaba use text analysis and text mining techniques to ask about the shift from 'humanities computing' to 'digital humanities'; Svensson takes a cultural studies approach looking at the actors and discourse.

2. Data and Methods

Like Wang and Inaba our method is to use DH methods like correspondence analysis, but we analyzed a different, messier and fuller text, the archives of the HUMANIST listserv. HUMANIST is a discussion list that was started by Willard McCarty when he was at the Centre for Computing in the Humanities at the University of Toronto in 1987. The first message introducing the list started with, 'HUMANIST is a Bitnet/NetNorth electronic mail

network for people who support computing in the humanities.'[6] The list is still running, and, except for an interlude of a few years, is still moderated by Willard. HUMANIST has limitations that will be discussed in the full paper, but it provides a rich text for understanding the field in the English-speaking world.

Voyant, formerly called Voyeur, is a suite of text analysis tools that can be used individually or combined into 'skins' on texts that you upload or link to.[7] Voyant is a second generation of web analytical tools that benefited from lessons learned by HyperPo and TAPoR. Our first generation of tools could only handle smaller texts (up to about 500K or the equivalent of a book or two) and therefore weren't suitable for the study of large corpora. These tools were also developed as close (re)reading tools in a context of literary study of a single text or small corpus. In this project we adapted Voyant to handle the sort of corpus that allows us to do diachronic analysis and distant reading; we added mining tools that facilitate the formation and validation of hypotheses.

Part of our methodological practice is to develop a series of hypotheses about the last 25 years of Humanist. These are then tested, adapting and adding features to Voyant as needed. The tool development is thus deeply integrated with the analytical process. We will use the hypotheses and unexpected themes to discuss our flight through the corpus.

3. From Humanities Computing to Digital Humanities

We started with the expectation that there had been a shift from 'humanities computing' to 'digital humanities' as discussed by Wang and Inaba. While we certainly found 'digital humanities' taking off in 2004-5, we were surprised that 'humanities computing' continues to be a popular phrase. What is harder to do is to pinpoint why the term started to be used. One of the limitations of a public list is that it is suitable for epidemiological studies that tell you what happened and what correlated with what happened, but it is harder to tell what motivated people to change how they named the field.

Figure 1: Digital humanities and humanities computing

4. The Passing of Centers

A second hypothesis we started with was that the 1980s and 1990s had been a time of centers that made computing available in labs, but that the field had moved on to a more distributed project model. We did not find evidence for this. While the word project becomes more popular and it is true that some centres has closed, we found others had started up. Perhaps our particular experience with certain centers closing biased us into thinking this was a larger phenomenon.

5. The Turn of the Web

The most dramatic evidence of change was so obvious that we didn't think about it or hypothesize it until we developed a tool that could show clusters of years and keywords using correspondence analysis.

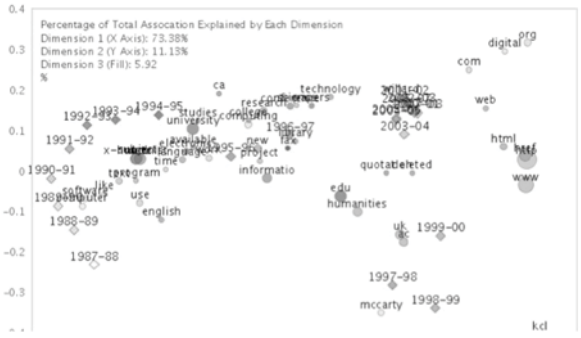


Figure 2: Scatter plot from correspondence analysis

The scatterplot display of the major dimensions generated showed a strong pull of patterns like 'web', 'www', and 'html.' There seem to be three phases in the data studied:

- 1987-95: a phase when humanities computing is interested in computers, software, hardware, texts, and English.
- 1996-2000: a transitional period which may be an artifact of the corpus.

- 2001-present: a shift to digital web services and collaborative projects.

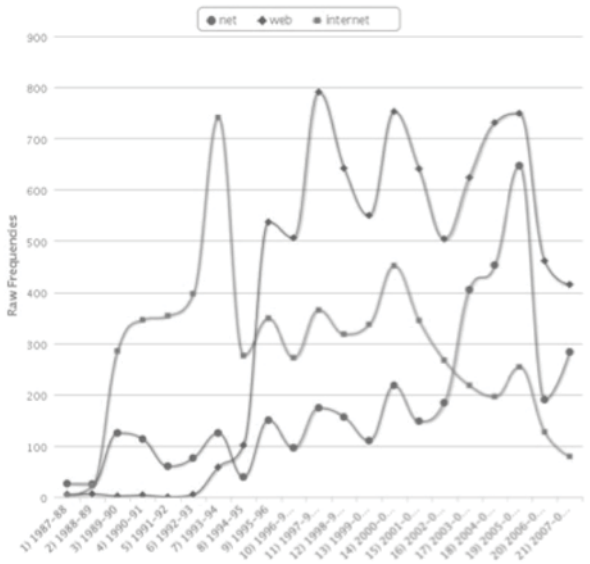


Figure 3: Net, web and internet

We believe the use of the web by humanists in the mid 1990s was a transformative for the field and may explain why less and less of our discussion was about hardware and software.

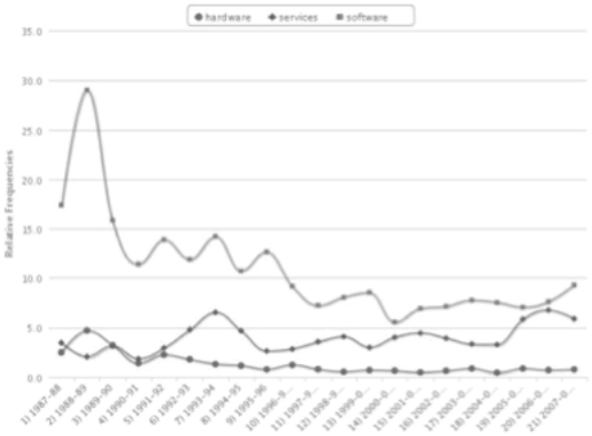


Figure 4: Software, services and hardware

As computers became ubiquitous in the 1970s and 1980s, humanities computing was concerned with supporting the new hardware and software. HUMANIST itself was initially conceived as a discussion list for those who supported others. Many humanists in 1986 were just beginning to use computers for word-processing and most didn't have email. With the web as a canvas for digital projects we started to pay less attention to 'processing' and more to 'methods.' We began to make our own tools after the hard work of developing scholarly electronic texts. Above all we returned to 'content' going beyond text to look at other 'media' and the 'social.'

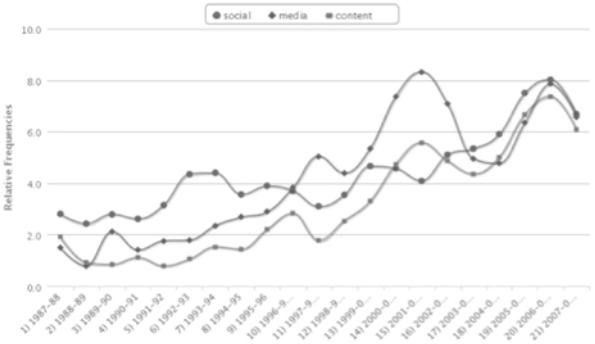


Figure 5: Social, media and content

6. Conclusion

The new-found traction of the digital humanities in academia is an opportunity to reflect on how the discipline has evolved and what it has become. Our contribution here is twofold: 1) a reading of HUMANIST that can help understand who we are and 2) a presentation of text analysis in practice. Of course, this is just our fly-through the corpus; with accessible web archives and tools like Voyant, the hall is open.

Notes

1. Humanist Discussion Group 15(1), see www.digitalhumanities.org/humanist/Archives/Virginia/v15/0000.html (`www.digitalhumanities.org/humanist/Archives/Virginia/v15/0000.html`).
2. Pannapacker, W. (2009). The MLA and the Digital Humanities. Brainstorm: The Chronicle Review's blog, Dec. 28, 2009 chronicle.com/blogPost/The-MLAthe-Digital/19468/ (`chronicle.com/blogPost/The-MLAthe-Digital/19468`).
3. Unsworth, J. (2010). The State of Digital Humanities, 2010. Address to the Digital Humanities Summer Institute, University of Victoria, Victoria, June 2010. The PDF is available at www3.isrl.illinois.edu/~unsworth/state.of.dh.DHSI.pdf (`www3.isrl.illinois.edu/~unsworth/state.of.dh.DHSI.pdf`).
4. Wang, X., and M. Inaba (2009). Analyzing Structures and Evolution of Digital Humanities Based on Correspondence Analysis and Co-word Analysis. Art Research, No. 9, Mar. 2009, pp. 123-134.
5. Svensson, P. (2010). The Landscape of Digital Humanities. Digital Humanities Quarterly 4(1). This is actually the first of a multi-part essay.
6. This is the opening line of the 'Welcome to HUMANIST' message, 14 May 1987. See www.digitalhumanities.org/humanist/Archives/Virginia/v01/8705.1324.txt (`www.digitalhumanities.org/humanist/Archives/Virginia/v01/8705.1324.txt`)
7. See voyant-tools.org (`voyant-tools.org`) or voyeurtools.org (`voyeurtools.org`)

The Digital Mellini Project: Exploring New Tools & Methods for Art-historical Research & Publication

Rodríguez, Nuria
nro@uma.es
University of Málaga, Spain

Baca, Murtha
mbaca@getty.edu
Getty Research Institute, USA

Albrezzi, Francesca
falbrezzi@getty.edu
Getty Research Institute, USA

Longaker, Rachel
rlongaker@getty.edu
Getty Research Institute, USA

1. Background

This paper will present the project *Digital Mellini: Exploring New Tools & Methods for Art-historical Research & Publication*, a joint initiative of the Getty Research Institute (Los Angeles) and the University of Málaga (Spain). This paper shall discuss several aspects and challenges related to the development of digital resources for art-historical research.

The main objectives for the *Digital Mellini* Project are:

1. To explore new methods and tools with which to reinvent the concept of scholarly work and critical publishing in the field of humanities, particularly in the context of art history, in which the convergence of text and image is essential and provides an interesting context for research – a context that has not yet been adequately investigated.[1]
2. To create a methodological model for developing collaborative digital publications that incorporate texts, digital facsimiles, images, computational tools for linguistic analysis and visual communication, and forums for exchanging ideas and sharing knowledge. The ultimate goal is that the international community of specialists can utilize this methodological model and apply to a variety of art-historical projects.

These objectives can be seen within the framework of the recent growing interest that many of us in the art-historical community have in response to the challenges offered by the digital society – we need to analyze and reflect critically on the 'digital status' of the discipline of art history from both the methodological and epistemological perspectives. We must be willing to open innovative and creative paths in the design of digital resources for art-historical research – paths that will lead to new lines of investigation and new methods of and attitudes toward information-sharing.

Art historians must participate in the development of tools in the Digital Humanities. In an article published in the *Chronicle of Higher Education* in 2009,

Johanna Drucker pointed out that many of the critical debates that have characterized the development of the humanities in recent decades – subjectivity of interpretation, multicultural perspectives, a recognition of the social nature of the production of knowledge, etc. – are absent in the design of digital resources. Instead, more emphasis has been placed on adapting to the characteristics of digital media than to the epistemological and interpretive exigencies of the various humanistic disciplines.[2] Lev Manovich made similar observations in a Digital Humanities conference held in June 2009 in Maryland, when he asked, 'Is it sufficient to borrow techniques from the fields of computer science, information visualization and media art – or do we need to develop new techniques specific to the humanities?' (Manovich 2009: xv).

Regarding the hope that Digital Humanities will assume this new role, we perceive the need to build environments that are intellectually productive and will facilitate interpretive and critical studies above and beyond simple information systems designed for the storage and retrieval of data.

The *Digital Mellini Project* is a response to these challenges, as the aim of the project is to contribute to the development of virtual research environments (VRE) to promote critical research in the field of art history.

The *Digital Mellini Project* includes another desideratum that has emerged in recent years: the production of knowledge based on aggregation and collaboration. The culture of open, shared knowledge, which was already being promulgated by UNESCO as early as 2005 as a fundamental way to transition from the information society to the knowledge society, along with the possibilities offered by Web 2.0 technologies, is an essential factor within the evolution of digital scholarship.

There are already several projects oriented toward the development of collaborative environments (NINES[3], The VRE-SDM Project[4] , The Transcribe

Bentham Project[5], etc.); and there is a burgeoning bibliography on this topic (Deegan & McCarty 2011). Proof of this interest has also been shown by the high percentage of presentations that dealt with this problem in the last Digital Humanities conference, held at Stanford University in June 2011.

Electronic publications and digital editions are not new in the realm of Digital Humanities; the first reflections on what an electronic edition might be appeared in Finneran (1996) and Bornstein and Tinkle (1998). There are also reflections and proposals relating to collaborative editions of unpublished manuscripts in which a dispersed community of scholars could add annotations, as in the case of the Peirce Project (Neuman et al. 1992) and the Codex Leningradensis (Leningrad Codex Markup Project 2000).

Nevertheless, as Susan Hockey (2004) states in her summary of the history of Digital Humanities, 'The technical aspects of this are fairly clear. Perhaps less clear is the management of the project, who controls or vets the annotations, and how it might all be maintained for the future'.

Similarly, the *Digital Mellini Project* intends to continue this line of research, addressing the questions that still remain open and launching new ones, cultivating debate and reflection.

2. Digital Mellini Project: development and intellectual questions

The Digital Mellini Project focuses on a particular text, an unpublished manuscript written in 1681 by Pietro Mellini, a Roman nobleman. This text belongs to the Getty Research Library, Special Collections, found under the title *Relatione migliori di Casa delle pittura Melini*. Pietro Mellini poetically described the best works as he inventoried the collection of his family's paintings. This text carries a self-interest for art historical research due to its uniqueness – it is the only document known to date that has a hybrid composition of ekphrastic description in poetic form with a factual descriptive inventory. In addition, it is an important source for studies related to collecting, the market valuation of art, the provenance of collections, etc. Most of the works are unidentified and to analyze their possible allocations through the descriptions provided by Pietro Mellini is a very attractive task for art historians. Lastly, this text gives us an opportunity to explore the possibilities that digital media offers to deepen the relationships between text and image. One of the objectives of the critical edition is to establish hypotheses about the identity of the works described within the manuscript.

To carry out this critical collaborative edition, the Digital Mellini Project is developing a computational prototype. Exploring the possibilities of Drupal, a digital environment is being designed and conceived specifically to this analytic objective. Currently, the project team has completed an alpha version of the digital workspace. The first prototype of a beta version is expected to be finished by the end of 2012.

This prototype was constructed around the object of the critical edition, Mellini's text, and it includes:

- text itself - A diplomatic transcription encoded in XML-TEI, translations into English and Spanish, and a digital facsimile – provided in an interface that allows comparative analysis [Fig. 1];

Figure 1

- A representative image gallery, consisting of images of works proposed by experts as possible identifications of the paintings described by Mellini. Additionally, provisions within the environment make it possible to establish comparisons between images and with the linguistic descriptions, so scholars can deepen the analysis of the word-image relationship;
- Tools to complement or extend into other areas of research. For example, a repository of thematic texts and/or those structurally related to the Mellini's text.
- Tools for aggregating critical annotations to the text. These annotations can be written by each participating scholar, according to a methodology of action and interaction developed for this purpose. Also, these annotations can be arranged as a response or reply to other entries, thus generating a critical debate among the community of scholars participating [Fig. 2].
- Space for debate and reflection on issues related to the text where it can be more broadly discussed.

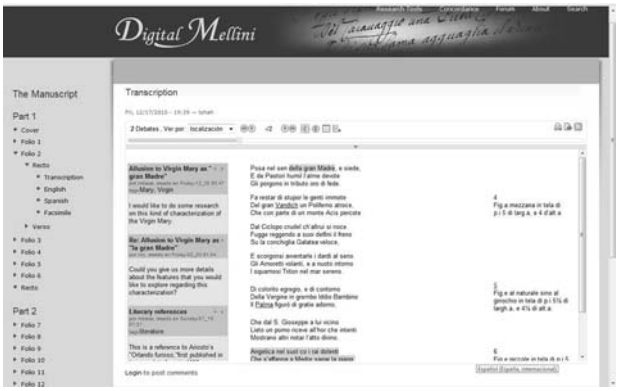


Figure 2

Note, therefore, that what is proposed as a digital critical edition is far from the conventional academic model of a print publication, which would exist in a static format as the end result of an interpretive study. The result may have been made by one or several specialists, but in the end, what we find is the proposal of one particular interpretation. It is also far from the digital editions that fit this model (eg Van Gogh Letters Project) or electronic repositories focused mainly on the analysis and exploration of texts (The Orlando Project-History of British Women's Writing). The proposed model is multifaceted. Digital Mellini simultaneously integrates multiple and sometimes contradictory and paradoxical views. It is open and dynamic. Due to processing system records and the dialogue between specialists in the critical annotation process, we can reconstruct the interpretive process itself, as aggregation and contributions grow over time. Therefore, the digital edition is not only the end product of a study and/or research, but also the intellectual and interpretive process through which it unfolds.

In the digital edition, in addition to showing this prototype and its performance, we will address in more detail the intellectual and methodological questions that confront us while developing a resource of this kind, taking into consideration the goal of advancing digital scholarship in the field of art history and the Humanities in general. These topics include:

1. How to show different viewpoints contributed by specialists, and displaying through visual means the diversity of interpretations and perspectives that converge simultaneously on a text from different contexts – geographical, ideological, theoretical, methodology, etc.?
2. How to deal with the problem of linguistic variability in a multilingual community of specialists without circumscribing to the most prevalent language?
3. How to face the problem of authorship? How to maintain the legitimate intellectual property rights of each participant in a publication that emerges from the aggregation of many?
4. What are the real behaviors and research practices of art historians using such environments? Do these practices lead to the emergence of a new art-historical knowledge of a digital nature or not?
5. How to develop specific methodologies and protocols for action to guide, assist, and coordinate the participation and interaction of specialists in this type of environment?
6. How to encourage the participation of specialists in the context of an academic system that does not 'recognize' the work developed in these digital environments? For example, what kind of scientific production – monographs, articles, papers – are contributions that fit within critical annotations to a digital edition when referenced in an academic CV? Should we think, then, of the incorporation of new categories of scientific production in line with the specificities of digital media?

Finally, the conclusive question: how can the discipline of art history move away from single authorial models toward more open, collaborative models of research and publication?

References

Bornstein, G., and T. Tinkle (1998). *The Iconic Page in Manuscript, Print, and Digital Culture*. Ann Arbor: U of Michigan P.

Deegan, M., and W. McCarty (2011). *Collaborative Research in the Digital Humanities*. Surrey: Ashgate.

Drucker, J. (2009). Blind Spots. Humanist must plan their digital future. *The Chronicle Review* 55, is. 30. `http://chronicle.com/free/v55/i30/30b00601.htm` (acessed 15 October 2011).

Finneran, R. J. (1996). *The Literary Text in the Digital Age*. Ann Arbor: U of Michigan P.

Hockey, S. (2004). The History of Humanities Computing. In S. Schreibman, R. Siemens, and J. Unsworth, eds (2004), *A Companion to Digital Humanities*. Oxford: Blackwell. `http://www.digitalhumanities.org/companion/` (accessed 15 October 2011).

Leningrad Codex Markup Project (2000). *Project 'EL': The XML Leningrad Codex*. `http://www.leningradensis.org` (accessed 15 October 2011).

Manovich, L. (2009). Activating The Archive, or: Data Dandy Meets Data Mining. In *Digital*

Humanities 2009. Conference abstract. Maryland: Maryland Institute for Technology in the Humanities, p. xv.

Neuman, M., M. Keeler, C. Kloesel, J. Ransdell, and A. Renear (1992). The Pilot Project of the Electronic Peirce Consortium. In *ALLC-ACH92 Conference Abstracts and Program*. Oxford, pp. 25-27.

Rodríguez Ortega, N. (2010). Digital Resources for Art-Historical Research: Critical Approach. In *Digital Humanities Annual Conference 2010. Books of Abstracts*. London: Office for Humanities Communication – Centre for Computing in the Humanities, King's College, pp. 199-201.

Rodríguez Ortega, N. (2010). La cultura histórico-artística y la Historia del Arte en la sociedad digital. Una reflexión crítica sobre los modos de hacer Historia del Arte en un nuevo contexto. *Museo y Territorio* 2-3: 9-26.

Zweig, R. W. (1998). Lessons from the Palestine Post Project. *Literary and Linguistic Computing* 13: 89–97.

UNESCO (2005). *Hacia las sociedades del conocimiento*. París: Ediciones UNESCO.

Notes

1. Nevertheless, interesting research has been carried out on linking images to text. See Zweig 1998
2. 'Many humanities principles developed in hard-fought critical battles of the last decades are absent in the design of digital contexts (...) For too long, the digital humanities, the advanced research arm of humanistic scholarly dialogue with computational methods, has taken its rules and cues from digital exigencies' (Drucker 2009).
3. `http://www.nines.org/` (accessed 15 October 2011).
4. Virtual workspace for the study of Ancient Documents, University of Oxford. `http://bvreh.humanities.ox.ac.uk/VRE-SDM` (accessed 15 October 2011).
5. `http://www.ucl.ac.uk/transcribe-bentham/` (accessed 15 October 2011)

Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research

Roe, Glenn H.
glenn.roe@mod-langs.ox.ac.uk
University of Oxford, UK

The ARTFL Project
University of Chicago, USA

1. Introduction

The relationships between texts are multifaceted and complex, ranging from directly attributed quotations to indirect influences and faint allusions. Tracing these links at all levels is a core humanistic endeavor that can illuminate the meaning and reception of texts through the identification and examination of references, commonplaces, borrowings, re-workings, and even plagiarism. To aid in this sort of intertextual discovery, computational methods for examining literary text reuse can draw upon sequence alignment techniques developed primarily for gene sequencing programs in bio-informatics and plagiarism detection systems, with specific adaptations to suit the needs of humanities scholars. This paper examines just such an approach, first describing an open-source software package developed by the ARTFL Project at the University of Chicago for text reuse discovery in digitized text collections, and then offering a concrete application of this technique for literary critical research.

The software, named 'PAIR: Pairwise Alignment for Intertextual Relations' (Horton et al. 2010), applies sequence alignment techniques to large collections of literary and historical texts over several languages. The PAIR system first indexes documents by breaking them into overlapping sequences of words, called 'n-grams' or 'shingles,' and then creates a database of the occurrences of each shingle in a given corpus. This index allows for the discovery of text reuse by looking for occurrences of the same word sequences shared between documents or within different parts of the same document. Through various tuning parameters, this technique can be made flexible enough to find sequence matches with minor differences in word order, missing words, orthographic variations, misrecognized characters, and other textual variants. The first part of this

paper will thus examine the philosophy behind the PAIR approach, the algorithmic design, its implementation as an open-source Perl module, and its eventual application to a variety of tasks relevant to humanities and social science research.

2. Background

While technology is a relatively new entrant to the domain of textual studies, it nonetheless offers scholars important tools for tracing the currents of literary text reuse. Indeed, this form of 'intertextuality' can be considered a specific case of the more general problem of sequence alignment; that is, the task of identifying regions of similarity shared by two strings or sequences, often thought of as the longest common substring problem. This technique is widely applied in the field bio-informatics, where it is used to identify repeated genetic sequences, and for plagiarism detection in texts or computer programs (Clough et al. 2002; Lyon et al. 2001; Bourdaillet & Ganascia 2007).

In creating PAIR, we attempted to adapt existing techniques to suit the particular needs of humanities scholarship. Many algorithms such as BLAST exist for identifying duplicated DNA, for example, but these tend to emphasize speed over completeness. In text analysis, our corpora are small enough and our interest deep enough that we can emphasize retrieving as many hits as possible, since computation time is cheap and literary scholars are not traditionally in a hurry. The scholar's time, however, is quite valuable, so we want to return results that are of maximal interest, avoiding a preponderance of banal commonplaces such as 'Your devoted, humble servant.'

PAIR is thus based on *k-tuple* heuristics that provide a suitable balance of efficiency against completeness. Applied to text data, this involves the generation of 'shingles' (or n-grams), which are overlapping sequences of words. Preprocessing, such as removal of function and short words and reduction of orthographic variants (accents, spelling changes, case, etc.), is performed during shingle generation. This has the effect of folding numerous shingles into one underlying form for matching purposes, thus eliminating minor textual variations, which makes matching more flexible or 'fuzzy.' It also somewhat reduces the overall number of unique shingles, which aids speed of search (Seo & Croft 2008).

Once identified, the shingles within a defined window surrounding the shared shingle in each document are compared and evaluated. PAIR allows the user to set criteria for the acceptability of matches, such as the minimum overlap in shingles between the two sets, the minimum length of a shared shingle sequence, or the maximum number of consecutive gaps allowed between matching sequences in either set. If the criteria are met, the match is expanded, examining wider contexts in each document, until the criteria are violated, at which point the match is terminated and recorded. Furthermore, user configurable parameters for match retention and expansion allow for the fine-tuning of results, which is particularly important given the often 'noisy' information space of many humanities text collections.

3. Use Case: Voltaire and the Encyclopédie

As a concrete use case for the PAIR approach outlined above, we examined the intertextual relationships of two of the 18th-century's most important text collections: Denis Diderot and Jean d'Alembert's philosophic war machine, the *Encyclopédie* (28 *in-folio* volumes, published from 1751 to 1772 – digital edition provided by the ARTFL Project, University of Chicago), and the Complete Works of Voltaire (over 100 volumes, data provided by the Voltaire Foundation, University of Oxford). Both resources represent monuments of Enlightenment thought as well as model digital humanities databases: highly curated collections of historically significant texts built using the open-source PhiloLogic search and analysis software developed at the University of Chicago. By comparing these two data sets using the PAIR sequence alignment approach, we can come to a better understanding of the multifaceted and at times problematic relationship – one of influence, anxiety, and intertextuality – between the French Enlightenment's most emblematic writer and its most widely-read text.

Though initially enthusiastic about the promise and ambitions of Diderot and d'Alembert's project, Voltaire nonetheless contributed only 45 articles to the enterprise (Voltaire 1987). This curious lack of engagement on the part of one of the leading *philosophes* in the most important publication of the mid-18th century has led many to conclude that this diminished role was due to philosophical differences between Voltaire and the *Encyclopédie*'s editors (Jacob 2006). Or, as Jonathan Israel has recently contended, as a leading figure of the 'mainstream' brand of enlightenment (essentially Lockean-Newtonian in nature), Voltaire was in no way eager to follow the *Encyclopédie* as it steadily moved in the direction of a more 'Radical Enlightenment,' fundamentally Spinozist-materialist in inspiration (Israel 2006). The validity of this interpretation, however – which relies on a superficial reading of Voltaire as 'author' in the *Encyclopédie* rather than as 'authority' – begins to break down once one is presented with the results of the PAIR comparison. Indeed, even a cursory

examination of the more than 10,000 matching sequences between Voltaire's Complete Works and the *Encyclopédie*, demonstrates the preponderance of Voltaire's textual presence as an authority over and against his relatively restrained role as an encyclopedic author; a fact that without sequence alignment would have likely gone largely unnoticed or at least greatly underestimated.

Nowhere is this interaction between Voltaire and the *Encyclopédie* more pronounced than in his last, longest, and perhaps least known work, the *Questions sur l'Encyclopédie* (1770-74). Here, Voltaire revisits the 'encyclopedic moment' of the 1750s and recasts many of the concerns still relevant to he and his fellow *philosophes* some 20 years later. Adopting the same textual strategies as the *encyclopédistes* before him – indirect citation, playful borrowings, veiled references, etc. – Voltaire's reassessment of the encyclopedic texts is a treasure-trove of hidden intertextual associations. Our exploration of the *Questions* using the PAIR system will thus, for the first time, give us a more general idea of the scope and scale of Voltaire's prolonged engagement with the *Encyclopédie* and its contributors. Finally, this extended vision (or version) of Voltaire's 'encyclopedism' – made possible through the application of the sequence alignment techniques outlined above – implies a certain continuity of Enlightenment thought (at least by its main protagonists) from 1750 to 1775. By way of this 'intertextual' continuity, we thus arrive at a more comprehensive understanding of the French Enlightenment than is currently reflected by recent attempts (Jacob 2006; Israel 2006) at dividing its participants into binomial 'mainstream' and 'radical' camps.

References

Altschul, S. F., W. Gish, W. Miller, et al. (1990). Basic Local Alignment Search Tool. *The Journal of Molecular Biology* 215: 403–10.

Bourdaillet, J., and J.-G. Ganascia (2007). Alignment of noisy unstructured data, *IJCAI-2007*, Hyderabad, India - January 8, 2007.

Clough, P., R. Gaizauskas, S. S. L. Piao, and Y. Wilks (2002). METER: MEasuring TExt Reuse. *Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics,* pp. 152-159.

Lyon, C., J. Malcolm, and B. Dickerson (2001). Detecting Short Passages of Similar Text in Large Document Collections. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pp. 118-125

Horton, R., M. Olsen, and G. Roe (2010). Something Borrowed: Sequence Alignment and the Detection of Similar Passages in Large Text Collections. *Digital Studies - Le Champ numérique* 2.1.

Israel, J. (2006). *Contested Enlightenment: Philosophy, Modernity, and the Emancipation of Man 1670-1752*. Oxford: Oxford UP.

Jacob, M. (2006). *The Radical Enlightenment: Pantheists, Freemasons and Republicans*. London: Cornerstone Publishing.

Seo, J., and B. W. Croft (2008). Local text reuse detection. *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* New York, NY, pp. 571-578.

Voltaire (1987). *Œuvres complètes. Volume 33*. Oxford: Voltaire Foundation.

Digital resources

The ARTFL *Encyclopédie* Project (University of Chicago): `http://encyclopedie.uchicago.edu/`

Voltaire électronique database (Voltaire Foundation, University of Oxford): `http://www.lib.uchicago.edu/efts/VOLTAIRE/`

PAIR: `http://code.google.com/p/text-pair/`

Engaging the Museum Space: Mobilising Visitor Engagement with Digital Content Creation

Ross, Claire Stephanie
claire.ross@ucl.ac.uk
University College London, UK

Gray, Steven
steven.gray@ucl.ac.uk
University College London, UK

Warwick, Claire
c.warwick@ucl.ac.uk
University College London, UK

Hudson Smith, Andrew
a.hudson-smith@ucl.ac.uk
University College London, UK

Terras, Melissa
m.terras@ucl.ac.uk
University College London, UK

The following paper presents the results of the QRator research project which aims to understand how digital technologies, such as interactive labels and smart phones, create new ways for users to engage with museum objects; investigate the value and constraints of digital sources and methods involving cultural content; and demonstrate how crowdsourced digital interpretation may be utilised as a research source.

Museum exhibitions have been transformed by the addition of digital technology to enhance the visitor experience. Ubiquitous mobile technologies offer museum professionals new ways of personally engaging visitors with content, creating new relationships between museums and their users. Museums and other cultural institutions have made significant investments in developing and disseminating digital content in the physical museum space to reach and engage users, marking a shift in how museums communicate publicly their role as custodians of cultural content and their attitude towards cultural authority. Despite recent technical advances in collections access and interpretation, a number of key issues still remain. Does the rapidly changing technological environment provide a more engaging and participatory visitor experience?

The QRator project explores how mobile devices and interactive digital labels can create new models for public engagement, visitor meaning-making and the construction of multiple interpretations inside museum spaces. This project is located within the emerging technological and cultural phenomenon known as 'The Internet of Things'; the technical and social shift that is anticipated as society moves to a ubiquitous form of computing in which every device is 'on', and connected to the Internet. QRator is based on 'Tales of things' (`http://www.talesofthings.com`) which has developed a 'method for cataloguing physical objects online which could make museums and galleries a more interactive experience' (Giles 2010). QRator takes the technology a step further, allowing users to take part in content creation on digital interactive labels- static iPads and their own mobile phones: a sustainable model for two-way public interaction in museum spaces. The QRator project uses iPads installed in the UCL Grant Museum of Zoology to provide a fully interactive experience where visitors respond to questions posed by the curators, contribute to discussions, and leave comments about individual exhibits. Visitors' comments are synchronised with the QRator website (`http://www.qrator.org`) to allow them to contribute to the continuing discussion away from a museum setting. The application provides each exhibit with a QR code, a matrix barcode that embeds information such as text or an URL within a graphic that users can read using mobile devices, which links the physical exhibit with the associated conversations. When scanned these codes allow users to discover information about an object and join the conversation from their own mobile device. The unpredictable, multiple forms of interpretation produced by the use of mobile devices and interactive labels make us to reconsider ways in which museums provide information about objects and collections and should also allow museums to become more engaging for visitors.

1. Methodology

Data from the ten QRator iPads was collected by archiving contributions from March to July 2011; each individual visitor contribution is simultaneously uploaded to the master database on the Tales of Things website, followed by the QRator website pulling the data about each case label (current question) from the master database and integrates these comments within QRator online. These comments are then aggregated together based on the current questions originally asked by the museum. This resulted in a corpus of 1463 visitor contributions, totalling 13,308 words and 2,708 unique tokens, providing a rich dataset for the analysis of visitor experience.

Visitor contributions were categorized qualitatively using open coded content analysis where each comment was read and categorized. Contributions

were divided into three basic categories; about the current question or topic, about the museum, or noise. Despite the apparently simplistic categorisation it is possible to discover patterns of use and begin to understand how visitors are relating to and interpreting the exhibitions, and making meaning from their experience.

For the purpose of this study, various quantitative measures were used such as analysing the frequency of comments according to date and time, comparing comment rate between the ten iPad's and suitable text analysis tools were used to interrogate the corpus. One of these purpose built tools emulates the popular Wordle[1] visualisation where the frequency of words within the whole corpus is sized in relation to the words font size. This tool provided more stylistic control over the visualisation than other web-based services like Wordle. In addition, the corpus was analysed using a Sentiment analysis tool, SentiStrength[2] developed by Thelwall et al, (forthcoming), in order to automatically measure emotion in the visitor comments, which provides an indication of a positive or negative museum experience.

2. Findings

The largest proportion of the comments in the corpus fell into the category of topic (42%), triggered predominately by the QRator interface and questions posed by the museum curators, suggesting that visitors are inspired to share their own experiences, thus co-constructing a public multiple interpretation of museum objects. This is very pleasing, since this was exactly what the museum professionals had hoped might happen.

The lack of spam and inappropriate commenting is surprising (22%). Many museums have been hesitant to open up communication to greater participation by visitors. There is an ingrained fear in the museum profession that visitors will leave inappropriate comments when there is no moderation or intervention by the museum (Russo et al. 2008) despite research showing that museum visitors want to engage with complex, controversial topics by making comments or talking to staff and other visitors (Kelly 2006). The QRator project and the Grant Museum have, however, adopted the concept of 'radical trust' in the visitor community:

'Radical trust is about trusting the community. We know that abuse can happen, but we trust (radically) that the community and participation will work. In the real world, we know that vandalism happens but we still put art and sculpture up in our parks. As an online community we come up with safeguards or mechanisms that help keep open contribution and participation working ' (Fichter 2006).

Inherent in the term is the suggestion of a previous lack of trust shown by museums towards visitors, but also the admission that such trust is regarded as new and perhaps dangerous. Nevertheless, the QRator data suggests that 'radical trust' in visitors does indeed work: spamming and inappropriate commenting does not appear to have happened to a significant extent in the Grant Museum.

Interestingly, many of the visitor comments focused on opinions of the museum as a whole (36%). This raises the question of whether a digital technology used in this way promotes of an opportunity for visitors to make meaning from their whole experience, rather than engage with the exhibit specific content and interpret the exhibitions themselves.

The length of comment may also be used as an indicator of engagement- if we assume that those who are interested in an issue or topic may wish to write at greater length. Indeed the average length of comment increased significantly between categories. The noise category had an average of 4.071429 words, comments on the museum had 7.431599 words and visitor contributions on topic had an average of 15.36672 words. This is pleasing, since it suggests that visitors were inspired by the questions to engage with topics in a relatively complex fashion. Additionally when compared to the SentiStrength results, which classifies for positive and negative sentiment on a scale of 1 (no sentiment) to 5 (very strong positive/negative sentiment), highlights that the comments on the museum were in average more positive in sentiment (2.04 positive) whereas the comments on topic had an equal positive to negative response (1.52 positive; 1.55 negative). Suggesting more engaged texts often contain a mix of positive and negative sentiment, in contrast to less engagement which is more likely to produce a single sentiment result.

3. Conclusion

Digital technologies are becoming more embedded, ubiquitous and networked, with enhanced capabilities for rich social interactions, context awareness and connectivity. This has led to unprecedented changes in the provision of digital museum resources, which are beginning to transform the experience of visiting museums. The QRator project represents a shift in how cultural organisations act as trusted and authoritarian institutions; communicate knowledge to the community; and integrate their role as keepers of cultural content with their responsibility to facilitate access to content. It also suggests that

users are willing to take part in a dialogue, and express their views about their visit and individual object via digital technologies. It further suggests that in most cases they can be trusted to do so in a thoughtful, serious fashion. The challenges that digital technology and participatory media bring to museums demonstrate a change from a one to many transmission to a many to many interaction, in which museums use their own voice and authority to encourage participatory communication and content creation with visitors. The growing emphasis on the interactional and informal nature of learning in museums provides the perfect opportunity to showcase digital interactive technologies as important resources for engaging visitors in exhibits and more generally in museums as a whole (Thomas & Mintz 1998; Marty & Burton Jones 2007; Heath & vom Lehm 2010).

Acknowledgments

The authors of this paper would like to acknowledge the other members of the QRator team: Jack Ashby and March Carnall (UCL Grant Museum of Zoology), Sally MacDonald (UCL Museums and Public Engagement), Sussanah Chan (UCL) and Emma-Louise Nicholls (UCL Grant Museum of Zoology). QRator was funded through Beacons for Public Engagement Innovation Seed funding and we would like to thank Hilary Jackson (UCL Public Engagement) for her support and advice throughout the project.

References

Fichter, D. (2006). Web 2.0, library 2.0 and radical trust: A first take. April 2. Accessed 27th October 2011 at http://library2.usask.ca/~fichter/blog_on_the_side/2006/04/web-2.html

Giles, J. (2010). Barcodes help objects tell their stories. *New Scientist* (17th April, 2010).

Heath, C., and D. vom Lehn (2010). Interactivity and Collaboration: new forms of participation in museums, galleries and science centres. In R. Parry (ed.), *Museums in a Digital Age*. Abingdon: Routledge, pp. 266-280.

Kelly, L. (2006). Museums as sources of information and learning: The decision-making process. *Open Museum Journal* 8, accessed at http://archive.amol.org.au/omj/volume8/volume8_index.asp

Marty, P., and K. Burton Jones (2007). *Museum Informatics: People, Information, and Technology in Museums*. New York: Routledge.

Russo, A., J. Watkins,L. Kelly, and S. Chan (2008). Participatory communication with social media. *Curator* 51(1): 21-31.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas (forthcoming). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*.

Thomas, S., and A. Mintz, eds. (1998). *The Virtual and the Real: Media in the Museum*. Washington, DC: American Association of Museums.

Notes

1. http://www.wordle.net/
2. http://sentistrength.wlv.ac.uk/

Aiding the Interpretation of Ancient Documents

Roued-Cunliffe, Henriette
henriette.roued@classics.ox.ac.uk
Centre for the Study of Ancient Documents,
University of Oxford, UK

How can Decision Support System (DSS) software aid the interpretation process involved in the reading of ancient documents? This paper discusses the development of a DSS prototype for the reading of ancient texts. In this context the term 'ancient documents' is used to describe mainly Greek and Latin texts and the term 'scholars' is used to describe readers of these documents (e.g. papyrologists, epigraphers, palaeographers). However, the results from this research can be applicable to many other texts ranging from Nordic runes to 18th Century love letters.

In order to develop an appropriate tool it is important first to comprehend the interpretation process involved in reading ancient documents. This is not a linear process but rather a recursive process where the scholar moves between different levels of reading, such as 'understanding the meaning of a character' or 'understanding the meaning of a phrase' (Youtie 1963; Terras 2006). This realization has been paramount to the development of the DSS prototype.

1. Aiding, not making decisions

When examining whether a tool such as DSS can aid the reading of ancient documents it is important first to realize what such a tool can and should do and what it cannot and should not do. I did not set out to develop a tool that could automate the process of reading ancient documents without human interaction. Firstly, I did not wish to see humans removed from this interpretational process and replaced with robotics.

Good reading is a reflective as well as a sensational process which engages our abstract mental faculties, and we give it up at our peril (Deegan & Sutherland 2009: viii).

Secondly, interpretation is a part of the 'higher levels of analytic and creative functions which are the prerogative of the human mind' (Busa 1980: 89) and the symbol-processing powers of a computer should not be mistaken for this (Jonscher 2000: 149).

Instead this research has examined how the use of DSS technology can support the interpretative process, among other things, by remembering the complex reasoning that forms a large part of this interpretation process. Remembering decisions and the complex reasoning leading up to decisions is something that DSS is good at. The DSS prototype has demonstrated how this is possible.

Furthermore the prototype has demonstrated other aspects of reading ancient documents, apart from the interpretation process, where IT tools can provide support such as searching huge datasets and the automated publishing of editions.

2. The DSS Prototype

The prototype was developed as a proof of concept. First I used existing research on reading ancient documents (Bowman & Brady 2005; Terras 2006; Tarte 2011) to model the network of interpretation (Roued-Cunliffe 2010), and then I developed the DSS prototype to demonstrate how a DSS can aid the interpretation process based on this model.

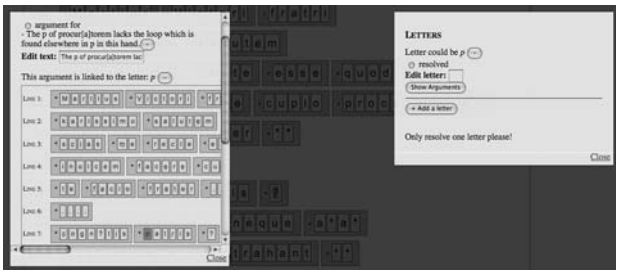


Figure 1: Screenshot from the DSS prototype (Roued-Cunliffe 2012: 87)

The prototype (Fig. 1) demonstrates how it would be possible to input data and edit the structure and meaning of a document on different levels. Scholars can begin by entering basic structure such as line or character elements and work back and forth between this level and higher levels such as editing the meaning of a letter or word. Furthermore, the prototype enables scholars to use the letters and character elements they have input to search relevant corpora databases for possible parallels.

The prototype uses the concept of argumentation, which can be defined as 'the process of constructing arguments about propositions, and the assignment of statements of confidence to those propositions based on the nature and relative strength of their supporting arguments' (Fox et al. 1996: 428). In the prototype these statements of confidence are merely whether the user believes the reading to be right or wrong. It does not require users to signify how much they believe in a certain decision. Furthermore, the arguments are either **for** or **against** the interpretation in question. No matter how many arguments are **for** or **against** any one interpretation the prototype will not try to persuade

the user to make a certain decision. The prototype is first and foremost flexible and user-driven and its role is purely that of a good assistant, reminding users of their own decisions and reflections.

3. Future DSS

The DSS prototype, presented in this paper, has been successful in illustrating how this technology can aid such a complex interpretation process, as is the reading of ancient documents. However, in future this prototype will need to be re-developed to contain, among others, a more advanced user interface. With this in hand it would be possible to conduct user testing, which would give a basis for the development of a fully working DSS. I have a suggestion for a future layer-based system that would be user-friendly and could work well in conjunction with a DSS and other useful tools for aiding the reading of ancient documents.

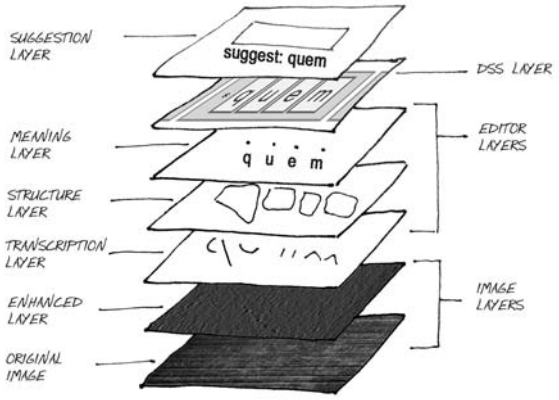


Figure 2: Drawing of the layer-based DSS concept (Roed-Cunliffe 2012: 132)

The layer-based system would contain four types of layers that users can swap and switch between so that they can edit the different layers at any time and edit one layer while others are visible below (Fig. 2). The first types of layers are the image layers. These image layers could be the original image of an object, PMT images of an object (Earl et al, 2011) or enhanced images (e.g. with wood-grain removed or the illumination corrected (Tarte et al, 2011)). The second type of layers are the editor layers which can include a layer for hand-drawn transcriptions of the text, a layer in which to edit the structure of the text (i.e. lines, words, characters and spaces in the text) and a layer in which to add meaning to the text (i.e. this letter is *b*). The third type of layers is the DSS layer which would resemble the DSS prototype. The fourth and final type of layers are the suggestion layers from which it would be possible to retrieve for example suggested words from a word search engine based on letters input in the earlier editor layer.

4. Conclusion

The research presented in this paper aims to illustrate how DSS software could aid the interpretation of ancient documents and has been successful in doing this through the development of the DSS prototype. The DSS prototype covered different areas of the interpretation process from input to argumentation **for** and **against** different decisions. It has illustrated how DSS software can assist scholars by remembering their complex reasoning and suggesting readings along the way. Furthermore, the research describes an idea for a fully formed future DSS based on the finding from the development of the prototype.

Acknowledgements

This research is a part of the author's thesis: 'A Decision Support System for the Reading of Ancient Documents' at the Faculty of Classics, University of Oxford. It is funded by an AHRC Doctoral Studentship attached to the e-Science and Ancient Documents (eSAD) project (http://esad.classics.ox.ac.uk/). I wish to thank Prof. Alan Bowman, Prof. Sir Michael Brady, Dr. Melissa Terras, Dr. Segolene Tarte, Dr. Charles Crowther, Margaret Sasanow and John Pybus for their support throughout my research.

References

Bowman, A. K., and M. Brady (2005). *Images and Artefacts of the Ancient World*. London: British Academy.

Busa, R. (1980). The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities* 14: 83-90.

Deegan, M., and K. Sutherland (2009). *Transferred Illusions: Digital Technology and the Forms of Print*. London: Ashgate.

Earl, G., P. Basford, A. Bischoff, A. Bowman, C. Crowther, J. Dahl, M. Hodgson, L. Isaksen, E. Kotoula, K. Martinez, H. Pagi, and K. E. Piquette (2011). *Reflectance Transformation Imaging Systems for Ancient Documentary Artifacts*. http://ewic.bcs.org/upload/pdf/ewic_ev11_s8paper3.pdf (accessed 07.03.2012)

Fox, J., P. Krause, and M. Elvang-Gøransson (1996). Argumentation as a General Framework for Uncertain Reasoning. In D. Heckerman and E. H. Mamdani (eds.), *UAI '93: Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence, July 9-11, 1993, The Catholic University of America, USA*. San Fransisco: Morgan Kaufmann Publishers, pp. 428-434.

Jonscher, C. (2000). *Wired Life: Who Are We in the Digital Age?* London: Anchor.

Roued-Cunliffe, H. (2010). Towards an Interpretation Support System for Reading Ancient Documents. *Literary and Linguistics Computing* 25(4): 365-379.

Roued-Cunliffe, H. (2012). *A decision Support System for the Reading of Ancient Documents*. D.Phil Thesis, Classics Faculty, University of Oxford.

Tarte, S. (2011). Papyroligical Investigations: Transferring Perception and Interpretation into the Digital World. *Literary and Linguistics Computing* 26(2): 233-247.

Tarte, S., M. Brady, A. K. Bowman, and M. Terras (2011). Image Capture and Processing for Enhancing the Legibility of Incised Texts. In V. Vahtikari, M. Hakkarainen, and A. Nurminnen, *Eikonopoiia. Digital Imaging of Ancient Textual Heritage, Technological Challenges and Solutions*. Commentationes Humanarum Litterarum 129. Helsinki: Academy of Finland, pp. 173-88.

Terras, M. (2006). *Image to Interpretation: Intelligent Systems to Aid Historians in the Reading of the Vindolanda Texts*. Oxford Studies in Ancient Documents. Oxford: Oxford UP.

Youtie, H. C. (1963). The Papyrologist: Artificer of Fact *GRBS* 4: 19-32.

The Twelve Disputed 'Federalist' Papers: A Case for Collaboration

Rudman, Joseph
jr20@heps.phys.cmu.edu
Department of English, Carnegie Mellon University, USA

1. Introduction

This paper discusses the controversy over the authorship of the *Federalist*papers as seen and studied by traditional historians and by over 100 non-traditional authorship attribution practitioners.

The *Federalist* papers were written during the years 1787 and 1788 by Alexander Hamilton, John Jay, and James Madison. These 85 'propaganda' tracts were intended to help get the U.S. Constitution ratified. They were all published anonymously under the pseudonym, 'Publius.' The authorship of certain of the *Federalist* essays was disputed from the beginning. Both Hamilton and Madison produced lists that claimed some of the same papers. There followed a series of lists, some claiming authorship for Madison and some for Hamilton.

The consensus of traditional scholarship, seconded by Mosteller and Wallace, allocates the papers: Hamilton 51 (1, 6-9, 11-13, 15-17, 21-36, 59-61, 65-85); Madison 29 (10, 14, 18-20, 37-58, 62, 63); Jay 5 (2-5, 64).

In 2005, I presented a paper at the ACH/ALLC conference, 'The Non-Traditional Case for the Authorship of the *Federalist* Papers: A House Built on Sand?" After 7 more years of 'traditional" research and the addition of over 70 non-traditional studies of the *Federalist*, I am able to remove the question mark and put forth a reasoned argument that many, if not all, of the twelve disputed papers are a collaboration and not written solely by Madison, as the consensus of traditional scholarship and non-traditional authorship studies claim.

In 1964, Mosteller and Wallace, building on the earlier unpublished work of Frederick Williams and Frederick Mosteller, published their non-traditional authorship attribution study, *Inference and Disputed Authorship: The* Federalist. It is arguably the most famous and well respected of all of the non-traditional attribution studies. Since then, well over a thousand papers have cited the Mosteller and Wallace work and over 100

non-traditional practitioners have analyzed and/or conducted variations of the original study.

Mosteller and Wallace set the boundary conditions for the subsequent non-traditional work – e.g., not using the Jay articles as a control. Most of these later practitioners do not select or prepare the input text as carefully as Mosteller and Wallace – and their selection and preparation was not as rigorous and complete as it should have been – as we will see.

2. Problems with the non-traditional case

There are many problems with the Mosteller and Wallace study and with the over 90 other non-traditional studies that cast strong doubts on their results:

(1.) The Federalist Papers

A crucial step of any non-traditional authorship study is to obtain a 'starting text.' As a rule, the closest text to the 'final' holograph should be found and used. Every step away from that holograph introduces systematic errors.

(2.) The Hamilton Texts

Mosteller and Wallace go outside of the *Federalist* papers to construct their block of Hamilton tracts.

(3.) The Madison Texts

Mosteller and Wallace also go outside of the *Federalist* papers to construct their block of Madison tracts. In the case of the Madison block, there is a 20 year difference in their production dates. And what is worse – some of these outside essays have been shown to be not by Madison.

(4.) The Control Texts

There is not one non-traditional study that uses any meaningful controls. The Mosteller and Wallace and Wachal use of a 'training set' is not a control. The closest test to a necessary control is found in the 'validation' of four Hamilton papers (79, 80, 82, 85) and other sets by Mosteller and Wallace.

3. More problems – Arriving at the analysis texts

(A) Background and Definitions

No attribution practitioner should question the fact that valid texts are needed if valid results are to be obtained. No matter how sophisticated the statistical analysis is, a bad text invalidates the results.

(B) Unediting

- Editorial Interpolation
- Printer Interpolation

(C) De-editing

- Quotes
- Plagiarism
- Collaboration
- Graphs and Numbers
- Guide Words
- Foreign Languages
- Translations

(D) Editing

- Encoding the Text
- Regularizing
- Lemmatizing

There are many significant weaknesses in the text preparation part of the Mosteller and Wallace study. For example, they state that:

- They ignore the extent of the editing done by the other man [i.e. Hamilton or Madison] and by all of the 'outside' editors.
  - They did not disambiguate words – e.g. the personal pronoun 'I' is teated the same as the Roman numeral 'I' – the noun 'abuse' is treated exactly like the verb 'abuse'.
  - They do not publish complete details of their 'little book of decisions' or the rational behind any of these decisions. –
  - They do not use the newspaper versions of the first 77 papers.
  - They typed the text onto cards to be read by the computer, but for reasons of 'economics' they used hand counts for much of their study. They write about the many problems this introduces but do not assign any systematic errors – e.g. (1) they show the differences in the hand counts vs. the machine counts, (2) they tell us that their proofreaders missed missing words, repeated lines, and single word repetitions.
  - Missing bibliographical sources – e.g. Smyth and Wachal.

One of the guiding principles of any scientific study is 'reproducibility.' Any other practitioner should be able to reproduce a given study and get identical results. None of the over 90 *Federalist* studies mentioned in this paper give anywhere near the information needed – a fatal flaw.

Time will not allow for a detailed explication of these studies. In essence, they are all fatally flawed – many

do not indicate which edition they used, most either did no unediting, de-editing, or editing – or they fail to say if they did anything to the text.

4. The case for 'collaboration'

There can be no doubt that the *Federalist* project was a collaboration – a collaboration on many levels – but the depth of that collaboration is what is in question.

We know that Publius' *Federalist* series was the product of three men – Hamilton, Madison, and Jay. No one disputes this, as long as specific papers are not discussed.

Most scholars agree that *Federalist* 18, 19, and 20 were written jointly by Hamilton and Madison. Exactly what parts were by which man is not agreed upon.

There is strong evidence that Hamilton and Madison also had joint hands in many other numbers of the *Federalist* – including the twelve disputed papers.

The evidence comes from traditional and non-traditional authorship attribution methodologies.

Traditional

- 1840 – Renwick's *Lives of John Jay and Alexander Hamilton*.
- 1894 – Whitaker's, 'A Problem in Authorship: Who Wrote "The *Federalist*?"'
- 1978 – Smyth's, 'The *Federalist*: The Authorship of the Disputed Papers.'
- 1984 – Cary's, 'Publius – A Split Personality.'
- 1984 – Furtwangler's, *The Authority of Publius: A Reading of the Federalist Papers*.
- 1999 – Kesler's 'Introduction to the *Federalist Papers.'*

Non-Traditional

The Collins et al. study, 'Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in the *Federalist* Papers' confirms the premise of a deeper collaboration. Most of the non traditional authorship studies do not agree with each other. A list will be shown that reveals how many times each of the twelve disputed papers have been attributed to Hamilton in the over 90 other non-traditional studies.

This paper concludes with a discussion of the following:

- Acceptance of Results by Non-Traditional Practitioners
- Acceptance of Results by History Scholars
- Do the Multiple Flaws in the Non-traditional Studies Invalidate the Results

The bibliography for this paper contains well over 300 entries. The following is just a sample.

References

Adair, D. (1944). The Authorship of the Disputed Federalist Papers. *The William and Mary Quarterly* (Third Series) 1(2): 97-122.

Bosch, R. A., and J. A. Smith (1998). Separating Hyperplanes and the Authorship of the Disputed Federalist Papers. *The American Mathematical Monthly* 105(7): 601-607.

Carey, G. W. (1984). Publius – A Split Personality? *The Review of Politics* 46(1): 5-22.

Collins, J., et al. (2004). Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in *The Federalist* Papers. *Computers and the Humanities* 38(1):15-36.

Cooke, J. E., ed. (1956).*The Federalist*. Cleveland: Meridian Books, The World Publishing Company.

Davis, G. (2003). `RE: Gutenberg Edition of *Federalist*. Private E-mail, 20 November 2003, 18:46:51.

Farringdon, M. G., and A. Q. Morton (1990). Fielding and the *Federalist*. *CS Report Series* University of Glasgow CSC 90/R6.

Forsyth, R. S. (1995). Stylistic Structures: A Computational Approach to Text Classification. Diss. University of Nottingham.

Fung, G. (2003). The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization. *Proceedings of the 2003 Conference on Diversity in Computing*. Atlanta, Georgia, pp. 42-46.

Gerritsen, C. M. (2003). Authorship Attribution Using Lexical Attraction. M.S. Dissertation, Massachusetts Institute of Technology.

Hilton, M. L., and D. I. Holmes (1993). An Assessment of Cumulative Sum Charts for Authorship Attribution. *Literary and Linguistic Computing* 8(2): 73-80.

Holmes, D. I., and R. S. Forsyth (1995). The *Federalist* Revisited: New Directions in Authorship Attribution.*Literary and Linguistic Computing* 10(2): 111-127.

Jockers, M. L., and D. M. Witten (2010). A Comparative Study of Machine Learning Methods for Authorship Attribution. *Literary and Linguistic Computing* 25(2): 215-224.

Khmelev, D. V., and F. J. Tweedie (2001). Using Markov Chains for Identification of Writers. *Literary and Linguistic Computing* 16(3): 299-307.

Kjell, B. (1994). Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers. *Literary and Linguistic Computing* 9(2): 119-124.

Koppel, M., N. Akiva, and I. Dagan (2006). Feature Instability as a Criterion for Selecting Potential Style Markers. *Journal of the American Society for Information Science and Technology* 57(11): 1519-1525.

Levitan, S., and S. Argamon (2006). Fixing the Federalist: Correcting Results and Evaluating Attribution.' Paper presented at *Digital Humanities 2006*. Paris, Sorbonne 2006 (Paper courtesy of authors.)

Martindale, C., and D. McKenzie (1995). On the Unity of Content Analysis in Authorship Attribution: The *Federalist*. *Computers and the Humanities* 29: 259-270.

Marcus, L. S. (2000). Afterword: Confessions of a Reformed Uneditor. In A. Murphy (ed.), *The Renaissance Text: Theory, Editing, Textuality*. Manchester: Manchester UP, pp. 211-216.

Marcus, L. S. (1996). *Unediting the Renaissance: Shakespeare, Marlow, Milton*. London: Routledge 1996.

McColly, W., and D. Weire (1983). Literary Attribution and Likelihood-Ratio Tests: The Case of the Middle English Pearle-Poems. *Computers and the Humanities* 17: 65-75.

Merriam, Th. (1989). An Experiment with the *Federalist* Papers. *Computers and the Humanities* 23(3): 251-254.

Mosteller, F., and D. L. Wallace (1964). *Applied Bayesian and Classical Inference: The Case of* The Federalist *Papers*. New York: Springer 1984. CSLI Publications published a reprint of the second edition in 2007 with a new introduction by John Nerbonne.

Mustafa, T. K., et al. (2010). Dropping Down the Maximum Item Set: Improving the Stylometric Authorship Algorithm in the Text Mining for Authorship Investigation. *Journal of Computer Science* 6(3): 230-238.

Oaks, M. (2004) Ant Colony Optimisation for Stylometry: The Federalist Papers. *Proceedings of the 5th International Conference on Recent Advances in Soft Computing,* pp. 86-91.

Pennebaker, J. W. The *Federalist*. Unpublished preliminary work – courtesy of the Author.

Piaia, J. [For Frederick Mosteller] Private E-mail, Tuesday, 22 July 2003, 10:57:38.

Project Gutenberg `http://promo.net/pg/` [9-30-2003].

Rokeach, M., et al. (1970). A Value Analysis of the Disputed Federalist Papers. *Journal of Personality and Social Psychology* 16(2): 245-250.

Scigliano, R., ed. (2000). *The Federalist*. New York: The Modern Library.

Smyth, L. Qu. (1978). The *Federalist*: The Authorship of the Disputed Papers. Ph.D. University of Virginia.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2001). Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities* 35: 193-214.

Stillinger, J. (1991). Multiple Authorship and the Question of Authority. In. D. C. Greetham and W. Speed Hill (eds.), *Text: Transactions of the Society for Textual Scholarship (5)*. New York: AMS Press, pp. 283-293.

Wachal, R. S. (1966). Linguistic Evidence, Statistical Inference, and Disputed Authorship. Dissertation, University of Wisconsin.

Waugh, S., A. Adams, and F. Tweedie (2000). Computational Stylistics Using Artificial Neural Networks. *Literary and Linguistic Computing* 15(2): 187-197.

Yang, A. C. C., et al. (2003). Information Categorization Approach to Literary Authorship Disputes. *Physica A* 329(3-4): 473-483.

Writing with Sound: Composing Multimodal, Long-Form Scholarship

Sayers, Jentery
jentery@uvic.ca
University of Victoria, Canada

Historically speaking, the fields of digital humanities and media studies have remained parallel at best, with the former anchored more in computational practices and technical competencies than the latter. However, recent work – such as Sharon Daniel's 'Public Secrets' (2007), Kathleen Fitzpatrick's *Planned Obsolescence* (2011), Matthew Kirschenbaum's *Mechanisms* (2008), Kari Kraus's 'Conjectural Criticism' (2009), and Lev Manovich's cultural analytics (2012) – suggestively troubles this parallelism. From the vantage of media studies, digital humanities allows scholars to shift from commenting *about* new media and technologies to constructing arguments *with* and *through* them (McPherson 2009: 120). Informed by claims from experience and anchored in embodied acts of building, digital humanities arguments necessarily become 'hands on,' and scholarly distance from technologies no longer holds. Meanwhile, media studies investments in cultural criticism and situated knowledge-making are increasingly important to today's digital humanities practitioners, involved such as they are in multimodal communication (e.g., interactive visualizations, geospatial representations, rich exhibits, and gaming). For instance, Alan Liu argues that 'digital humanities should enter into fuller dialogue with the adjacent fields of new media studies and media archaeology so as to extend *reflection* on core instrumental technologies in cultural and historical directions' (Liu 2012: 501, emphasis added). This fuller dialogue would enhance the field's awareness of how work with technologies and data intersects with the relevant social, economic, and political issues of our time. Assuming those conversations are inevitable (and that media studies and digital humanities should continue to overlap and intersect), a key question thus emerges: what would be an appropriate platform for such forms of scholarly communication? How would it function, and under what assumptions about how digital humanities and media studies are practiced?

With these questions in mind, this talk uses the author's ongoing book project (tentatively titled, *How Text Lost Its Source: Magnetic Recording Cultures*) to present his findings on the yet-to-be-released Scalar platform. Supported by the Andrew W. Mellon Foundation and the U.S. National Endowment for the Humanities, Scalar is designed for authoring and publishing multimodal books. Built using PHP ARC2 (a MySQL-based Semantic Web/RDF framework), HTML, CSS, jQuery, XLST, Dublin Core, and other ontologies including SIOC, it enables users to assemble media (e.g., audio files) from multiple sources and juxtapose them with their own writing. The platform particularly facilitates work with visual materials; however, it also lends itself to audio, which is central to the history articulated in *How Text Lost Its Source*. Throughout the talk, 'multimodal' is preferred over 'multimedia' because the term stresses systems of 'sensory or perceptual experience' over the 'means of conveying [or storing] a representation' (Anastopoulou, Baber & Sharples 2001). And in the particular case of a Scalar book, it assumes that: (1) attention behaviors such as reading, watching, and listening are not inherent to or determined by a medium, (2) digital communication frequently entails blending approaches to composition (e.g., through images, audio, video, text, and databases), (3) layers of digital content are materially distinct and yet function relationally, and (4) scholarly interpretation demands several sensory modes (e.g., listening closely, scanning, and clicking).

The author's preliminary findings on Scalar show that, most importantly, it encourages reflexive approaches to the computational processes involved in digital humanities research. Influenced by the work of LaDona Knigge and Meghan Cope (2006) as well as Mei-Po Kwan (2002), the author organizes these findings according to how Scalar fosters: (1) exploratory research across multiple archives, (2) iterative and recursive argumentation, (3) an oscillation between abstract and concrete expressions, and (4) multiple interpretations of media and cultural history. Throughout, an emphasis is placed on juxtaposing critical writing with the critical use of audio in Scalar, including a few demonstrations of *How Text Lost Its Source*.

In the case of exploratory research across multiple archives, the platform is persuasive because it affords composites of text, audio, video, and images drawn from the various histories of magnetic recording, enabling the author to represent audio across the material specificities of multiple media (rather than reducing audio to simply sound or text). It also allows historical evidence to be modeled and exhibited independently of the author's writing (e.g., audiences can navigate all audio files collected for the history without reading the author's interpretations). However, any given audio file in a Scalar book can be annotated through discrete, time-stamped

commentary by the author, and this commentary is displayed within the medium's own temporality. While such features are now typical in visual culture (e.g., annotating lexia in Commentpress or tagging an image in Flickr), few such mechanisms exist for the scholarly treatment of sound.

Additionally, Scalar facilitates iterative and recursive expression because it not only keeps a version history of all contributions and makes that history public; it also allows authors to duplicate media content (e.g., a magnetic tape recording) in another context and then re-evaluate it. Evidence in a scholarly argument thus becomes subject to constant re-visitation and re-use (e.g., the same tape recording is interpreted several times), underscoring the fact that media histories emerge from multiple (and often conflicting) perspectives, worldviews, and accounts. Importantly, these media – together with the author's writing – are presented throughout a Scalar book both in the aggregate and in the particular. Using RDF/XML and D3.js dynamic visualizations, any single instance (i.e., anything with a URL) in the book can be situated in relation to the balance of the content. Consequently, the book's historical materials are structured and expressed in such a way that they can be studied in isolation but cannot exist independently. And when a book is modeled effectively, its audiences can study complex cultural relationships (e.g., between two audio files) established by the author. Importantly, such relationships are frequently constructed and encoded based on what has not been written or recorded in media history. While RDF/XML and D3.js help scholars organize and convey it, no computational practices exist (as of yet) for mining or visualizing implicit content, especially in the case of sound.

Finally, the Scalar platform excels at fostering various interpretations of media history by allowing audiences to – in the fashion of context-sensitive design – select how a book's content is viewed. These views range from 'text-only' and 'media-emphasis' to a radial visualization, a force-directed graph, and a history browser. Again, this array of perspectives brushes against any totalizing account of media history, the cultural history of magnetic recording included. It also destabilizes a scholar's authority over an audience's interpretation as it allows them to arrange and re-arrange content. Such an emphasis on the constructedness of arguments is especially key in a moment when visualizations are gaining traction in the field. The dynamic juxtapositions of writing and media in Scalar – D3 visualizations among them – actively demonstrate how our data and content are taken, not given (Drucker 2011). They also facilitate robust and often contradictory accounts of material histories, where scholars can: (1) argue with and through the media they study; (2) present material relationally and discretely, either with or without commentary; (3) author within a medium's temporality; (4) exhibit the otherwise ignored processes of research and revision; (5) duplicate and interpret media in multiple contexts; and (6) underscore the contingent character of historical artifacts, to such an extent that any complete description of a given media object is difficult to say the least.

References

Anastopoulou, S., C. Baber, and M. Sharples (2001). Multimedia and Multimodal Systems: Commonalities and Differences. *Proceedings of the 5th Human Centred Technology Postgraduate Workshop*. University of Sussex. http://www.syros.aegean.gr/users/manast/Pubs/Pub_conf/C03/C03.pdf.

Daniel, S. (2007). Public Secrets. *Vectors* 2(2). http://vectors.usc.edu/index.php?page=7&projectId=57.

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* 5(1). http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Fitzpatrick, K. (2011). *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York: New York UP.

Kirschenbaum, M. (2008). *Mechanisms: New Media and the Forensic Imagination*. Cambridge: MIT Press.

Knigge, LaDona, and M. Cope (2004). Grounded Visualization: Integrating the Analysis of Qualitative and Quantitative Data through Grounded Theory and Visualization. *Environment and Planning A* 38(11): 2021-2037.

Kraus, K. (2009). Conjectural Criticism: Computing Past and Future Texts. *Digital Humanities Quarterly* 3(4). http://www.digitalhumanities.org/dhq/vol/3/4/ 000069/000069.html.

Kwan, M.-P. (2002). Feminist Visualization: Re-envisioning GIS as a Method in Feminist Geographic Research. *Annals of the Association of American Geographers* 92: 645-661.

Liu, A. (2012). Where Is Cultural Criticism in the Digital Humanities? In M. K. Gold (ed.), *Debates in the Digital Humanities*. Minneapolis: U of Minnesota P, pp. 490-509.

Manovich, L. (2012). Trending: the Promises and the Challenges of Big Social Data. In M. K. Gold (ed.),

Debates in the Digital Humanities. Minneapolis: U of Minnesota P, pp. 460-475.

McPherson, T. (2009). Media Studies and the Digital Humanities. *Cinema Journal* 48(2): 119-123.

Intra-linking the Research Corpus: Using Semantic MediaWiki as a lightweight Virtual Research Environment

Schindler, Christoph
schindler@dipf.de
German Institute for International Educational Research, Germany

Ell, Basil
basil.ell@kit.edu
Karlsruhe Institute of Technology, Germany

Rittberger, Marc
rittberger@dipf.de
German Institute for International Educational Research, Germany

In recent years, virtual research environments (VREs) emerged as a topic referring to the established research field in the digital humanities: enabling research practices with digital tools. First projects in this area are realized and discussed by the community (Carusi 2010; Dunn 2009; Neuroth et al. 2009). In the humanities, researchers point out that the so-called '*data deluge*' (Hey et al. 2003), which has influenced several national and supranational information policy agendas in the sciences, does not cover the full range of aspects of research practices in the humanities. While digital libraries and archives offer a new plurality of research resources in the humanities, the '*complexity deluge*' (Dunn 2009) formulates an opposite agenda addressing the sometimes fuzzy, interfering and dispersed practices of humanities research.[1]In this paper we want to address this tension between research data, metadata and collaborative action in the design of research corpora carried out within a VRE. Therefore we will focus the ongoing corpus re-arrangement and the potentials of *Social Semantic Media* technologies to expand metadata creation and use in qualitative research. Tools for qualitative research target the flexible coding system, i.e. allowing researchers to annotate research resources according to a classification system that may evolve over time. However, the tools have been criticized for the limited metadata interoperability of resources and research findings (Corti et al. 2011). In our paper, we outline the aspects of interoperability in qualitative corpora research and focus on researchers' capabilities to intra-link the corpus. We use the term *intra-linking* [2]to address

a main aspect in qualitative research: to create and to describe entities while allowing for the ongoing re-arrangement of entities and their properties in the research process. These capabilities will be discussed and exemplified within the scope of the project *Semantic MediaWiki for Collaborative Corpora Analysis* (*SMW-CorA*) which aims to reconfigure *Semantic MediaWiki* as a lightweight virtual research environment.[3]

The field of corpora centered research in the digital humanities offers interesting insights into the design of VREs. In the early 1990s, Biber pointed out main aspects of corpus design by problematizing *a priori* determinations of its boundaries and formal specifications. He recommends the selection of relevant objects and the formal description to be realized as a cyclic or iterative process of corpus work (Biber 1993: 256). While a linguistic approach mainly aims at a statistical 'representation' in relation to a target population, qualitative corpus research, which is focused here, pursues a so called qualitative selection, i.e. a typification of yet unknown properties in research (Bauer 2000: 20). We argue that this indeterminacy of entities and properties in qualitative research emphasizes the affordance of a VRE enabling researchers to intra-link the corpus – it means to give them the ongoing capabilities to create, modify and re-arrange entities and properties while doing research. This topic of qualitative corpus research addresses the research and design desideratum of qualitative annotations (Juola 2008) and a demanded shift to further capabilities for the researcher to control the data (Smith 2008: 178).

While the *SMW-CorA* project targets the re-use of its VRE infrastructure in different research contexts in the mid-term, its initial design is subject to a co-operation with a research project in the history of education, involving a major library in this field. The educational research project encompasses the analysis of a corpus of 25 educational lexica dating from 1774 to 1942, reconstructing the development of educational science. Discourse, field and content analysis are supported and applied to grasp the networked relationships in this scholarly field. The Library for the History of Education (BBF) hosts a large part of the lexica at the digital library *Scripta Paedagogica Online* (*SPO*).[4] The collection amounts to nearly 22,000 articles. Each lexicon is bibliographically described and accessible online as image files. Participatory and agile design approaches are used to offer an adequate shared space for the different stakeholders to articulate possible potentials and boundaries of the VRE.

The *SMW-CorA* project builds on a *Social Semantic Media* technology which in turn is based on the Web 2.0 software *MediaWiki* (*MW*),[5] which is used at the well-known online encyclopedia Wikipedia, and the extension *Semantic Media Wiki* (*SMW*).[6] The latter enables the use of semantic annotations and integration within the Semantic Web through import and export of semantic data and linking to external entities. While other VREs[7] have been realized using this wiki technology, the *SMW-CorA* project promotes research on a corpus by configuring the facilities of *MW* and *SMW* and by developing further extensions to offer a configurable and lightweight VRE.

As such, the basic framework *MediaWiki* offers some facilities for creating and typifying entities in a corpus for research (in our case: lemmas, authors, institutions etc.). The online encyclopedia Wikipedia demonstrates, besides the collaborative creation of texts, the capability to store documents and digital objects. Therein a wiki page can be related to other entities by hyperlinks and arranged within a hierarchical category system. *SMW* extends this by offering an increased granularity for describing and linking the corpus by adding metadata descriptions used in the Semantic Web. The unspecific hyperlinks between entities can be typified by metadata and thereby entities can be enriched with attributes or semantic relations to other entities. Furthermore, it is possible to import and export metadata in the Semantic Web standard RDF.

Within the scope of the project a set of use-cases is explored comprising the scholarly work in the research life cycle from importing research resources, coding, classifying and analyzing these to the export for re-use. The supported and envisaged use cases have in common that entities can be integrated, created, modified and interlinked (i.e. intra-linking). This functionality is focused within the project and supported by tools to enable researchers to carry out an ongoing re-arrangement of their corpora. While the aspects of importing lexica from a digital library and of exporting the bibliographic data in RDF are discussed in (Schindler et al. 2011), further use cases exemplify these capabilities. Besides this import functionality, the enrichment of entities such as editors, authors, and related affiliations with properties by using *Semantic Web Browsing* technologies is a further example.[8] Thereby, semantic properties from digital archives, libraries or further collections can be integrated and adapted to the locally used metadata schema in the VRE. This enables researchers for example to add biographical data of authors or editors from authority files (e.g. German National Library GND). Furthermore the import and collaborative development of taxonomies (e.g. classification or coding schema) are supported by interlinking entities within the VRE and linking entities into the Web of

Data. These links to the world external to the corpus, together with the data export facilities of the VRE, enable the reuse of the content created within the VRE. Additionally, the VRE provides functionalities for creating and re-arranging metadata schema as well as a bottom-up task management to allow supervision of the research process.

To summarize, we identified the need of researchers to create, manage and intra-link entities and metadata objects in a research corpus. This functionality is relevant for multiple use cases where researchers perform a qualitative analysis on a corpus of resources such as digital/digitized documents and images. Our main contribution is the development of a lightweight collaborative and adaptive VRE that enables researchers to perform these tasks as well as to enable export of the created data and the content's sharing and reuse. Since the VRE is based on a flexible Open Source platform it can be tailored by the researchers towards their specific needs. Therefore this lightweight environment may serve as a starting point for further re-uses and re-configurations in unforeseen research settings and functionalities required in the future.

Funding

This work was supported by the German Research Foundation (DFG) in the domain of 'Scientific Library Services and Information Systems' (LIS). The funded project is entitled: 'Entwicklung einer Virtuellen Forschungsumgebung für die Historische Bildungsforschung mit Semantischer Wiki-Technologie – Semantic MediaWiki for Collaborative Corpora Analysis' [INST 367/5-1, INST 5580/1-1]

References

Barad, K. (2003). Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society* 28(3): 801-831.

Bauer, M. W., and B. Aarts (2000). Constructing a research corpus. In M.W. Bauer and G. Gaskell (eds), *Qualitative researching with text, image and sound: a practical handbook*. Thousand Oaks: Sage, pp. 19-37.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing* 8(4): 243-257.

Carusi, A., and T. Reimer (2010). *Virtual Research Environment. Collaborative Landscape Study. A JISC funded project.* http://www.jisc.ac.uk/media/documents/publications /vrelandscapereport.pdf (accessed 30 October 2011).

Corti, L., and A. Gregory (2011). CAQDAS Comparability. What about CAQDAS Data Exchange? *Forum: Qualitative Social Research* 12. http://www.qualitative-research.net/index.php/fqs/article/viewArticle/1634 . (accessed 30 October 2011).

Dunn, S. (2009). Dealing with the complexity Deluge – VREs in the arts and humanities. *Library Hi Tech* (27)2: 205-216.

Edwards, P. N., M. S. Mayernik, A. L. Batcheller, et al. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science* 415): 667-690.

Hey, T., and A. Trefethen (2003). The Data Deluge: An e-Science Perspective. In F. Berman, G. Fox, and T. Hey (eds.), *Grid Computing: Making the Global Infrastructure a Reality*. Chichester: John Wiley & Sons, pp. 809-824.

Huvila, I. (2008). Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management. *Archival Science* 8(1): 15-36.

Juola, P. (2008). Killer Applications in Digital Humanities. *Literary and Linguist Computing* 23: 73-83.

Neuroth, H., F. Jannidis, A. Rapp, and F. Lohmeier (2009). Virtuelle Forschungsumgebungen für e-Humanities. Maßnahmen zur optimalen Unterstützung von Forschungsprozessen in den Geisteswissenschaften. *Bibliothek. Forschung und Praxis* 2: 161-169.

Schindler, C., C. Veja, M. Rittberger, and D. Vrandečić (2011). How to teach digital library data to swim into research. *Proceedings of I-SEMANTICS 2011: 7th International Conference on Semantic Systems,* Sept. 7-9, 2011, Graz, Austria New York: ACM International Conference Proceedings Series (2011), 142-149

Smith, N., S. Hoffmann, and P. Rayson (2008). Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations. *Literary and Linguist Computing* 23: 163-180.

Notes

1. It should be mentioned that a similar field of tension is articulated for the sciences as 'science friction' (Edwards et al. 2011) by addressing the problems of different disciplines working on the same phenomena and trying to interoperate.
2. This term refers to the concept of 'intra-action', which describes the interrelations and re-configurations of research apparatuses in respect of 'locally stabilized phenomena' (Barad 2003: 817).

3. This project is realized in co-operation between the German Institute for International Educational Research (DIPF), the Karlsruhe Institute of Technology (KIT), the Library for Research on Educational History (BBF) and educational researchers mainly of the Georg-August-University Göttingen. See http://www.dipf.de/en/projects/virtual-research-environment-for-research-in-the-history-of-education-smw-cora.
4. The lexica are available at http://bbf.dipf.de/digitale-bbf/scripta-paedagogica-online/digitalisierte-nachschlagewerke.
5. See http://www.mediawiki.org
6. The extension Semantic MediaWiki is described at http://www.semantic-mediawiki.org and offers further extensions at http://semantic-mediawiki.org/wiki/Help:Extensions.
7. Some interesting cases using MW or SMW as a VRE are http://www.docupedia.de/ , a reference work in the area of historical research, http://wiki.digitalclassicist.org and in the context of archives (Isto, 2008). Some further examples of using SMW are listed on the webpage http://smw.referata.com/wiki/Special:BrowseData/Sites?Data_type=Science.
8. A prototype of this SMW-based extension Semantic Web Browser is developed in collaboration with KIT – Benedikt Kämpgen, Anna Kantorovitch, and Denny Vrandečić – and is accessible at http://www.mediawiki.org/wiki/Extension:Semantic_Web_Browser.

Corpus Coranicum: A digital landscape for the study of the Qu'ran

Schnöpf, Markus
schnoepf@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

1. Introductiom

In 2007 begun the project Corpus Coranicum located at the Berlin-Brandenburg Academy of Science and Humanities with an estimated duration of twelve years. A goal of the project is a holistic documentation of the holy text. The project consists of different modules: Collection of early manuscripts, documentation of environmental texts to the Quran, documentation of alternate writings and finally a commentary on each sura of the Quran. In the last years the technological infrastructure has been set up and data was collected in a SQL database. The commentary of the was realised in XML and is stored in a XML-database. The website of the project thus combines SQL and XML in an integrated information system. Lately, a bibliography consisting of 8000 references was added to the system. Further investigations are directed more on a scientific dating, analysing the materiality of early written documents. They are scheduled for 2012/13 within a French-German joint research project. Another new module is a glossary of the and early Arabic literature. For overcoming troubles in the presentation of early Arabic script a special font has been developed: The Coranica.

The project Corpus Coranicum began at the Berlin-Brandenburg Academy of Science and Humanities in 2007 with a planned duration of 12 years. The project aims at both a holistic edition of the Quran and also an extensive commentary. In addition to the Al-Azhar Quran edition from 1923/1924, this project will provide the reader with early written testimonies as well as oral reading variants that are manifested in early islamic literature.

2. Manuscripta

The project sees the module Manuscripta Coranica as following the tradition of G. Bergsträßers planned Apparatus Criticus, which due to Bergsträßers early death could never be realised. Thus the project aims at a new approach to the text of the Quran. Bergsträßer collected the oldest Quran manuscripts

by travelling the Arabic world in the 1920s with a 35-mm camera.

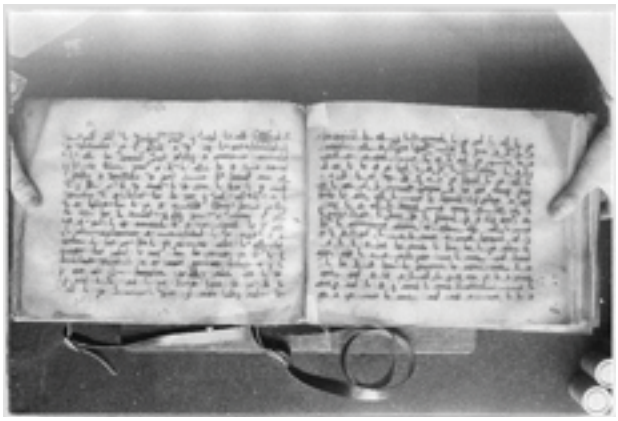


Figure 1: A page from the Kodex Meknes from the collection of Bergsträßer

These photographs remained lost until members of the Corpus Coranicum found the filmstrips in 2008 (Higgins 2008). Immediately we began to digitize the images in levels of grey. The project can therefore present digitizations from Quran manuscripts in the module Manuscripta Coranica that are lost today. A database containing metadata that includes the range of verses in one image has been developed and filled. The early manuscripts have a high degree of ambiguity, as vocals and diacritics were used irregularly and different from today. Thus these manuscripts document the continuity as well as the development of the Arabic writing system.

Figure 2: Same page in the module Manuscripta Coranica

The Quran was revealed to Mohammed from God via the angel Gabriel beginning in 610 (western dating). Until 632, the year of Mohammed's death, the divine revelations were mainly orally memorized and written on palm leaves. The Quran consists of 114 chapters that are called suras. Each sura consists of a different number of verses with a total number of 6326 verses. The length of the verses degrees from sura one to 114. The Corpus Coranicum aims at documenting the double string of the quranic textual history, as it was handwritten and orally (which was codified in later islamic sources) passed on next generations. So, beneath the module Manuscripta Coranica, which documents the handwritten tradition of text, the oral reading tradition of the Quran has to be documented as well.

3. Lectiones

This is the module Lectiones Coranicae, in which we present the different readings of the Quran. In this module only these variants that have phonetic effects, are collected. Each variant is assigned to one or more reader or the bequeathing person. A synoptical view on these variants looks like a score of the Quran.

Figure 3: Lectiones Coranicae

4. Commentary

The commentary on each sura contains different parts. First, texts from the environment of the Quran like the bible, Talmud and early Arabic inscriptions, papyri and others have to be mentioned. We call these texts intertexts. Here one encounters a high diversity of different writing systems, such as Greek, Syrian, Pahlawi and others. Sources are therefore found in many different languages. With this collection of heterogeneous texts connected to each verse of the Quran the user will be able to draw a map of traditions that were mentioned in the holy text.

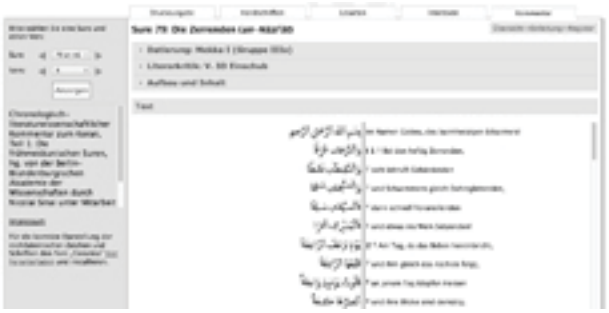


Figure 4: The module commentary on the Quran

The grammar in the suras is the theme of the second part of the commentary. A word can be explained through comparisons with other manifestations of the word in the Quran. In addition, the identification of Quranic loan words must be mentioned. In order to facilitate this work, a database Glossarium Coranicum has been developed. The formal structure of each sura is the subject of another part of the commentary. The literal appearance of oaths, revelations and

other structural phenomena are recorded in this section, thus documenting the formal history of Quranic speech. After having captured the lexical, grammatical and literal characteristics of the sura, it is possible to identify insertions, in so far as they can be proofed following the interpolation hypothesis. The interpretation is the last part of the commentary for each sura. On the one hand each text can be assembled in different ways, reflecting the chronology of the sura's formation. On the other hand the differences to the intertexts can be used to show the uniqueness of the sura.

5. The digital landscape

Coming to the digital environment of the project, we try to organize the information in different databases that provide the backbone of the Corpus Coranicum:

- Manuscripta Coranica
- Lectiones Coranicae
- Bibliographia Coranica
- Glossarium Coranicum
- Intertexts
- Quran itself

Figure 5: The module intertexts

The connection between these databases is the sura and the verse. It is thus possible to show all connected information for one verse: the manuscripts showing the verse, the intertexts of the verse and the loan words used in the verse. The commentary for each sura is stored in XML, following and extending the transcription rules of the Text Encoding Initiative. We therefore, have a heterogeneous digital information landscape, where different and yet separated information systems are bundled under one umbrella. Generally stated, the different cultures the project deals with are combined within this information system. The usage of different writing systems and languages are thus another challenge in the project. As available fonts were unable to represent ancient Arabic writings (the vocalisation points are orientated on modern Arabic writing systems) a special font had to be developed within the project, called Coranica, which we offer as well for download as open source.

Arabic digital philology is a young and emerging field in the digital humanities. Speech technologies like the Hopper of the Perseus project or the Buckwalter ensive arabic dictionary, remain unfulfilled. However, freshly published resources like the Lane dictionary can be used to broaden the perception of Arabic texts. In my presentation I will demonstrate the techniques and web-contents of the Corpus Coranicum, as the first results of the project goals are published in the first half of 2012. Thus the Quran will be presented for the first time as a digital reference system not only for the Quran itself, but as well integrating other sources from the early Arabic and old Ancient world. Thus, philology must make sense to the text itself (Pollock 2009).

References

Bergsträsser, G. (1993). *Nichtkanonische Koranlesarten im Muhtasab des Ibn-Ginnī / von G. Bergsträsser*. Egelsbach: Hänsel-Hohenhausen.

Higgins, A. (2008). The Lost Archive. *Wall Street Journal*. Available at: `http://online.wsj.com/article/SB120008793352784631.html` (accessed 25 March 2012).

Pollock, S. (2009). Future philology? The fate of a soft science in a hard world. *Critical Inquiry*. 35(4): 931-961.

Spitaler, A. (1935). *Die Verszählung des Koran nach islamischer Überlieferung*. München: Bayer. Akad. d. Wissenschaften.

Wansbrough, J. (2004). *Quranic Studies: Sources and Methods of Scriptural Interpretation*. London: Prometheus Books.

The MayaArch3D Project: A 3D GIS Web System for Querying Ancient Architecture and Landscapes

Schwerin, Jennifer von
jvonschw@unm.edu
University of New Mexico, USA

Richards-Rissetto, Heather
heathmrr@hotmail.com
University of New Mexico, USA

Agugiaro, Giorgio
agugiaro@fbk.eu
FBK Trento, Italy

Remondino, Fabio
remondino@fbk.eu
FBK Trento, Italy

Girardi, Gabrio
gabrio.girardi@graphitech.it
GraphiTech Trento, Italy

In this paper, we highlight the need in humanities research for 3D tools with enhanced analytical functionality and present an innovative 3D Web GIS system that we are developing – called QueryArch3D – for studies of ancient architecture and landscapes. We explore the role this tool is playing in the development of new methodologies and formulating and addressing research questions in humanities research. Specifically, we demonstrate how research and teaching on archaeology and art history can be dramatically assisted by a Virtual Research Environment that offers the ability to search and query segmented 3D models (both CAD and reality-based) that are linked to attribute data stored in a spatial database in the context of an online VR landscape – in this case, the eighth-century Maya kingdom of Copan, Honduras.

Modern sensor and computing technologies are changing the practice of art history and archaeology because they offer innovative ways to document, reconstruct, and research the ancient world (e.g., El-Hakim et al. 2008; Reindel & Wagner 2009). 3D digital models and virtual reality (VR) environments allow for remote viewing of objects, as well as multiple iterations of hypothetical reconstructions, and offer the sense of space and experience that researchers now desire (e.g., Barcelo et al. 2000). But, as has been pointed out, there is a common perception that while 3D models are good for education or data conservation, they are not useful for research because they are often deemed as purely illustrative and not useful for analysis (e.g., Frischer & Dakouri-Hild 2008). One promising opportunity offered by 3D models is to use them as visualization 'containers' for different kinds of information. Given the possibility to link their geometry to external data, 3D models can be analyzed, split into sub-components, and organized following proper rules. Powerful 3D visualization tools already exist, but often they implement no or only limited query functionalities for data retrieval, and very few of these are web-based. In contrast, Geographic Information Systems (GIS) include queries as standard functions and allow for complex spatial analyses – and therefore are well-suited to research (e.g., Bodenhamer et al. 2010; Conolly & Lake 2006).

Along these lines, archaeologists are using GIS to perform quantitative analyses, such as visibility, accessibility, and network studies, to explore the structure of ancient societies and relationships between anthropogenic and natural phenomena. GIS software, however, falls short when dealing with detailed and complex 3D data. It is comprised of 2.5D data, which are not ideally suited to approaches such as performance studies, phenomenology and aesthetics, the relationship of architecture and landscape, and archaeoastronomy that are becoming increasingly common in archaeology and art history. 3D models of architecture are simply more appropriate for these kinds of investigations, but most 3D digital models of architecture are single objects, removed from the context of their place in the landscape and are not linked to scientific data.

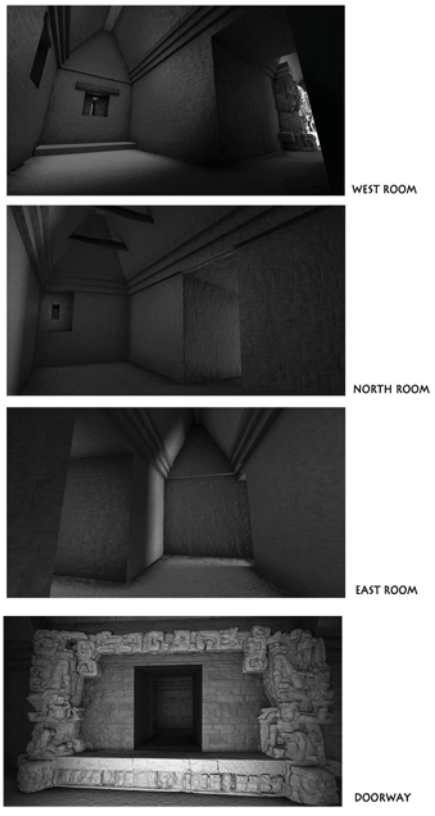
In 2009, the MayaArch3D Project (http://mayaarch3d.unm.edu) was begun to explore the possibilities of integrating GIS and 3D digital tools for humanities research. This interdisciplinary, international project was funded largely by two Digital Humanities Start-Up Grants from the National Endowment for the Humanities (USA) and brings together art historians, archaeologists, and cultural resource managers with experts in remote sensing, photogrammetry, 3D modeling, and virtual reality. The project was founded by Jennifer von Schwerin and Heather Richards-Rissetto at the University of New Mexico working with data from the UNESCO World Heritage site and ancient Maya city in Copan, Honduras, when they realized that their research goals would be better served if they could join their separate sets of data in a tool that linked GIS and 3D data in a virtual reality environment. Von Schwerin, an art historian, had turned to digital 3D tools in order to test her reconstructions of an eighth-century temple at

Copan (Figure 1). But other aspects of her research seek to analyze the temple within its larger urban context and the role that space plays in human experience – something for which GIS is well-suited. Richards-Rissetto, an archaeologist, created a GIS for Copan to study the visual and spatial relationships between built forms and natural landscape features, but soon realized that her research could benefit from a 3D perspective, as the 2D perspective of GIS maps limited her interpretations (Figure 2**).**

Working together, von Schwerin and Richards-Rissetto asked themselves: What are the real research possibilities for 3D models? How can we create an online resource for researchers of Maya architecture where they can compare and study geo-referenced 3D models and attribute data? How can we perform quantitative and qualitative comparisons with other Maya structures and analyze architecture in larger spatial and temporal contexts?

To address these questions, the MayaArch3D Project developed a new computing pipeline in order to build a prototype tool for an online, searchable repository – called *QueryArch3D* – that brings together GIS data, 3D models, and virtual environments for teaching and research on ancient architecture and landscapes. This pipeline uses PostgreSQL as spatial repository back-end and allows for the import and export of standard GIS formats, as well as 3D models as triangulated meshes via the common obj format. Like traditional databases, this tool can curate, query, and compare 2D digital objects (such as drawings, maps, diagrams, text, photographs, and videos). However, what is unique and technologically cutting-edge is that QueryArch3D enables users to 1) integrate and edit 2D and 3D data of *multiple resolutions*, 2) to perform attribute and *spatial* queries of archaeological data, and 3) to visualize, compare and analyze *3D* buildings and artifacts – all in a single *online*, navigable virtual reality landscape. Developed in 2010 in collaboration with Fabio Remondino and Giorgio Agugiaro at the Bruno Kessler Foundation (FBK) in Trento, Italy, and Gabrio Girardi at Graphitech, in Trento, Italy, the QueryArch3D tool stores both low resolution models, high resolution reality-based, hybrid models and all ancillary attribute data in an open source spatial database, and then makes them queryable via a virtual reality environment that runs on the Unity 3 game engine. Currently the contents of the digital resource include collections on ancient Copan (Figure 3). Overall, our preliminary research results indicate the QueryArch3D tool is poised to offer a new level of collaborative work in any field that works with 3D models, GIS, and large data sets, because it allows researchers to bring 2D and 3D datasets from separate projects into a single environment and work in real-time using the online query and analysis capabilities.

This paper presents this new tool and the results of the beta-testing that was carried out in fall 2011 with researchers, students, and educators in the humanities. Beta testers unanimously were enthused about the tool, particularly at having the ability to (1) navigate online through a virtual model of an ancient Maya city, (2) access higher resolution models of objects, (3) query the archaeological database via the model – and vice versa. Suggestions for improvement centered on decreasing the initial download time, improving the user interface (changes to navigation commands, adding a text search box), and adding a broader range of spatial queries. Because the project is still in its initial stages, the paper also summarizes current research problems, and the tasks to be solved in the project's next stage of development, in collaboration with the German Archaeological Institute and the University of Heidelberg. The paper concludes with a critical assessment of the possibilities that 3D Web GIS systems currently offer scholars and educators in terms of organizing, searching, and visualizing data, and identifying patterns over space and time.



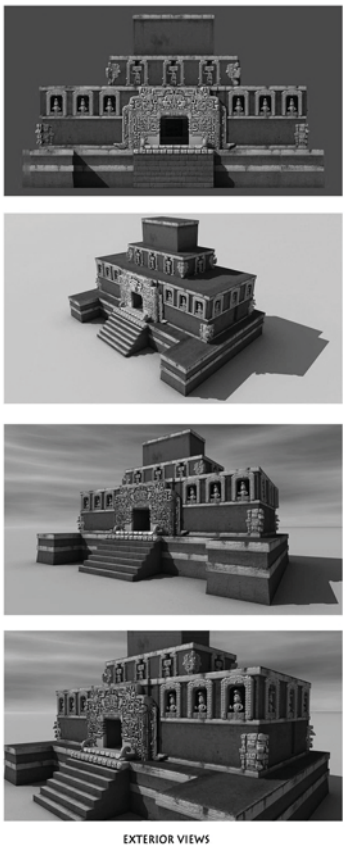


Figure 1a and 1b: Working hypothetical reconstruction and 3D model of Temple 22 created by the MayaArch3D Project (Graphic: Raul Maqueda, from von Schwerin et al. 2011)

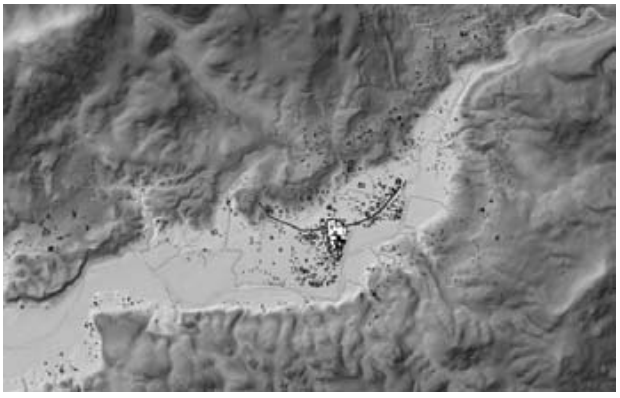


Figure 2: Copan GIS (Graphic: Heather Richards-Rissetto 2011)

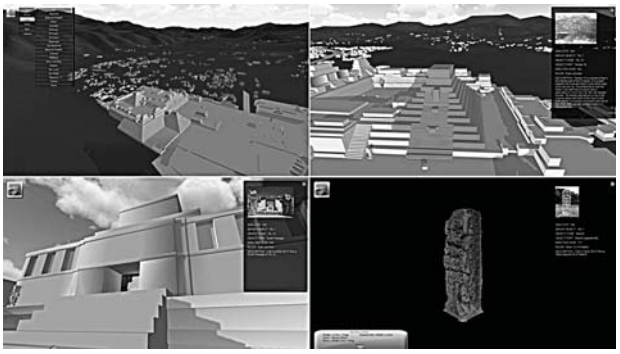


Figure 3: QueryArch3D tool showing different data visualization modes and levels of detail. Graphic: Giorgio Agugiaro

References

a) Selected MayaArch3D Project Publications:

Agugiaro G., F. Remondino, G. Girardi, J. von Schwerin, H. Richards-Rissetto, and R. de Amicis (2011). QueryArch3D: Querying and visualising three-dimensional archaeological models in a web-based interface. In *Geoinformatics*, Faculty of Civil Engineering, Czech Technical University, ISSN 1802-2669.

Remondino F., A. Gruen, J. von Schwerin, H. Eisenbeiss, A. Rizzi, S.Girardi, M. Sauerbier, and H. Richards-Rissetto (2009). Multi-sensor 3D Documentation of the Maya Site of Copan. *Proceedings of the 22nd CIPA Symposium, 11-15 October, 2009, Kyoto, Japan (2009).* `http://cipa.icomos.org/KYOTO.html`.

Richards-Rissetto, H. (2012). Studying Social Interaction at the Ancient Maya Site of Copán, Honduras: A Least Cost Approach to Configurational Analysis. In D. A. White and S. Surface-Evans (eds), *Least Cost Analysis of Sociocultural Landscapes: Archaeological Case Studies,* Accepted by University of Utah Press, January 2011.

von Schwerin, J. (2011). The Sacred Mountain in Social Context: Design, History and Symbolism of Temple 22 at Copán, Honduras. *Ancient Mesoamerica* 22(2): 271-300.

von Schwerin, J., H. Richards-Rissetto, F. Remondino, G. Agugiaro, M. Forte, and R. Maqueda (2011). *The MayaArch3D Project; Digital Technologies for Research in Maya Archaeology*. Final Performance Report and White Paper for NEH Digital Humanities Level II Start- Up Grant.

b) Related Publications:

Barcelo, J. A., M. Forte, and D. Sanders D. (2000). *Virtual Reality in Archaeology*. BAR International Series 843, Oxford.

Bodenhamer, D. J., J. Corrigan, and T. M. Harris (2010). *The spatial humanities: GIS and the future of humanities scholarship*. Bloomington : Indiana UP.

Conolly, J., and M- Lake (2006). *Geographical Information Systems in archaeology*. Cambridge: Cambridge UP.

Dear M., J. Ketchum, S. Luria, and D. Richardson (2011). *GeoHumanities Art, History, Text at the Edge of Place*. New York: Routledge.

El-Hakim, S., J.-A. Beraldin, F. Remondino, M. Picard, L. Cournoyer, and M. Baltsavias (2008). Using terrestrial laser scanning and digital images for the 3D modelling of the Erechteion, Acropolis of Athens. *Proceedings of the DMACH Conference, Digital Media and its Applications in Cultural Heritage*. Amman, pp. 3-16.

Frischer, B., and A. Dakouri-Hild (2008). *Beyond illustration: 2d and 3d digital technologies as tools for discover in archaeology*. Oxford: Archaeopress.

Llobera, M. (2006). What you see is what you get?: Visualscapes, visual genesis and hierarchy. In P. Daly and T. Evans (eds.), *Digital Archaeology: Bridging Method and Theory*. New York, London: Routledge, Taylor and Francis, pp. 148-167.

Manferdini, A. M., F. Remondino, S. Baldissini, M. Gaiani, and B. Benedetti (2008). 3D modeling and semantic classification of archaeological finds for management and visualization in 3D archaeological databases. *Proceedings of 14th Int. Conference on Virtual Systems and MultiMedia* (VSMM 2008), pp. 221-228

Reindel, M., and G. A. Wagner, eds. (2009). New Technologies for Archaeology: Multidisciplinary Investigations in Palpa and Nasca, Peru. Natural Science in Archaeology. Heidelberg/Berlin.

Remondino, F., S. El-Hakim, S. Girardi, A. Rizzi, B. Benedetti, and L. Gonzo (2009). 3D Virtual reconstruction and visualization of complex architectures-The 3D-ARCH project. *IAPRS&SIS* Vol.38(5/W1)

Robertson, E., J. Seibert, D. Fernandez, and M. Zender, eds. (2006). *Space and Spatial Analysis in Archaeology*. University of New Mexico and University of Calgary Press: Albuquerque and Calgary.

Multi-dimensional audio-visual technology: Evidence from the endangered language documentation

Sharma, Narayan P.
narayan.sharma57@gmail.com
SOAS, University of London, UK

The audio-visual documentation of endangered language provides multi-dimensional data which can effect outcomes of research such as understanding the complexities of linguistic diversity and transmitting memory and knowledge. Audio-visual documentation also allows for observing the unique cultural heritage embedded in the culture, and also for understanding the environmental, multilingual and socio-cultural context of the speech community.

This study investigates how the digital audio-video capturing can record several types of information simultaneously and how it has become a successful method for not only documenting the language but also observing the cultural uniqueness.

Linguists such as Daniel L. Everett (2001) prefers monolingual field method for studying the languages, but others like George Cowan (1975) claims that the monolingual approach is a serious barrier to maintaining the good-will of the community members. A common obstacle that linguists face needs to understand the lingua franca, which is essential for practical and administrative purposes of studying a language (Newman & Ratliff 2001). As most of the speakers in this study are multilingual, the language of elicitation and the medium of interview are also influenced by this factor and multilingualism is becoming a social phenomenon among Pumas. The medium of conducting interview and the elicitation for this study are in a 'lingua franca', Nepali language. Interviewees in the study speak Nepali as a second or first language. The findings reveal that monolingual research is only contextual and it is not sufficient to study endangered languages.

In 2010 field research was conducted in Nepal to study the 'Morpho-syntax of Puma, which is an endangered language. During eight months in the field site, a total of 45 informants both male and female were selected and interviewed from all Puma speaking Village Development Committees. A total hour of audio and video recording was 08:35:03 in the natural settings. The recorded corpus is divided into 83 different sessions and all of them

have been transcribed in Puma, mostly by native speakers and then translated into Nepali, and English by the researcher. The software/ digital tools and major recording equipments used in this work are as follows:

(a) Software

- Transcriber (for transcription from audio and/or video data) .trs file
- Toolbox (for glossing/segmenting, adding sound files into texts and creating a dictionary) .txt file
- ELAN (for linking audio-visual data with texts; video annotation) .eaf
- Handbrake to convert original video format (MTS/ AVCHD) into MP4
- Kigo Video converter to convert original video format (MTS/ AVCHD) into MP4

(b) Recording equipments

- Zoom-4Hn Recorder for audio recording
- Canon Legaria HF S200 for video recording
- Audio-Technica AT8022 stereo microphone
- Audio Technica Pro 70 Lavalier
- Rod NTG2 microphone
- Canon Powershot Digital Camera

WAV, MP4 and text files of the data have been created and can be archived and disseminated for use. In the case of many endangered languages, such materials are essential for developing pedagogical materials in mother tongue education and language revitalization. Csató and Nathan (2003) emphasize that whereas linguists working in the community earlier contributed through traditional academic publishing, linguists in recent years create multimedia resources to highlight their relationships with the community.

This paper attempts to explore fundamental questions of direct elicitation.

Based on complex verbal morphology and paradigms in Puma, without direct elicitation, it is not possible to gather authentic data on verbal paradigms, deixis, compound verbs and among others. Identifying morphemes and their multiple usages are challenging in the Puma language. Using questionnaires and basic sentences are still the most widely and commonly used method for linguistic research. Dixon (2010) states that the only way to understand the grammatical structure of a language is to analyse recorded texts but not elicited sentences. The findings in the study pose that elicited sentences are quite essential and significant especially for eliciting Puma verb paradigms and figuring out deictic categories. We can never neglect direct elicitation, and the established questionnaires and basic sentences (i.e Dahls' 1985 questionnaire) can be recorded in Puma.

The recording of speech from different genres such as narratives, conversations, myths and stories, life histories, songs, poetry and daily accounts is the core of linguistic fieldwork. Both recording of text and direct elicitation are essential tools that have their own variety of uses. Since none of them are sufficient for all linguistic analysis, both of them should not be overlooked (Mithun 2001). The text collection, its transcription and translation, and glossing are important tools for understanding and learning languages.

Metadata is the additional but essential information needed for archiving and managing language documentation and can be referred to as meta-documentation, which includes information about the identity of stakeholders, the attitudes of language contributors and the methodology (Austin 2010; Nathan 2010). The linked-metadata should be kept and all files name such as text file, WAV file, MP4 file, ELAN file and transcriber file must be the same otherwise there would be great problem not only for future researchers but also for archiving.

Linguists cannot avoid cultural aspects of speakers while studying their language, as culture is an integral part of the language community. The study shows evidence that certain unique genres of ritual speech, like the *hopmacham*, a chant praising the forces of creation during marriages (CPDP 2004-2008), are still well-known and in practice in the Puma community while it has already disappeared in other neighboring ethnic groups. In this study *hopmacham* was initially recorded like other texts but its cultural heritage and uniqueness was unknown until it was glossed and translated into Nepali. The recordings allowed for deeper analysis of culture in the community.

Integrating linguistic and cultural resources is a crucial endeavour for documentary linguistics. The audio-visual data include many different types of information which can record several types of information simultaneously such as linguistic and cultural diversity, ethnography of speaking traditions of both everyday and ritual language, socio-cultural aspects of the community such as rites and rituals. It also captures some anthropological considerations such as religion, settlement, clans, emigration, and the socio-linguistic aspects, and most importantly videos and photographs of informants and local features. Thus, the analysis of the study makes clear that digital audio-video method is the best method for documenting endangered language and integrating linguistic and cultural data for archive, dissemination, and mobilization in the humanities.

References

Austin, P. K. (2010). Current Issues in Language Documentation. In P. Austin (ed.), *Language Documentation and Description*. London: SOAS, vol. 7, pp. 12-33.

Chintang and Puma Documentation Project (CPDP) (2004-2008). http://www.uni-%09leipzig.de/~ff/cpdp/ .

Cowan, G. (1975). The monolingual approach to studying Amuzgo. In A. Healey (ed.), *Language Learner's Field Guide*. Ukarumpa, Papua New Guinea: SIL, pp. 272-276.

Csató, É. Á., and D. Nathan (2003). Multimedia and documentation of endangered languages. In P. Austin (ed.), *Language Documentation and Description*. London: SOAS, vol. 1, pp. 73-84.

Dahl, Ö. (1985). *Tense and Aspect Systems*. Oxford: Blackwell.

Dixon, R. M. W. (2010). *Basic Linguistic Theory*. Oxford: Oxford UP.

Everett, D. L. (2001). Monolingual field research. In P. Newman and M. Ratlif (eds.), *Linguistic Fieldwork*, Cambridge: Cambridge UP, pp. 166-88.

Mithun, M. (2001). Who shapes the record: the speaker and the linguist. In P. Newman and M. Ratlif (eds.), *Linguistic Fieldwork*, Cambridge: Cambridge UP, pp. 34-54.

Nathan, D. (2010). Language documentation and archiving: from disk space to my space. In P. Austin (ed.), *Language Documentation and Description*. London: SOAS, vol. 7.

Newman, P., and M. Ratliff (2001). *Linguistic Fieldwork*. Cambridge: Cambridge UP.

Contours of the Past: Computationally Exploring Civil Rights Histories

Shaw, Ryan Benjamin
ryanshaw@unc.edu
University of North Carolina at Chapel Hill, USA

Historical insight is only achieved when the contours of our view of the past are as clear as possible.

(Frank Ankersmit)

Each history provides a unique view of the past by picking out a path through the field of possible events (Veyne 1984: 36). And because histories respond to earlier histories, their paths intersect. At these intersections lie events which become taken for granted as 'sites to be visited' on 'a pre-arranged itinerary marking out the recommended scenic route (and the beaten track) from one major point of interest to the next' (Rigney 1990: 37). Some of these itineraries are then reified as periods.

If we accept these itineraries too readily, we risk missing opportunities for new insights. Historical insight results not just from the accumulation of new facts about the past, but from the development of new views upon the facts we already know. The more stories we have about some piece of the past, 'the deeper our insight into it will be [...] because only the presence of other stories enables us to draw the contours and to recognize the specificity of the view of the past presented in each one' (Ankersmit 1983: 219).

In 1983, Frank Ankersmit theorized a procedure for drawing these contours by clustering stories that contain overlapping sets of propositions. By taking a set of stories about a person, place, or period, transforming each story into a list of propositions, and aligning them in such a way that the lists could be compared, Ankersmit posited that 'certain classificatory patterns [would] automatically appear' (Ankersmit, 1983: 145).

The *Contours of the Past* project is testing Ankersmit's theory by building tools for comparing and contrasting histories of the civil rights movement. The historiography of the civil rights movement exemplifies how periods coalesce around dominant stories. The first generation of civil rights scholars 'conceived of the civil rights struggle as primarily a political movement that secured legislative and judicial triumphs' (Lawson 1991). The

chronology of this movement began with *Brown v. Board of Education of Topeka* in 1954 and ended with the Voting Rights Act of 1965. Due in part to intense media coverage, by the end of the twentieth century this chronology had become a well-beaten path not only among scholars but also in popular understanding.

In the past decade, a new generation of scholars has sought to broaden this itinerary, telling 'the story of a 'long civil rights movement' that took root in the liberal and radical milieu of the late 1930s, was intimately tied to the 'rise and fall of the New Deal Order,' accelerated during World War II, [and] stretched far beyond the South' (Hall 2005). The concept of the Long Civil Rights Movement (LCRM) is about more than simply replacing one story with a new one. By widening the scope of the civil rights movement, the LCRM opens space for a greater diversity of stories, enabling greater insight but making it more difficult to grasp the movement as whole. We aim to show how computational analysis might complement the work of scholars evolving new periodizations and perspectives, by providing tools for comprehending narrative patterns in a more complex whole.

We are applying cutting-edge text analysis techniques to two corpuses: eighty-seven books made available by the UNC Press through their Publishing the LCRM project, and transcripts of approximately 350 interviews conducted by the Southern Oral History Program as part of their LCRM initiative.[1] The specific techniques we are applying are *event parsing* and *narrative clustering*.

Event parsing involves identifying sentences that communicate some event, for example a strike, a protest, a bombing, or a legislative act. Specifically, it involves identifying *frames*, conceptual structures that describe particular events along with their participants and settings.[2] Typically event parsing has been used for detection and tracking of topics in news media, automated question answering, text summarization, and the production of structured data from unstructured text.

We apply event parsing to different ends. We do not treat events as 'facts' that can be consumed independently of the histories from which they were 'extracted.' Historical knowledge inheres within the narrative form, so the 'extraction' metaphor is a poor fit for tools that aim to enhance access to historical knowledge (Shaw 2010). Instead, we use the results of event parsing as features for comparing stories though narrative clustering: treating stories as 'bags of events' and applying statistical techniques for grouping together similar 'bags.'

We envision two forms of comparison. First, along the lines of Ankersmit's original proposal, we can highlight the specificity of a given history by showing which events it recounts that are not recounted by similar histories. In conjunction with information about the history such as when, where, and by whom it was produced, such comparisons could provide a powerful means of assessing the depth and scope of a given collection of histories.

Second, given a group of histories recounting overlapping sets of events, we can compare their 'speeds.' Roland Barthes noted that an event that takes up dozens of pages in one history may be covered by just one in another – a phenomenon he called *acceleration*.[3] Because we propose to bring together histories that recount the same events, and because those events are parsed from the actual texts of those histories, we can potentially compare how different histories speed up and slow down time.

Contours of the Past is comparable to recent projects applying topic modeling to historical sources such as diaries and newspaper archives (Blevins 2011; Nelson). Topic modeling is a statistical technique for discovering independent topics in some collection of documents, where a 'topic' is defined as a group of words that tend to appear in the same documents (Blei, Ng and Jordan 2003). Topic modeling has found favor among digital humanists for quickly identifying themes in a large collection of documents without having to specify some set of themes ahead of time.

However, topic modeling as it is usually applied directly to the words used in historical documents will mainly reflect patterns of word usage. This is exactly what is desired for investigations of diction and style, but for identifying common patterns of historical narration, this may not be the best approach. An oral history containing a firsthand account may use language that is very different from that found in a scholarly monograph, even if both sources are describing the 'same' event. Yet two sentences may evoke the same semantic frame even if they do not have any words in common. Thus we can potentially find common patterns of historical narration across different kinds of narrative source by applying clustering techniques, not at the surface level of language (the specific words used), but at the level of frame-semantic representation.

References

Agirre, E., and P. G. Edmonds (2006). *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer `http://www.wsdbook.org/`.

Ankersmit, F. R. (1983). *Narrative Logic: A Semantic Analysis of the Historian's Language*. The Hague: M. Nijhoff.

Barthes, R. (1981). The Discourse of History. Translated by S. Bann. In E. S. Shaffer (ed.), *Comparative Criticism: A Yearbook* 3. Cambridge: Cambridge UP.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation *The Journal of Machine Learning Research* 3: 993 http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

Blevins, C. (2011). Topic Modeling Historical Sources: Analyzing the Diary of Martha Ballard. *Proceedings of Digital Humanities 2011*. Stanford, CA, 19-22 June 2011. http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-173.xml.

Gildea, D., and D. Jurafsky (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics* 28(3): 245 http://dx.doi.org/10.1162/089120102760275983.

Hall, J. D. (2005). The Long Civil Rights Movement and the Political Uses of the Past. *Journal of American History* 91(4): 1235 http://www.jstor.org/stable/10.2307/3660172.

Lawson, S. F. (1991). Freedom Then, Freedom Now: The Historiography of the Civil Rights Movement. *The American Historical Review* 96(2): 456 http://www.jstor.org/stable/2163219.

Nelson, R. K. Mining the *Dispatch*. Digital Scholarship Lab, University of Richmond http://dsl.richmond.edu/dispatch/.

Rigney, A. (1990). *The Rhetoric of Historical Representation: Three Narrative Histories of the French Revolution*. Cambridge: Cambridge UP.

Shaw, R. (2010). From Facts to Judgments: Theorizing History for Information Science. *Bulletin of the American Society for Information Science and Technology* 36(2): 13 http://dx.doi.org/10.1002/bult.2010.1720360207.

Veyne, P. (1984). *Writing History: Essay on Epistemology*. Translated by M. Moore-Rinvolucri. Middletown, Connecticut: Wesleyan UP.

Notes

1. See the Publishing the LCRM project at http://https://lcrm.lib.unc.edu/blog/ https://lcrm.lib.unc.edu/blog/ and the LCRM Initiative at http://www.sohp.org/content/our_research/the_long_civil_rights_movement_initiative/
2. More specifically, it involves two tasks: first identifying the frames invoked by particular words in a text (a form of *word sense disambiguation*) and then assigning entities to the various roles in each frame, known as semantic role labeling. For an overview of the former, see Agirre and Edmonds (2006). For the latter, see Gildea & Jurafsky (2002).
3. Barthes (1981: 9) hypothesized that 'the nearer we are to the historian's own time [...] the slower the history becomes.'

Notes from the Collaboratory: An Informal Study of an Academic DH Lab in Transition

Siemens, Lynne
siemensl@uvic.ca
University of Victoria, Canada

Siemens, Raymond
siemens@uvic.ca
University of Victoria, Canada

1. Introduction

Digital Humanities (DH) as a discipline is highly collaborative and as such requires a departure from typical humanities work patterns with its focus on the lone scholar (Siemens 2009; Siemens et al. 2011). In particular, DHers must develop new skills and knowledge and negotiate new ways of conducting research and organizing people, financial resources, space and other factors (Scholars' Lab 2011). In this regard, the Sciences and Applied Sciences can provide guidance and suggest several models upon which DH can draw to achieve project objectives.

The Sciences use two primary models of organisation for conducting research. The first model is the typical faculty-directed laboratory where the lead researcher creates a research vision and hires staff to conduct the research. This individual is ultimately responsible for research results, including publications, patents, discoveries, and use of funds, to their department, faculty, university, academic discipline, funders and others. Collaboration occurs at the level of research vision execution, rather than the development of that vision (Haynes et al. 2006). Ultimately, the lead researcher is evaluated by traditional academic measures, including the number of articles, books, patents, lines of software code, and level of grant funding (Cantwell 2011). Many books and manuals are available to guide the new scientist on now to set up and manage a lab, people and resources (Cohen & Cohen 2005; Howard Hughes Medical Institute & Burroughs Wellcome Fund 2006).

Alternatively, Big Sciences are using collaboratories to create 'centres without walls' where researchers work together to create common access to data and instruments, such as supercomputers, telescopes, and global history databases within an umbrella research area, but do not necessarily work on the same research projects. These collaborations are most successful when everyone contributes to the same degree that they draw upon the common instruments and data (de Moor & van Zanden 2008; Wulf 1993). Various studies have been conducted to evaluate the productivity of these different models (Finholt 2003; Haynes et al. 2006).

Alongside these well-evolved and understood models of collaboration – that is those that follow patterns of the single-researcher directed 'collaborat-ory' and that of the multiple-researcher directed co-laboratory (Glasner 1996: 111) – opportunities exist to create hybrid organisations and apply within the Sciences and beyond to the Humanities and Social Sciences. But how do these models work within DH? Can they be applied directly to DH or does this community need collaborative research models that differ from those typically found in the Sciences?

This paper will examine the experiences of one academic DH lab as it adapts these models, develops collaborative structures, meets its research objectives and produces outputs that can be evaluated by traditional academic measures. It will conclude with recommendations for other Digital Humanists who are setting up their own labs or are collaborating on the creation of new DH centres.

2. A Case Study

The case study focuses on the evolution a DH lab as it experimented with different organisational forms to support collaboration over a seven year period of growth, and three distinct stages of organisation and operation. It initially operated as a single faculty-directed lab with graduate research assistants, postdoctoral fellows, programmers, and others as staff in its first four years. At that point, those who worked in the lab underwent a research visioning exercise, the results of which suggested that a hybridised 'collaborator' model would be the most appropriate organisational form, one that would allow a fuller and more active participation of faculty, staff and contractors in visioning, consensual decision-making, and leadership. Finally, the lab returned to a more standard model, with a plan for more structured growth in the future.

Over all these transitions, the impacts of these various internal operational models were tracked against standard academic and funding agency benchmarks that included measurable research resource intake, provision of teaching and service, and research outputs, such as books, articles, conference papers, and other types of production, more DH-oriented research outputs in the form of tools and prototype development, and further issues – some measureable, such as documented internal and external complaints, and some less so. As other labs have found, not meeting these targets can lead to reduced grant funding and lack of research space

(Cantwell 2011) and, so, meeting such targets is directly tied to the ongoing operation of a lab.

Particularly notable among the results discussed is that the lab's experimentation with a hybridised collaboratory model unintentionally introduced a diffused accountability structure; many internal mechanisms and accountability structures that allowed earlier successes were inadvertently removed as part of the process of hybridising the two models. As a result, with the checks and balances that ensured consideration of basic collaboration principles removed, lab productivity by all measures fell drastically and almost-immediately, and team functionality became severely handicapped. Formal accountability structures themselves became viewed as contrary to collaborative principles and, without those accountability structures in place, it became increasingly difficult for team members to follow through on the development and implementation of project plans; acts of planning became less meaningful in this context as well. Without a set structure that ensured that those contributing resources and those who were accountable to outside structures (Department, the Faculty, university and funding agencies) enacted a leadership role, few sanctions for non-performance existed (de Moor et al. 2008) and 'free-riding' became the chief mode of interaction among staff and contractors. Further, an informal work culture developed that was contrary to local and university policies as well as, in some cases, funding agency guidelines; internal lab groups formed, striated, clique-ified, and could not work together. This situation contributed to a dramatic increase in personnel complaints.

In the end, the experiment could be judged to be a failure as measured by many common benchmarks. Despite attempts to fuse what the lab felt to be the most desirable features of the single-researcher directed 'collaborat-ory' and that of the multiple-researcher directed 'co-laboratory', the lab became neither. Upon the feedback from the lab's advisors, the research lead is moving towards [1] a return to the model of the single-researcher directed 'collaborat-ory,' with [2] promise to implement plans to become, more formally, a multiple-researcher directed 'co-laboratory' – all the while retaining the strong internal operational structures and hierarchies that were in place before the experiment.

3. Conclusion

This case study contributes to ongoing discussions in DH about appropriate organization forms and accountability mechanisms (de Moor et al. 2008; Dormans & Kok 2010; Warwick 2004) , roles, contributions, and status within projects (#alt-academy, 2011) , and the nature of collaboration (Siemens 2009) . It provides several lessons for consideration. First, Science models for collaboration and the creation of well-functioning labs can be applied within DH (Bland & Ruffin IV 1992; Haynes et al. 2006) . Second, some hierarchy is necessary, particularly when one individual is responsible for sourcing the money and other resources which sustains the lab. This lead researcher is accountable for these funds and 'must have ultimate authority' (Lawrence 2006; Rogers-Dilon 2005: 449). Having said this, consensus and active participation in discussions around ways to achieve a lab's research direction can create an exciting and intellectually productive environment and working relationships, as measured by staff satisfaction and academic metrics. These roles must be backed up with clear objectives, tasks, timelines and consequences for non-performance. Lastly, this case study serves as a cautionary tale for new DH researchers who will be tasked with setting up their own labs, managing staff and research, creating partnerships, while remaining accountable to stakeholders inside and outside the university.

References

#alt-academy (2011). #alt-academcy: A media commons project `http://mediacommons.futureofthebook.org/alt-ac/` [November 1, 2011].

Bland, C. J., and M. T. Ruffin IV (1992). Characteristics of a productive research environment: Literature review. *Academic Medicin,* 67(6): 385-397.

Cantwell, B. (2011). Academic in-sourcing: International postdoctoral employment and new modes of academic production. *Journal of Higher Education Policy & Management* 33(2): 101-114.

Cohen, C. M., and S. L. Cohen (2005). *Lab dynamics: Management skills for scientists*. Cold Spring Harbor, New York: Cold Spring Habor Laboratory Press.

de Moor, T., and J. L. van Zanden (2008). Do ut des (i give so that you give back): Collaboratories as a new method for scholarly communication and cooperation for global history. *Historical Methods* 41(2): 67-80.

Dormans, S., and J. Kok (2010). An alternative approach to large historical databases. *Historical Methods* 43(3), 97-107.

Finholt, T. A. (2003). Collaboratories as a new form of scientific organization. *Economics of Innovation and New Technology* 12(1): 5-25.

Glasner, P. (1996). From community to 'collaboratory'? The human genome project and the

changing culture of science. *Science and Public Policy* 23(2): 109-116.

Haynes, L., S. Pfeffer, J. M. Boss, et al. (2006). Lab management: Insights for the new investigator. *Nature Immunology* 7(9), 895-897.

Howard Hughes Medical Institute & Burroughs Wellcome Fund (2006). *Making the right moves: A practical guide to scientific management for postdocs and new faculty* (2nd ed.). Research Triangle Park, North Carolina: Howard Hughes Medical Institute, Burroughs Wellcome Fund.

Lawrence, K. A. (2006). Walking the tightrope: The balancing acts of a large e-research project. *Computer Supported Cooperative Work: The Journal of Collaborative Computing* 15(4): 385-411.

Rogers-Dilon, R. H. (2005). Hierarchical qualitative research teams: Refining the methodology. *Qualitative Research* 5(4): 437-454.

Scholars' Lab (2011). The praxis program at the scholars' lab `http://praxis.scholarslab.org/` [September 12, 2011].

Siemens, L. (2009). It's a team if you use 'Reply all': An exploration of research teams in digital humanities environments. *Literary & Linguistic Computing* 24(2): 225-233.

Siemens, L., R. Cunningham, W. Duff, et al. (2011). A tale of two cities: Implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities. *Literary & Linguistic Computing* 26(3): 335-348.

Warwick, C. (2004). 'No such thing as humanities computing?' An analytical history of digital resource creation and computing in the humanities. *Paper presented at the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing*. tapor.humanities.mcmaster.ca/html/Nosuchthing_1.pdf [May 25, 2006].

Wulf, W. (1993). *National collaboratories: Applying information technology for scientific research*. Washington, DC: National Academies Press.

XML-Print: an Ergonomic Typesetting System for Complex Text Structures

Sievers, Martin
sievers@uni-trier.de
Trier Center for Digital Humanities (Kompetenzzentrum), Germany

Burch, Thomas
burch@uni-trier.de
Trier Center for Digital Humanities (Kompetenzzentrum), Germany

Küster, Marc W.
kuester@fh-worms.de
Worms University of Applied Sciences, Germany

Moulin, Claudine
moulin@uni-trier.de
Trier Center for Digital Humanities (Kompetenzzentrum), Germany

Rapp, Andrea
rapp@linglit.tu-darmstadt.de
TU Darmstadt, Germany

Schwarz, Roland
schwarzr@uni-trier.de
Worms University of Applied Sciences, Germany

Gan, Yu
gany2d01@uni-trier.de
Trier Center for Digital Humanities (Kompetenzzentrum), Germany

1. Introduction

The software *XML-Print* is used to typeset arbitrary XML files. The joint research project is funded by the German Research Foundation (DFG) for the period starting 1 March 2010 to 28 February 2014. It is a key component of the TextGridLab, which has been under continuous development since 2008.[1]

XML-Print supports users in formatting their semantically annotated data rulebased and in outputting a high-quality PDF document. Based on existing standards like XSL (XSLT and XSL-FO), Unicode and OpenType, a modern graphical user interface offers different kinds of layout options. Those are then processed by a newly developed typesetting engine using the functional programming language F#.

2. Scholarly Typesetting

XML-Print targets in particular specific challenges from the typesetting of scholarly texts such as critical editions, multilingual synoptic editions or scholarly dictionaries.

Taking the example of critical editions as a key product of philological research, they regular attempt to fully describing alternative existing witnesses of the text (old manuscripts, early prints, etc.) and fully covering the genesis of the text (authorial or scribal additions, deletions, comments, etc.).

Critical editions in print use specific layout conventions. The editor notes witnesses in one apparatus and often adds their explanations in a critical apparatus which constitutes a fourth flow on the page. More complex layouts can include more critical apparatus and/or annotations in margins. Other challenges beyond the 'normal' typesetting tasks include in particular synoptic prints, marginalia with complex references, unusual or non-standard characters and symbols, including those not (yet) present in the Unicode specifications etc.

3. State of the Art

Many humanities scholars have started to encode their work as XML files. The TEI guidelines have been a major contribution to that. However, when it comes to the stage of publication often problems arise: What tool should be used for that? There are of course well-known solutions: Open-source typesetting engines like TeX (Knuth, 1986) and TUSTEP (Ott, 1979) can be 'programmed' to convert large amounts of XML data somehow into their own markup language and then into an output format, typically PDF. However both batch systems need an experienced and skilled user in order to get high-quality results. Very often individual and highly specialized extensions have to be added, in particular in view of the challenges of non-standard typesetting requirements such as multiple apparatus. In addition these systems suffer from the problem of content mixed up with formatting information.

Apart from batch systems many proprietary 'WYSIWYG' software came up through the years, e.g. the Critical Edition Typesetter (CET) and the Classical Text Editor (CTE). These are, however, mostly isolated solutions with data – once input and annotated – 'getting lost' in a proprietary format.

From the commercial 'desktop publishing' sector Adobe InDesign is a reasonable choice using its own XML format as a medium step, but it lacks in the implementation of scientific printing. The same is valid for office suites which are not meant for high-quality typesetting of scientific content.

More about the requirements of scientific typesetting and existing solutions can be found in Küster and Ludwig (2008).

4. Functionality

The following use case is a typical example for the publication of an XML file using *XML-Print*: A scholar has a critical edition encoded in XML and wants to present a first printed version to their colleagues. In order to do so they have to perform the following steps:

- Identify different structures and think about corresponding ways of formatting, e.g. for chapters, sections, footnotes, paragraphs etc.
- Create a *format* for each of the identified structures. Set sizes, spaces, text decorations and other attributes as needed. Modify the *standard format* if necessary.
- Use the *mapping* dialog to select pairs of XML elements and *formats*.
- Start the integrated typesetting engine to get a PDF document.
- If necessary, make changes to the XML source and/or alter formats and mappings.

Figure 1 illustrates the overall data flow.

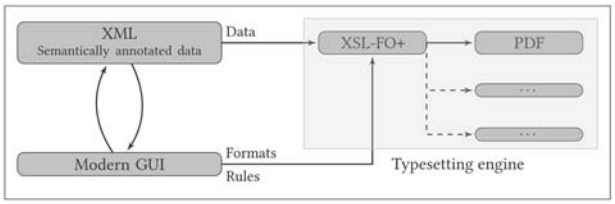


Figure 1: Data flow in XML-Print: The user decides on the formatting of an annotated XML text with the help of a GUI (front-end). Those information are then merged into a XSL-FO+ file and converted into an internal intermediate format for typesetting (back-end). At last the renderer outputs a PDF file (or any other supported format)

4.1. Graphical User Interface (GUI)

A user-friendly and modern GUI is essential for the acceptance by the 'community' (cf. Nielsen 1993; Warwick et al. 2011). To make integration into the toolbox of the TextGridLab easier, the front-end has been implemented as an Eclipse plug-in using the Rich Client Platform (RCP) technology. It offers ways to select different layout options which are then incorporated into a dynamic XSLT stylesheet to generate an *XSL-FO+* file. This format is based on the standard XSL-FO, but continuously extended where needed.[2]

Figure 2 shows the user's view on the *XML-Print* GUI.

Figure 2: The user's view on the GUI of XML-Print is tripartite: On the left the XML tree can be expanded while the formats are listed on the right. The middle part links those both together by mappings

Formats

Each *format* is a set of *XSL-FO+* attributes. They determine the concrete rendering later on by the typesetting engine. The attributes are divided into categories inspired by the XSL-FO terminology (block, inline, footnote etc.) to easily navigate through. This idea was derived from similar software familiar to many users. This way the user can set all appropriate values as needed.

New *formats* can be created and existing ones may be copied, edited or deleted. The complete set of *formats* can then be saved and transferred to another scholar or can be used for any other *XML-Print* project.

Mappings

Each XML element can be *mapped* to a *format*. Different attribute values can be considered as well, e.g. to distinguish <note type="footnoteA"> from <note type="footnoteB">. The selection of corresponding structures can be easily done directly on the XML tree. Alternatively an arbitrary XPath expression can be used.

Mappings can not only be created, edited and deleted, but also deactivated for testing purposes. A rank order is also established to allow the scholar to use overlapping *mappings*, i.e. XML structures belonging to more than one *mapping*.

4.2. Typesetting Engine

The disadvantages of existing approaches stated in section 3 led to the decision for a completely new typesetting engine. Established algorithms like the one by Knuth and Plass (1981) are combined with new ideas (cf. Brüggemann-Klein et al. 2003) and open-source libraries like Hunspell and iText are incorporated by a functional approach using the programming language F# as part of the .NET framework and its implementation both under Microsoft .NET and its open-source, cross-platform counterpart Mono.

Functional programming coexists with imperative programming for quite some time[3], but was often considered to be too 'academic'. This has, however, changed over the last years not least because of some important advantages arising from the different approach on modern multicore systems that are available everywhere: no side effects, better evaluation techniques and strong abilities for modularization and parallelism (see e.g. Hughes 1990).

Of course, all essential features of recent typesetting engines are to be implemented, i.e. good line and page breaking, support of OpenType fonts, floating objects, tables, lists etc. For the underlying access to OpenType fonts and PDF generation we leverage existing cross-platform, open-source libraries such as iTeXtSharp. The main focus, however, has been on those algorithms which substantially improve the quality compared to existing programs. Thus we have already implemented an interface to typeset an arbitrary number of footnotes and apparatus. Other important and requested features are multi-column layout, especially parallel text, and marginals.

The resulting program can be run stand-alone (batch mode) or integrated in the Eclipse plug-in. A Web service will be offered as well.

Figure 3 shows an example of an *XML-Print* output.

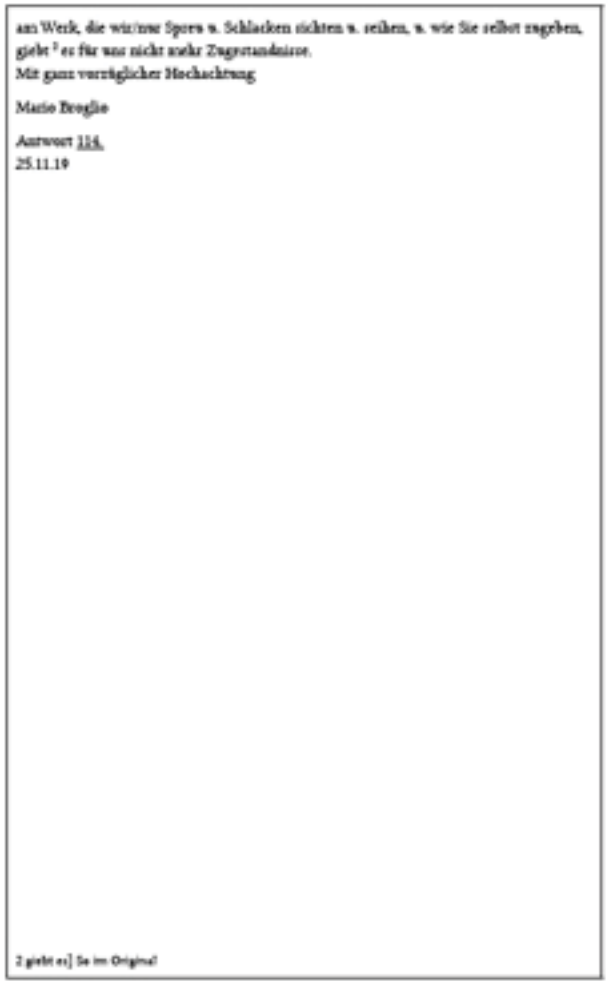
VALORI PLASTICI
RASSEGNA D'ARTE
DIREZIONE E AMMINISTRATIONE
10, Via Ciro Menotti
ROMA (49)

Rom, XI 1919

Sehr geehrter Herr Schwitters.
Haben Sie vielen Dank für Ihre Zeilen, die nur aus einem Künstlerglauben, der Schöpferglauben ist, entspringen können und Zeugnis davon ablegen, daß auf dem Gebiet geistigen Wirkens und Waltens jedes ferne Erbeben der Stringenz verstanden wird.
Die Fotografien Ihrer Arbeiten habe ich mit viel Freude und Interesse betrachtet u. will gerne einiges davon veröffentlichen.
Mein Plan ist es, jeder Nummer von Valori Plastici ihren Typus zu verleihen, und ich hatte in's Auge gefaßt, im Lauf der nächsten Jahres eine Nummer den hervorragendsten deutschen Künstlern zu widmen a). Während der Winters komme ich nach Deutschland u. hoffe, daß sich die Gelegenheit bieten wird uns kennen zu lernen und näheres zu vereinbaren. –
Was den Text anbelangt, ist Valori Plastici ausschließlich Abhandlungen über die bildende Kunst gewidmet: die Kubistennummer hat eine Ausnahme gebildet, die sich nicht wiederholen darf b). –
Ich halte es für meine Pflicht, Ihrem liebenswürdigen Entgegenkommen zu entsprechen und Ihnen einige Fotos unserer Hauptmitarbeiter für den „Zweemann" c) zu senden, u. bitte Sie, mir näheres über die Remuneration der Wiedergabe mitzuteilen, um Ihnen die Zustimmung meiner Freunde erwirken zu können. Ich würde Sie dann bitten, mit Erlaubnis der Verlages „Valori Plastici" unter die Reproduktionen setzen zu lassen. Mit Holzschnitten, u.s.w. ist es hierzulande schlecht bestellt u. was die schriftlichen Beiträge anbelangt, muß ich gestehen, daß ich selbst viel zu kämpfen habe; meine Pultfächer sind übervoll, u. doch finde ich nichts, was ich Ihnen mit klassisch-italienischem Künstlergewissen empfehlen könte [1]. – Wir sind wenige Freunde

1 empfehlen könte] So im Original

a) deutschen Künstlern zu widmen] In Valori Plastici II, 7-8 (1920) erschien ein Artikel über George Grosz und Paul Klee (S. 88f.), jedoch keine Reproduktionen von Werken dt. Künstler, sondern ausschließlich solche des italien. Künstlers Giorgio di Chirico. KS wird mit keinem Wort erwähnt und es erschienen auch in späteren Ausgaben weder Artikel von oder über ihn noch Abbildungen seiner Werke in Valori Plastici.
b) die Kubistennummer...wiederholen darf] Valori Plastici I, 2-3, (Feb./März 1919) war dem franz. Kubismus in bildender Kunst und Poesie gewidmet, obwohl sich die Zeitschrift eigentlich ausschließlich mit den bildenden Künsten auseinandersetzte.
c) Fotos unserer Hauptmitarbeiter... „Zweemann"] Im => Zweemann wurden keine Bilder der „Hauptmitarbeiter" von Valori Plastici – italienische Künstler des Magischen Realismus – veröffentlicht. Allerdings schrieb Theodor Däubler (Triest/Österreich 1876-1934 St. Blasien/Schwarzwald), einer der Hauptautoren von Valori Plastici, auch für den Zweemann.



am Werk, die wir/nur Spuren u. Schlacken sichten u. sehen, u. wie Sie selbst zugeben, giebt [2] es für uns nicht mehr Zugeständnisse.
Mit ganz vorzüglicher Hochachtung

Mario Broglio

Antwort 114.
25.11.19

2 giebt es] So im Original

Figure 3: Resulting output of the typesetting engine: Apart from paragraph handling and the use of different fonts the two different types of 'notes' (<note type="footnoteA"> <note type="footnoteB">)are both set as footnotes. They are sorted and numbered automatically according to their corresponding format

5. Further Enhancement

XML-Print has been released as an alpha version and is tested by researchers of different projects, e.g. for a volume of selected letters of Kurt Schwitters (cf. project's web site, 2011 and figure 3). The typesetting engine as well as the GUI are continuously enhanced, taking into account new user requirements. By extending the attributes for formatting and user-friendly ways, e.g. to generate registers, tables or the type area, *XML-Print* will offer more and more features to researchers. The development process continues bipartite: On the one hand the support of all standard XSL-FO elements and basic typesetting functionality has to be completed and maybe adapted to a new specification or other output formats, on the other hand advanced algorithms for more complex problems will be developed and implemented.

Within the next months the following tasks will be targeted:

- Continuous work on typographic requirements based on 'real life examples'
- Integration of the open-source tool xindy for indexes
- Implementation of more output formats, especially PDF variants like PDF-X and PDF-A
- Development of a preview mode to allow faster response to minor changes

With the integration into TextGridLab another large group of new users will get in touch with the software and help to improve it.

References

Brüggemann-Klein, A., R. Klein, and S. Wohlfeil (2003). On the Pagination of Complex Documents. In R. Klein, H. Six and L. Wegner (eds.), *Computer Science in Perspective*. Berlin and Heidelberg: Springer, pp. 49-68.

Hughes, J. (1990). Why Functional Programming Matters. In D. Turner D. (ed.), *Research Topics in Functional Programming*. Reading: Addison-Wesley, pp. 17-42. `http://www.cs.kent.ac.uk/people/staff/dat/miranda/whyfp90.pdf` (accessed 13 March 2012). First published 1989: 10.1093/comjnl/32.2.98

Knuth, D. E. (1986). *The TeXbook*. Reading: Addison-Wesley.

Knuth, D. E., and M. F. Plass (1981). Breaking paragraphs into lines, *Journal Software: Practice and Experience* 1111: 1119–1184.

Küster, M. W., and C. Ludwig (2007). *Publishing*. `http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid-R1_4_Publishing.pdf` (accessed 13 March 2012).

Nielsen, J. (1993). *Usability Engineering*. London: Academic Press.

Ott, W. (1979). A Text Processing System for the Preparation of Critical Editions. *Computers and the Humanities* 13(1): 29–35.

Warwick, C., M. Terras, P. Huntington, N. Pappa, and I. Galina (2006). *The LAIRAH Project. Log Analysis of Digital Resources in the Arts and Humanities*. `http://www.ucl.ac.uk/infostudies/claire-warwick/publications/LAIRAHreport.pdf` (accessed 13 March 2012).

Wie Kritik zur Kunst wird. Project web site, `http://www.avl.uni-wuppertal.de/forschung/projekte/wie-kritik-zu-kunst-wird.html` (accessed 13 March 2012).

Notes

1. Version 1.0 of the TextGridLab was released in July 2011.
2. There are many attributes intended for XSL-FO 2.0 (e.g. fo:marginalia) and other structures not even specified there. We had to add attributes, e.g. for the placement of footnote apparatus. See the requirements and the latest version of the XSL-FO working draft for more examples.
3. Lisp was the first functional-flavored language in the late 1950s. Modern examples apart from F# are Scala, Haskell and XSLT.

Federated Digital Archives and Disaster Recovery: The Role of the Digital Humanities in Post-earthquake Christchurch

Smithies, James Dakin
james.smithies@canterbury.ac.nz
University of Canterbury, New Zealand

The Canterbury region, in the South Island of New Zealand, experienced two major earthquakes during 2010 and 2011. On September 4, 2010 a magnitude 7.1 quake struck at 4.35 am, causing widespread damage and two serious injuries. Significant aftershock sequences followed. On February 22 2011 a 6.3 magnitude quake hit at 12.51 pm. This earthquake caused severe damage and resulted in the loss of 181 lives, making it the second worst natural disaster in New Zealand history. Like the first, the second quake has been followed by thousands of aftershocks, including two significant earthquakes on June 13th 2011.

The University of Canterbury CEISMIC Canterbury Earthquake Digital Archive draws on the example of the Centre for History and New Media's (CHNM) September 11 Archive, which was used to collect digital artefacts after the bombing of the World Trade Centre buildings in 2001, but has gone significantly further than this project in its development as a federated digital archive. The new University of Canterbury Digital Humanities Programme – initiated to build the archive – has gathered together a Consortium of major national organizations to contribute content to a federated archive based on principles of openness and collaboration derived directly from the international digital humanities community. Two primary archive 'nodes' have been built by the Ministry of Culture and Heritage ('QuakeStories') and the University of Canterbury ('QuakeStudies') to collect content from the public and researchers respectively, and a 'front window (www.ceismic.org.nz) has been provided by the University of Canterbury to bond the Consortium, raise funds, and provide a platform for future aggregated search functions, which will be powered by New Zealand's bespoke cultural heritage schema maintained by Digital NZ. Other nodes in the federation include The Museum of New Zealand Te Papa Tongarewa, the National Library, Christchurch City Libraries, NZ On Screen, and the Canterbury Museum. The aim is to create a permanent record of digital objects for both present and future generations. To this end the technical requirements for QuakeStudies have been reviewed by the National Digital Heritage Archive with a view to ingesting significant subsets of content (if not creating a complete dark archive) for long-term preservation. Significant attention has been paid during the design process to multi-cultural and multi-lingual requirements, to ensure content from a broad range of New Zealand communities can be ingested and researched. Future development aims to create a bi-lingual interface in English and Māori.

The story behind the UC CEISMIC Canterbury Earthquake Digital Archive goes somewhat further than other similar digital archives. Not only is it being used to initiate New Zealand's first Digital Humanities programme, but it hopes to fulfil an important role in the cultural and intellectual recovery of the Canterbury region following the earthquakes of 2010 and 2011. New Zealand is a country with significant levels of technology uptake, and the vast majority of content produced following the earthquakes was created in digital form. As the central focus of the recovery efforts was, of necessity, focussed on the physical and spiritual well-being of the Canterbury public, it was quite possible that large amounts of valuable content would be lost to future generations. This altered somewhat after the initial phase of critical response ended, only to be replaced with new issues. Various institutions began gathering digital content into their separate repositories, but no co-ordinated approach was taken, creating a situation where disparate 'nodes' of content might be stored with little possibility of sharing and reuse. It was becoming possible that, although terabytes of content would be captured, future generations of citizens and researchers would need to go to myriad different archives, each with their own metadata standards, in order to get a complete picture of events. Aside from the obvious inconvenience of this, such a situation would seriously constrain the possibility of sophisticated downstream data analysis and content reuse.

The digital humanities ethos of sharing and open collaboration has had a significant positive effect in this context. Consistent recourse to the digital humanities' message of collaboration has fostered a culture of trust that has in turn allowed an extremely broad Consortium to be initiated. Although there is little chance that the resulting federation will be technically seamless, this has allowed potential conflicts of interest to be put aside and technical discussions to start at a relatively early stage in proceedings, significantly enhancing the chances of developing a highly functional distributed archive. Additionally, the digital humanities' emphasis on open communication and community engagement has fostered a healthy culture across the federation,

which has contributed significantly to the success of the project. This is represented most forcefully in the use of not only crowd-sourcing techniques, but a mobile recording studio fitted out with video and audio equipment, that has been taken to the suburbs of Christchurch to record public reaction to the earthquakes. This pro-active approach, coupled with robust attention to project structure, governance and human ethics, has created not only a digital archive, but a community of friends and partners, and a vibrant new digital humanities programme.

The project is also unusual for a digital humanities project in it becoming a flagship project for the broader university. Although the project and research teams are predominantly from the arts and humanities, close collaboration is also occurring with computer scientists, health researchers, social scientists and economists. As with the interest from New Zealand's national heritage agencies, digital humanities principles of collaboration and sharing, combined with well-considered metadata ontologies and system architecture, has prompted the project to occupy a central position in the post-earthquake recovery landscape. More than just an IT project, the CEISMIC Canterbury Earthquake Digital Archive is providing local, national and international public and researchers with a forum for discussion, organization and collaboration as well as a heritage asset in itself.

This paper will outline the project and present a model that will hopefully allow our approach to be reproduced in similar post-disaster recovery situations. Key to this model is the conscious use of digital humanities methodologies such as crowd-sourcing, community building and attention to open metadata ontologies and open access principles to create a robust and functional federated archive system. The model has several benefits, including the ability to develop a 'distributed nodal network' of archives and repositories independently, thus reducing the need for centralisation that would encumber development, but it requires a long-term vision and a strong governance framework to ensure the federation holds together and organizations feel comfortable sharing content. Similarly, while it offers excellent potential for teaching and research across the humanities as a whole, the relatively advanced nature of the project provides limited opportunity to involve students in system development. Instead, the project has created internships that will see students working as 'curators' on the research node in the federation, uploading content and taking responsibility for metadata quality and the integrity of manual procedures.

References

Adams, W. R. (2009). Archiving Digital Materials: An Overview of the Issues. *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve* 19: 325-35.

Berry, D. (2011). The Computational Turn: Thinking About the Digital Humanities. *Culture Machine* 12.

Cloonan, M. (2007). The Moral Imperative to Preserve. *Library Trends* 55(3): 746-55.

Conway, P. (2010). Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *Library Quarterly* 80(1): 61-79.

Day, M. (2008). Toward Distributed Infrastructures for Digital Preservation: The Roles of Collaboration and Trust. *The International Journal of Digital Curatin* 3(1): 15-28.

Gourley, D., and P. Viterbo (2010). A sustainable repository infrastructure for digital humanities: the DHO experience *Digital Heritage,* pp. 473-81.

Granger, S. (2002). Digital Preservation and Deep Infrastructure. *D-Lib Magazine* 8(2). `http://www.dlib.org/dlib/february02/granger/02granger.html` (Accessed 28 February 2012).

Bellardo, T. H. (2008). Mass digitization: implications for preserving the scholarly record. *Library Resources & Technical Services* 52(1): 18-26.

Lehmann, K. (1996) Making the Transitory Permanent: The Intellectual Heritage in a Digitized World of Knowledge. *Daedalus* 125(4): 307-29.

Martin, J., and Coleman, C. (2002). Change The Metaphor: The Archive as an Ecosystem. *Journal of Electronic Publishing* 7(3).

Neri, M. (2001). Putting the Cart Before the Horse: Understanding the Pros and Cons of Digital Preservation. *Library & Archival Security* 17: 59-64.

Scheinfeldt, T. Stuff Digital Humanists Like: Defining Digital Humanities by Its Values. *Found History*, n.d. `http://www.foundhistory.org/2010/12/02/stuff-digital-humanists-like/` (Accessed 28 February 2012).

Modeling Medieval Handwriting: A New Approach to Digital Palaeography

Stokes, Peter
peter.stokes@kcl.ac.uk
King's College, London, UK

As many hundreds of thousands of medieval manuscripts are now being digitised, with many millions of pages becoming available, the question of how to find specialised content in this material is becoming increasingly urgent. In this paper I present a new conceptual model for the description and therefore retrieval of features of handwriting in Western medieval script. Digital Humanities requires first a theoretical model which outlines all of the features of a given domain and the relationships between them (McCarty 2004), and this is the focus of the present paper. However, the implications of this work are very much wider: just as the TEI has lead to 'a new data description language that substantially improves our ability to describe textual features' (Renear 2004: 235), so the formal model of handwriting presented here sharpens and could even resolve long-standing problems in palaeographers' own terminology and practice.[1]

1. The Problem

Common research questions in palaeography include finding examples of a particular way of writing a given letter, or scribes who constructed letters in particular ways. For example, it has been asserted that a specific form of **t** was used only by English scribes in the ninth century (Dumville 1984: 249-250), and that 'Square' features in Anglo-Caroline minuscule of the late tenth century is distinctive of Canterbury (Bishop 1972: xxii). These assertions have generally been made with little supporting evidence, in part because the corpus of surviving manuscripts is too large to support systematic analysis, but one should now be able to overcome this by producing an online resource for searching and viewing scribal practices and examples of script. This is an objective of the DigiPal project,[2] which will provide an online catalogue of about 1,200 samples of English handwriting from the eleventh century, along with annotated digital images of about half of these. However, this objective is made very difficult by several factors. Image retrieval is very well studied, with approaches ranging from Content Based Image Retrieval (CBIR) such as that used by Google Images, to tagging with complex taxonomies like IconClass, as well as intermediate combinations of the two such as that provided by font identification software.[3] However, CBIR cannot easily be tuned to scholars' criteria if these do not match the machine's perception of 'content'. Similarly, systems like IconClass are very large, very intimidating, and require specialised training to use effectively. IconClass also depends on a shared vocabulary and this is famously absent from palaeography: the *Comité international de paléographie latine* was established in 1953 partly for this purpose but it has not produced even a draft document some fifty-eight years later.[4] This can partly be overcome by thesauri which map equivalent terms from different practices, but an even more fundamental lack is that of a rigorous conceptual model for handwriting itself. The textual community has been debating 'what text is' for some time, partly because of the TEI (Pierazzo & Stokes 2011: 399-401), but no discussion like this has yet arisen in palaeography (though see Stokes 2010: 1228-1230). With neither a clear model nor a common vocabulary, it is hardly surprising that the few attempts at digital catalogues to date have been unsuccessful.[5]

2. The Model

To illustrate the difficulties, consider the following list of possible research questions:[6]

- Show me a list of all manuscripts containing the 'Insular' form of a (Fig. 1a).
- Show me images of all occurrences of the 'Caroline' form of s written by scribes who otherwise wrote Insular minuscule (Fig. 1b).
- Tell me which scribe(s) habitually wrote the 'rounded' form of a in combination with a hooked top-stroke of t (Fig. 1c).
- Give me images of all letters with forked ascenders (Fig. 1d).

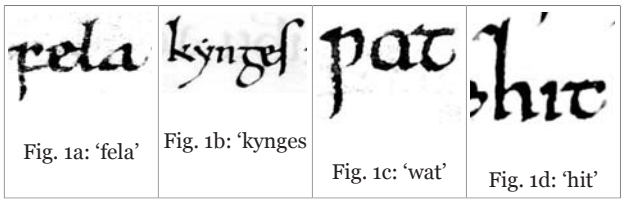


Fig. 1a: 'fela' | Fig. 1b: 'kynges' | Fig. 1c: 'wat' | Fig. 1d: 'hit'

Figure 1: Examples of Medieval Handwriting

Although these questions may look similar, they have a range of important differences which are largely concealed by the ambiguous word 'letter'. The challenge is therefore how to model handwriting in sufficient detail to represent these nuances, considering not only a scribe's general practice but

also its specific instances. It requires first a clear distinction between *script* ('the model which the scribe has in mind's eye when he writes'), and *(scribal) hand* ('what he actually puts down on the page': Parkes 1969: xxvi). The word 'letter' is also ambiguous, and so further terms are useful here. Starting with the most concrete, a *graph* is an instance of the letter as written on the page. In contrast, an *idiograph* is 'the way (or one of the ways) in which a given writer habitually writes' a given letter, and an *allograph* is a recognised way of writing the given letter, as shared by a group rather than being distinctive of any given individual (Davis 2007: 254-255; cf. also Emiliano 2011).[7] A *grapheme* is a letter-entity as an idealized, abstract, discrete, unit of a writing system. Also useful is Unicode's *character* to refer to the intermediate level between grapheme (which by definition has no physical form) and allograph. Thus the grapheme <a> has (at least) two characters, 'capital' **A** and 'small' **a**. The second of these has many allographs, one of which is Insular **a**, this is normally found in the script known as Insular cursive minuscule; in Figure 1a and 1c we have two distinct allographs of **a**; and so on. Grapheme and character are both 'emic', whereas allograph, idiograph and graph are all 'etic'.

Two further terms are required to describe specific details of letters. *Component* refers to the basic parts which make up characters, such as ascenders, descenders, serifs and so on; components may themselves have further components. *Features* are descriptive labels which may apply to one or more components: thus a descender may be straight or curved, long or short, and so on. Finally, we also need general stylistic features which may not refer to any specific component, such as the thickness and angle of the pen.

From these terms, a formal model has been developed which captures these entities and the relationships between them; this is represented in a simplified UML diagram in Figure 2 below.

According to this model, a grapheme has one or more characters, each character can be represented by an arbitrary number of allographs which in turn are represented by any number of idiographs and those in turn by graphs; a scribal hand comprises a set of graphs, scribal practice a set of idiographs, and so on. Each character, allograph, idiograph and graph can be made up of components, every component has at least one feature, and so on.

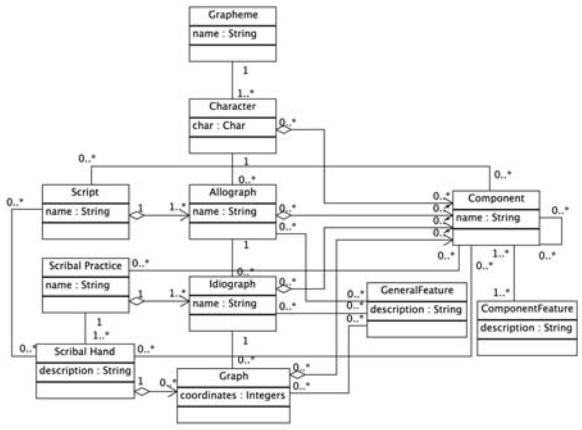


Figure 2: A Simplified UML Diagram showing the conceptual model of medieval handwriting

3. Significance and Usage

This model allows us to phrase palaeographical research questions much more precisely. It also allows structured descriptions of both habitual and specific writing-practices, thereby allowing complex queries. For example, we can specify:

1. A set of possible **characters** and the **components** that each character must have.
2. A **script**, namely a set of allographs which are normally written together, including their **components** and **features**.
3. A **scribal hand**, namely the set of idiographs which a scribe normally uses.
4. A set of **graphs**, for example those in an image of a particular manuscript page.

This model is also significant for capturing not only forms of letters but also specific stylistic features which apply across letters: not only particular forms of **b**, for example, but also the type of wedges on ascenders of any letter, or indeed wedges on ascenders of all letters *except* **b**.

An important further advantage is that if one level is fully defined then many details of the levels below that can be inferred automatically and so only exceptions need be recorded. For example:

1. The character **b** has an ascender and a bowl.
2. Insular minuscule script normally features wedges on ascenders.
3. It follows from (1) and (2) that allographs of **b** in Insular minuscule will normally have wedges on ascenders.

This allows rapid data capture despite the highly detailed model – a crucial point in a context where many thousands of graphs and idiographs will need to be described in order to be worthwhile. One

limitation of the model is that it allows one simply to record the presence of a feature, rather than (for example) numerical values to indicate 'fuzzy' cases where a feature is only partially present. The model could easily be adapted to accommodate this, but the increased time in data-entry would be prohibitive. Instead, in the implementation used for DigiPal, many features have n-ary distinctions instead: for example, ascenders can be wedged or clubbed or barbed or deeply-split or without decoration, and so on (see further Stokes 2011, Part v).

This paper will present the model and also demonstrate its use in practice through an annotation tool and search interface which are being developed for the DigiPal project. Reference will also be made to alternative uses which are being planned in the near future, including (we hope) application to Hebrew and Cuneiform script. The paper will also reflect on the process as an example of Digital Humanities bringing rigour to an existing discipline. McCarty has argued that modeling and the computer's 'demand for complete consistency' leads directly 'to the epistemological question *of how we know what we know*' (2004). However, this question is itself fundamental to palaeography, where one of the biggest challenges has been to make explicit the apparently subjective judgments of experts. As Derolez wrote of these, 'their great deficiency … lies in the difficulty of putting them into words', and, more generally, he asked 'whether morphological features [i.e. the shapes of letters] can be described in an unambiguous way' (2003: 7). If a model such as this can achieve what fifty-eight years of the *Comité* could not then that would surely be a significant achievement in a field which has often viewed quantitative approaches and the Digital Humanities with significant mistrust (Derolez 2003: 7-9; Stokes 2009).

References

Bishop, T. A. M. (1971). *English Caroline Minuscule*. Oxford: Clarendon Press.

Davis, T. (2007). The Practice of Handwriting Identification. *The Library*, 7th series, 8: 251-276.

Derolez, A. (2003). *The Palaeography of Gothic Manuscript Books*. Cambridge: Cambridge UP.

Dumville, D. N. (1983). Motes and Beams. *Peritia* 2: 248-256.

Emiliano, E. (forthcoming 2011). Issues in the Typographic Representation of Medieval Primary Sources. In Y. Kawaguchi and M. Minegishi (eds.), *Corpus Analysis and Diachronic Linguistics*. Amsterdam: John Benjamins.

McCarty, W. (2004). Modeling: A Study in Words and Meanings. In S. Schreibman, R. Siemens and J. Unsworth, *A Companion to Digital Humanities*. Oxford: Blackwell. 254-270.

Parkes, M. B. (1969). *English Cursive Bookhands 1250-1500*. Oxford: Oxford UP.

Pierazzo, E., and P. A. Stokes (2010). Putting the Text Back into Context. In F. Fischer et al. (eds.), *Kodikologie und Paläographie im Digitalen Zeitalter*. Norderstedt: Books on Demand, vol. 2, pp. 397-430 http://kups.ub.uni-koeln.de/4360/.

Renear, A. (2004). Text Encoding. In S. Schreibman, R. Siemens and J. Unsworth, *A Compa nion to Digital Humanities*. Oxford: Blackwell, pp. 218-239.

Stokes, P. A. (2007/08). Palaeography and Image Processing. *Digital Medievalist* 3 http://www.digitalmedievalist.org/journal/3/stokes/.

Stokes, P. A. (2009). Computer-Aided Palaeography: Present and Future. In M. Rehbein et al. (eds.), *Kodikologie und Paläographie im Digitalen Zeitalter*. Norderstedt: Books on Demand, pp. 309-338 http://kups.ub.uni-koeln.de/volltexte/2009/2978/pdf/KPDZ_I_Stokes.pdf.

Stokes, P. A. (2010). Scripts. In A. Classen (ed.), *Handbook of Medieval Studies*. Berlin: de Gruyter, vol. 3, pp. 1217-1233.

Stokes, P. A. (2011). Describing Handwriting (Parts i–). London: King's College http://digipal.eu/blogs/blog/describing-handwriting-part-i/.

Stokes, P. A. (forthcoming). Palaeography and the "Virtual library". In B. Nelson and M. Terras (eds.), *Digitizing Medieval and Early Modern Material Culture*. Arizona: Center for Medieval and Renaissance Studies.

Unicode Consortium (2011). *Glossary of Unicode Terms* http://unicode.org/glossary/.

Notes

1. For initials expression of the ideas in this paper see the DigiPal blog (Stokes 2011).
2. http://digipal.eu
3. http://www.iconclass.nl; http://www.identifont.com/; http://new.myfonts.com/WhatTheFont
4. http://www.palaeographia.org/cipl/ciplGen.htm. See further Derolez 2003: 6-9, and Stokes 2010: 1229-1230.
5. See the Palaeographic Catalogue at http://www.arts.manchester.ac.uk/mancass/C11database/letter_catalogue.php , and that by Stokes (2007/8: §§24-26). The shortcomings of both are discussed by Stokes (forthcoming).
6. These emerged from a workshop for the DigiPal project at King's College London, 6 September 2011.

7. Idiograph should not be confused with ideograph, a 'symbol that primarily denotes an idea or concept' (Unicode 2011).

A Digital Geography of Hispanic Baroque Art

Suárez, Juan-Luis
jsuarez@uwo.ca
The University of Western Ontario, Canada

Sancho-Caparrini, Fernando
fsanchocaparrini@gmail.com
Universidad de Sevilla, Spain

In *Toward a Geography of Art*, Thomas DaCosta Kaufmann, stated that his research would 'investigate how notions of place, of the geographical, have been inflected into writing about change through time as it has been and is still discussed in art history' (DaCosta 2004). He goes back to some of the ideas of this book in his contribution to the multi-volume Catalogue of the 2010-2011 international exhibition *Painting of the Kingdoms*. There he insists on the fact that political geography and artistic geography do not coincide as countries, viceroyalties, native areas, and notions of center and periphery superpose one another in different research works and cataloguing efforts. Da Costa Kaufmann also emphasizes the need of a theory of diffusion that help explain the movements of creators, paintings and features from territory to territory and the effects that this transfers have in the spatial organization of art that experts carry out.

Here, we present the results of a multi-disciplinary collaboration in Digital Humanities, Computer Science and Art History that focuses in proposing a digital geography of Hispanic Baroque art. By digital geography we imply the various possible organizations of the place of art by digital means in a manner that connects various types of data about authors and art-works with different notions of space. This digital geography of art also takes advantage of recent advances in data mining and visualization to offer multiple views of the space of Hispanic Baroque art, as related to geography, movement through territories, transfers over time and cultural borders, clusters of artistic centers (as opposed to centers and peripheries), and movements of works from their places of origin due to contemporary practices of collection by museums and private collectors.

The results shed light on the different ways in which social practices – from creation to circulation to collection – affect the spatial organization of art beyond political territories. The paper also shows how culture – defined as information that affects humans' behavior and represented here by the case

of Hispanic baroque paintings – organizes different real and symbolic 'places' in different times. We argue that the study of large-scale cultural systems such as the Hispanic Baroque is better suited by a combination of tools and concepts that deal with the complex and evolving nature of the system and can study it through multi-scale techniques that reduce that complexity to a minimum, offering new ways of arranging the space in which that system unfolded over time. Finally, we argue that this methodology can be extended to other projects in Digital Humanities.

Over the last few years, we have collected an online BaroqueArt (http://baroqueart.cultureplex.ca/) database of more than 12,000 paintings and more than 1,500 creators associated with the territories of the Hispanic Monarchy from the 16th to the beginning of the 19th centuries. The database also contains around 400 series, 200 schools and 2500 geographical locations[1]. On top of the data stored under a traditional entity-relation model, we implemented a system of annotations that would allow to work on the objects stored in the database and that would provide enough flexibility to describe all aspects of any artwork, defining a hierarchy in a structure similar to an ontology. From a set of more than 200 descriptors we carried out a manual semantic annotation of all artworks (with an average of 5.85 descriptors/work and peaks of 14 per work).

To analyze the resulting dataset we represent it as a graph in which artworks are nodes and that relations among them are established as a function of the descriptors shared by the works (for example, if an artwork is described through 7 descriptors and another work is also using the same set of descriptors, then we say that these artworks are connected through a weighted link with weight 7). We limited our experiment to the period 1550-1850 and divided the global graph into 12 sub-graphs in order to study the temporal evolution, each of them covering a period of 25 years (see Fig.1).

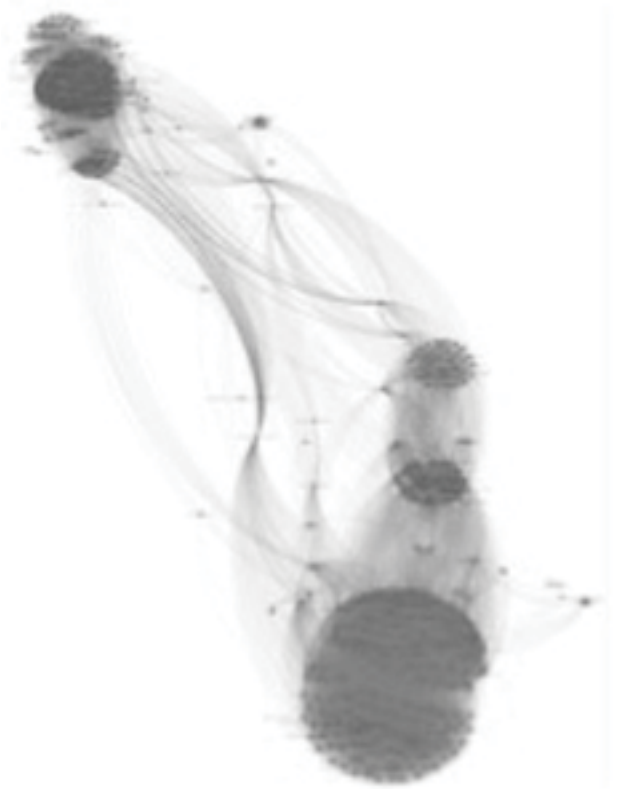


Figure 1: Graph obtained from one of the periods

Then, for each of the periods of our data we determine which are the clustering classes, that can be considered bags of 'similar artworks', and calculate the distances between classes by measuring the frequency of use of descriptors in the artworks contained in the cluster. We apply our own algorithm to distribute those classes in a 2D space so that their relative positions represent the relative distances among them (the closer the clusters are, the more similar descriptors they use). We are aware that 100% accuracy is impossible because of the size of the descriptors pool we are using, that would require a higher dimensional space.

Once these clusters are organized in our space with a size proportional to the amount of artworks they have, we go back to the descriptors they contain and we generate the areas of influence of each descriptor as a potential field. As it is normal that, due to the ontological organization of the descriptors, some of these areas contain other areas or sections of them, we also represent the borderlines of the areas to show how these intersections play out. This allows us to generate different views of the art-space, taking into account elements such as time, descriptors by modularity class, or specific descriptors closely related to current discussions by art historians specializing in the period (see Figure 2). These different views provide as many different facets of a digital geography of Hispanic Baroque Art.

Also, we calculate distances between similarity classes in different time periods so that we can infer which class evolve from previous ones and draw the semantic evolution of the artworks. This is fundamental for a better understanding of the generation of families of artworks and the variants that this evolution produces, what would help us to connect this process with explanations in political, artistic or economic discourses.

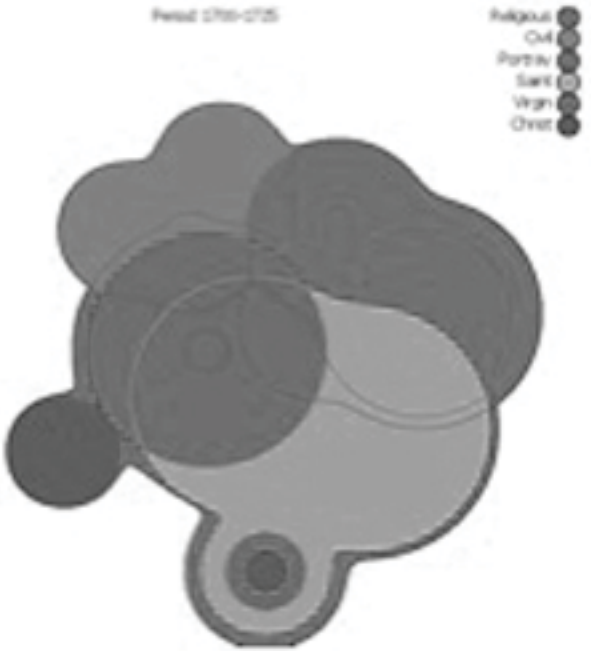


Figure 2: Art-space of one period from the point of view of main descriptors

Finally, and from the geographical information available for a subset of artworks (where original location and current location has been determined) we can make a representation of the movement of the artworks along time, and obtain information about how the museums (the main repositories of artworks currently) and other collectors have collected artworks from specific areas or attending to the semantic groups we have discovered from previous analysis.

This methodology allows us to address different issues related to the political, geographic and cultural aspects of art production, reception, and consumption. Some of these issues are: are paintings local, regional, or national?; how different visualizations affect the clustering of art-works and artists?; are there differences between political and artistic territories?; what is the transmission of features across time and space like?; which is the effect of flows of art-works away from their place of origin due to market forces?; how different clusters of art behave and what its effect is on center-periphery debates?. The result is a digital geography of Hispanic Baroque art that will contribute to a better understanding of art history from a spatial point of view and will also shed light on cultural transfers in complex systems.

Funding

This work was supported by the Social Sciences and Humanities Research Council of Canada and by the Canada Foundation for Innovation.

References

DaCosta Kaufmann, T. (2004). *Toward a Geography of Art*. Chicago: The University of Chicago Press, p. 6.

Suárez, J. L., F. Sancho, and J. de la Rosa (2011). The art-space of a global community: the network of Baroque paintings in Hispanic-America. *Proceedings of the I nternational Conference on Culture and Computing 2011 at Kyoto University, Japan*, 2011, pp. 45-50.

Suárez, J. L., F. Sancho, and J. de la Rosa (2012; forthcoming). Sustaining a Global Community: Art and Religion in the Network of Baroque Hispanic-American Paintings. Special Section of *Leonardo Journal* (MIT Press).

Notes

1. For a detailed explanation of the methodology, please see Juan Luis Suárez, Fernando Sancho and Javier de la Rosa (2011 and 2012).

Approaching Dickens' Style through Random Forests

Tabata, Tomoji
tabata@lang.osaka-u.ac.jp
University of Osaka, Japan

1. Outline of this study

This paper describes a new approach to Dickens' style, applying a state-of-art machine-learning classification technique in an effort to distinguish Dickens' texts from a reference corpus of texts. This study makes use of Breiman's (2001) 'Random Forests' in order to spotlight lexical items Dickens consistently used or avoided in his texts in comparison with the control set.

By demonstrating how we can identify Dickensian stylistic markers, this study also proposes Random Forests as a more powerful alternative for traditional 'key word' analysis, a popular method in corpus linguistics for extracting a set of words that characterize a particular text, a particular register/ (sub)corpus, or a particular diachronic set of texts, etc. from others.

In a typical key word analysis, the significance of difference between two sets of texts in the frequency of a word is calculated based on log-likelihood ratio (LLR, henceforward; Tun- ning 1993; Rayson & Garside 2000). However, LLR alone provides little information about proportion of difference between the sets, not to mention whether or not a particular word is distributed evenly throughout each set of texts. LLR tells us only about the 'significance' (NOT the 'degree' of) of discrepancy between two sets of texts with respect to frequency of a partic- ular lexical variable. Thus, a potential drawback of the method is seen in a case where a word occurring with exceptionally high frequency in a single text can over-represent the entire set it belongs to. Using Random Forests, this study will address such an issue and provide means to identify a more reliable set of markers of Dickens' style.

2. The drawbacks of traditional 'key' word analysis

Textual analysis often begins with identifying key words of a text on the assumption that key words reflect what the text is really about and that they are likely to reflect stylistic features of the text as well. Key words as a corpus linguistic terminology are defined as words that 'appear in a text or a part of a text with a frequency greater than chance occurrence alone would suggest' (Henry & Roseberry 2001: 110). A popular method to measure 'keyness' is to calculate an LLR to assess the significance of difference between the expected frequency and the observed frequency of a word in a text. However, one drawback of this approach emerges when we compare Dickens' texts with, for example, a set of texts by his contemporary Wilkie Collins.

Rank	Word	Frequency	In Files	Proportion	LLR
1	upon	12,990	24	0.27%	8871.058
2	and	176,688	24	3.65%	8215.058
3	mr	31,312	24	0.65%	5151.845
4	very	14,312	24	0.30%	3639.361
5	so	20,986	24	0.43%	3479.701
6	a	109,288	24	2.26%	2666.071
7	but	26,202	24	0.54%	2580.812
8	said	30,698	24	0.63%	2457.399
9	pickwick	2,198	2	0.05%	2373.409
10	great	6,975	24	0.14%	2015.407
11	much	7,268	24	0.15%	1912.053
12	nicholas	1,740	4	0.04%	1878.858
13	they	17,630	24	0.36%	1863.395
14	tom	1,846	21	0.04%	1712.521
15	dombey	1,420	1	0.03%	1533.321
16	replied	3,875	24	0.08%	1447.446
17	john	2,113	24	0.04%	1432.579
18	pecksniff	1,250	1	0.03%	1349.755
19	gentleman	4,601	24	0.10%	1344.429
20	or	16,102	24	0.33%	1278.461
21	king	1,687	21	0.03%	1275.183
22	joe	1,172	17	0.02%	1251.145
23	martin	1,121	7	0.02%	1210.460
24	boffin	1,105	1	0.02%	1193.183
25	being	6,904	24	0.14%	1181.222
26	though	3,841	24	0.08%	1180.777
27	sam	1,210	4	0.02%	1143.636
28	many	4,225	24	0.09%	1141.551
29	down	8,408	24	0.17%	1088.151
30	weller	992	2	0.02%	1071.165
31	dorrit	971	2	0.02%	1048.489
32	'em	1,331	21	0.03%	1047.537
33	old	9,624	24	0.20%	1041.342
34	nickleby	939	1	0.02%	1013.936
35	clennam	938	1	0.02%	1012.856
36	were	17,745	24	0.37%	989.771
37	indeed	2,503	24	0.05%	966.672
38	such	7,351	24	0.15%	961.150
39	florence	1,083	4	0.02%	940.432
40	squeers	861	1	0.02%	929.711

Table 1: 40 most significant key words of Dickens compared with Wilkie Collins (sorted according to log-likelihood ratio (LLR))

Although proper names such as *Dombey*, *Pecksniff*, and *Boffin* are ranked among the top 40 Dickensian 'key' words, none of these would normally be counted as important at least from a stylistic perspective, apart from the fact these represent a few peculiar Dickensian characters, who appear only in a single text. As we go down the list, it turns out that 19 out of the top 100 are those which occur only in a single text. As we turn our eyes to words listed as Collins's key words, we see 35 out of the top 100 are from a single text. If we include words that occur only in a very small number of texts, the proportion becomes even greater.

Another (though less obvious) drawback comes from the mathematical fact that LLR em- phasizes high-frequency items: in fact, high-frequency words tend to predominate towards the top of the list in Table 1, with words in lower frequency-strata not tending to be highlighted easily. A better alternative is needed.

3. Random Forests as a stylometric tool

Random Forests (henceforth, RF) are a very efficient classification algorithm based on ensemble learning from a large number of classification trees (thus 'forests') randomly generated from the dataset. As RF builds a classification tree from a set of bootstrap samples, about one-third of the cases are left out for running an internal unbiased estimate of the classification error each time. The process is iterated until *n*-th (500th by default) tree is added to the forests, there being no need for cross-validation or a separate test (Breiman and Cutler, ND). One of the most prominent features of RF is its high accuracy in respect to classification of data sets. In the experiments reported by Jin and Murakami (2007), RF was unexcelled in accuracy among other high-performance classifiers, such as *k* Nearest Neighbor, Support Vector Machine, Learning Vector Quantification, Bagging and AdaBoosting, in distinguishing between 200 pieces of texts written by 10 modern Japanese novelists. In my own experiments, RF constantly achieved accuracy rates as high as 96–100% in distinguishing Dickens's texts from the control set of texts.

RF is capable of handling thousands of input variables and of running efficiently on large database.[1] The first series of experiments with RF in this study were on a set of 24 Dickens's texts versus a comparable set of 24 Wilkie Collins' texts with from more than five thousand to as few as fifty word-variables. The results were fairly consistent with over 96% accuracy. The best of the results was seen when 300 input variables are used (Table 2). The out-of-bag (OOB) estimate of error rate was 0%, or 100% accuracy in distinguishing between Dickens's and Collins's texts.

Call:
randomForest(formula = dat$AuthGroup ~ ., data=dat[,2:301], proximity=T, importance=T, ntry=20)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 20
OOB estimate of error rate: 0%

Confusion matrix:

	Collins	Dickens	class.error
Collins	24	0	0
Dickens	0	24	0

Table 2: A result of running Random Forests

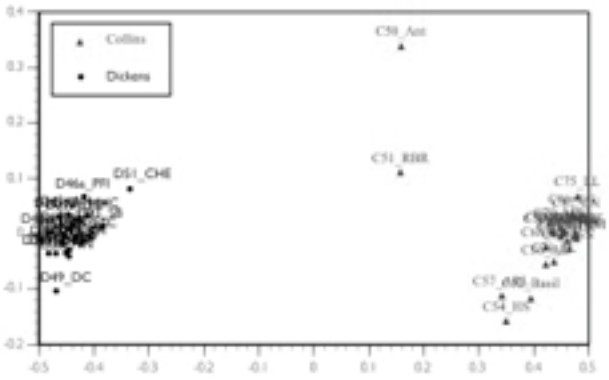


Figure 1: A multi-dimensional scaling diagram based on the proximity matrix generated by RF: Dickens versus Collins

RF computes proximities between pairs of cases that can be visualized in a multidimensional scaling diagram as in Figure 1. Dickens's texts and Collins's texts, respectively, form distinct clusters, with two unusual pieces by Collins shown as outliers: *Antonina* (1850), which has its setting in ancient Rome, and the travel book *Rambles beyond Railways* (1851).

Of further merit is the ability to highlight the lexical items that contribute most strongly to authorial classifications. RF shows the importance of variables in two measures: Mean-DecreaseAccuracy and MeanDecreaseGini. The former indicates the mean of decrease in the accuracy of classification when a particular variable is excluded from analysis. The latter shows the mean of decrease in Gini index, an index of uneven distribution of a particular variable between the groups, when the variable in question is excluded from analysis. The two indices are comparable to each other. Table 3 list marker words of Dickens and Collins, respectively, in the order of importance (sorted according to the mean decrease in Gini index).

Dickens markers:
very, many, upon, being, much, and, so, with, a, such, indeed, sir, off, but, would, down, great, there, up, or, were, head, they, into, better, quite, brought, said, returned, rather, good, who, came, having, never, always, ever, replied, boy, where, this, sir, well, gone, looking, dear, himself, through, should, too, together, these, like, an, how, though, then, long, going, its
Collins markers:
first, words, only, end, left, moment, room, last, letter, to, enough, back, answer, leave, still, place, since, heard, answered, time, looked, person, mind, on, woman, at, told, she, own, under, just, ask, once, speak, found, passed, her, which, had, me, felt, from, asked, after, can, side, present, turned, life, next, word, new, went, say, over, while, far, london, don't, your, tell, now, before

Table 3: Import variables: Dickens markers and Collins markers (in the ordner of importance)

A close comparison of Table 3 with Table 1will show how RF helps identify words with high discriminatory power. Proper nouns and words distributed unevenly in each set now have effectively made way for words which are consistently more frequent in one author than the other. Although Table 3 in itself is worth scrutiny in that it reveals how Dickens markers are contrasted with Collins markers, it is necessary to compare Dickens texts

with a larger and more representative corpus of writings in order to spotlight Dickens's stylistic features in a wider perspective.

4. Comparing and contrasting Dickens with a reference corpus of 18th-and 19th-Century authors

In the following analysis, the set of 24 Dickens' texts was compared with a reference corpus consisting of 24 eighteenth-century texts and 31 nineteenth-century texts.

When 300 input variables were used, the OOB estimate of error rate fell down to 1.28%, or 98.72% of runs correctly distinguishing Dickens from the reference set of texts. Fig. 2 distinguishes between the Dickens cluster and the control cluster. One seeming anomaly is the position of *A Child's History of England* (1851), which finds itself as an outlier. This history book for children is considerably different in style from other Dickensian works. Therefore, it is not unexpected for this piece to be found least Dickensian, a phenomenon in consistent with previous multivariate studies based on other linguistic variables, such as collocates of *gentleman* (Tabata 2009: 272), *–ly* adverbs (Tabata 2005: 231) and part-of-speech distribution (Tabata 2002: 173).

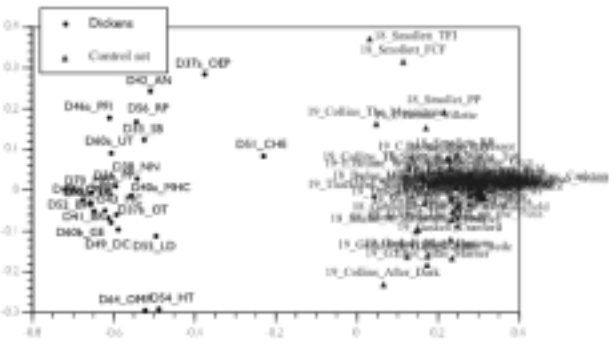


Figure 2: A multi-dimensional scaling diagram based on the proximity matrix generated by RF: Dickens versus reference corpus

Table 4 Important variables: Dickens versus the reference corpus (in the order of importance)

Positive Dickens markers
eyes, hands, again, are, these, under, right, yes, up, sir, child, looked, together, here, back, it, at, am, long, quite, dies, better, mean, why, turned, where, do, face, new, there, dear, people, they, door, cried, in, you, mrs, very, way, man
Negative Dickens markers
lady, poor, less, of, things, leave, love, not, from, should, can, last, saw, now, next, my, having, began, out, letter, had, I, money, tell, such, to, nothing, person, be, would, those, fat, miss, life, called, found, with, how, must, more, herself, well, did, but, much, make, other, whose, as, own, take, go, no, gave, shall, some, against, wife, since, first, them, word

Table 4: Important variables: Dickens versus the reference corpus (in order of importance)

Table 4 arrays major Dickens markers. Although relationships among these words are com- plex enough to defy straightforward generalization, one could see the predominance of words related to description of actions – typically bodily actions – or postures of characters rather than words denoting abstract ideas. Words like *eyes, hands, saw, looked, back* in particular have caught eyes of critics such as Hori (2004) and Mahlberg (2007a, 2007b), which suggests the present methodology is well-grounded. Stubbs (2005) states:

[E]ven if quantification only confirms what we have already know, this is not a bad thing. Indeed , in developing a new method, it is perhaps better not to find anything too new, but to confirm findings from many years of traditional study, since tis gives confidence that the method can be relied on (Stubbs 2005: 6).

In order to determine local textual functions (Mahlberg 2007a, 2007b) each of these words performs, it is of course necessary to go back to texts and examine the words in local contexts with the help of other tools such as concordance, collocation, and *n*-grams. However, the present method opens up a promising pathway to deeper textual analysis. The results of this study will also point to the effectiveness of this approach in cases of disputed authorship.

5. Appendix

No.	Author	Texts	Abbr.	Category	Date	Word-tokens
1	Dickens	*Sketches by Boz*	(SB)	Sketches	1833–6	187,474
2	Dickens	*The Pickwick Papers*	(PP)	Serial Fiction	1836–7	298,887
3	Dickens	*Other Early Papers*	(OEP)	Sketches	1837–40	66,939
4	Dickens	*Oliver Twist*	(OT)	Serial Fiction	1837–9	156,869
5	Dickens	*Nicholas Nickleby*	(NN)	Serial Fiction	1838–9	321,094
6	Dickens	*Master Humphrey's Clock*	(MHC)	Miscellany	1840–1	85,831
7	Dickens	*The Old Curiosity Shop*	(OCS)	Serial Fiction	1840–1	217,375
8	Dickens	*Barnaby Rudge*	(BR)	Serial Fiction	1841	233,979
9	Dickens	*American Notes*	(AN)	Sketches	1842	101,623
10	Dickens	*Martin Chuzzlewit*	(MC)	Serial Fiction	1843–4	335,462
11	Dickens	*Christmas Books*	(CB)	Fiction	1843–8	154,410
12	Dickens	*Pictures from Italy*	(PFI)	Sketches	1846	72,497
13	Dickens	*Dombey and Son*	(DS)	Serial Fiction	1846–8	343,987
14	Dickens	*David Copperfield*	(DC)	Serial Fiction	1849–50	355,714
15	Dickens	*A Child's History of England*	(CHE)	History	1851–3	162,883
16	Dickens	*Bleak House*	(BH)	Serial Fiction	1852–3	354,061
17	Dickens	*Hard Times*	(HT)	Serial Fiction	1854	103,263
18	Dickens	*Little Dorrit*	(LD)	Serial Fiction	1855–7	338,076
19	Dickens	*Reprinted Pieces*	(RPR)	Sketches	1850–6	91,468
20	Dickens	*A Tale of Two Cities*	(TTC)	Serial Fiction	1859	136,031
21	Dickens	*The Uncommercial Traveller*	(UT)	Sketches	1860–9	142,773
22	Dickens	*The Great Expectations*	(GE)	Serial Fiction	1860–1	184,776
23	Dickens	*Our Mutual Friend*	(OMF)	Serial Fiction	1864–5	324,891
24	Dickens	*The Mystery of Edwin Drood*	(ED)	Serial Fiction	1870	94,014
					Sum of word-tokens in the set of Dickens texts:	4,841,357

Table 5: Dickens component of ORCHIDS

No.	Author	Texts	Date	Word-tokens
1	Defoe	*Captain Singleton*	1720	118,916
2	Defoe	*Journal of Prague Year*	1722	83,494
3	Defoe	*The Military Memoirs of Captain George Carleton*	1728	80,617
4	Defoe	*Moll Flanders*	1724	138,094
5	Defoe	*Robinson Crusoe*	1719	232,453
6	Fielding	*A Journey from this World to the Next*	1749	45,024
7	Fielding	*Amelia*	1751	212,339
8	Fielding	*Jonathan Wild*	1743	70,086
9	Fielding	*Joseph Andrews*	1742	126,342
10	Fielding	*Tom Jones*	1749	347,219
11	Goldsmith	*The Vicar of Wakefield*	1766	63,076
12	Richardson	*Clarissa*	1748	939,448
13	Richardson	*Pamela*	1740	438,562
14	Smollett	*Peregrine Pickle*	1751	330,557
15	Smollett	*Ferdinand Count Fathom*	1753	157,032
16	Smollett	*Humphrey Clinker*	1771	150,281
17	Smollett	*Sir Launcelot Greaves*	1760	89,010
18	Smollett	*Roderick Random*	1748	191,559
19	Smollett	*Travels through France and Italy*	1766	121,032
20	Sterne	*A Sentimental Journey*	1768	41,028
21	Sterne	*Tristram Shandy*	1759–67	184,429
22	Swift	*A Tale of a Tub*	1704	44,225
23	Swift	*Gulliver's Travels*	1726	103,806
24	Swift	*A Journal to Stella*	1710–3	191,740
		Sum of word-tokens in the set of 18th Century texts:		**4,693,540**

Table 6: 18th Century component of ORCHIDS

No.	Author	Texts	Date	Word-tokens
1	A.Brontë	*Agnes Grey*	1847	68,352
2	Austen	*Emma*	1815	160,899
3	Austen	*Mansfield Park*	1814	159,921
4	Austen	*Northanger Abbey*	1803	77,810
5	Austen	*Persuasion*	1816 (1818)	83,380
6	Austen	*Pride and Prejudice*	1813	121,874
7	Austen	*Sense and Sensibility*	1811	119,793
8	C.Brontë	*Jane Eyre*	1847	188,092
9	C.Brontë	*The Professor*	1857	88,281
10	C.Brontë	*Villette*	1853	193,819
11	Collins	*After Dark*	1882	136,356
12	Collins	*The Moonstone*	1868	196,506
13	Collins	*The Woman in White*	1859	246,917
14	E.Brontë	*Wuthering Heights*	1847	117,344
15	G.Eliot	*Adam Bede*	1859	215,253
16	G.Eliot	*Brother Jacob*	1864	16,693
17	G.Eliot	*Daniel Deronda*	1876	311,400
18	G.Eliot	*Middlemarch*	1871–2	317,975
19	G.Eliot	*Silas Marner*	1861	71,449
20	G.Eliot	*The Mill on the Floss*	1860	207,505
21	Gaskell	*Cranford*	1851–3	71,037
22	Gaskell	*Mary Barton*	1848	161,098
23	Gaskell	*Sylvia's Lovers*	1863	191,176
24	Thackeray	*Barry Lyndon*	1844	125,986
25	Thackeray	*Vanity Fair*	1848	303,530
26	Trollope	*Rachester Towers*	1857	197,691
27	Trollope	*Can You Fogive Her*	1865	316,349
28	Trollope	*Doctor Thorne*	1857	220,867
29	Trollope	*The Eustace Diamonds*	1873	269,981
30	Trollope	*Phineas Finn*	1869	263,393
31	Trollope	*The Warden*	1855	72,068
		Sum of word-tokens in the set of 19th Century texts:		**5,292,795**

Table 7: 19th Century component of ORCHIDS

References

Breiman, L. (2001). Random forests. *Machine Learning* 45: 5-23.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-74.

Henry. A., and R. L. Rooseberry (2001). Using a small corpus to obtain data for teaching genre. In M. Ghadessy, A. Henry and R. L. Roseberry (eds.), *Small Corpus and ELT*. Amsterdam,Philadelphia, Pa.: John Benjamins, pp. 93-133.

Hori, M. (2004). *Investigating Dickens' Style: A Collocational Analysis*. New York: Palgrave Macmillan.

Jin, M., and M. Murakami (2007). Random forest hou ni yoru bunshou no kakite no doutei (Authorship Identification Using Random Forests), *Toukei Suuri* (*Proceedings of the Institute of Statistical Mathematics*) 55(2): 255-268.

Mahlberg, M. (2007a). Corpus stylistics: bridging the gap between linguistic and literary studies. In M. Hoey et al. (eds.), *Text, discourse and corpora. Theory and analysis*. London: Continuum, pp. 217-246.

Mahlberg, M. (2007b). Clusters, key clusters and local textual functions in Dickens. *Corpora* 2(1): 1-31.

Rayson, P., and R. Garside (2000). Comparing Corpora Using Frequency Profiling. *Proceedings of the Workshop on Comparing Corpora, Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* (ACL 2000), 1-8 October 2000, Hong Kong, 1-6. Available online at `http://www.comp.lancs.ac.uk/computing/users/paul/phd/phd2003.pdf`.

Stubbs, M. (2005). Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature* 14(1): 5-24.

Tabata, T. (2002). Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution. In T. Saito, J. Nakamura and S. Yamasaki (eds.), *English Corpus Lin- guistics in Japan*. Amsterdam: Rodopi, pp. 165-182.

Tabata, T. (2005). Profiling stylistic variations in Dickens and Smollett through correspondence analysis of low frequency words. *ACH/ALLC 2005 Conference Abstracts*, Humanities Computing and Media Centre, University of Victoria, Canada, pp. 229-232.

Tabata, T. (2009). More about gentleman in Dickens. *Digital Humanities 2009 Conference Abstracts, University of Maryland, College Park, June 22–25, 2009*, The Association for Literary and Lin- guistic Computing, the Association for Computers and the Humanities, and the Society for Digital Humanities – Société pour l'étude des médias interactifs, pp. 270-275.

Notes

1. Breiman, L., and A. Cutler, *Random Forests*. Online resource. (Last accessed 20 October 2011.) `http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm`

Interfacing Diachrony: Visualizing Linguistic Change on the Basis of Digital Editions of Serbian 18th-Century Texts

Tasovac, Toma
ttasovac@humanistika.org
Center for Digtial HumanitiesBelgrade, Serbia

Ermolaev, Natalia
n.ermolaev@rutgers.edu
Program in Library and Information Science,
Rutgers University, USA

Analyzing, indexing and marking-up raw data and providing various types of annotations and metadata, named entities and other contextual information is essential for effective searching and retrieval of cultural heritage content (Borin et al.; Borin et al. 2007; Schreiber et al. 2008; Christopher 2011). With commercial search engines dominating our day-to-day interaction with information and framing our data access with non-transparent page-ranking mechanisms, it is especially important for institutions of learning and cultural preservation to use available technologies to encourage meaningful and profound engagement with digital content. In this respect, the interface must not only function as the frame for direct access and data retrieval, but also as the platform for the user's cognitive, aesthetic and performative interaction with digital objects as such. DH projects, however, often approach interface design in a haphazard way, more as an afterthought than as an essential system component (Warwick et al. 2008). They also tend to be dominated by task-oriented and efficiency-driven paradigms of machine engineering that disregard the performative nature of the interface as a reading space (Drucker 2011a, 2011b).

Ongoing research in digital humanities, human-computer interaction and interface design is focusing on ways to explore textual data by visualizing relations between different sets of data and identifying patterns such as word trends, named entities, collocations etc. (Bederson & Schneiderman 2003; Unsworth 2005; Don et al. 2007; Fry 2007; Greengrass & Hughes 2008; Rockwell et al. 2010). Collins (2010) has explored interactive interfaces that provide direct access to both the online natural language processes, such as statistical translation, keyword detection, and parsing, and the outcomes of sophisticated linguistic analysis. So far, however, no attempts have been made to visually link the user's subjective experience of archaic texts with a graphic representation of diachronic changes in language. In this paper, we describe our design and development of an interface for representing and highlighting linguistic change on the basis of our annotated digital editions of Serbian 18th-century literary texts (Tasovac & Ermolaev 2011). We treat the interface not only as a dynamic framework for engaging with the text, but also as a 'provocation to cognitive experience' (Drucker 2011a: 9).

Unlike many European languages, which underwent relatively uninterrupted linguistic and cultural evolutions, the modern Serbian literary language, as codified by Vuk Stefanović Karadžić in the 19th century, was largely based on the vernacular of his time and had little in common with the literary standards of the previous epoch (Ивић 1990). As a consequence of this radical caesura, 18th-century texts pose a formidable challenge to the modern reader who is unaccustomed to these archaic (Old Church Slavonic) and foreign (largely Russian) imprints (Albin 1970). That is why the Digital Library of Serbian Cultural Heritage of the 18th Century – a joint project of the Belgrade Center for Digital Humanities and the National Library of Serbia – employs a host of mark-up and annotating strategies to make these texts more accessible. The texts are encoded as a word-aligned corpus of TEI XML documents in two versions: one using traditional 18th-century orthography, including the graphemes that have since disappeared from Serbian, and one using modernized and standardized Serbian spelling that increases the legibility and, to a certain degree, searchability of these texts for modern users. Furthermore, the corpus contains lexical and semantic annotations that introduce a large number of modern-day equivalents to the largely archaic vocabulary of the corpus, as shown in the following table:

Original orthography	Modernized orthography	Modern-day equivalent	Type of change
богъ	бог	бог	orthographic
любовъ	љубов	љубав	phonetic
утѣшенїе	утешење	утеха	morphological
благодѣтель	благодетељ	доброчинитељ	lexical
любезница	љубезница	драга особа (женског пола)	conceptual (term no longer lexicalized)

By applying basic techniques of cross-lingual information retrieval to a historical dimension of one language, and making provisions for multiple indexing and annotations, our project exposes a notoriously difficult chapter in the development of the Serbian language to a wider audience, without sacrificing the edition's scholarly potential. The

user can search the corpus not only using modern standardized spelling, but also the above-mentioned range of modern-day equivalents. This type of extended search, which results in a larger pool of data than a search based on original, non-standardized orthographic forms alone, has the potential of radically opening up the text for the modern reader and leading to discoveries of previously unnoticed thematic similarities and correspondences across different works and authors.

The interface is built dynamically using XQuery and advanced JavaScript on top of the TEI XML files which are stored in the eXist database. We offer three basic views of the text: a *static view*, a *user-driven dynamic view*, and the *system-driven dynamic view*. The *static view* is the most basic: the user can choose to read the text in either the original orthography (OO) or its modernized version (MO). In either version, each word is also a hyperlink: clicking on a word reveals a pop-up window which lists the form of the corresponding orthographic version as well as the lexical or semantic annotation , when appropriate. In the static view, the user can also choose to *juxtapose* (place side by side), *interpolate* (view line by line) or *superimpose* (merge and differentiate by means of color) the two orthographic versions of the text.

The *user-driven dynamic view* stresses the evolution of linguistic change: it allows the reader to treat the original 18th-century text as a temporal snapshot – a frozen expression of a certain time and age – that can be, using the familiar UI device (slider) 'updated' (or 'fast-forwarded') to its other more modern linguistic forms. The stages of transformation include: modernized spelling that keeps the linguistic peculiarities of the original, phonetic changes morphological changes, lexical changes and, finally, conceptual reformulations. Each stage of transformation can be viewed separately or in combination with others. And each change takes place inside the text by means of in-line animated transformations of individual word forms.

Finally, the *system-driven dynamic view* is the most experimental: it creates a unique reading environment in which annotated words change their linguistic form in-line, but each at its own, randomly assigned speed. In this playful space, the linear reading act becomes predicated upon a chance encounter with a traditional or modern form, an exercise in Benjaminian translation (Benjamin 1972) where the site of alterity (in our case temporal alterity) quite literally (and never in quite the same sequence of transformations) disturbs the text's imaginary stability. Reading becomes a performance in which the reader is performing the text, and vice versa.

By providing these various visual interfaces in our Digital Library of Serbian Cultural Heritage of the 18th Century, we address for the first time in the framework of visual DH, the subjective perspective of the reader's process of decoding archaic language. We treat the linguistic expression of the 18th-century Serbian culture not as a stable and self-evident entity, but as a locus of meaning that can (and should) be transformed by the reader. While the interface provides access to systematic linguistic, cultural and historical annotations, it also creates an opportunity for playful interaction with the text as a space of linguistic and cultural mutability.

The more possibilities at a user's disposal to engage with a digital cultural heritage object, the greater its impact will be at the levels of both individual and institutional experience. The framework of the Digital Library of Serbian 18th-Century Texts can serve as a model for other smaller, lesser-resourced languages struggling with the quest to keep their cultural heritage alive after historical ruptures, linguistic caesuras, changes of alphabet etc.

References

Albin, A. (1970). The Creation of the Slaveno-Serbski Literary Language. *The Slavonic and East European Review* 48(113): 483-91.

Bederson, B., and B. Schneiderman (2003). *The Craft of Information Visualization: Readings and Reflections*. San Francisco, CA.: Morgan Kaufmann.

Benjamin, W. (1972). Die Aufgabe des Übersetzers. *Gesammelte Schriften IV/1*, 9-21. Frankurt Main: Suhrkamp.

Borin, L., M. Forsberg, and D. Kokkinakis (2010). *Diabase: Towards a diachronic BLARK in support of historical studies*. `http://spraakbanken.gu.se/personal/lars/pblctns/lrec2010-diabase.pdf` . Accessed 15 March 2010.

Borin, L., D. Kokkinakis, and L.-J. Olsson (2007). Naming the past: Named entity and animacy recognition in 19th century Swedish literature. *Workshop on Language Technology for Cultural Heritage Data*, pp. 1-8.

Christopher, A. (Cal) Lee (2011). A framework for contextual information in digital collections. *Journal of Documentation* 67(1): 95-143.

Collins, Ch. M. (2010). *Interactive Visualizations of Natural Language*. University of Toronto.

Don, A., et al. (2007). Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining With Visualization. *Proceedings of the*

Sixteenth ACM Conference on Information and Knowledge Management, pp. 213-22.

Drucker, J. (2011a). Humanities Approaches to Interface Theory. *Culture Machine*: 121-20.

Drucker, J. (2011b). Performative Materiality and Interpretative Interface. *Digital Humanities 2011: Book of Abstracts*, pp. 39-40.

Fry, B. (2007). V*isualizing Data: Exploring and Explaining Data With the Processing Environment*. Sebastopol: O'Reilly Media.

Greengrass, M., and L. Hughes, eds. (2008). *The Virtual Representation of the Past*. Aldershot Hants, England; Burlington VT: Ashgate.

Rockwell, G., et al. (2010). Ubiquitous Text Analysis. *Poetess Archive* 2(1): 1-19.

Schreiber, G., et al. (2008). Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. *Web Semantics: Science, Services and Agents on the World Wide Web* 6(4): 243-49.

Tasovac, T., and N. Ermolaev (2011). Encoding Diachrony: Digital Editions of Serbian 18th-Century Texts. *Lecture Notes in Computer Science*, 6966497-500.

Unsworth, J. (2005). *New Methods for Humanities Research. The 2005 Lyman Award Lecture. November 11. National Humanities Center. Research Triangle Park, NC.* `http://www3.isrl.illinois.edu/~unsworth/lyman.htm` . Accessed June 10 2011.

Warwick, C., et al. (2008). If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. *Literary and Linguistic Computing* 23(1): 85-102.

Ивић, Милка (1990). О *језику Вуковом и вуковском* . Нови Сад: Књиж. заједница Новог Сада.

Promise and Practice of Enhanced Publications to Complement Conventionally-Published Scholarly Monographs

Tatum, Clifford
clifford@tatum.cc
eHumanities Group, Royal Netherlands Academy of Arts and Sciences, The Netherlands

Jankowski, Nicholas
nickjan@xs4all.nl
eHumanities Group, Royal Netherlands Academy of Arts and Sciences, The Netherlands

Scharnhorst, Andrea
andrea.scharnhorst@dans.knaw.nl
eHumanities Group, Royal Netherlands Academy of Arts and Sciences, The Netherlands

---

The authors of this paper have been involved in preparing a series of enhanced publications as part of a national initiative to introduce 'enhanced publishing' to scholars in the Netherlands. One of these publications is the scholarly monograph *Virtual Knowledge,* in production at MIT Press. This paper reviews different approaches to enhanced publications. In particular, it reports the intentions, challenges, achievements and tensions in the actual workflows in preparation of the *Virtual Knowledge* enhanced publication version of the manuscript. An enhanced publication can generally be described as scholarly text and related materials presented in a Web environment with interlinking of document 'objects' (Jankowski et al. 2011). These objects may include: research data, instrumentation, additional analyses, post-publication addendums, and exchanges between book authors and readers. These objects are often identified in a manner facilitating access and involving utilization of uniform standards of identification. Basic enhancements to this particular volume involve preparation of supplementary resources (e.g., links, blogs, appendices), chapter visualizations (e.g., animations, figures, tables), hyperlinks to materials in and external to the book, and opportunities for conducting searches within and external to the book. The enhancements also emphasize integration of informal scholarly communication (e.g., social media) with traditional academic publishing. The enhanced component also emphasizes integration of informal scholarly communication (e.g., social media) with traditional

academic publishing, and the efforts to achieve this aspect are highlighted in the paper.

The overall design strategy for this project was oriented toward introducing book content on the web with two primary objectives in mind: 1) to facilitate compatibility with contemporary Web-based discursive practices, and 2) to facilitate content interoperability in both Web 2.0 and Semantic Web formats. Challenges involved with the first objective are related to varied work practices among participating authors. In addition to variation in skills, competencies, and preferences in working with the web medium, the benefits from the required additional work remain uncertain. The individual career benefit from publishing a book is in most cases largely the same with or without the Web-based enhancements. The second objective entails integration of two distinct content structures. The resulting design is a hybrid platform that leverages Web 2.0 participatory modes of scholarly communication combined with formalized content structures imposed by Semantic Web formats. Differences in the affordances provided by these two content schemas brings into focus contemporary tensions related to scholarly communication in the digital era and raises interesting questions about the emerging role of digital media in scholarly communication.

The expected contribution of this approach to enhanced publications pertains to consumption of the book content by both humans and machines. The platform facilitates situating a published book among related scholarship, institutional settings, and with regard to prior work from contributing authors. This is seen as particularly useful for interdisciplinary research, and especially within a still emerging field. This sort of contextualization is by itself envisioned as value added from the standpoint of consumers of book content and for those engaging in Web-based intertextual discourses. Another contribution is the ways in which the book content is structured in the Web environment. Facilitated through links, categories and tags, and the co-location of related materials, contextualization of content also creates object relationships, e.g. between an author and with regard to intertextual relationships, which are readable by search engines. Fundamental to achieving these objectives is an understanding of the Web as an environment that is dynamic and evolving, which is far removed from not only the printed book, but also from the environment of digital repositories.

References

Jankowski, N. W., C. Tatum, Z. Tatum, and A. Scharnhorst (2011). Enhancing scholarly publishing in the humanities and social sciences: Innovation through hybrid forms of publication. *Paper, PKP conference*. Available at: http://pkp.sfu.ca/ocs/pkp/index.php/pkp2011/pkp2011/paper/view/326/185

Culpeper's legacy: How title pages sold books in the 17th century

Tyrkkö, Jukka Jyrki Juhani
jaytyrkko@me.com
University of Helsinki, Finland

Suhr, Carla Maria
carla.suhr@helsinki.fi
University of Helsinki, Finland

Marttila, Ville
ville.marttila@helsinki.fi
University of Helsinki, Finland

Nicholas Culpeper (1616–1654) was the best known name in seventeenth century medical publishing in London and is listed in the *English Short Title Catalogue* (ESTC) as the author of more than 230 books and as the translator of dozens more. An apothecary, man-midwife, and astrophysician, Culpeper is best remembered as the translator and editor of the *London Dispensatory* (1649), an unlicensed and best-selling translation of the *Pharmacopoeia Londinensis*, the official medicine book of the Royal College of Physicians. However, modern scholarship tells us that Culpeper was only partly responsible for his prodigious and lasting success. Much of his fame can be attributed to the efforts of his many publishers and printers, who over several decades turned the name Culpeper into a commercial brand by reprinting, reissuing, and frequently misrepresenting the author's relatively few authentic works (see McCarl 1996; Furdell 2002). During his lifetime Culpeper became one of the first names in scientific writing that could sell books. Books bearing his name were widely published throughout the eighteenth century, and sporadically to the present day.

This paper takes the case of Culpeper as a pilot study of title pages as a form of advertisement. Extra copies of title pages were commonly printed as flyers and posted on booksellers' stalls, hung on cleft sticks, or tacked to walls (Shevlin 1999: 48). In the seventeenth century, the title page – including the title of the work – was largely the domain of the bookseller and printer (McKerrow 1928: 91; Shevlin 1999: 52), making title pages a part of what Genette calls 'publisher's peritext' (1997: 16). In this paper we investigate the typographic and text-structural features of the title page in books attributed to Culpeper. The work builds on an earlier pilot study (Tyrkkö 2011) that identified the systematic nature in which printers and publishers made commercial use of not only the name Culpeper, but also the paratextual features of his previous books in an effort to emulate the style of his authentic works.

To enable the structural and typographical analysis of these title pages, the title pages of all books listed from the period 1649-1700 that mention Culpeper's name and are available at the British Library, Cambridge University Library or Wellcome Trust Library were transcribed and annotated for structural parts and named entities, as well as for visual features such as layout, graphic elements, and different typefaces and font sizes thereof. The annotation process started with the taking of close measurements, down to one fifth of a millimeter, of the aforementioned elements from original artefacts, and was completed using digital facsimiles from *Early English Books Online* (EEBO).

The annotation scheme is based on the TEI P5 *Guidelines for Electronic Text Encoding and Interchange*, using elements from the *Core, Default text structure*, and *Names, Dates, People, and Places* modules to annotate the textual structure and named entities, and elements from the *Core*, *Representation of primary sources* and *Tables, Formulæ, and Graphics* modules to annotate the visual layout of the title page. For the purposes of annotating the typographical layout with sufficient accuracy and consistency, we have developed an experimental system for annotating the different typefaces and their sizes, using the 'rend' attribute and a controlled value set. This system – which is based on relating the size of the different typefaces used on the title page to the absolutely measured 'base type' of the text – is intended to combine the benefits of absolute and relative measurement and to alleviate the difficulties caused by working with digital facsimiles, such as unknown scaling factors and distortion of proportions.

The quantitative analysis related the visual and structural features of the title pages to the bibliographic and sociohistorical parameters of the texts – such as stated target audience, format, identity and geographic location of the publisher and printer, publication year (whether before or after Culpeper's death) and the known relationship of the text to Culpeper. This data is imported into a database together with the abovementioned bibliographic and sociohistorical data, obtained from the ESTC, the *British Book Trade Index* (BBTI), the *Oxford Dictionary of National Biography* (ODNB) and earlier book historical research.

The database of paratextual, bibliographic and sociohistorical data will be queried using methods of multivariate analysis to identify relationships between the physical features of the title pages and the variables of their production histories. The

analyses will reveal diachronic trends in the design of title pages bearing the name Culpeper, and bring to light the underlying factors which influenced the decisions regarding the physical presentation of the texts and to highlight the different means used by printers and publishers to market their products. More specifically, this allows us to reconstruct a timeline of how the commercial brand of Culpeper was created and identify which features of the title page were specific to the brand, which were typical for the time and which were specific to particular publisher's or printer's house style. The findings will be examined in light of broader book historical scholarship on such features in an effort to distinguish features specific to the corpus of Culpeper books.

References

Early English Books Online (EEBO). http://www.eebo.chadwyck.com/ .

English Short Title Catalogue (ESTC) . http://www.estc.bl.uk/ .

Furdell, E. L. (2002). *Publishing and Medicine in Early Modern England*. Rochester: U of Rochester P.

Genette, G. (1997). *Paratexts: Thresholds of Interpretation*. Tr. by J. E. Lewin. Cambridge: Cambridge UP.

McCarl, M. R. (1996). Publishing the Works of Nicholas Culpeper, Astrological Herbalist and Translator of Latin Medical Works in Seventeenth-Century London. *Canadian Bulletin of Medical History / Bulletin canadien d'histoire de la médecine* 13(2): 225-276.

McKerrow, R. B. (1928). *An Introduction to Bibliography for Literary Students*. 2nd impression with corrections. Oxford: Clarendon Press.

Oxford Dictionary of National Biography (ODNB). http://www.oxforddnb.com .

Shevlin, E. F. (1999). "To reconcile book and title, and make 'em kin to one another": The evolution of the title's contractual functions. *Book History* 2(1): 42-77.

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 1.9.1. Last modified 5 March, 2011. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ . (Accessed 31 October, 2011).

Tyrkkö, J. (2011). Selling Culpeper: A case study into the use of title pages in seventeenth century commercial publishing. *Presentation at SHARP 2011*, Washington D.C., July 14-17, 2011.

The Differentiation of Genres in Eighteenth- and Nineteenth-Century English Literature

Underwood, Ted
tunder@illinois.edu
University of Illinois, Urbana-Champaign, USA

Sellers, Jordan
eseller2@illinois.edu
University of Illinois, Urbana-Champaign, USA

Auvil, Loretta
lauvil@illinois.edu
University of Illinois, Urbana-Champaign, USA

Capitanu, Boris
capitanu@ncsa.illinois.edu
University of Illinois, Urbana-Champaign, USA

The title of this paper has changed, because our effort to answer one question (about linguistic register) exposed a larger question with broad relevance to literary study, and indeed to the definition of 'literature.'

To start where the process of inquiry actually began: what happened to English poetic diction around 1800? William Wordsworth's claim to have brought poetry back to 'the language of conversation in the middle and lower classes of society' in *Lyrical Ballads* has long been represented as a turning point in literary history. Given the weight attributed to this claim, it is surprising that scholars haven't tried to test it. Did the language of poetry actually become more formal or specialized in the eighteenth century? And if so, did the change reverse itself around 1798? Finally, was this phenomenon restricted to poetry, or was it a broader transformation of diction that affected other genres as well?

We are increasingly in a position to answer questions like these. True, we can't ask eighteenth-century English speakers to demonstrate 'the language of conversation in the middle and lower classes of society.' Moreover, standard contemporary tests of difficulty (like the Flesch-Kincaid Readability Test) are not very applicable to earlier periods, because they rely on sentence length. Practices of punctuation have changed over time, making the average sentence steadily shorter from the seventeenth century through the twentieth.

It is more practical to assess the formality of diction. This assessment is particularly easy to make

in English because of an important peculiarity of its history: English was for two hundred years (1066-1250) almost exclusively a spoken language, while French and Latin were used for writing. Any English word that survived this period had to be the kind of word that gets used in conversation. Words that entered the language afterwards were often borrowed from French or Latin to flesh out the learned vocabulary. Even today, the distinction between these two parts of the lexicon remains an important aspect of linguistic register. For instance, Laly Bar-Ilan and Ruth Berman have shown that contemporary spoken English is distinguished from writing by containing a higher proportion of words from Old English. Moreover, this differentiation between writing and speech increases as students enter high school, where they also learn to use a greater proportion of words from French and Latin in formal expository prose than they do in written narrative (Bar-Ilan & Berman 2007).

If learned and informal registers were distinguished this way in the thirteenth century, and the same thing holds true today, then one can reasonably infer that it held true in the eighteenth and nineteenth centuries as well (for further evidence, see DeForest & Johnson 2001). Thus an etymological approach to diction can show us how the 'register' of a given genre changed across time, becoming more conversational or more formal.

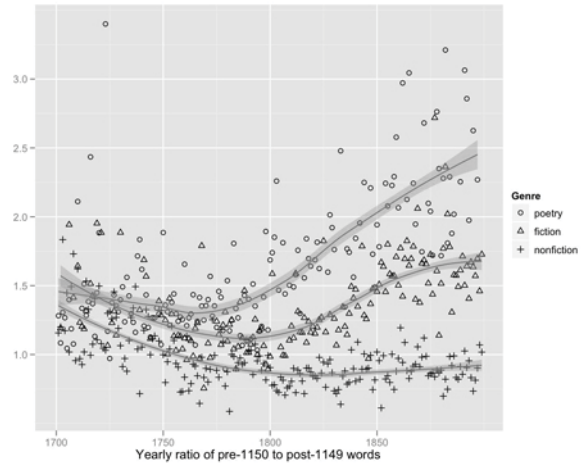


Figure 1

We have explored this question in a collection of 3,724 volumes. The eighteenth-century part of the collection was manually keyed by ECCO-TCP; the nineteenth-century part of the collection was digitized by OCR, but has been corrected with a fuzzy-matching script that has a machine-learning component and is extensively optimized for nineteenth-century OCR (our strategy was based on Lasko & Hauser 2002). More importantly, the comparative logic of this inquiry largely factors out the false negatives produced by imperfect OCR.

Instead of distinguishing 'Germanic' and 'Latinate' diction, we have used the first attested date for each word, choosing 1150 as a dividing line because it's the midpoint of the period when English was not used in writing. But date-of-entry of course correlates strongly with the Germanic/Latinate division. One can in fact simply measure the average length of words and produce very similar results (the correlation between the pre/post-1150 ratio and average word length is usually -.85 or lower). We exclude a generous list of stopwords (determiners, prepositions, conjunctions, pronouns, and the verb to be). The reason is that, as Bar-Ilan and Berman point out, 'register variation is essentially a matter of *choice*' (15). There is usually no alternative to stopwords, so they may not reveal much about register. We also exclude abbreviations, proper nouns, and words that entered the language after 1699.

The results of this inquiry do suggest that the register of poetry took a new turn in the late eighteenth century. In the course of the nineteenth century older, pre-1150 words became dramatically more common in poetry. It is reasonable to infer that poetic diction became more familiar and less overtly learned. But this particular detail is hardly the most striking fact about Fig. 1. What's salient is rather a broader process of generic differention from 1700 to 1899 that affected prose fiction and nonfiction as well as poetry. In the year 1700, these genres each had their own peculiarities of diction, but they didn't belong to sharply distinguished registers of the language. By 1899, they did. The ratio of pre- to post-1150 words became almost twice as high in prose fiction as in nonfiction, and almost three times higher in poetry. This suggests that the story usually told about Wordsworth may be misleading: far from rebelling against specialized poetic diction, he was producing a manifesto for a style that was – less recondite, to be sure – but also *more* sharply differentiated from prose.

This result matters, more broadly, because it bears an interesting relationship to the emergence of 'literature' as a distinct cultural category in the same period. In the early eighteenth century, 'literature' could encompass anything read by the middle and upper classes, emphatically including nonfiction. By the end of the nineteenth century, 'literature' was a category sharply distinguished from nonfiction prose, and valued for special aesthetic qualities. If this conceptual shift was also accompanied by a systematic transformation of diction, we may be able to learn something about the nineteenth-century concept of 'literature' by paying close attention to the way diction changed.

To start a debate on this topic, we will present an argument with two parts. First, we will show that the differentiation of genres was not merely a matter of linguistic register. When we compare genres in a more general way (using metrics like Spearman correlation and cosine similarity) it remains true that poetry, fiction and, to some extent, drama became steadily less like nonfiction prose in the period 1700-1899 (for the rationale behind these metrics, see Kilgarriff 2001).

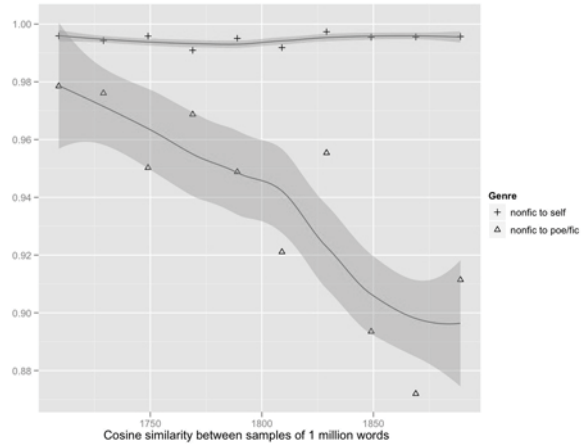


Figure 2

Second, we will begin to interpret the meaning of the divergence, by examining lists of words that became relatively more (or less) overrepresented in literary genres over this period. Briefly, we will suggest that the transformation of literary diction dramatizes a transformation of the logic of cultural capital. In the eighteenth century, literary diction explicitly thematized the social status associated with fine writing ('muse,' 'pomp,' 'taste,' 'applause,' 'genius,' 'merit,' 'talents.') In the nineteenth century, literary diction became more concrete, to be sure, and less explicitly learned – but it also tended to disavow the social in favor of a pure subjectivity embodied, for instance, in nouns and verbs of perception ('listened,' 'heard,' 'looked,' 'felt,' 'dream,' 'eye'). We propose that this emphasis on subjectivity and immediacy in fact dramatizes a mode of cultural capital associated with literature's autonomy from other institutions (Ross 1998). We don't pretend to have proven that hypothesis. There is room for a great deal more study and argument. We do, however, claim to have shown that the diction of poetry, fiction, and nonfiction prose differentiate from each other over the period 1700-1899 – a puzzle that will need *some* kind of explanation. We offer the puzzle itself as an example of the way text-mining is already beginning to shape debates in literary history and critical theory.

Acknowledgments

The arguments and methods in this paper have been deeply shaped by conversation with Natalia Cecire, Tanya Clement, Katherine Harris, Ryan Heuser, Natalie Houston, Matt Jockers, Benjamin Schmidt, John Unsworth, and Scott Weingart.

Funding

This work was supported by the Andrew W. Mellon Foundation grant, *Expanding SEASR Services*.

References

Bar-Ilan, L., and R. A. Berman (2007). Developing register differentiation: the Latinate-Germanic divide in English. *Linguistics* 45: 1-35.

DeForest, M., and E. Johnson (2001).The Density of Latinate Words in the Speeches of Jane Austen's Characters. *Literary and Linguistic Computing* 16: 389-401.

Lasko, T. A., and S. E. Hauser (2002). Approximate String Matching Algorithms for Limited-Vocabulary OCR Output Correction. *US National Library of Medicine* `http://archive.nlm.nih.gov/pubs/hauser/Tompaper/tompaper.php`

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics* 6: 97-133.

Ross, T. (1998). *The Making of the English Literary Canon*. Montreal: McGill-Queen's UP.

Wordsworth, W. (1798). Advertisement. *Lyrical Ballads, with a Few Other Poems,* by William Wordsworth and S. T. Coleridge. Bristol: Biggs & Cottle.

Digital editions with eLaborate: from practice to theory

van Dalen-Oskam, Karina
karina.van.dalen@huygens.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Dutch Academy of Arts and Sciences, The Netherlands

van Zundert, Joris Job
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Dutch Academy of Arts and Sciences, The Netherlands

1. Introduction

The aim of this paper is to show how the development and use of a web based scholarly environment for the creation of digital editions leads to meta reflection and further theorizing on what a digital edition could or should be. This effect of deploying and using virtual environments in turn may lead to improved versions of such working environments or to improvements to other tools facilitating the creation of digital editions.

However creative and inventive scholars may be, it is extremely difficult to contemplate all possibilities and consequences of digital editions without actual hands on experience in a digital environment in which such editions can be created and experimented with. For this and for other reasons, *eLaborate* was developed by the Huygens Institute for the History of the Netherlands, an institute of the Royal Netherlands Academy of Arts and Sciences. eLaborate is an online research environment in which scholars can create an edition of a text, individually or in a group. They can upload scans of the manuscript or printed book they want to edit, transcribe the text using a zooming and panning tool, write annotations to the text, categorize these according to their own needs, and publish the edition online with minimal interference of IT developers. eLaborate version 3 was released in the Fall of 2011. eLaborate3 is built on a new IT architecture, because the wishes and needs of users of eLaborate1 (2005) and eLaborate2 (2009) could not be accommodated in the technological framework that was chosen as the base at the start of the first project in 2003.

2. How eLaborate changes editorial practice

For more information about the changing organizational practices that arise in a collaboratively made edition, we refer to (Beaulieu, van Dalen-Oskam & van Zundert [to be published] and Pierazzo 2008). In this paper, we will focus on the edition itself, taking as the subject of our case one already published online digital edition made with eLaborate2.

The example we take is Marjolein Hogenbirk's edition of the Middle Dutch Arthurian romance *Walewein ende Keye* ('Gawain and Kay'), published online at `http://www.waleweinendekeye.huygens.knaw.nl/` (Hogenbirk 2009). The most prominent representation of the edited text is per column in the manuscript. See for example 182ra (`http://www.waleweinendekeye.huygens.knaw.nl/path/editie/kolom/182r/182ra`). eLaborate shows the facsimile of this column, the transcription, and at the far right all the annotations to the transcription are listed. The user can adjust the placement of the facsimile in its panel and zoom in, change the width of the panels (cf. Fig. 1), or click panels away to give the other panels more space.

Figure 1: The eLaborate *Walewein ende Keye* edition opened on colum 182ra

The freedom of customization available to the user goes even beyond this. The key here is the 'Options' button in the top right of the screen, where the user can adjust the display of abbreviations: show these in italics, or show these in roman font (in the example they are shown in italics). Another choice concerns line numbers: the user can choose the line numbers of the current edition as well as that of an earlier edition, or choose not to see any of these (in the example only the line numbers of the current edition are shown). The other Options-tab enables the user to choose which of the annotations to see (Fig. 2). The

edition of *Walewein ende Keye* has three different annotation categories, which are marked with three different colours in the transcription panel. Hovering over them with the mouse makes the content and categorization of the specific annotation visible. All annotations are visible in the right hand panel. But the user can deselect those categories that need not be visible for now. When deselected, they immediately disappear as coloured markings in the transcription panel and from the list in the right hand panel.

Figure 2: The eLaborate *Walewein ende Keye* options tab on 'Annotatiecategorieeën' ('Annotation categories')

The editor has of course added markings of abbreviations, line numbers, and annotation categories in the work environment. This leads us to a very important new basic point of departure or principle for digital editions derived from current practice: everything that is explicitly marked in the edition should be easily made invisible on the fly by the user, whereas it is impossible for the user (as well as for the editor himself) to add non-systematical information on the fly. The occurrence of medieval abbreviations e.g. is unpredictable and therefore cannot be added by current technology correctly automatically. The same goes for the general presentation of the edition. The edition was made column for column, and chapter headings were marked. Therefore it can also be viewed chapter by chapter ('Overzicht per hoofdstuk') or as a complete text ('Volledige tekst'). Should the editor have chosen for a complete text only, the column presentation would not have been possible.

3. How practice leads to new theory

Simple as this new guideline may seem, it leads to a more theoretical reflection on the nature of digital editions in general and on its possible future(s). For example, we could also apply the principle to more fundamental levels. Based on a.o. funding possibilities, an editor of medieval texts until now had to choose for which audience he would prepare the edition. A diplomatic edition for linguistic and literary scholars, with the exact spelling as in the manuscript or print and with abbreviations marked in italics? Or a critical edition for students, with adaptation of spelling, interpunction and capitalization, and with solved abbreviations? With limited funding the last option usually seemed best. But a critical edition closes the edition down for certain kinds of research. Spelling and interpunction are changed and therefore cannot be researched anymore. This also makes it much more difficult to establish the probable date and location of writing. Such a 'closed-down' edition has only limited usability and limited influence on the general progress of knowledge. It can even introduce difficult-to-check mistakes that could do harm (cf. van Dalen-Oskam & Depuydt 1997).

Can we address both scholarly users and students in one and the same digital edition without closing down the edition? The basic approach seems to be that the text in a form as close as possible to the source is the basis of all types of representations and that almost all kinds of interpretations by the editors are kept separate from this basic text. The editor transcribes the words or signs diplomatically, for instance the word *auenture* 'adventure'. The word could receive an annotation that the critical representation of this word is *aventure* (generated automatically if the editor has run a parser for lemmatization and part-of-speech tagging, having confirmed or corrected the results). The form of the added knowledge would be very different from now, but still depends on the expertise of the editor. Ofcourse we do need a user-friendly way to add all these layers of interpretation to the base text. It is clear that the stand-off mark-up approach as advocated by a.o. (Schmidt 2010) is a necessity to make this work. And then perhaps another problem can also be addressed. What if we have several versions of a text and we want to reconstruct the version as it may have functioned at a certain point in time? If we make one edition with one reconstruction, other kinds of research would again be excluded. So we need to experiment whether we can add a separate layer to editions based on the different manuscripts, with its own annotation categories, which could generate the text in the selection and the order the scholar needs. The user could then toggle between a representation of the original sources and the reconstructed text (van Dalen-Oskam, under review).

Should this be possible, text and interpretation will be separate for ever, making sure all kinds of research are still possible on (different representations of) the same digital edition. But the role of the editor as analyzing, reconstructing, and validating knowledge

remains as firm as ever – with a much higher level of reusability.

References

Beaulieu, A., K. H. van Dalen-Oskam, and J. J. van Zundert (to be published). Between tradition and Web 2.0: eLaborate as social experiment in humanities scholarship. In T. Takseva (ed.), *Social Software and the Evolution of User Expertise: Future Trends in Knowledge Creation and Dissemination*. To be published by IGI Global.

eLaborate: eLaborate1 (2005). Not available anymore.

eLaborate: eLaborate2 (2009). http://www.e-laborate.nl/en/ .

eLaborate: eLaborate3 (2011). http://www.elaborate.huygens.knaw.nl/ .

Hogenbirk, M., ed. (2009). *Walewein ende Keye*. Een dertiende-eeuwse Arturroman overgeleverd in de *Lancelotcompilatie*. Digitale editie, bezorgd door Marjolein Hogenbirk, met medewerking van W. P. Gerritsen. Eerste editie. November 2009, online op: http://www.waleweinendekeye.huygens.knaw.nl .

Pierazzo, E. (2008). Editorial Teamwork in a Digital Environment: The Edition of the Correspondence of Giacomo Puccini. *Jahrbuch für Computerphilologie* 10, http://computerphilologie.tu-darmstadt.de/jg08/pierazzo.html .

Schmidt, D. (2010).'The inadequacy of embedded markup for cultural heritage texts. *Literary and Linguistic Computing* 25(3); 337-356.

van Dalen-Oskam, K., and K. Depuydt (1997). Lexicography and philology. In K. H. van Dalen-Oskam, K. A. C. Depuydt, W. J. J. Pijnenburg, and T. H. Schoonheim (eds), *Dictionaries of Medieval Germanic Languages. A Survey of Current Lexicographical Projects*. Turnhout: Brepols, pp. 189-197. (Selected Proceedings of the International Medieval Congress, University of Leeds, 4-7 July 1994).

van Dalen-Oskam, K. (under review). All in one, one for all? A possible world of digital editions. Under review for *Variants*.

Delta in 3D: Copyists Distinction by Scaling Burrows's Delta

van Zundert, Joris Job
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands - Royal Dutch Academy of Arts and Sciences, The Netherlands

van Dalen-Oskam, Karina
karina.van.dalen@huygens.knaw.nl
Huygens Institute for the History of the Netherlands - Royal Dutch Academy of Arts and Sciences, The Netherlands

This paper reports on our ongoing stylistiscs research and tool development in the specific case of authorship attribution for the Middle Dutch Arthurian novel *Roman van Walewein* (Romance of Gawain). First we introduce the use case, and we concisely recapture our prior research. We then progress to the description of the re-issued improved software and algorithm we developed and applied to our use case. We conclude with an overview of the first results we derived from applying our tools and approach.

The *Roman van Walewein* (ca. 1260, Middle Dutch) was written by two authors, Penninc and Vostaert. Only one manuscript containing the complete text, explicitly dated as copied in the year 1350, is left to us. Some fragments of another, probably somewhat younger manuscript contain about 400 lines. The text in the complete manuscript consists of 11,202 lines of rhyming verse. The manuscript was written by two copyists. The first seems to have written the lines 1-5.781 and the second the lines 5,782-11,202.

The second author, Vostaert, explicitly claims to have added about 3,300 lines to Penninc's text. Several scholars of Middle Dutch have pointed out on various grounds that Vostaert (or the copied text) can not be correct in this claim. We decided to apply current state of the art authorship attribution techniques to find whether this would point us to a specific line in the text where the text before that line and the text after that line contrast the most. In prior research we have used a lexical richness measure, Udney Yule's Characteristic K, and Burrows's Delta to try to determine the position in the text where Vostaert may have started to complete Penninc's text. We thus found that applying Burrow's Delta indeed gives relatively strong grounds to assume that Vostaert simply stated a truth. Figure 1 can be read as a support of this finding, showing a significant change

of stylistic stability occurring around verse line 7,882 (cf. van Dalen-Oskam & van Zundert 2007). New knowledge that we derive from this result is that Vostaert may well have reworked the last portion of his predecessor's work. However, the most surprising result was that our analysis pointed towards the possibility that Delta is a better indicator for a change of copyists than anything else (cf. the clear local maximum in the graph of figure 1, which is located around verse line 5,783 and reflects the change of hands that takes place in the manuscript).

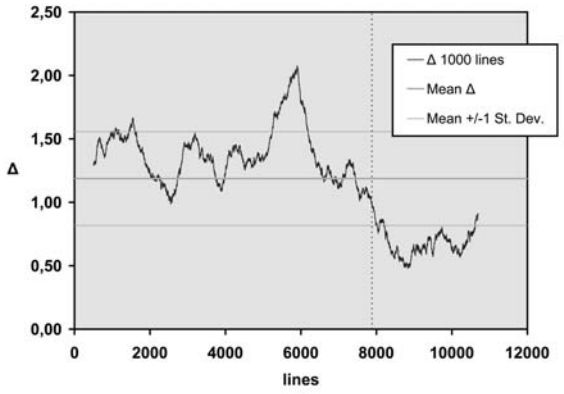


Figure 1: 'Walking' Delta. The text corresponding to lines 8,000 - 11,000 in Vostaert's part of the romance was used as a subset for the procedure. From this subset the top most 50 frequently used words were used for comparison with the text in a 'sliding window' of each consecutive 1,000 lines. Taken from (van Dalen-Oskam & van Zundert 2007)

In our research until 2007 we were only able to tentatively look into this copyist distinguishing ability of the Delta measure, and it remains to be investigated until now how effective Delta is as a procedure for copyist distinction. Although Kestemont and van Dalen-Oskam (Kestemont & van Dalen-Oskam 2009) verified the difference between the two scribes of the extent Walewein manuscript using a lazy machine learning technique on other features such as character n-grams and several simple rations, they did not apply a form of Delta procedure to this problem. In our 2007 article we stated that the current application cannot easily and automatically generate and compare a large collection of Delta graphs based on an iterative subselection of a word frequency list' (van Dalen-Oskam & van Zundert 2007: 360). This is the point where we resume our research endeavor.

We have further developed our 'sliding window' approach to measuring the stylistic consistency of texts (i.e. the ability to establish the stability of the value of Delta throughout a single text). We will release the MIT licensed open source Ruby code for this version of the software to Github in June 2012. This version will include a REST-full interface to accommodate polling this tool as a web enabled programmatically negotiable service, a web GUI for human-computer based interaction with the tool, and user documentation as well as format description specifics required for files to be parsed with this tool.

Our analyses until 2007 relied on stepping a 'window' through a text, computing the Delta for that window of text using the complete text as a statistical background and a sizeable portion of the part unanimously attributed to Vostaert as a Delta comparison base. The graphs thus computed showed the consistency (or non consistency) of Delta throughout the text tested. Having improved scalability and performance of the algorithm we are now able to run several Delta measuring windows through a text at higher speeds, enabling near instant analysis (typically <5s analysis time for a 65k tokens sized text, analyzed through an approximately 8k token sized window).

This improved approach has enabled us to return to the research question at hand: how well does Delta discriminate copyists? With the current software we are able to determine how stable the characteristic is for Delta that we found in our first computations (see Fig. 1.) throughout the word frequency spectrum of the text. Our algorithm computes the Delta characteristic of a text in an iterative subselection of the word frequency list. For instance, it computes this characteristic first in the top 50 highest frequency spectrum, then it does the same for the top 10 to 60 highest frequencies, then for the top 20 to 70 highest frequencies, then 30–80, 40–90, etc... The size and 'pace' of the frequency subset window can in fact be chosen arbitrarily, though sensible heuristics need to be applied to yield meaningful results. In essence then, the current algorithm not only slides a window for Delta measurement through the text, but also slides a measurement window through the frequency spectrum of the vocabulary of the text. In the case of *Walewein* this yields typically a result as depicted in Fig. 2.

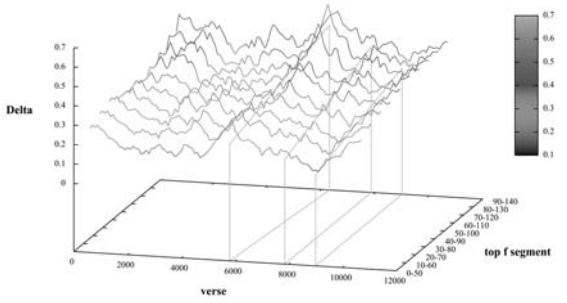


Figure 2: Sliding the 'walking' Delta through a consecutive subset of the full vocabulary frequency spectrum of the text of Walewein

There are several conclusions we can draw from this result. The general characteristic we found in our earlier attempts to establish the stylistic consistency throughout the text of *Walewein* holds for any subspectrum of high frequency vocabulary we choose. We conclude that the reproduceability of this characteristic in different frequency spectra, means that indeed this is a viable method of tracing stylistic consistency in a text. We also note that the value for Delta on average rises slightly when our window on the subset of high frequency word usage descends through the full spectrum of vocabulary frequencies. This in itself confirms that Delta is a valid indicator for stylistic similarity, or at least for similarity in high frequency word usage. But more important: because the peak of stylistic inconsistency around verse 5,782 is stable our findings also confirm our tentative assumption that 'semi-high' frequency word usage is a better indicator for copyist change than high frequency word use – at least in the specific case of the *Walewein*. We can in the case of *Walewein* also infer which part of the high frequency word use spectrum is the best indicator of copyist change: the 80 to 130 most frequently used words. Discussing the lexical characteristics of this part of the vocabulary as well as what this may indicate about copyist behavior will be part of our paper.

Further testing on known use cases of copyist change need to establish whether this result is generalizable. If so, we may have found in this particular application of the Delta procedure an adequate identification procedure for copyist change.

References

Burrows, J. F. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing* 22: 27-48.

Burrows, J. F. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17: 267-287.

Hoover, D. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing* 19: 453-475.

Van Dalen-Oskam, K. H., and J. J.van Zundert (2007). Delta for Middle Dutch: Author and copyist distinction in "Walewein". *Literary and Linguistic Computing* 22: 345-362 (First published June 2, 2007: 10.1093/llc/fqm012).

Kestemont, M., and K. H. van Dalen-Oskam (2009). Predicting the Past: Memory-Based Copyist and Author Discrimination in Medieval Epics. *Proceedings of the twenty-first Benelux conference on artificial Intelligence (BNAIC, 2009),* Eindhoven, The Netherlands, pp. 121-128.

Smith, P. W. H., and W. Aldridge (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. *Journal of Quantitative Linguistics* 18: 63-88 (First published February 24, 2011: 10.1080/09296174.2011.533591).

Stein, S., and S. Argamon (2006). A Mathematical Explanation of Burrows's Delta. *Proceedings of Digital Humanities 2006*, Paris, France, July 2006, p. 207 < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.8771&rep=rep1&type=pdf > (accessed 14 March 2012).

Johnson, D. F., and G. H. M. Claassens (2000). *Dutch Romances I: Roman van Walewein*. Cambridge: D.S. Brewer.

Wiki Technologies for Semantic Publication of Old Russian Charters

Varfolomeyev, Aleksey
avarf@psu.karelia.ru
Petrozavodsk State University, Russian Federation

Ivanovs, Aleksandrs
aleksandrs.ivanovs@du.lv
Daugavpils University, Latvia

The term 'semantic publication' denotes an electronic text publication that is provided with additional information layers, which represent knowledge about the text in a formalized way suitable for automatic processing. In the modern Web environments, semantic publications, especially in digital libraries and electronic journals, have become quite topical (Baruzzo et al. 2009; Shotton et al. 2009; de Waard 2010). Their advantages are rather obvious. Firstly, semantic publications provide better facilities for searching for information, since such publications draw together search algorithms used by humans and computers. For instance, if in an ontology is mentioned that a definite term has a synonym, searching for the term and its synonym can be performed simultaneously; and it is not necessary to mention the synonym in the request. Secondly, since formalized knowledge can generate new knowledge, semantic publications can be used as a knowledge base in order to advance hypotheses for further research by means of automatic inference.

It seems that semantic publications of historical records have one more advantage different from the above-mentioned common advantages: semantic layers in the publications of historical records reflect researchers' interpretations, which can be verified by means of formalized, computer-based procedures.

In order to reveal the advantages of semantic editions in research of medieval documentary records, this paper presents a prototype of the semantic publication of the 13th century Old Russian charter corpus, which forms a constituent part of the vast collection of medieval and early modern records 'Moscowitica–Ruthenica' kept in the Latvian State Historical Archives, a structural unit of the Latvian National Archives (Riga, Latvia). This collection of documents provides historians with firsthand information about relations of Old Russian and Byelorussian lands and towns (Smolensk, Novgorod, Pskov, Polotsk, etc.), as well as Lithuania (later – Poland-Lithuania) with Riga, Livonia, Hanseatic League and some German towns in the late 12th – early 17th centuries. For the first time, 'Moscowitica–Ruthenica' – a historical name of this document collection – was mentioned in the archival inventories dated back to the 1630s. Although 'Moscowitica–Ruthenica' as a department of the Latvian State Historical Archives does not exist any more, its documents constitute a natural complex of historical records, which should be studied as a whole (Ivanovs & Varfolomeyev 2005).

In the prototype of the semantic publication, five interconnected charters have been used (Charters nos. 1, 3a, 4, 5, and 6, see Ivanov & Kuznetsov 2009). Actually, there should be mentioned twelve charters that reveal the course of relations between Riga and Smolensk in the 13th – first half of the 14th century, however, the presentation has its limits, therefore the basis of the prototype has been reduced. In the centre of the semantic network represented in the prototype, there is the Missive of Archbishop of Riga Johann II to Fedor Rostislavich, Prince of Smolensk, blaming inhabitants of Vitebsk for unjustified complaint against Rigans (Charter no. 6, 1285–1287).

In order to provide the texts with additional descriptive metadata, information about persons, sites, documents, etc. mentioned in the charters is revealed and linked with the corresponding data extracted from different specialized ontologies. In the last years, a great number of different ontologies have been created, including those intended for historical and source studies, e.g. ontology CIDOC CRM. In this ontology, there are classes and relations that can be used in description of historical persons, sites, and historical events, which are related to museum objects (Doerr 2003). On the basis of this ontology, a number of specialized ontologies for description of definite historical aspects have been elaborated. There can be mentioned ontologies created within Pearl Harbor Project in the USA (Ide & Woolner 2007) or CultureSampo Project in Finland (Ahonen & Hyvönen 2009). Unfortunately, such ontologies can not meet all the requirements of the semantic publication of charter corpora, since they do not reflect the specificity of written historical records. Therefore, the authors of the paper propose a document-oriented approach to creation of ontologies (in contrast with event-oriented approaches accepted in the above-mentioned ontologies).

The semantic publication of the charters constitutes two kinds of semantic links. First, there are links between historically and thematically interconnected charters (these interconnections emerged when the charters were drawn up, in the course of documenting of relations between Smolensk and Riga in the 13th century). Thus, diverse links between information reflected in the charters within

this complex of historical records can be revealed. Second, there are links with other historical records, which do not belong to the complex 'Moscowitica–Ruthenica'. In this case, information extracted from the complex of charters 'Moscowitica–Ruthenica' (it can be called 'internal information') is linked with 'external' information, provided either directly (by other historical records), or indirectly (by research papers, specialized ontologies, etc.).

Within the semantic publication, relations between its objects are described using triplets: 'a charter is written by a person', 'a charter is sent to a person', 'a charter mentions a person'. As it is commonly done in different ontologies, inverse relations can also be introduced, e.g. 'a person is mentioned in a charter'. It should be noted that this publication is partly based on hypothetical data; hypothetic nature of some relations may be reflected using definite combinations of words ('probably refers to' instead of 'refers to').

However, production of semantic publications on the basis of ontologies, which are recorded using Semantic Web technologies – RDF or OWL, is time-consuming. It seems that opportunities and tools provided by semantic Wiki-systems can facilitate this process. For instance, Semantic MediaWiki (Krötzsch et al. 2007) offers specialized, rather simple markup tools that can be used to indicate different objects (place-names, persons' names, etc.) in the texts of the charters and to supply the texts with meta-information. The principle feature of this system is the use of typified hyperlinks between pages. These pages constitute the objects of the semantic network, but hyperlinks – denote relations between the objects.

On the site http://histdocs.referata.com, the text of the Charter no. 6 is presented. The text has been translated from Old Russian into English and published applying Semantic MediaWiki. In the published text, hyperlinks to other pages of this semantic network have been marked out. These pages contain texts of the charters of the complex 'Moscowitica–Ruthenica', as well as data related to historical persons, places, etc. Below the text, the facts related to a definite charter are mentioned (e.g. 'Charter 6 mentions Helmich', 'Charter 6 probably refers to Charter 4', etc.) These facts are linked with the text, and this linkage is based on researcher's interpretation of the document.

It should be noted that within the semantic network different facts about the objects, which are recorded by means of Semantic MediaWiki tools, can be automatically transformed into RDF triplets. Therefore, Wiki-systems can be used for production of semantic publications of charters and other written documents.

However, some shortcomings of Wiki-systems can be mentioned. For example, non-standard fonts can not be used in transcription of the texts; the texts of the charters can not be linked with raster images of the documents; the texts can not be marked up on the basis of XML markup standard (e.g. in accordance with TEI or CEI markup schemes). Therefore, a specialized Wiki-system for charters' editing purposes should be developed. In the presentation of the paper, some possible solutions to the problems mentioned above are examined.

References

Ahonen, E., and E. Hyvönen (2009). Publishing Historical Texts on the Semantic Web – A Case Study. *Proceedings of the Third IEEE International Conference on Semantic Computing (ICSC2009)*. Berkeley, pp. 167-73.

Baruzzo, A., et al. (2009). Toward Semantic Digital Libraries: Exploiting Web2.0 and Semantic Services in Cultural Heritage *Journal of Digital Information* 10(6). `http://journals.tdl.org/jodi/article/viewArticle/688/576` (accessed 14 March 2012).

de Waard, A. (2010). From Proteins to Fairytales: Directions in Semantic Publishing. *IEEE Intelligent Systems* 25(2): 83-88.

Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata *AI Magazine* 24(3): 75-92.

Ide, N., and D. Woolner (2007). Historical Ontologies. In K. Ahmad, C. Brewster and M. Stevenson (eds.), *Words and Intelligence II: Essays in Honor of Yorick Wilks*. Dordrecht: Springer, pp. 137-52.

Ivanov, A., and A. Kuznetsov (2009). *Smolensko-rizhskie akty, XIIIv. – pervaia polovina XIVv.: Dokumenty kompleksa Moscowitica–Ruthenica ob otnosheniiakh Smolenska i Rigi* [Treaties between Smolensk and Riga: 13th – First Half of the 14th Century: Documents of the Complex Moscowitica–Ruthenica about Relations between Smolensk and Riga]. Riga.

Ivanovs, A., and A. Varfolomeyev (2005). Editing and Exploratory Analysis of Medieval Documents by Means of XML Technologies. In *Humanities, Computers and Cultural Heritage*. Amsterdam: KNAW, pp. 155-60.

Krötzsch, M., et al. (2007). Semantic Wikipedia. *Journal of Web Semantics* 5(4): 251–61.

Shotton, D., et al. (2009). Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a

Research Article. *PLoS Computational Biology* 5(4): e1000361. http://dx.doi.org/10.1371/journal.pcbi.1000361 (accessed 14 March 2012).

L'histoire de l'art à l'ère numérique – Pour une historiographie médiologique

Welger-Barboza, Corinne
corinne.welger@sfr.fr
université Paris 1 Panthéon-Sorbonne, France

Le tournant numérique en Histoire de l'art se prend plus difficilement que dans les autres disciplines des Humanités. L'une des raisons est la transparence dans laquelle ont longtemps été tenus les instruments de l'étude et de la recherche. Et bien que l'image soit considérée comme la pierre angulaire de la méthodologie de la discipline, ses dimensions de médiation et de médiatisation sont rarement abordées, les dispositifs complexes auxquels les images participent, pas davantage. Mais on constate que le tournant numérique donne une nouvelle impulsion aux études qui s'attachent à penser la technologie de la discipline, à ses différentes époques. Plus encore que la quantité, les orientations de ce regain de curiosité en font l'intérêt.

La clarification des conditions d'exercice de l'enseignement, de l'étude et de la recherche, à l'ère du numérique, nécessite d'étudier les conditions passées (techniques, supports et modes de circulation des images des oeuvres). En d'autres termes, je fais l'hypothèse que le tournant numérique requiert l'élaboration d'une historiographie médiologique de l'histoire de l'art afin d'appréhender les singularités d'une pratique numérique des images ainsi que son impact sur cette discipline.

1. Les tendances d'une historiographie médiologique

Pour ce qui concerne le régime photographique de la reproduction, en partie associé à la culture imprimée, on dispose déjà d'une historiographie conséquente : d'abord, les écrits sources relatant les usages de la technique et les réflexions épistémologiques des grandes figures de la discipline des 19ème et début 20ème siècles sont désormais restituées ;un balisage rendu aisé par W. Freitag (1979-1980), 1998). Quant aux mérites du nouveau support, au regard des différentes techniques de gravure, l'ouvrage de William M. Ivins (1953) fait date. Le fait que de la photographie accompagne l'établissement de la discipline académique et de ses méthodes, est unanimement reconnu.

Pour une bonne part, ces écrits sont traversés par les thématiques de l'exactitude et de l'interprétation, – le thème de la fidélité au référent ou à 'l'esprit de l'artiste' – en écho aux débats conflictuels entre graveurs, lithographes et photographes de la seconde moitié du 19ème (Renie 2001). Au moment du tournant numérique, on retrouve cette perméabilité entre les préoccupations de l'histoire de l'art et des démarches relevant de l'esthétique ou de la théorie de l'art – le concept d'aura de Benjamin (1935-1939) dont on peut douter qu'il concerne directement les méthodes de la discipline. L'appropriation du procédé technique par les historiens de l'art du 19ème et du 20ème siècle, est elle-même intriquée à leur 'vision' de l'art (Wölfflin 1896, 1897, 1915).

Mais des travaux plus récents visent à saisir une plus grande complexité médiologique: à ce titre, on citera la notion 'd'économie visuelle' qui, outre les rapports entre les moyens de reproduction coexistant à une époque donnée (2nde moitié du 19ème siècle), leurs coût et efficacité, rend compte des modes de réception (Bann 2001); ou encore, les quatre colloques *Photo archives and the Photographic Memory of Art History*, qui abordent, pour la première fois fois, le fait archivistique même dans toutes ses dimensions, au-delà de l'aptitude du support à rendre compte de son référent (Caraffa 2011).

Ces déplacements de la recherche historiographique concernent également l'étude des périodes antérieures et manifestent que ce mouvement de réflexivité excède le régime documentaire que l'on est en train de quitter. Des études récentes portent sur les 'musées de papier', des recueils d'antiquaires jusqu'aux sommes illustrées d'histoire de l'art et soulèvent des enjeux épistémologiques associés à l'environnement médiologique.

2. De l'expérience à la culture visuelles

Ainsi, on tente de restituer certains aspects de la pratique des images. Les collections de dessins et de gravures ne donnent que partiellement lieu à des volumes reliés et laissent les feuilles libres; aussi, par exemple, le *Museo Cartaceo*, premier recueil du genre, favorise-t-il l'examen visuel, '*ispezione oculare*' en donnant lieu à des manipulations diverses de confrontation des images, engageant la combinaison du travail de l'œil et du geste (Bickendorf 2010).

L'agencement même des images représentant les œuvres fait l'objet d'attentions nouvelles. Ainsi, par exemple, les planches comparatives de Séroux d'Agincourt (1810-1823), mettent en oeuvre un programme didactique voué à démontrer la corruption du goût antique, la décadence du Moyen Age (Mondini 2005).

Ces travaux partagent une même ambition épistémologique : la réhabilitation d'une histoire de l'art par l'image, à même de réévaluer le poids respectif du visuel et du discursif dans l'historiographie de la discipline. Plus amplement, ces musées de papier s'inscrivent dans une enquête épistémologique qui associe arts et sciences, dans les transformations de la culture visuelle au 17ème siècle, époque d'instrumentation de la vision par le télescope et le microscope.

La 'fabrique du regard' (Sicard 1998) est illustrée par une première figuration d'objets observés au-delà du visible à l'œil nu. La fameuse abeille grossie, réalisée par Francesco Stelluti, en 1625, devient un repère signifiant (Bickendorf 2010). Cette innovation participe de la culture visuelle en gestation dans laquelle baigne Cassiano del Pozzo.

C'est encore la fabrique d'un nouveau regard qui soutient la constitution de *La République de l'œil*, (Griener 2010). L'auteur vise à rassembler les éléments matériels et immatériels, intellectuels, cognitifs, de la culture visuelle du 18ème siècle. L'*instrumentarium* de l'historien de l'art s'y compose aussi bien des techniques de reproduction des œuvres (dessins, gravures) que des régimes de perception et de mémorisation qui s'instaurent. Les travaux de Locke aussi bien que la dépose des pigments d'un support à un autre forment les conditions de possibilité de la 'lecture des œuvres'; l'image se détache ainsi de l'oeuvre et prend forme dans un espace mental qui fonctionne en corrélation entre les espaces d'exposition (Salons, musées) et celui du livre.

Ces approches ont pour point commun de réinsérer le livre d'images dans un dispositif complexe: contexte intellectuel, pratiques cognitives ou sociabilité des acteurs impliqués.

3. En écho avec les préoccupations actuelles de l'histoire de l'art numérique

L'étude des remédiations successives de l'image recèle un potentiel heuristique. Ainsi, la multiplicité des montages numériques d'images fait pièce à la réduction méthodologique liée au photographique (modèle binaire de la double projection, standards de l'imagerie imprimée); la liberté des agencements dans les livres anciens peut inspirer.

Par exemple, la grille comme organisation structurelle d'images montre sa permanence. (Fig. 1) Mais alors que dans le régime imprimé, la grille comparative soutient une proposition didactique

achevée, (Cf. supra), avec le numérique, elle favorise la prévisualisation de corpus d'images (Welger-Barboza 2011) (Fig. 2) Ici, l'agencement est indissociable de la notion d'agentivité ; c'est en termes d'affordance des interfaces que l'on doit analyser les conditions sémiotiques et cognitives de la manipulation des images, de l'action conjuguée de l'œil et du geste (Rueckert, Radzikowska & Sinclair 2011).

Figure 1: Jean-Baptiste Louis Georges Séroux d'Agincourt (Vol. IV, planche LXIV : Tableau historique et chronologique des frontispices des temples, avant et durant la décadence de l'art

Figure 2: Grille de vignettes –
Gestionnaire personnel d'images (Picasa)

Autre changement d'envergure, le cadre d'expérience de la pratique visuelle contemporaine, est lié aux dimensions variables du dispositif numérique [navigateur, bases d'images et éditions numériques produites par d'autres opérateurs, individuels ou institutionnels]. Les partitions entre les différentes situations de travail et de communication (personnel/collectif, privé/public, institutionnel/ non-institutionnel), dans l'accès et l'utilisation des images, sont fluides, les délimitations poreuses. La culture numérique en gestation, avec ses outils et ses pratiques, forme le cadre de l'expérience visuelle pour la discipline, en transforme les méthodes comme les objets, la culture académique aussi bien que les modes de socialisation.

Dans la brève histoire de l'art numérique, des déplacements font pendant à ceux que l'on remarque dans l'historiographie médiologique de l'histoire de l'histoire de l'art. Les premières réflexions se sont naturellement portées sur les propriétés du document numérique versus le document photographique, sur l'impact de l'accessibilité de grands corpus d'images, sur la réactualisation du Musée imaginaire étendu à la 3D – Piero Project (1994, Princeton) est emblématique. Désormais, l'approche de la complexité du dispositif numérique peut se nourrir de l'étude des supports du passé de l'histoire de l'art.

References

Bann, S. (2001). Photographie et reproduction gravée. *Études photographiques*, [En ligne], mis en ligne le 10 septembre 2008. URL: `http://etudesphotographiques.revues.org/index241.html` Consulté le 12 mars 2012.

Benjamin, W. (1935-1939). L'œuvre d'art à l'époque de sa reproductibilité technique. *Œuvres* T. III. Paris: Gallimard 2000.

Bickendorf, G. (2004). Dans l'ombre de Winckelmann : l'histoire de l'art dans la 'république internationale des Lettres' au 18ème siècle. *Revue de l'Art* n°146/2004-4, *L'histoire de l'histoire de l'art*, pp. 7-20.

Bickendorf, G. (2010). Musées de papier – Musées de l'art, des sciences et de l'histoire. In *Musées de papier*: *l'Antiquité en livres, 1600-1800*. Paris: Le Louvre/Gourcuff Gradenigo – Catalogue d'exposition, Musée du Louvre, 25-09-2010 au 03-01-2011

Caraffa, C., ed. (2011). *Photo Archives and the Photographic Memory of Art History*. Berlin and Munich: Deutscher Kunstverlag.

CIHA London (2000). Thirtieth International Congress of the History of Art, *Art History for the Millenium: Time.,* Section 23, Digital Art History Time, London, 3-8 September 2000. AHWA

Archive en ligne: http://www.unites.uqam.ca/AHWA/Meetings/2000.CIHA/index.html , Consulté le 12/03/12.

Decultot, E. (2010). Genèse d'une histoire de l'art par les images – Les recueils d'antiquités et la naissance du discours historique sur l'art, 1600-1800. In *Musées de papier*: *l'Antiquité en livres, 1600-1800*. Paris: Le Louvre/Gourcuff Gradenigo – Catalogue d'exposition, Musée du Louvre, 25-09-2010 au 03-01-2011.

Falguieres, P. (1996). Les Raisons du catalogue. *Du Catalogue, Cahiers du musée national d'art moderne* nº 56/57, pp. 5-20.

Freitag, W. M. (1979-1980). Early Uses of Photography in the History of Art. *Art Journal* 39(2): 117-123.

Freitag, W. M. 1(998). Histoire de l'art et nouvelles techniques de reproduction. In *Histoire de l'histoire de l'art, Tome II. XVIIIe et XIXe siècles, (Dir. Edouard Pommier) Actes du colloque du musée du Louvre. (1994-1995)*. Paris: Klincksieck/Musée du Louvre.

Greenhalgh, M. (2004). *Art History, A Compagnion to Digital Humanities*. Oxford: Blackwell http://www.digitalhumanities.org/companion , Consulté le 12/03/12.

Griener, P. (2010). *La République de l'œil – L'expérience de l'art au siècle des Lumières*. Paris: Ed. Odile Jacob.

Haskell, F. (1987). *La difficile naissance du livre d'art*. Paris: RMN 1992.

Ivins, W. M. (1969 [1953]). *Prints and Visual Communication*. Cambridge, Mass.: MIT Press, 1969 (1953). (Londres: Routledge & Kegan Paul).

Mondini, D. (2005). *Mittelalter im Bild. Séroux d'Agincourt und die Kunsthistoriographie um 1800*. Zürich: Zurich InterPublishers.

Recht, R. (1996). La Mise en ordre, Note sur l'histoire du catalogue. *Cahiers du musée national d'art moderne* nº 56/57, pp. 21-35.

Renie, P. L. (2001). Guerre commerciale, bataille esthétique: la reproduction des œuvres d'art par l'estampe et la photographie, 1850-1880. V. Goudinoux et M. Weemans (dir.), *Reproductibilité et irreproductibilité de l'œuvre d'art*. Bruxelles: La Lettre volée, pp. 59-81.

Ruecker, S., Radzikowska, M., and S. Sinclair (2011). *Visual Interface Desing for Digital Cultural Heritage – A guide ro Rich-Prospect Browsing*. Farnham: Ashgate.

Seroux d'Agincourt, J. B. L. G. (1810-1823). *Histoire de l'art par les monumens, depuis sa décadence au IVe siècle jusqu'à son renouvellement au XVIe*. Paris: Treuttel et Würtz, 6 vol.

Sicard, M. (1998). *La fabrique du regard (XV-XXe siècles)*.Paris: Ed. Odile Jacob.

Welger-Barboza, C. (2011). Pratiques numériques de l'image. *Observatoire critique, Etude des resources numériques pour l'histoire de l'art* http://observatoire-critique.hypotheses.org/

Wölfflin, H. (1896, 1897, 1915). *Comment photographier les sculptures*. Présentation, traduction et notes par Jean-Claude Chirollet. Paris: L'Harmattan 2008.

Benefits of tools and applications for a digitized analysis of Chinese Buddhist inscriptions

Wenzel, Claudia
claudiawen@googlemail.com
Heidelberger Akademie der Wissenschaften, Germany

The research project 'Buddhist stone scriptures in China' hosted at the Heidelberg Academy of Sciences and Humanities and headed by Prof. Lothar Ledderose is devoted to the documentation of holy Buddhist scriptures that were carved onto rocks, boulders and stones from the sixth to the eighth century in China. The digital processing and presentation of these stone carved texts offers remarkable benefits with regard to facilitation of the workflow, incorporation of spatial relations, and transparency.

1. Application of digital tools for enhancing the workflow of transcribing and editing the texts to establish the corpus of carved stone sutras

Traditionally, pre-modern Chinese texts engraved on durable materials like metal or stone were first 'copied' by brushing a moistened sheet of paper into the cavities of the carved surface, and applying black ink on all outstanding parts, leaving the traces of the chisel in white on the paper. These transportable rubbings were collected from the eighteenth century on by Chinese epigraphers, who studied their ancient scripts and transcribed them for further publication. However, the rubbings are in no way a faithful copy of what was originally written into stone. The very process of taking them from the stone cut characters already involves a reading and interpretation of less well preserved glyphs.

To avoid such flaws and short-comings of the traditional Chinese procedure of text documentation, our project relies solely on the carved text to provide the lemmas for the corpus of stone sutras. The original stone is documented either by photographs alone, or by both scans and photographs. Then it is collated with several text witnesses, including different versions of rubbings, and different versions of the text extant either in manuscripts, or printed editions of the Buddhist canon, as well as the digital edition of the Buddhist canon, edited by the Chinese Buddhist Electronic Text Association (CBETA). As a result, the original stone text can be transcribed by reconstructing lost or damaged characters, and can be discussed in text critical annotations.

Digital research and publication helps making this editing process more transparent and replicable. The web application documents each inscription on five levels by providing (1) a scan or photograph of the original stone, (2) one or several paper rubbings of the stone, (3) a transcription in Unicode characters reflecting the original layout of characters in vertical columns on the stone, (4) a transcription in Unicode characters with modern punctuation and reading direction in horizontal lines from left to right, and (5) an English translation of the text including text critical annotations. In this way, text editing is significantly enhanced in terms of correctness as well as transparency, since the user can follow the transcribing of each stone character step by step.

The basis for the transcription process is the documentation of the original stone, which is done ideally by 3D laser scans. After processing, the inscriptions can be viewed either in 2 D or 3 D view by means of an applet developed for the project by the i3 Mainz (University of Applied Sciences, Mainz).[1] This applet was already implemented in the workflow of reading and transcribing the text in modern Unicode characters, but it is also available online. It provides various options like zooming, adjusting the light position, or applying colors for better legibility of the carved characters, thereby empowering users to not only better comprehend the editing process, but also to continue research by themselves. The use of this digital tool exceeds by far the possibilities of traditional print media.

Another main challenge of the edition of the medieval stone sutras is their transcription in modern Unicode characters. Many of the ancient carved texts use glyphs not yet encoded in the Unicode standard. Although more Chinese ideographs are being added currently to the Unihan database, the problems of mapping ancient glyphs to Unicode characters are complex, and cannot be solved completely by encoding more glyphs. By programming and implementing a 'snippet tool' that is able to search and line up images of the same character connected to a certain Unicode character, we not only provide a tool for an analysis of writing styles, but also bring to light the complexity of the issue of mapping glyphs to unified characters.

2. Incorporation of spatial relations

For a comprehensive understanding of the historical background behind the creation of stone carved sutras in China, each inscription site has to be seen in its particular topography in which it is embedded. The choice of the site is of religious importance, and infers particular ritual uses. Inter-spatial relations between multiple sites are also of interest. [2] For this reason, all inscriptions and sites were GIS documented and can now be addressed on interactive maps.[3] In cooperation with the Institute of Geography of Heidelberg University, more tools for a GIS based analysis have been developed to clarify the spatial history of the creation of inscription sites, the related spreading of Buddhist sutras, and the migration movements of the monks responsible for it.[4]

Right from the start, the project has generated 3D models of larger inscription sites that have been integrated into the project's web-application. At first, we made use of these models to provide map-like views of the general layout of inscriptions at particular sites. In one case, the course of a pilgrim's path was simulated in a flash animation. More recently, the visualization with 3D models has become even more rewarding for a particular type of inscription sites, namely those hosting stone sutras inside caves. Now the user can find the inscription site on a map, zoom in, select a particular cave, and by means of a panoramic view, approach the inscriptions inside the digital cave and read them herself by means of high-resolution photographs. In this respect, digital space is not only a synthesis of traditional documentation with rubbings, photographs, and printing, but offers a new level of close-to- reality documentation.[5]

3. (Future[6]) digital publication of research results and their transparency for the community

All data of the project are stored in an open source XML database called eXist (exist-db.org); the mark-up of texts follows **TEI** (Text Encoding Initiative) standards. On the basis of the same set of data, a web application as well as a print publication is generated. For the latter, the data is being transformed by XSLT. The digital as well as the print publication are set up similarly. The scheme of documenting each inscription on the aforementioned five levels pervades the digital and the print publication. While the purpose of this structure is to provide maximum transparency about the editing process for the reader/user, the electronic publication is in this respect even more efficient. By integrating the aforementioned digital tools (Mainz applet, snippet-tool, GIS analysis tools) into the project's web-application, the (future) user is enabled to replicate the editing of stone scriptures, and to continue research herself.

Funding

This work was supported by the Heidelberg Academy of Sciences and Humanities (Heidelberger Akademie der Wissenschaften), Germany. The development of digital tools mentioned was supported by the German Ministry of Education and Research (BMBF) within a program on interdisciplinary research between natural and social sciences, namely the research project '3D-Sutras: A web based atlas of laser scanned Buddhist stone inscriptions in China,' co-operated by the Heidelberg Academy of Sciences, the Department of Geography of Heidelberg University, and the Institute for Spatial Information and Surveying Technology, i3mainz, of the University of Applied Sciences in Mainz.

References

Auer, M., B. Höfle, S. Lanig, A. Schilling, and A. Zipf (2011). 3D-Sutras: A web based atlas of laser scanned Buddhist stone inscriptions in China. *14th AGILE International Conference on Geographic Information Science*, Utrecht, The Netherlands 18-21 April 2011.

Lanig, S., A. Schilling, M. Auer, B. Höfle, N. Billen, and A. Zipf (2011a). Interoperable integration of high precision 3D laser data and large scale geoanalysis in a SDI for Sutra inscriptions in Sichuan (China). *Geoinformatik 2011– Geochange*, Münster, Germany.

Lanig, S., B. Höfle, M. Auer, A. Schilling, H. Deierling, and A. Zipf (2011b). Geodateninfrastrukturen im historisch-geographischen Kontext – Buddhistische Steinschriften in der Provinz Sichuan/China. AGIT 2011. *Symposium Angewandte Geoinformatik*, Salzburg, Österreich.

Ledderose, L. (2012, in print) Fünf Perspektiven auf Steinerne Sutren. In I. Reichle and V. Lepper (eds.), *Perspektiven*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften 2012.

Schmidt, N., R. Schütze, and F. Boochs (2010a). 3D-Sutra – Interactive Analysis Tool for a Web-Atlas of Scanned Sutra Inscriptions in China. *Proceedings of the ISPRS Commissi on V Mid-Term Symposium Close Range Image Measurement Techniques*, Newcastle upon Tyne, United Kingdom.

Schmidt, N., F. Boochs, and R. Schütze (2010b). Capture and Processing of High Resolution 3D-Data of Sutra Inscriptions in China. http://www.springerlink.com/content/978-3-642-16872-7/ Digital Heritage, http://www.springerlink.com/content/0302-9743/ Lecture Notes in Computer Science 6436: 125-139, DOI: 10.1007/978-3-642-16873-4_10, Limassol, Zypern

Tsai, S., and C. Wenzel (2009). The Stone Inscriptions of the Six Mountains of Zoucheng. Die Steinschriften der Sechs Berge von Zoucheng. In Museum für Ostasiatische Kunst Köln und Forschungsstelle der Heidelberger Akademie der Wissenschaften, Buddhistische Steinschriften in Nord-China (eds), *The Centenary of the Museum of East Asian Art in Cologne. The Heart of Enlightenment. Buddhist Art in China 550-600. 100 Jahre Museum für Ostasiatische Kunst in Köln. Das Herz der Erleuchtung. Buddhistische Kunst in China 550-600*. Köln, pp. 24-37.

Notes

1. Schmidt et al. 2010b: 3-9; Schmidt et al. 2010a: 2-5.
2. Tsai & Wenzel 2009: 24.
3. Auer et al. 2011: 2-5; Lanig et al. 2011 b: 3-4.
4. Lanig et al. 2011a: 4-5; Lanig et al. 2011b: 7-8.
5. Ledderose 2012.
6. According to the agreement of the Chinese-German co-operation project, a print publication has first priority; the digital publication has to follow afterwards. For this reason, the project's web-application at http://www.stonesutras.org is to date still password protected. Access will be granted by request; please contact Claudia.Wenzel@urz.uni-heidelberg.de.

The ARTeFACT Movement Thesaurus: toward an open-source tool to mine movement-derived data

Wiesner, Susan L.
slw4w@virginia.edu
University of Virginia, USA

Bennett, Bradford
bcb3a@virginia.edu
University of Virginia, USA

Stalnaker, Rommie L.
rstalnaker81@gmail.com
Kennesaw State University, USA

1. Problem

The 20th century was a period of vast growth in dance especially in Western cultures, with multiple genres being created and codified techniques being developed. Along with the explosion of new works (danced texts) came an upsurge of research into dance and its acceptance as a scholarly discipline. However, research into movement, and movement-based arts, depends greatly on the ability to peruse documentation beyond static written texts and photographic (still) images. Thus, as visual capture technologies developed, the preferred means of recording and studying a dance work is film and/or video (i.e. visual data). As beneficial as access to film has been to the discipline, this method of preserving and accessing dance contains its own challenges. The current practice of viewing hours of film hinders researchers' abilities to (a) find movement-derived data (b) find that data quickly (c) find data accurately described and (d) reuse the data. Further, while there are standards for preserving video, there are no standards for providing access and any attempt at mining data from a moving image is fraught with difficulty. Therefore, a new model is required, one that exploits advances in computer software and hardware and can enhance research and innovation into movement-based research in the humanities.

2. Our Solution

With funds from an NEH Level II Digital Start-up Grant, the ARTeFACT Movement Thesaurus (AMT) uses motion capture technologies to study movement patterns through a corpus of movement-derived data. In the third phase of the ARTeFACT project, a multi-disciplinary project first developed at the

University of Virginia, the AMT includes over 200 movements derived from codified techniques: ballet, jazz, modern dance, and tai chi. Prior to motion capture of movements, we defined and categorized each movement 'STEP' in order to develop an ontology (saved as xml files). An eight-camera Vicon system captures individual movements and movement phrases typically seen in the studio and on stage. Using custom Matlab software, 3-D data of individual movements are quantified through mathematical interpretation of joint positions, using the ground truth data of the VICON motion capture system for its input to develop the algorithms. In future, the program, idMove, (developed for the DH SUG) will be modified to use only 2-D data.

The second planned future component is the development of algorithms to convert 2-D images from the video into files of position data (in practice the second component, creating the position data from video would be applied first to generate the data for movement identification). These two components will be worked on in parallel. In addition, work will be undertaken to examine and improve the robustness of the algorithms when data sets are incomplete. We will validate that the code works with dancers of different morphologies and levels of ability. Also there are often times in dances when body parts are obscured by other dancers or by the dancer him or herself and we will develop our algorithms to work when sections of movements can not be seen. Finally, we will consider movement phrases, a series of dance moves, (the current software is designed for films of a single dance move) and will develop the ability to identify the individual moves within a string of moves.

As choreographers rarely create dances based solely on individual STEPS, and prefer to use them to create a new vocabulary per the requirements of the dance work, we are moving beyond the codified technical movements to incorporate conceptualized movements into the AMT. At this time, we are using Lakoff and Johnson's work on conceptual metaphor as a basis for pattern recognition of embodied semantics, and we plan to utilize corpus linguistics methods as a basis to formulate a statistical analysis of the STEPS (words) 'spoken' in a dance work. This approach is admittedly problematic, in that a movement phrase does not parallel written phrases; however, we are continuing to work with the ideas of statistical analysis against distinct movement vocabularies created as representative of a concept. Thereby we are creating a lexicon of dance based both upon technical description of the moves (STEPS) as well as theme based moves. We are striving toward a future in which researchers will be able to upload videos and have the dance "annotated" by the AMT software for data mining of movement-based texts.

3. Rationale

Dance movement, as a non-verbal language, cuts across cultures without the need of 'translation.' The body speaks through a kinesthetic voice. While appreciating that there may be cultural differences at work in choreography, in western theatre dance there is generalized understanding of movement techniques and vocabulary. Therefore, the response to a work, especially as to the meaning of a dance, allows most viewers to understand the work. In other words, there is a set of movements that can be read either by an understanding of the technical form or through mutual conceptual frameworks. That said, the most common verbal languages used in dance are English and French with steps codified to such an extent that dancers and researchers the world over understand a passé, a fondue, a frappe, a fouette, a flat back, a brush knee, etc. Thus, we have begun loading the AMT with codified movements. This will allow researchers to view these movements, performed by a subject-matter-expert, via the step name or by individual movements of body parts (at this time, the knee and foot). By extending the AMT to include conceptual movements, we will enable researchers to search based on an idea (the first conceptual set we are including incorporates movements based on the conceptual metaphor Conflict).

4. Conclusion

We will present the NEH funded portion of the ARTeFACT project: the AMT. This is a major step toward providing access to movement-derived data through sophisticated data mining technologies. By using motion capture technologies we are developing a sophisticated, open source tool that can help make film searchable for single movements and movement phrases. By bringing together engineers, movement specialists, and mathematicians we will forge ahead to break new ground in movement research and take one step closer to the creation of an automated means of mining danced texts and filmed movements.

References

Ahmad, K., A. Salway, J. Lansdale, H. Selvaraj, and B. Verma (1998). (An)Notating Dance: Multimedia Storage and Retrieval. Conference Proceedings, *International Conference on Computational Intelligence and Multimedia Applications*, World Scientific. Singapore, p. 788.

Bailey, H., M. Bachler, S. Buckingham Shum, A. Le Blanc, S. Popat, A. Rowley, and M. Turner (2009). Dancing on the Grid: Using e-Science Tools to Extend Choreographic Research.

Philosophical Transactions of the Royal Society A (13 July 2009) 367 (No. 1898): 2793.

Coartney J., and S. Wiesner (2009) Performance as digital text: Capturing signals and secret messages in a media-rich experience. *Literary and Linguistic Computing* 24: 153.

Lakoff, G., and M. Johnson (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenges toWestern Thought*. New York: Basic Books.

Lakoff, G., and M. Johnson (1980). *Metaphors We Live By* Chicago: U of Chicago P.

Starkweather, J. A. (2003). Overview: Computer-Aided Approaches to Content Recognition. In G. Gerbner et al. (eds.), *The Analysis of Communication Content*. New York: John Wiley & Sons, p. 339.

Turner, V. (1974). *Dramas, Fields, and Metaphors* London: Cornell UP.

Wiesner, S. L. (2011) *Framing Dance Writing: A Corpus Linguistics*. Saarbrücken: Lambert Academic Publishing.

The electronic 'Oxford English Dictionary', poetry, and intertextuality

Williams, David-Antoine
david.williams@uwaterloo.ca
University of Waterloo, Canada

1. Overview

This paper will discuss recent research carried out in the context of two grant projects: 'Poetry and Contingency' (funded by a Social Sciences and Humanities Research Council of Canada [SSHRC] Insight Development Grant – Digital Economy Priority Area) and 'Applying search and stylometry techniques to *OED2* and poetic text corpora' (funded by a UWaterloo SSHRC Institutional Grant). It will discuss the processes and methods developed for analyzing and comparing large, heterogeneous text datasets, as well as preliminary results of these analyses and their applications in and implications for traditional literary studies. With implications for digital humanities research, computational approaches to language and literature, linguistics, lexicography, poetics, and literary criticism, the paper will address conference themes of 'cultures, languages, and methods' and several of their intersections.

2. Context

The *Oxford English Dictionary* (*OED*) is widely considered to be the greatest philological and lexicographical achievement in English. The core of the work is its 2.5 million quotations, a significant portion of them from poetic and other literary texts, which both shape and illustrate the various sense definitions of roughly 600,000 English words and word forms. Conversely, since its publication, poets have relied on the *OED* to guide their deployments and arrangements of English words in poems. This reciprocal intertextuality has led to two striking facts which have received insufficient scholarly attention: 1) that the *OED*'s definitions of English words depend to a significant degree on poetic language, which is striking because by any standard account, poetic usage tends away from the denotative or definitional and towards the connotative and metaphorical; and 2) that much English poetry of the last hundred years contains a philological, etymological, and lexicographical dimension, informed by the *OED*.

Although the Second Edition of the *OED* (Murray et al. 1989) was among the earliest large books to be presented to public and academic communities in digitized, marked-up form, and despite the ongoing comprehensive rewriting of the Dictionary (Simpson & Weiner 2000-) no version has ever been marked for quotation genre, meaning that until now the reciprocal influences between dictionary and author have been difficult to identify, and impossible or impracticable to quantify with reference to specific literary genres (e.g. poetry or verse drama, etc.). My projects use the 1989 electronic *OED2*, digitized at the University of Waterloo, alongside electronic corpora of poetry, derived from Project Gutenberg, Chadwyck-Healey, and other datasets, to generate quantitative and qualitative assessments in two broad fields of inquiry: 1) What has been the influence of poetry on the English language's most comprehensive lexicographical work? and 2) What influence has the *OED* had on English-language poetry?

3. Work in Progress

The first prong of the project involves creating a parallel *OED2* in which poetic quotations are marked for genre, to allow for advanced search and comparison of poetic quotations. Marking *OED2* for quotation genre will allow for fast, comparative analysis of the influence of poetic writing on the compilers of the *OED*. I will discuss the challenges of this task and the methods developed to achieve it, as well as the search and comparison scripts developed to query the new resource. I will give examples and discuss the kinds of queries and comparisons made possible by this new resource, and their implications for literary studies and lexicography.

If the first prong is designed to generate questions and answers about poetic influence on *OED2*, the second prong investigates a more intricate and subtle problem: the influence of the *OED* on poetry. For instance, it is a trivial thing for a critic to suppose that Geoffrey Hill's talk of '*wrinching and spraining the text*' is a reference to Milton, since it is contains nonstandard spelling, is italicized, and appears in a poem with Milton in the title (Hill 2008: 5). But *OED2* also quotes exactly those words from *Of Reformation* (Milton 1698: 269) under both 'wrench, v.', and 'sprain, v.' suggesting a second possible source text for Hill. Testing an individual poem for potential *OED2* intertextuality requires a set of text comparison techniques plus an appropriate method of applying these. The process can also be applied systematically, to an electronic corpus (or several corpora) of poetry written since c.1884, the year *OED1* fascicles began to be published. Relatively simple text-similarity approaches to the two corpora (including, e.g., string, n-gram, functional n-gram, and low-probability statistical approaches) will yield hundreds if not thousands of instances such as these, without depending on the poet to flag his or her source. More sophisticated stylometric techniques will give more promising indications of potential poetic intertextuality with various dictionary entries, including their etymology, definition, and quotation fields. Using these tools, we can quickly identify a very large number of candidates for intertextuality in English poems. I will discuss the comparison methods developed, as well as preliminary results of these, and their implications for literary studies.

A few facts make *OED2* intertextuality a special case, overlapping with cases of attribution or allusion, but not identical to these. For one, *OED2* is already a multiauthored and intertextual text, written and compiled by hundreds of lexicographers over more than a hundred years, following varying practices and relying on thousands of sources comprising millions of quotations. Secondly, though *OED2* carries linguistic information (such as pronunciations, etymologies, definitions, etc.) as well as historical usage information (in the quotations) about every English word that is likely to occur in a poem or anywhere else, it is not often the only text to carry any one subset of this information. This means that finding *OED2* in poems may point to another source for a poetic passage than (just) *OED2* itself. Comparing the etymology field of a word in *OED2* to its occurrence in a poetic text, for instance, might point to an etymological play on words, without conclusively attributing this to *OED2*. Or comparing the quotation fields may suggest an allusion to a text that *happens to be* quoted in *OED2*, even if *OED2* is not itself the source of the allusion. But it may also point to true influence, in the form of a poetic allusion or reference which has been occasioned by *OED2 and not the original text*. I will discuss, with the help of preliminary results, the value and implications of each type of discovery, and ways of differentiating among these possibilities when appropriate. This is, I will argue, a discussion which takes up the crucial question of how, and to what extent, computing technologies can benefit the field of literary criticism, as a species of literary scholarship with its own goals and commitments.

References

Brewer, C. (2008). *Treasure-House of the Language: The Living OED*. New Haven: Yale.

Brewer, C. (2009). Literary Quotations in the *OED*. *Review of English Studies* 61: 93-125.

Brewer, C. (2011). 'Happy Copiousness'? *OED's* Recording of Female Authors of the Eighteenth Century. *Review of English Studies* 62: 86-117.

Burchfield, R. W. (1989). *Unlocking the English Language*. London: Faber.

Forstall, C. W., and W. J. Scheirer (2010). Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2). http://https://letterpress.uchicago.edu/index.php/jdhcs/article/view/56/67 (accessed March 2012).

Heaney, S. (1996). *Opened Ground: Poems 1966-1996*. London: Faber.

Hill, G. (2008). *A Treatise of Civil Power*.London: Penguin.

Hollander, J. (1981). *The Figure of Echo*. Berkeley: U of California P.

Holmes, David I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13(3): 111-17.

Irwin, W. (2001). What is an Allusion? *The Journal of Aesthetics and Art Criticism* 59(3): 287-297.

Koppel, M., and J. Schler (2004). Authorship Verification as a One-Class Classification Problem. *Proceedings of the 21st International Conference on Machine Learning*, pp. 489-95.

Koppel, M., J. Schler, and E. Bonchek-Dokow (2007). Measuring Differentiability: Unmasking Pseudonymous Authors *Journal of Machine Learning Research* 8: 1261-1276.

Literature Online . http://lion.chadwyck.com (accessed March 2012).

Milton, J. (1698). *Historical, Political, and Miscellaneous Works*. Amsterdam.

Muldoon, P. (2001). *Poems, 1968-1998*. London: Faber.

Murray, J., et. al., eds. (1989). *Oxford English Dictionary*. 2nd ed., compiled by J. A. Simpson and E. S. C. Weiner, 20 vols. Oxford: Oxford UP.

Neumann, P. Statistical metalinguistics and Zipf/Pareto/Mandelbrot. http://www.csl.sri.com/users/neumann/#12a (accessed March 2012).

Project Gutenberg. http://gutenberg.org (accessed March 2012).

Ricks, C. (2002). *Allusion to the Poets*. Oxford: Oxford UP.

Ruthven, K. K. (1969). The Poet as Etymologist. *Critical Quarterly* 11(1): 9-37.

Simpson, J., and E. S. C. Weiner (2000-). *OED Online*. 3rd ed., rev. J. A. Simpson et al. Oxford: Oxford UP.

Tompa, F., E. Blake, E., and T. Bray (1991). *Shortening the OED: Experience with a grammar-defined database*. Waterloo: UW Centre for the New Oxford Dictionary and Text Research.

Trillini, R. H., and S. Quassdorf (2010). A 'Key to All Quotations'? A Corpus-Based Parameter Model of Intertextuality. *Literary and Linguistic Computing* 25(3): 269-86.

Vogel, C., and G. Lynch (2008). Computational Stylometry: Who's in a Play? *Lecture Notes in Computer Science* 5042: 169-186.

Reasoning about Genesis or The Mechanical Philologist

Wissenbach, Moritz
moritz.wissenbach@uni-wuerzburg.de
University of Würzburg, Germany

Pravida, Dietmar
dpravida@goethehaus-frankfurt.de
University of Frankfurt, Germany

Middell, Gregor
gregor@middell.net
University of Würzburg, Germany

1. Introduction

Dating manuscripts is one of the most demanding tasks of textual scholarship (see e.g. Bockelkamp 1982; Tyson 1987). While there is a vast body of literature on dating and constituting witness stemmata for ancient and medieval codices (West 1973; Bischoff 2009), there seems to be hardly any systematic discussion on the issue of dating modern manuscripts. There have been successful approaches to date older material by computational phylogenetics (Robinson et al. 1998) with the consecutive import of quantitative methods and tools into the field, but to our knowledge, there is no approach that uses formal logics.[1]

Usually, the tools at the modern philologist's disposal are very basic: A pencil or marker pen and a maximal writing surface (*Figure 1*: Reconstruction of the genesis of Faust manuscript VH2). In this paper, we examine the use of a formal knowledge representation to facilitate the dating of a large corpus of inscriptions on manuscripts. The knowledge representation is to be a

- point where diverse and possibly conflicting research sources can be integrated
- working tool for the philologist to be queried and modified during research
- basis of facts for a formal reasoning calculus
- database for users.

The quality of the results of automatic reasoning will be of special interest. Will they be incorrect or correct, and in case of the latter, will they be largely expected or to some degree unexpected?

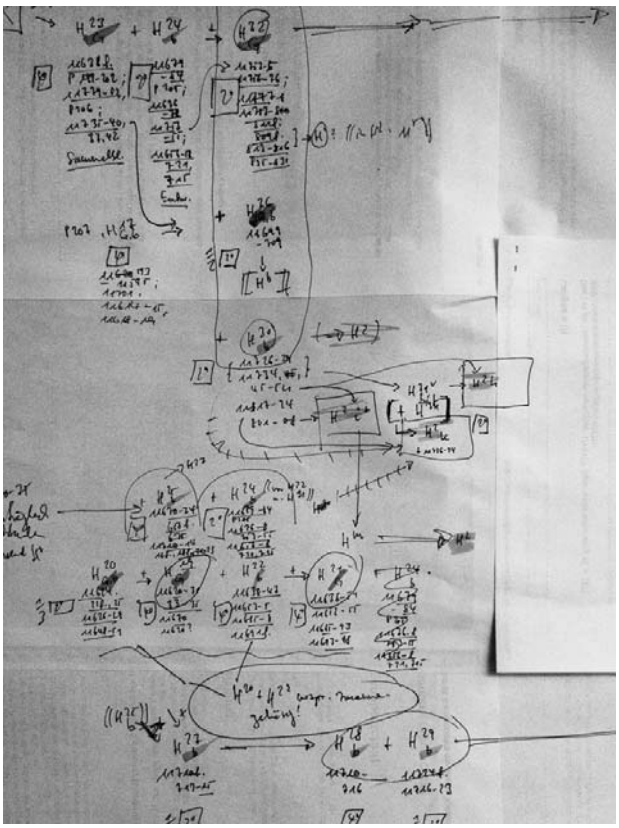


Figure 1

1.1. Relativedating of modern manuscripts

By *inscription* we understand any portion of written text on a manuscript in one specific phase of writing. If there is no *absolute chronology* at hand, the only possibility left is to try to give a *relative chronology*. We believe that a considerable part of conventional dating procedures can be reduced to a combination of more elementary steps that allow a more explicit and even formal approach.

1.2. Inductive vs. deductive reasoning

Relations between inscriptions can be used as predicates in a formal logic and reasoned upon by a suitable calculus. In order to reason deductively, a general rule or a set of rules is required. It is imaginable for this set to be induced automatically by a machine learning algorithm. In this study, however, the set of rules is manually defined. Thus, the presented approach is purely deductive.

1.3. Dealing with contradictions

Some assertions are taken from various sources of research, others are calculated automatically from existing data and even others are established manually. There will inevitably be contradictions. Possible solutions are:

1. Do not try to resolve and thus conceal the contradictions.
2. Try to specify explicit rules that prioritize one class of assertion over the other.
3. Use a method of approximate reasoning.

We have decided for the second approach, as it preserves the advantages of the first while still enabling a very unambiguous specification and interpretation.

2. Reasoning rules for relative dating

To our knowledge there is no extensive list of typical modes of philological reasoning available, so we will try to give some elementary suggestions (following Pravida 2005:58f.).

2.1. Relations between inscriptions

Chronological anteriority

The relation that we are interested in as a result is a temporal one: Which inscriptions *predate* other inscriptions:

$$I <_{pre} J$$

Syntagmatic precedence

Inscription I *syntagmatically precedes* a second inscription J if the text of J follows the text of I with respect to a larger text that contains them both:

$$I <_{syn} J$$

Paradigmatic relationship

Inscription I *is paradigmatically related* to a second inscription J if they share text:

$$I =_{para} J$$

Text-genetic anteriority

Inscription I *text-genetically precedes* an inscription J, if I contains an earlier stage of a text contained by J:

$$I <_{tgen} J$$

Text-genetic anteriority by definition implies chronological anteriority.

Exclusive containment

The overall syntagmatic interval of one inscription, I, is said to be exclusively contained by the overall syntagmatic interval of another inscription J, if some part of the text of inscription I lies within the overall syntagmatic interval of inscription J, where they have a larger coherent textual neighbourhood than in inscription I:

$$I <_{con} J$$

The notion of exclusive containment captures the case of additional amplification of a passage. Exclusive containment is some special case of text-genetic posteriority with regard to some portion of an inscription. (Exclusive containment is very common in the working manuscripts of Brentano; we take it as a matter of course that it will be necessary to refine our set of relations as soon as we will extend our approach to other authors.)

2.2. Rules

Rules or axioms can be defined as a basis for deduction. The language we use is that of predicate logic. The transitivity of the *predates* relation, e.g., can be expressed as

$$I <_{pre} J \wedge J <_{pre} K \Rightarrow I <_{pre} K.$$

Other properties of relations can be modelled for the respective relations accordingly.

Precedence

Rules that might potentially lead to contradictions need to be ordered, so that it is clear which rule has precedence over another. Consider the two rules

$$(a)\ A \Rightarrow C$$
$$(b)\ B \Rightarrow \neg C$$

In order to subordinate rule b under rule a, we add a conjugation of the negated antecedent of the first rule:

$$(b') \; B \wedge \neg A \Rightarrow \neg C$$

In this way, all rules can be ordered and thus the set of rules be made logically consistent. In the following, we will list the rules ordered by priority from lowest to highest. The term pi is used to designate priority.

Syntagmatic precedence

$$I <_{syn} J \wedge p1 \Rightarrow I <_{pre} J$$

Paradigmatic relations

The paradigmatic relation is symmetric. For dating, another hint (term c) must be considered (material, elementary genetic classification, textual):

$$I =_{para} J \wedge c \wedge p2 \Rightarrow I <_{pre} J$$

Containment

$$I <_{con} J \wedge p3 \Rightarrow I <_{pre} J$$

Text-genetic anteriority

$$I <_{tgen} J \Rightarrow I <_{pre} J$$

3. An Example from Clemens Brentano's *Romances of the Rosary*

The twelfth canto of the *Romances of the Rosary* serves as an example that our default assumptions and their formalisation do in fact yield adequate results (cf. Pravida 2005; Brentano 2006). We consider the drafts (inscriptions) J1-6.

From the data which can be obtained from manuscript descriptions, we gather the relations shown in *Figure 2*:

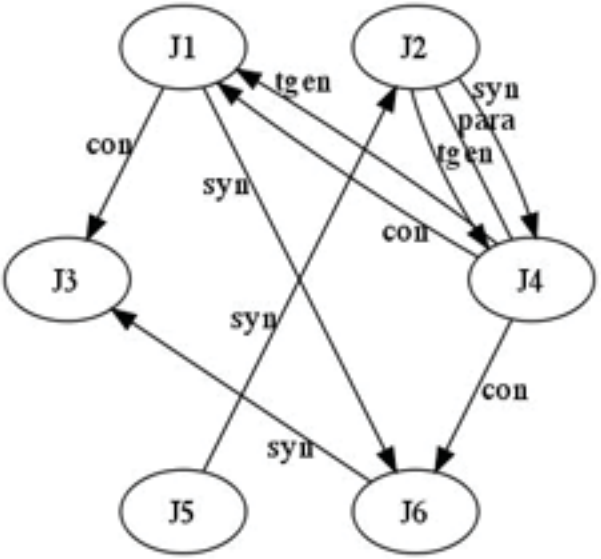


Figure 2

From textual evidence we know that:

$$J2 <_{tgen} J4 \quad (*)$$

$$J4 <_{tgen} J1 \quad (**)$$

(**)

Our reasoning machine concludes:

$$J1 <_{pre} J3 \quad (3)$$
$$J1 <_{pre} J6 \quad (1)$$
$$J2 <_{pre} J4 \quad (1),(2),(*)$$
$$J4 <_{pre} J1 \quad (**),(3)$$
$$J4 <_{pre} J6 \quad (3)$$
$$J5 <_{pre} J2 \quad (1)$$
$$J6 <_{pre} J3 \quad (1)$$

Which establishes a total order (*Figure 3*):

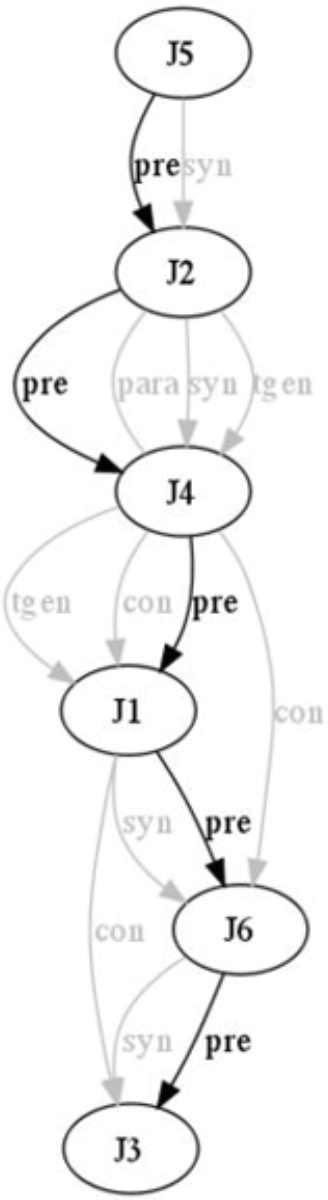


Figure 3

4. Application to the Faust edition / Outlook

The above example is well suited for our purpose, perhaps exceptionally so. Other texts may prove to be more difficult. Goethe's habits of composing, for example, are much less linear. Yet, we can show that the genesis of the last act of *Faust II* can be captured in a way that resembles our example (Pravida, forthcoming). We expect that the genesis of the whole work will permit an analysis closely following these lines. As there is much more electronically encoded material, a computer-assisted approach seems very promising to us.

We used first-order logic and a theorem prover (Schulz 2002). The required expressivity of the logical formalism is probably lower than first-order logic (e.g. it will not need quantification), but more expressive than commonly used Semantic Web logics (e.g. OWL/SWRL) for the purpose of rule prioritisation. The adequacy of a non-monotonic logic, such as default reasoning (Reiter 1980; Delgrande & Schaub 1997), or multi-valued logics and approximate reasoning is to be evaluated. As many facts as possible should be automatically extracted or calculated from existing data, e.g. by means of automatic collation (Stolz & Dimpel 2006), etc.

References

Birdsall, J.N. (1992). The Recent History of New Testament Textual Criticism (from Westcott and Hort, 1881, to the present). In *Aufstieg und Niedergang der römischen Welt. Geschichte und Kultur Roms im Spiegel der neueren Forschung*. Vol. II,26,1. Berlin: de Gruyter, pp. 99-197.

Bischoff, B. (2009). *Paläographie des römischen Altertums und des abendländischen Mittelalters*. 4th ed. Berlin: Erich Schmidt.

Bockelkamp, M. (1982). *Analytische Forschungen zu Handschriften des 19. Jahrhunderts*. Hamburg: Hauswedell.

Brentano, C. (2006). *Romanzen vom Rosenkranz. Frühe Fassungen*. Ed. by Dietmar Pravida. Stuttgart: Kohlhammer 2006.

Delgrande, J. P., and T. Schaub (1997). Compiling Reasoning with and about Preferences into Default Logic. *Proceedings of the International Joint Conference on Artificial Intelligence* 1, pp. 168-175.

Greg, W. W. (1927). *The Calculus of Variants. An Essay on Textual Criticism*. Oxford: Clarendon.

Lachmann, C. (1842). *Testamentum Novum Graece et Latine*. Vol. 1. Berlin: Reimer.

Pravida, D. (2005). *Die Erfindung des Rosenkranzes. Untersuchungen zu Clemens Brentanos Versepos*. Frankfurt: Peter Lang.

Pravida, D. (forthcoming). Die Entstehung von *Faust II*, 5. Akt (1. Fassung). To appear in *Jahrbuch des Freien Deutschen Hochstifts (2012)*

Quentin, H. (1922). *Mémoire sur l 'établissement du texte de la Vulgate. 1re partie: Octateuque*. Paris: Gabalda.

Reiter, R. (1980). A Logic for Default Reasoning. *Artificial Intelligence* 13: 81-132.

Robinson, P., A. Barbrook, N. Blake, and C. Howe (1998). The Phylogeny of 'The Canterbury Tales'. *Nature* 394: 839-840.

Schulz, S. (2002). E – A Brainiac Theorem Prover. *Journal of AI Communications* 15(2/3): 111-126.

Stolz, M.. and F. M. Dimpel (2006). Computergestütztes Kollationieren und dynamische Textpräsentation – Ein Werkstattbericht aus dem Parzival-Projekt. http://nbn-resolving.de/urn:nbn:de:kobv:b4360-1004567

Tyson, A. (1987). *Mozart: Studies of the autograph scores*. Cambridge, Mass.: Harvard UP.

West, M. L. (1973). *Textual criticism and editorial technique, applicable to Greek and Latin texts*. Stuttgart: Teubner.

Notes

1. Our purpose of mechanizing the operation of manuscript dating has obvious parallels in various approaches (e.g. Quentin 1922; Greg 1927) of charting relationships between different manuscript readings by means of a formula system or a decision procedure (conveniently summarized in Birdsall 1992: 153sq). Dom Quentin, Greg and their followers could build on a highly sophisticated tradition of theoretical reflection on establishing manuscript genealogies without the intervention of subjective interpretation (Lachmann 1842: v) for which there is no analogue in our particular field of interest.

The Digital Daozang Jiyao – How to get the edition into the Scholar's labs

Wittern, Christian
cwittern@gmail.com
Kyoto University, Japan

1. Introduction

The Daozang jiyao project aims not only at providing a new text-critical edition of an important collection of Daoist works, but also to serve as a research hub to any research concerning the collection and indeed Daoism in China since the 18th century[1].

2. The Daozang jiyao 道藏輯要

The Daozang jiyao was first put together as a shorter compendium of the enormous Ming Dynasty Daozang, a collection that features some 1500 texts in thousands of scroll, hence its name, which translates as 'Essential of the Daozang'. On closer examimnation, only about 60 percent of the text are from the earlier collection however, with much contemporary material added by the group around Jiang Yuanting 蔣元庭 (style Yupu 予蒲, 1755-1819), who reportedly added 79 texts not contained in the Daozang of the Ming during the Jiaqing era (1796-1820). This edition is still available at some libraries around the world, but inspection of these holdings revealed that almost none of these editions are identical; the texts included have been reshuffled, added or removed, woodblocks for missing pages have been recarved and new material has been added. To complicated matters even further, a reprint was done in Chengdu, Sichuan at the beginning of the 20th century, for which all text have been recarved according to the standards of the time, which resulted in a great number of changes to the text and the characters used for writing it. The collection now contains slightly more than 300 texts.

The whole complexity of the textual tradition came only gradually to the light through bibliographical research done by Monica Esposito and other project members, but it become also clear that the digital edition should not only be able to reflect the textual history accurately but should also be done in a way that allows modern day scholars to make full use of the digital edition, which meant that a modern text with normalized characters and modern interpunction had to be produced as well.

In addition to the digital text itself, a compendium with introductions to the texts in English and Chinese, as well a detailed catalog that details the content of the different editions are planned.

3. Towards a digital edition

Given the rather limited resources of the project, an efficient editing routine had to be established, some of which have been already reported in earlier presentations (Wittern 2009, 2010a, 2010b, 2010c).

A multitude of workflows

When the project started in 2006, a first draft of the text files would be made by a partner institution, the Institute for Chinese Literature and Philosophy, at Academy Sinica in Taiwan. In Kyoto, these text files were transformed to TEI conformant XML markup and then proofread and collated in one single step.

It proved to be very slow, first of all, because proofreading and collation are very different activities that are not easily performed together, but also because the editing of the XML source proved to be challenging for the members of the project, who where mostly specialists in Daoist studies.

As a remedy, the workflow of collation and proofreading was separated and performed with a specialised web application, that would allow the operators to see both the digital facsimile and the transcribed text on the same page. This application was developed based on a new text model, that allowed convenient handling even of characters, that could not be distinguished by there character encoding; details of this can be found in Wittern (2010a).

However, as it turned out, this was not yet the end of the work. For further steps, including the collation and subsequent structural markup of the texts, the web based user interface proved to be too limiting and time consuming to use, so another solution had to be found. We were reluctant to go back to editing the XML source, because that made it less convenient to align the text with the digital facsimiles, so a specialieed editor was preferred.

As it happened, around that time Emacs 23 was released, the first version which truly supported Unicode (although for some of the more arcane characters, a patch was still needed), so it became feasible to provide a customized interface for collation and structural editing.

Instead of trying to directly add the results of the collation to the XML file, as has been tried at the beginning of the project, we used now the ideas presented in Wittern (2011) to use the distributed version control system (DVCS) git to maintain the different versions of the text in different **branches** of the system and could thus easily align them. The algorithms to collate the texts are then partly drawn from Schmidt & Colomb (2009), a working session with a text to be processed is shown in Figure 1.

Figure 1: Work on the CK-KZ branch of the Daozang jiyao in a specialized Emacs editing mode

Towards publication

Both stages of the workflow, the web application and the Emacs bases solution have been built not only for internal use, but as a prototype for the eventual publication. This will not only allow us to make more efficient use of scarce resources, but also gives plenty of time to collect experience and improve the interface.

The expected publication will thus have two faces:

- For the casual user, a web based digital edition will provide easy access to all features of the edition without the need to install specialized software or learn a new tool.
- For scholars, who want to incorporate the texts into their own digital library and work on them using their own tools, a git repository will provide access to all 'branches' of the edition and thus the full range of the established texts, including a normalized version.
- A master XML version, which documents the text using TEI conformant text-critical markup will also be available through a repository.

All of these editions will continue to be enhanced as the knowledge of the texts and their background increases. Since the texts are not simply made available for download, but are set up in a DVCS repository, not only can updates be integrated with local changes of users of the text, but they can make available other branches of the text through separate repositories.

4. Conclusions

Most of the text studied in the Humanities, and especially the material which is at the focus of this project, has come upon us through a long history of writing and re-writing of the text. To fully understand the text requires also an understanding of its tradition and the events that made it possible for us to hold the text in our hand, which might be excavations and spectacular finds or simply the steady and silent work of generations of curators and librarians that preserved the text for us. To realize its full potential, the digital text will thus not only try to reproduce one specific material edition, but allow the reader to look at versions of the text as they existed at certain points in history, so as to know what version of the text was current when it was read (and quoted) by such and such person. At the same time, the text will also be available together with other reference tools and research material in each of the Scholar's digital library or 'labs'.

Now, one of the real fundamental differences between digital text and printed texts as we have known it for centuries is the fact that the latter is a **single product**, whereas a text in the digital medium really wants to be seen in the context of all other texts, read in collaboration with other readers, in short requires a **platform** for scholarly editions, a 'Scholar's Lab', rather than simply a desk to be put on. How such a platform might be constructed and what it needs to do is a matter of heavy debate at the moment; I hope to voice my views at the presentation of this paper.

References

Burnard, L. and S. Bauman, eds (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford: TEI Consortium.

Esposito, M. (2009). The Daozang Jiyao Project: Mutations of a Canon. *Daoism: Religion, History and Society* 1: 95-153.

Esposito, M. (2011). The Daozang jiyao 道臧輯要 as Receptacle to the Three Teachings in Qing Daoism: Lay and Clerical Authorities Face to Face. In *Interactions between the Three Teachings*. K. Mugitani (ed.). Vol. II. Kyoto: Dokisha, pp. 431-69.

Robinson, P. (2009). Towards a Scholarly Editing System for the Next Decades. *In Sanskrit Computational Linguistics: First and Second International Symposia Rocquencourt, France, October 29-31, 2007 Prov- idence, RI, USA, May 15-17, 2008*. Revised Selected Papers. Springer London: Springer, pp. 346-57.

Schmidt, D., and R. Colomb (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies* 67(6): 497-514.

Shillingsburg, P. (2006). *From Gutenberg to Google : electronic representations of literary texts*. Cambridge: Cambridge UP.

Shillingsburg, P. (2009). How Literary Works Exist: Convenient Scholarly Editions. *DHQ: Digital Humanities Quarterly* 3(3).

Shillingsburg, P. (2010). How Literary Works Exist: Implied, Represented, and Interpreted. *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Willard Mc-Carty (ed). Cambridge: Open Book Publishers.

Wittern, Ch. (2009). Digital Editions of premodern Chinese texts: Methods and Problems – exemplified using the Daozang jiyao. In *Early Chán Manuscripts among the Dūnhuáng Findings – Resources in the Mark-up and Digitization of Historical Texts at University of Oslo, Sep. 28 to Oct. 3, 2009.*

Wittern, Ch. (2010a). Mandoku – An Incubator for Premodern Chinese Texts – or How to Get the Text We Want: An Inquiry into the Ideal Workflow. *Digital Humanities 2010*. Kings College London.

Wittern, Ch. (2010b). Rebirth of the Daozang Jiyao – The never-ending Process of Creating a Digital Edition. In *Cultural Crossings: China and Beyond in the Medieval Period Conference – Digital Projects in Asian Art and Humanities Workshop*. Charlottesville.

Wittern, Ch. (2010c). Some Remarks Concerning Digital Editions of Premodern Chinese Texts. In: *International Conference New Directions in Textual Scholarship*. University of Saitama.

Notes

1. The project was initiated and led by Monica Esposito (1962-2011), who's sudden parting came as a shock to all involved in the project. Like all outcomes of the project, this presentation is dedicated to her memory.

Posters

A Digital Approach to Sound Symbolism in English: Evidence from the Historical Thesaurus

Alexander, Marc
marc.alexander@glasgow.ac.uk
University of Glasgow, UK

Kay, Christian
christian.kay@glasgow.ac.uk
University of Glasgow, UK

This poster examines the phenomenon of sound symbolism, or phonaesthesia, in the history of English. Using digital humanities techniques applied to the database of the *Historical Thesaurus of English* (Kay et al. 2009; hereafter abbreviated to *HT*), it presents numerical evidence for the existence of five particular sound-clusters which carry specific meanings in English, and contrasts these with traditional analyses carried out manually in previous years by other scholars. The poster presents some word-initial phonaestheme clusters alongside examples of concepts realised by that cluster, alongside figures showing how many words in that conceptual category are phonaesthetically influenced.

1. Phonaesthesia

The general topic under study is the claim that certain sound-clusters in English convey independent meanings, and that these clusters influence the meaning of those words that contain them (see, *inter alia*, Hinton et al. 1994; Kay & Wotherspoon 2002; Reay 1991, 2006). For example, the cluster <*sw*-> is often mentioned as occuring at the start of words indicating movement through air (*swoop, sweep, swoosh, swash, swat*) in an imitative fashion, meaning that new words starting with that cluster will be affected by its meaning. As Michael Samuels states, the 'validity of a phonestheme is, in the first instance, contextual only: if it fits the meaning of the word in which it occurs, it reinforces the meaning, and conversely, the more words in which this occurs, the more its own meaning is strengthened' (Samuels 1972: 46).

This phenomenon is commonly discussed in lexical semantics and lexicography, and while there are theoretical arguments for and against its strength of effect in English, it is ripe for an empirical and digital investigation. Previous studies have relied on dictionary data, or on an analyst's own introspection. A new style of investigation has recently been made possible, however, by the completion of the *HT*, from which we take our data.

2. The Data

The *HT*, published in 2009, is the world's largest thesaurus and the most complete thesaurus of English, arranging into hierarchical semantic categories all 800,000 recorded meanings expressed in the language from Anglo-Saxon times to the present. These are put into very fine-grained semantic categories, specifying precisely the word's sense alongside attestation dates for that meaning. As an example, the word *broadsword* is recorded as being a type of sword and so is within that particular category (a category which exists seven layers of hierarchy down into the *HT* taxonomy). Moving upwards, all the words for *swords, knives, daggers*, etc exist within the larger category of side-arms, itself within the category of a sharp weapon (adjacent to club/stick and other blunt weapons), which is a sub-type of weapon, which is a form of military equipment, used in the enactment of armed hostility, which is a phenomenon which arises from society, one of the three top-level categories of the *HT*. All the recorded words in English are arranged in this way, permitting a fine-grained look at the relationship between word form and meaning. This makes the HT database ideal for a study of this type.

3. Methodology

The underlying *HT* database (see Kay & Chase 1987; Wotherspoon 1992, 2010), held at the University of Glasgow, is therefore a massive computational resource for analyzing the recorded words of English. The present paper presents data derived from a Python program, written by the authors, which searches through the *HT*'s word forms and produces data categorizing all word-initial consonant clusters according to their *HT* concept category, along with some figures about the size of that category and how many words it contains beginning with each consonant cluster. A combination of statistical filtering and manual analysis then resulted in a large set of English initial phonaesthemes; similar work can be undertaken in future on word-final phonaesthemes.

This set of data was then ranked in order of the putative strength of the phonaesthetic linkage (that is, the statistical preponderance in categories of a significant size). Five particular clusters were then identified as being of sufficient significance to be likely candidates as phonaesthemes:

<*wr*->, <*gr*->, <*sl*->, <*st*-> and <*fl*->.

The poster goes on to give examples of each throughout time, alongside figures of how many words make up the relevant semantic categories to which the clusters belong. For example, the <*wr*-> cluster is particularly associated with uncomfortable movement. It consists of 48% of the 145 words in the *HT* meaning a twisting movement (*writhing, wrenching, wresting, wringing, wreathing, wrying, writher, wriggle, wrinkle*, etc*)* and 15% of the 163 words meaning wrestling (including *wrestling, wraxling, wrestle, wristle,* and the dialectal *warsle).* It also appears in other related categories (13% of anger and 13% of misery, with words related to *wrath* and *wretchedness),* and in a significant metaphorical extension of the twisting sense above (15.1% of distortion or perversion of meaning).

4. Conclusions

By providing evidence of the sort outlined above, this poster describes data derived from applying digital humanities techniques to a new dataset for the study of English. It gives evidence which permits an empirical approach to old questions in linguistic theory, allowing us to move towards an analysis of phonaesthesia which focuses on what percentage of a given concept is realised by words beginning with a particular sound-cluster, and how this relates to similar occurrences in neighbouring semantic concepts.

Beyond the present poster, the diachronic dimension of the *HT* also allows us to plan future DH work examining the historical development of these patterns, and to link them to historical corpora – something which space does not allow on this poster. Such further research in this area would be able to use further datasets to address and investigate those instances across the history of English where phonaesthemes 'grow from minor coincidental identification between a few roots to much larger patterns' (Samuels 1972: 47).

References

Hinton, L., J. Nichols, and J. J. Ohala, eds. (1994). *Sound Symbolism*. Cambridge: Cambridge UP.

Kay, C., and T. J. P. Chase (1987). Constructing a Thesaurus Database. *Literary and Linguistic Computing* 2(3): 161-163.

Kay, C., and I. Wotherspoon (2002). Wreak, *wrack, rack, and (w)ruin:* the History of Some Confused Spellings. In T. Fanego, B. Mendez-Naya, and E. Seoane (eds.), *Sounds, Words, Texts and Change: Papers from 11 ICEHL*. Amsterdam: Benjamins, pp. 129-143.

Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford UP.

Reay, I. E. (1991). *A Lexical Analysis of Metaphor and Phonaestheme*. Ph.D. thesis: University of Glasgow.

Reay, I. E. (2006). Sound symbolism. In K. Brown (ed.), *Encyclopedia of Language and Linguistics*. Oxford: Elsevier, vol. 11, pp. 531–539.

Samuels, M. (1972). *Linguistic Evolution*. Cambridge: Cambridge UP.

Wotherspoon, I. (1992). Historical Thesaurus Database Using Ingres. *Literary and Linguistic Computing* 7(4): 218-225.

Wotherspoon, I. (2010). The Making of *The Historical Thesaurus of the Oxford English Dictionary*. In M. Adams (ed.), *'Cunning passages, contrived corridors': Unexpected Essays in the History of Lexicography*. Monza: Polimetrica, pp. 271-287.

Collaborative Video and Image Annotation

Arnold, Matthias
arnold@asia-europe.uni-heidelberg.de
Cluster of Excellence 'Asia and Europe in a Global Context', Heidelberg University, Germany

Knab, Cornelia
knab@asia-europe.uni-heidelberg.de
Cluster of Excellence 'Asia and Europe in a Global Context', Heidelberg University, Germany

Decker, Eric
decker@asia-europe.uni-heidelberg.de
Cluster of Excellence 'Asia and Europe in a Global Context', Heidelberg University, Germany

The Heidelberg Research Architecture (HRA)[1] is the Digital Humanities Section of Heidelberg University's Cluster of Excellence *Asia and Europe in a Global Context – Shifting Asymmetries in Cultural Flows*.[2] It brings together a team of IT scientists, software developers, and database architects with international researchers and students. In collaborative projects, the HRA aims to develop an integrated digital humanities environment for interdisciplinary studies of transcultural dynamics.

The core of the HRA research environment is Tamboti, the central metadata framework[3] which is used to store standards-based metadata information (e.g. MODS, VRA core 4, TEI, MADS) in an XML database[4]. Researchers can access their resources using a platform-independent browser interface to share and collaboratively work on metadata. In addition, an Open Annotation Collaboration (OAC)[5] module for comments and research annotations is under development. Flexible webservices and user-friendly interfaces for data input (using XForms) will help to further accumulate resources for research.

This poster presentation provides two examples of research projects focussing on visual resource material.

First, the Video Annotation Database[6] allows to annotate a film as a whole and to comment on individual scenes by using the Public Access Digital Media Archive (pad.ma)[7] video annotation software. It offers different types of annotation: transcripts to record the narration, descriptions to annotate scenes, and keywords to add subjects, identify locations, etc. All textual input can be searched. In addition, keywords may be geo-referenced and browsed or searched in a map-interface.

Second, for digital still images a separate module is being developed using the VRA core 4 XML schema[8]. Besides storing images with metadata its aim is to provide researchers with a set of useful tools to directly link the visual material with their research. Important features for its users will be the opportunity of collaborative metadata editing, a flexible system for sharing of resources, as well as tools for annotation and comments based on OAC. In addition, the module will allow to automatically extract embedded metadata on ingest via exif-tool. It will also be possible to mark, annotate and link parts of images, to group images on virtual lighttables, and to save and export image sequences ('visual itineraries').

Two ongoing research projects will exemplify the use of the databases:

Example 1: Global Politics on Screen

As an example for empirical historical research concentrating on visual source material, the project *Global Politics on Screen – A Japanese Film on the Lytton Commission in 1932* [9] tests the potential of video annotation tools by combining pad.ma with models of historical film interpretation. In this research based training project, a team of history students and researchers centres on analysing and commenting a silent Japanese propaganda film about a crucial moment of world history – the Japanese invasion of Manchuria in 1931 which led the way to the Second World War. While the League of Nations commission aimed at preventing further military disputes in the 'Far East' the focus of the film producers was to propagate their own interpretation of international politics. The analysis of the film is a starting point to approach the complex international diplomatic and economic relations of the early 1930s. The project builds up a platform to explore the global contexts and public representations of these critical historical incidents from a new perspective.

Example 2: Priya Paul Collection

The *Priya Paul Collection of Popular Art* [10]contains more than 4.200 illustrations from late 19th and 20th century India. It is one of the finest collections of ephemera such as old posters, calendars, postcards, commercial advertisements, textile labels and cinema posters. In a collaborative endeavor with *Tasveerghar – A Digital Network of South Asian Popular Visual Culture* [11]the collection was digitised and is now being annotated by experts from history, art history, visual and media anthropology, and ethnology. The subststantial research metadata of the collection is a showcase for the implementation of VRA Core 4 XML in the metadata framework and a user-friendly VRA metadata editor is under development. Content analysis is closely connected

with the intense study of the material which already resulted in fourteen research essays and extensive work of annotation. With the new digital still image module, Tamboti will serve as a platform to integrate archive and metadata storage with scientific discussion and analysis within a single framework.

Screenshot 1: Video Annotation Database - Editor

Screenshot 2: Image from the Priya Paul Collection – Image Viewer (HeidICON [12])

Coordination

Global Politics on Screen:

Prof. Dr. Madeleine Herren-Oesch, Cluster of Excellence "Asia and Europe in a Global Context", Heidelberg University;

Cornelia Knab, Cluster of Excellence "Asia and Europe in a Global Context", Heidelberg University;

Eric Decker, Heidelberg Research Architecture, Cluster of Excellence "Asia and Europe in a Global Context", Heidelberg University

Priya Paul Collection:

Prof. Dr. Christiane Brosius, Cluster of Excellence "Asia and Europe in a Global Context", Heidelberg University;

Sumathi Ramaswamy, Duke University

Yousuf Saeed, Tasveerghar, Delhi

Matthias Arnold, Heidelberg Research Architecture, Cluster of Excellence "Asia and Europe in a Global Context", Heidelberg University

Members

Global Politics on Screen:

Inci Bosnak, Sascha Herlings, Felix Nothdurft, Maya Okuda, Julian Wettengel, Dulip Withanage, Matthias Guth

Priya Paul Collection:

Laila Abu-Er-Rub, Suboor Bakht, Gerhard Schönfelder, Sridevi Padmahan, Simon Grüning, Tony Buchwald, Johannes Alisch

Project term

Global Politics on Screen: September 2010 – September 2012

Priya Paul Collection: January 2009 – September 2012

Funding

DFG - Cluster of Excellence "Asia-Europe in a Global Context", Heidelberg University

Cooperations

Cluster of Excellence "Asia-Europe in a Global Context", Heidelberg University;

History Department, Heidelberg University;

League of Nations Archive, UNOG Library, Geneva;

Japan Center for Asian Historical Records (JACAR) アジア歴史資料センター;

LONSEA Project, Heidelberg University;

Tasveerghar - A Digital Network of South Asian Popular Visual Culture, Delhi

Presentation

Cornelia Knab, Eric Decker, Matthias Arnold

Notes

1. Heidelberg Research Architecture, `http://hra.uni-hd.de`
2. Cluster of Excellence 'Asia and Europe in a Global Context', `http://www.asia-europe.uni-heidelberg.de`
3. Tamboti metadata framework, `http://tamboti.uni-hd.de`
4. eXist-db open source database management system, `http://exist-db.org`
5. Open Annotation Collaboration, `http://www.openannotation.org`
6. Video Annotation Database, `http://vad.uni-hd.de`
7. Public Access Digital Media Archive, `http://pad.ma`

8. VRA Core 4 XML standard, http://www.loc.gov/standards/vracore/
9. Global Politics on Screen, http://lytton-project.uni-hd.de
10. The Priya Paul Collection of Popular Art, http://priyapaulcollection.uni-hd.de/
11. Tasveerghar – A Digital Network of South Asian Popular Visual Culture, http://tasveerghar.net/
12. HeidICON – Die Heidelberger Bilddatenbank, http://heidicon.ub.uni-heidelberg.de

Le Système modulaire de gestion de l'information historique (SyMoGIH): une plateforme collaborative et cumulative de stockage et d'exploitation de l'information géo-historique

Beretta, Francesco
francesco.beretta@ish-lyon.cnrs.fr
Laboratoire de recherche historique Rhône-Alpes, Université de Lyon, France

Vernus, Pierre
pierre.vernus@ish-lyon.cnrs.fr
Laboratoire de recherche historique Rhône-Alpes, Université de Lyon, France

Hours, Bernard
bernard.hours@univ-lyon3.fr
Laboratoire de recherche historique Rhône-Alpes, Université de Lyon, France

Le but de ce poster est de présenter le Système modulaire de gestion de l'information historique (SyMoGIH), un projet né en 2007 qui a développé une méthodologie permettant la mise en place d'une plateforme collaborative et cumulative de stockage et d'exploitation de l'information géo-historique (cf. http://larhra.ish-lyon.cnrs.fr/Pole_Methodes/SyMoGIH_fr.php et http://halshs.archives-ouvertes.fr/halshs-00677658 . Il s'agira en particulier de présenter la méthode de modélisation adoptée par le projet SyMoGIH, méthode qui a permis la mise en place d'un système d'information collaboratif, ouvert à accueillir tout type d'information géo-historique.

Le projet SyMoGIH est né de la volonté d'utiliser les nouvelles technologies afin de mutualiser les données produites par les recherches individuelles des historiens, données souvent perdues après la publication des travaux qu'elles documentent, et aussi celles produites par des projets financés par l'Agence nationale de la recherche française, dont deux étaient en cours à l'époque au sein du Laboratoire de recherche historique Rhône-Alpes (CNRS-Université de Lyon). L'apprentissage du langage de modélisation ERD (entity-relationship diagrams) et la collaboration avec des collègues professeurs d'informatique a permis de mettre en place un système collectif d'alimentation de bases de

données. L'intégration d'un module de cartographie et d'analyse spatiale pour prendre en compte la dimension spatiale des données historiques et la volonté de disposer d'un système plus robuste ont conduit à l'adoption du système de gestion de bases de données (SGDB) PostgreSQL avec son extension PostGIS.

Le but du projet n'est toutefois pas de produire un nouveau logiciel mais de mettre à la disposition d'une communauté d'utilisateurs une plateforme ouverte et évolutive permettant le stockage collaboratif et cumulatif de l'information. Une trentaine d'utilisateurs et cinq projets collectifs sont hébergés actuellement dans la base de données du projet. Notre but est de mettre les outils digitaux au cœur de la recherche historique selon une démarche qui vise à élargir progressivemnent le nombre d'utilisateurs et qui met l'accent sur la formation des étudiants : une dizaine de travaux de master ont abouti grâce à l'utilisation de la méthode SyMoGIH, quelques doctorats sont en cours tandis que les enseignants et chercheurs du Laboratoire qui le souhaitent peuvent utiliser la plateforme commune pour héberger leurs propres données. Une charte d'utilisation, réglant les droits et obligations des utilisateurs, permet de gérer les questions délicates de la propriété et de l'exploitation des données.

Le poster se propose de présenter les fondements de la méthode de modélisation développée au sein du projet SyMoGIH, ainsi que les types d'exploitation auxquels peuvent être soumises les données collectées. Seront évoquées questions délicates que doit affronter l'historien souhaitant construire un système d'information utilisable pour sa recherche personnelle mais en même temps ouvert à un travail collaboratif : faut-il stocker des textes ou des données extraites des textes ? Faut-il enregistrer toutes les informations contenues dans un texte ou seulement celles liées à la recherche en cours ? Comment articuler la spécificité de la recherche individuelle avec la mutualisation des données et leur réutilisation pour d'autres recherches ?

Le système d'information mis en place comprend deux volets : l'un reproduit les informations telles qu'elles se trouvent dans les documents ; l'autre construit, par un affinement progressif et par le croisement des sources, des informations telles qu'elles se présentaient effectivement dans le monde historique étudié. En termes de choix technologiques, nous avons d'abord opté pour un système de bases de données relationnelles. Un travail de recherche important a été conduit pendant deux ans pour la mise en place d'un méta-modèle ouvert permettant de produire une modélisation documentée et perfectible de toute information géo-historique qu'on souhaite stocker. La mise en place de ce méta-modèle grâce à la modélisation ERD sera présentée en détail, ainsi que la distinction fondamentale introduite dans la construction des données entre un niveau 'objectif', visant le stockage collectif des informations, et un codage lié à la problématique de recherche individuelle ou d'un projet.

De plus, l'intégration de la dimension spatiale de la recherche a amené à la mise en place d'un *gazetteer* permettant de recenser et de localiser tout type de lieu ou de territoire, y compris dans son évolution diachronique. Enfin, la conception du système a visé une ouverture multidisciplinaire et multiculturelle, permettant de stocker toute information sous forme de texte typé par un code de langues selon la norme ISO 639-3.

Cette approche utilisant un SGBD s'est avérée particulièrement adaptée pour le volet 'reconstitution d'un monde historique' mais elle a montré ses limites pour ce qui concerne le stockage du contenu d'une source. Depuis deux ans nous avons par conséquent mis en chantier un couplage du SGBD avec l'encodage des textes en xml selon le schéma proposé par la *Text encoding initiative* (TEI), tout en utilisan les identifiants des objets tels qu'ils ont été définis dans la base de données comme attributs du balisage. Ce système est particulièrement adapté à l'encodage de textes qui sont destinés à une édition, sous forme papier ou digitale.

En s'appuyant sur la modélisation spécifique à chaque information stockée, tout utilisateur suffisamment formé au SQL peut extraire les informations qui lui sont accessibles grâce à des requêtes de base, voire produire de nouvelles connaissances grâce à des requêtes avancées. Les données ainsi produites sont exportées, habituellement sous format cvs, et sont ensuite visualisées et exploitées dans les logiciels existants de statistique, généalogie, analyse des réseaux, SIG, etc. Des formations spécifiques à ces logiciels sont dispensées régulièrement pour permettre aux étudiants et aux collègues d'exploiter les données qu'ils ont collectées.

Concernant la publication des données, il est possible de définir des populations d'objets propres aux différents projets hébergés, par exemple une population d'acteurs, ou d'institutions, dont on souhaite publier un choix de caractéristiques sur un site web, moyennant accord des 'propriétaires' des informations publiées. Des sites web dédiés à chaque projet peuvent ainsi être mis en place à partir de la base de données collective, ne publiant qu'une portion limitée d'informations. Actuellement, l'exemple le plus abouti est représenté par le projet de prosopographie du patronat français issu d'un financement de l'ANR (`http://www.patronsdefrance.fr/`). La méthode adoptée par SyMoGIH

permet de valoriser les données produites au cours d'un projet financé de durée limitée, en les rendant directement exploitables par les étudiants ou les chercheurs qui, en retour, continueront à alimenter et à enrichir les données même après la fin de la période de financement du projet.

Realigning Digital Humanities Training: The Praxis Program at the Scholars' Lab

Boggs, Jeremy
jkb2b@virginia.edu
University of Virginia, USA

Nowviskie, Bethany
bethany@virginia.edu
University of Virginia, USA

Gil, Alexander
colibri.alex@gmail.com
University of Virginia, USA

Johnson, Eric
ej9k@virginia.edu
University of Virginia, USA

Lestock, Brooke
bnl2ja@virginia.edu
University of Virginia, USA

Storti, Sarah
sas3ca@virginia.edu
University of Virginia, USA

Swafford, Joanna
jes8zv@virginia.edu
University of Virginia, USA

Praxis Program Collaborators
praxis2011@collab.itc.virginia.edu
University of Virginia, USA

In September 2011, the Scholars' Lab at the University of Virginia Library began the Praxis Program, an extra-curricular experiment at realigning graduate methodological training with the demands of the humanities in the digital age. The Praxis Program funds six humanities graduate students from a variety of disciplines to apprentice in the Scholars' Lab, receiving methodological training and collaborating on a shared digital humanities project for a full academic year. Our goal is to equip knowledge workers for emerging faculty positions or alternative academic careers at a moment in which new questions can be asked and new systems built – and along the way to analyze the role of library- and center-based practicum programs in the larger scene of digital humanities training.

This poster details the results of the pilot year of the Praxis Program, and explores approaches for developing and expanding the program.

1. Background and Goals

The Praxis Program represents a continuation of what John Unsworth (1999) dubbed 'the library as laboratory.' Because the Praxis Program is centered in a library-based digital humanities shop, the possibilities for exploring iterative, interdisciplinary processes in digital humanities training are greatly enhanced. The program builds on five years of experience with 20 interdisciplinary Graduate Fellows in a digital humanities fellowship program at the Scholars' Lab. The staff involved in the Praxis Program comes from a variety of humanities backgrounds and library departments.

2. Approaches

Our goal is to address methodological training in the humanities not just through workshops and courses, but by involving graduate students in digital projects from the ground up. The primary task of the 2011–12 Praxis Fellows is to build *Prism*, a web-based tool for 'crowdsourcing interpretation' of texts and images. For Praxis, this means learning by working with faculty, staff, and fellow students as colleagues, with all that entails: paying attention both to vision and detail; building facility with new techniques and languages not just as an academic exercise, but of necessity, and in the most pragmatic framework imaginable; acquiring the softer skills of collaboration (sadly, an undiscovered country in humanities graduate education) and of leadership (that is, of credible expertise, self-governance, and effective project management). All this also involves learning to iterate and to compromise – and when to stop and ship. In fact, collaboration and collective ownership of the work in Praxis Program is so important, the group's first task was to compose and publish a project charter, influenced by the work of Ruecker and Radzikowska (2004) and Siemens et al. (2009).

The curriculum for the Praxis Program is evolving and iterative, and draws on staff knowledge while adapting to the changing needs of a project actually in development. Over the course of the academic year, the program will cover a variety of topics: evaluating peer work in DH; programming and open source software development; prototyping, wireframing, and user experience design; HTML and CSS; database design; project management; and budget and grants management.

3. Results and Continued Development

Because Praxis Program participants value publicly sharing ongoing research, our project blog is highly active, and representative of both the successes and anxieties of open collaboration. By the end of the first year, the Praxis Program team will create and release a working version of *Prism* for public use and critique, and work on publishing the results of their research.

Recognizing that methodological training in the digital humanities is often absent or catch-as-catch-can at the graduate level, we are using the Praxis Program to experiment with an action-oriented curriculum live and in public, hoping to attract local allies as well as partners in labs and centers at other institutions (which could, in future, work as nodes in a larger Praxis Program network).

Above all, we want to situate our contribution to methodological training within a larger debate about the changing demands of the humanities in a digital age. To that end, we are partnering with the Scholarly Communication Institute in fostering conversation about methodological training in the digital humanities with professional societies, consortia of digital and traditional humanities centers, and other library- and center-based programs.

References

Gessner, G. C., E. J. Damon, J. Rutner, and K. Tancheva (2011). Supporting Humanities Doctoral Student Success: A Collaborative Project between Cornell University Library and Columbia University Libraries. *CLIR Report*, October 2011. Available: http://www.clir.org/pubs/ruminations/02cornellcolumbia/report.html

Leon, S. Project Management for Humanists: Preparing Future Primary Investigators. In B. Nowviskie (ed.), *#alt-academy*. Available: http://mediacommons.futureofthebook.org/alt-ac/pieces/project-management-humanists

Nowviskie, B. (2011). Where Credit is Due. May 31, 2011. Available: http://nowviskie.org/2011/where-credit-is-due/

The Praxis Program. University of Virginia Library's Scholars' Lab. Available: http://praxis.scholarslab.org

Ramsay, St. Care of the Soul. October 8, 2012. Available: http://lenz.unl.edu/papers/2010/10/08/care-of-the-soul.html

Ruecker, S., and M. Radzikowska (2004). The Iterative Design of a Project Charter for

Interdisciplinary Research. Available: http://mtroyal.academia.edu/MilenaRadzikowska/Papers/326958/The_Iterative_Design_of_a_Project_Charter_for_Interdisciplinary_Research

Scholarly Communication Institute. New-Model Scholarly Communication: Road Map for Change. Available: http://www.uvasci.org/current-institute/sci-9-report/

Siemens, L., et al. (2009). INKE Administrative Structure, Omnibus Document. Available: http://journals.uvic.ca/index.php/INKE/article/view/546/245

Unsworth, J. (1999). The Library as Laboratory. *Presentation at the Annual Meeting of the American Library Association*, 1999. Available: http://www3.isrl.illinois.edu/~unsworth/ala99.htm

Supporting the emerging community of MEI: the current landscape of tools for note entry and digital editing

Bohl, Benjamin W.
bohl@edirom.de
Universität Paderborn, Germany

Röwenstrunk, Daniel
roewenstrunk@edirom.de
Universität Paderborn, Germany

Viglianti, Raffaele
raffaele.viglianti@kcl.ac.uk
King's College London, UK

1. Introduction

The Music Encoding Initative (MEI), started as a 'one man show' at the University of Virginia. Perry Roland first designed an XML based encoding scheme for representing music notation according to principles and concepts that the Text Encoding Initiative was putting into practice for scholarly text encoding (Kepper, 2010). This effort addressed academic encoding approaches since the beginning, thus attracting a group of scholars interested in building a community around the MEI specification. In 2007, the format underwent its first cooperative improvements that added support for medieval notation, editorial interventions, and alignment with facsimile.

These personal initiatives prompted further involvement and soon could be channeled into the formation of a MEI study group in 2009. Since then, institutional support and public funding helped improving MEI. A project funded by the NEH and DFG led to the first community effort by making the first non-beta release of MEI publicly available in May 2010 (Röwenstrunk 2010). To date, the MEI has developed into an open and community-driven effort involving a council, a technical team and a growing user group.

This emerging community intends to improve the involvement of musicology within the Digital Humanities field. With this aim, some members have previously introduced examples of MEI usage to the DH community (see Kepper 2010; Viglianti 2010). Today, MEI's users are mostly involved in projects that employ the format and/or develop software to

support MEI use. This poster intends to demonstrate some of these efforts, particularly concerning the tool support for day-to-day work of the scholar for entering music notation and metadata, creating facsimile alignments and for the production of digital editions.

2. Note Entry

Transcribing or entering music notation with a computer is typically a more laborious task than transcribing text; this is because, unlike letters, music symbols themselves require complex codes to be represented. Many machine-readable formats, such as Humdrum or Lilypond, rely on ASCII-based structures which can be more or less intuitive to type. However, the most common note entry is done with WYSIWYG score editors, which allow the transcriber to enter the notation directly on a virtual score.

Note entry also lies at the heart of encoding music documents with MEI; however, encoding music notation in XML immediately poses a problem of overlapping hierarchies, because the score is organized as a grid that represents temporal sequence horizontally (one event comes after the other, such as a sequence of notes) and temporal co-occurrence vertically (events occur at the same time, such as a chord on the piano). The general complexity of music codes and the unavoidable workarounds for dealing with this condition makes hand-encoding a strenuous task.

Given these premises, it is evident that tools are necessary to enter music notation into the MEI format in a more intuitive manner. At this stage of MEI development as a community, there are some options available for simplifying note entry.

1. WYSIWYG score editors. In the past ten years, MusicXML has had outstanding success as an interchange format between different score editors. As a consequence, most specialized software now provide an export to MusicXML. The MEI provides XSLT stylesheets to convert from MusicXML 1.0 and 2.0 to MEI, which makes it possible to use score editors' exports to be converted into MEI. This, nonetheless, is not necessarily a straight-forward process. Given the graphical, non-linear organization of notation on the page, each score editor can be more or less precise about the graphical positioning of symbols on the page. Symbols such as directions, phrase marks, and dynamics are particularly affected as they can be attached either to a staff, a measure, or a specific event on the staff. These differences are seemingly innocuous to the score editor's users because their rendition on the page looks the same. However, when moving into a semantic representation system such as MEI, the associations of different symbols and their start/end points matter. Notwithstanding these difficulties, using score editors and MusicXML still simplifies note entry in MEI. Some post-processing, however, is still necessary.

2. MEI-specific score editors. As described above, MEI has existed for more than ten years, but has experienced a considerable community effort only in more recent times. Amongst several uplifting activities, a few focused on the creation of WYSIWYG editors able to export directly into MEI. The TextGrid component MEI Score Editor (MEISE) is the most complete to date.[1] Besides exporting notation directly to MEI, it is designed to be able to represent editorial aspects supported by the format, such as variants and alternative readings.

3. XML editors. The first two solutions may be sufficient for most encoding scenarios, especially if MEISE grows to be more comprehensive and/or is followed by other open source efforts for creating MEI exports for more common score editors. It is possible to imagine, however, that MEI will be used for deep-encoding projects, such as critical editions, diplomatic or genetic transcriptions, analyses, etc. For such scenarios, it is debatable that a generic tool would be able to cover all uses. Obtained as an heritage from Text Encoding Initiative, MEI's flexibility is both a blessing and a curse. It allows one to use the format as a framework within which it is possible to build specific encoding models through customizations; at the same time, though, it makes it impossible to build tools that can cater for all possible declinations of the format.[2] Nonetheless, tool development remains an essential component of the MEI's community efforts; particularly towards building tools tailored to specific tasks or in the form of libraries for manipulating MEI documents.[3] Those project that intend to use MEI at its full power, however, will eventually need to resort to the use of XML editors. Editing the XML manually is doable with a little training, but requires some imagination to see the score through the code. Alternatively, one could exploit many modern XML editors' visual support for text encoding. This is typically done through CSS-based previews that can be edited in real-time. A similar scenario is considerably harder to obtain with music notation; however experimentation is underway.

3. Supporting editorial work

MEI as a format was – amongst other goals – initially intended to 'represent the common expressive features of traditional facsimile, critical, and performance editions',[4] which opens the way for more and more critical edition projects to describe, preserve and publish their work with MEI.[5] Because of the very small or even non existing support of score

editors or renderers in the last years, most endeavors to create and publish digital music editions have been based on facsimiles and texts. Other formats could not bridge the gap, because of their limited representation of textual phenomena especially in the fields of editing and transcription.

With the latest possibilities for note entry and score rendering, digital editions are going to reach a higher level of quality. For example, searching and analysing music editions would become possible, similarly to what already happens in digital editions of literature. Moreover, readability and usability would be improved by selectively rendering scores in modern or old notation or in different keys.

The Edirom project started in 2006 with the aim to provide tools for creating and presenting historico-critical music editions. The project was one of the main contributors in moving MEI to a community-driven standard.[6] Given the lack of score writers for MEI, the Edirom tools kept using image based presentation techniques. But with emerging tools for MEI, more and more encoded music is being included for different purposes. Recent efforts include a score renderer by Thomas Weber which uses the web standard SVG to enable digital presentations with Edirom tools to show *ad hoc* generated notation and allow interaction with musical snippets like switching versions or highlighting specific symbols. Eventually, editions will no longer have to be constrained to providing one single edited text for scholars;[7] not every uncertainty will have to be resolved; and equitable versions could be presented simultaneously. Hence, the user might have a closer look at the underlying editorial work (see Bohl, Kepper & Röwenstrunk 2011).

In order to provide these functionality and level of knowledge, one needs a detailed encoding of music documents as described above, and a considerable amount of interlinking. Such data may then be used in the Edirom Editor for creating digital music editions. The Editor, for example, provides: a) mechanisms to link structural, musical information (like a movement or measure) to regions on facsimiles; b) functions for formally describing relationships or differences between these music documents. Some of the specific tasks in the workflow of creating digital editions are handled by specific tools outside or included within the Edirom Editor; such as semi-automated recognition of measure positions with Optical Music Recognition techniques or describing the metadata of a document, which is done with a included version of the Metadata Editor and Repository for MEI Data (MerMEId) from the Danish Centre for Music Publication.[8]

References

Bohl, B. W., J. Kepper, and D. Röwenstrunk (2011). Perspektiven digitaler Editionen aus Sicht des Edirom-Projekts. *Die Tonkust* 5.

Dabbert, J., and J. Veit (2010). MEISE: an editor for encoded/encoding music. *TextGrid Advisory Board Meeting*, Detmold. Available at `http://https://www.textgrid.de/fileadmin/TextGrid/konferenzen_vortraege/fachbeirat_0710/meise.pdf` (accessed November 2011).

Kepper, J. (2010). A Data Model for Digital Musicology and its Current State – The Music Encoding Initiative. *Digital Humanities 2010*, London. Available at `http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-817.html` (accessed November 2011).

Pierazzo, E. (2010). Editorial teamwork in a digital environment: the edition of the correspondence of Giacomo Puccini. *Jahrbuch für Computerphilologie* 10.

Röwenstrunk, D. (2010). Digital Music Notation Data Model and Prototype Delivery System. *Forum Musikbibliothek* 31.

Viglianti, R. 2010. Critical Editing of Music in the Digital Medium: an Experiment in MEI. *Digital Humanities 2010*, London. Available at `http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-819.html` (accessed November 2011).

Notes

1. A first beta release of MEISE is to be made available by the end of 2011.
2. '[H]umanities data have a very high rate of variation [...], and even when similarities do exist, the differences are important enough to imply that designing tools able to cope with such variety would be a very demanding task' (Pierazzo 2009).
3. See for example libmei `http://ddmal.music.mcgill.ca/libmei`
4. `http://ddmal.music.mcgill.ca/libmei`
5. The german projects OPERA (`http://www.opera.adwmainz.de/`), Reger-Werkausgabe (RWA, `http://www1.karlsruhe.de/Kultur/Max-Reger-Institut/de/rwa.php`) and Carl-Maria-von-Weber-Gesamtausgabe (`http://www.weber-gesamtausgabe.de`) are using the tools provided by the Edirom project.
6. The project organized a conference Digitale Edition zwischen Experiment und Standardisierung, 2007 in Paderborn (Germany), where Perry Roland was invited to present MEI. Later, members of the project co-applied for funding for organizing two more conferences where the MEI council was constituted (see Röwenstrunk 2010).
7. There will probably always be a normalized version of the edition for practical use.

8. http://www.kb.dk/en/kb/nb/mta/dcm/projekter/mermeid.html

'The Past Is Never Dead. It's Not Even Past': The Challenge of Data Provenance in the e-Humanities

Clark, Ashley M.
amclark4@illinois.edu
University of Illinois at Urbana-Champaign, USA

Holloway, Steven W.
hollowa2@illinois.edu
University of Illinois at Urbana-Champaign, USA

1. Summary

The humanities as a discipline has traditionally exhibited great care in documenting sources and establishing authentic chains of object transmission. Data provenance metadata, however, is rarely curated in digital humanities projects, perhaps due to the lack of interoperable standards for recording data provenance. Recent efforts by the W3C Provenance Working Group to create the PROV-DM (Provenance Data Model) and PROV-ASN (Provenance Abstract Syntax Notation) may answer the requirements of the e-humanist community for such a standard. We examine the provenance capture capabilities of two e-humanities virtual research environments (VRE), TextGrid and Meandre, which also serve as test-beds for PROV-ASN assertions. PROV-ASN provides an interlingua abstract enough to express a common set of data provenance assertions for both software environments. The data for these assertions must be culled manually from various directories and file-types in both TextGrid and Meandre. For data provenance metadata to become as normative a feature of the digital humanities as source provenance is for analog humanities, e-humanities developers need to aggregate provenance metadata, equip it with intuitive visualizations, and render it using interoperable standards capable of transcending a particular VRE or software platform.

2. Data Provenance and the Humanities

Like the source provenance of art and artifacts, 'data provenance' depends upon records to establish a chain of past events which provide context for a digital object at any point in time. However, data provenance traces not ownership changes to a static document, but transformations to a version

of a dataset, which then produces further versions. Records of data provenance might then include information on authorship, the name and version of software used to produce transformations, and descriptions of the transformations themselves.

For over fifteen years, the legal, business, computer science, e-science, and e-humanities communities have developed data provenance requirements specific to their disciplines, yet they share a need for interoperable standards. To date, little published research in e-humanities explicitly focuses on data provenance. Paradoxically, humanities scholarship flourishes on traceable reference. Humanities researchers traditionally perform close readings of object relationships, which requires judgments of source trustworthiness, authority, conceptual derivation, and class membership. It is a matter of concern, then, that the humanities, with its native affinity for historical thinking, should find itself unable to migrate provenance documentation methods into the digital realm.

If data provenance is to benefit the humanities, e-humanists must provide coherent metadata about their computing environment, records of state transformation, and file forensics no different than those entailed for e-science validation. Those who view the datasets should also have access to complete provenance information, ideally in human-readable form aggregated as supplemental resources. For this to occur, e-humanities researchers need tools which can either generate such provenance information automatically, or make it easy to manually gather and output. Familiar commercial authoring tools, such as Microsoft Word and Adobe Photoshop, support actionable data provenance (rollbacks to earlier states via the 'undo' function), but provide at best meager data provenance in the guise of autonomous metadata, susceptible to extraction and persistent storage. Because they cannot capture data provenance at a high-enough level of abstraction, open-source metadata streams for digital cameras (EXIF) and Adobe software (XMP), useful as they are, cannot serve the e-humanists' need for interoperable standards.

3. Provenance in e-Humanities Computing Environments

This poster will compare structured provenance metadata in two decidedly different e-humanities computing environments, TextGrid 1.0 and Meandre 1.4.9. By staging an identical literary 'experiment,' these tools illustrate typical provenance-capture shortfalls. This poster will also explore the advantages of machine-actionable, interoperable provenance metadata, using the W3C's PROV-DM standard as a basis for examining the current bedlam of e-humanities data provenance silos.

3.1. Meandre

The extensible Meandre software framework is a Java-based, open-source semantic web authoring and publishing tool with heavy emphasis on linguistic annotation and text analytics. The tool is intended to promote rapid software prototyping within two graphical user interfaces, the Meandre Infrastructure and Workbench. Meandre was developed by the Automated Learning Group, National Center for Supercomputing Applications at Urbana-Champaign, as part of an Andrew W. Mellon Foundation-funded text-mining initiative sponsored by the Software Environment for the Advancement of Scholarly Research (SEASR).

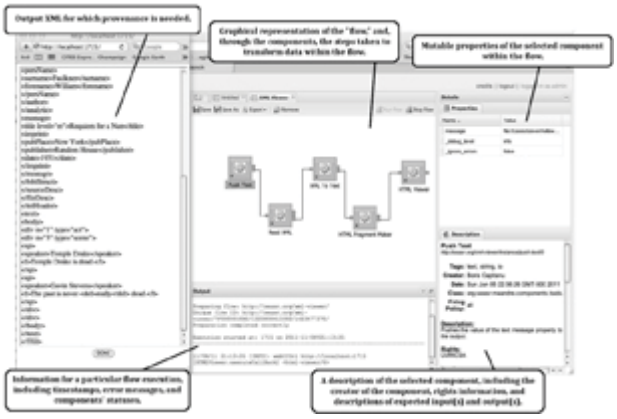


Figure 1: Provenance information as represented in Meandre

3.2. TextGrid

The Eclipse-based software TextGrid, funded by the Bundesministerium für Bildung und Forschung's D-Grid initiative, represents an attempt to create an e-humanities VRE, consisting of both a 'lab' that runs on local hardware, and a sophisticated repository back-end fostering preservation, discovery and publication. A set of purpose-made tools facilitate Text Encoding Initiative (TEI) XML markup of documents and document-images in a collaborative environment. TextGrid was designed for scholars pursuing traditional philological studies, but is intended to serve musicologists, epigraphists, linguists, and art historians.

Figure 2: Provenance information as represented in TextGrid

4. The W3C Provenance Data Model

Meandre and TextGrid serve as production work spaces, yet they fail to comprehensively log transformations of information in accessible ways. Though data provenance should be a requirement for e-humanities tools, solutions to provenance capture are a recent development. An emerging data provenance model, PROV-DM/-ASN, offers a means of capturing data provenance information at a level of abstraction that is hardware- and software-agnostic. For example, a passage from William Faulkner's *Requiem for a Nun* edited in both TextGrid and Meandre could be expressed with the same PROV-ASN assertions, despite radically different processing architectures:

Figure 3: Text transformations expressed through PROV-ASN assertions

In order to assemble comparable provenance metadata in TextGrid, it would be necessary, at a minimum, to manually open the five TEI revision files and view their descriptive and technical metadata in separate windows. A similar task in Meandre would require: saving five separate 'workflows' corresponding to the revisions e1-e5; viewing their associated documentation in both Workbench and Infrastructure; and manually accessing a variety of command-line log and server files.

The PROV-DM/-ASN specification makes allowances for recording system-specific details, such as software function names, component URIs, firing sequence, time of execution, and other technical metadata necessary for describing the process workflow. Since the PROV-O (Provenance Ontology) allows encoding in OWL2 web ontology language, the markup is machine-readable. In Meandre, assembling the data for the platform-specific provenance metadata would entail querying specific projects with command-line file utilities. Even then, some execution information would not persist after closing the program. In the TextGrid interface, a significant portion of execution information is simply unavailable. The experiment could be successfully re-run in Meandre and TextGrid using the saved workflow or project, respectively, but the full scope of the operating parameters remains concealed within software.

5. Concluding Remarks

Our poster shows that two prominent e-humanities VREs need considerable retooling to attain the benefits of comprehensive provenance metadata. Recent developments in provenance documentation standards promise the exchange of datasets with a common interlingua, facilitated descriptive analysis, and easier understanding of experimental protocols. Whether or not e-humanities developers adopt the PROV-DM/-ASN standard, the need for a systematic approach to data provenance-capture stands, as well as best practices which will benefit not only the users of e-humanities computing environments, but the larger digital humanities community.

Acknowledgments

This project was supported by DCEP-H, an initiative to extend the Data Curation Education Program to humanities. DCEP-H is based at the Center for Informatics Research in Science and Scholarship at and funded by IMLS Grant RE-05-08-0062-08.

References

Belhajjame, K., J. Cheney, D. Garijo, T. Lebo, S. Soiland-Reyes, and S. Zednik (2011). *PROV Ontology Model, W3C Working Draft 13 December 2011.* http://www.w3.org/TR/2011/WD-prov-o-20111213/ (accessed 9 March 2012).

Belhajjame, K., S. Cresswell, R. Golden, P. Groth, G. Klyne, J. McCusker, and S. Sahoo (2011). *PROV Data Model and Abstract Syntax Notation, W3C Working Draft 18 October 2011.* http://www.w3.org/TR/2011/WD-prov-dm-20111018/

Chao-fan, D., W. Tao, Z. Peng-cheng, and F. Yang-He (2010). A comparation of data provenance systems based on processing. *IEEE International Conference on Intelligent Computing and Intelligent Systems* 3: 374-379. Institute of Electrical and Electronics Engineers. http://dx.doi.org/10.1109/ICICISYS.2010.5658641

Freire, J., D. Koop, E. Santos, and C. T. Silva (2008). Provenance for computational tasks: A survey. *Computing in Science & Engineering* 10(3): 11-21. http://dx.doi.org/10.1109/MCSE.2008.79

Hasan, R., and M. Winslett (2009a). *Trustworthy history and provenance for files and databases.* Ph.D. thesis,

University of Illinois at Urbana-Champaign. http://search.proquest.com/docview/304899850?accountid=14553 (accessed 9 March 2012).

Küster, M., C. Ludwig, Y. Al-Hajj, and T. Selig (2011). TextGrid provenance tools for digital humanities ecosystems. *Proceedings of the 5th IEEE International Conference on Digital Ecosystems and Technologies 2011*. Daejeon, Korea, pp. 317-323. http://dx.doi.org/10.1109/DEST.2011.5936615 (accessed 9 March 2012).

Neuroth, H., ed. (n.d.). TextGrid: Home. http://www.textgrid.de/en.html (accessed 9 March 2012).

Moreau, L. (2009). *The foundations for provenance on the web*. doi:10.1.1.155.784 (submitted)

Simmhan, Y. L., B. Plale, and D. Gannon (2005). A survey of data provenance in e-science. *SIGMOD Record* 34(3), 31-36.

Software Environment for the Advancement of Scholarly Research. (2011). Meandre. http://seasr.org/meandre (accessed 9 March 2012).

The Social Edition: Scholarly Editing Across Communities

Crompton, Constance
ccrompto@uvic.ca
University of Victoria, Canada

Siemens, Raymond
siemens@uvic.ca
University of Victoria, Canada

The Devonshire MS Editorial Group

1. Introduction: Contexts for Social Editing

The integration of social tools into the electronic scholarly edition pushes the boundaries of authority, shifting power from a single editor, who shapes the reading of any given text, to a group of readers comprising a community whose interpretations themselves recast the edition form. This poster outlines a method that brings scholarly best practice, open access methodology, and social media technology to scholarly editing. As a test of the method, we, the Devonshire MS Editorial Group,[1] are leading the development of a social edition of the Devonshire Manuscript, a multi-authored verse miscellany compiled in the 1530s and early 1540s. Through the production of both a fixed and evolving edition of the text, we are exploring how social media can be used to change the role of the scholarly editor from the sole authority on the text to a facilitator who brings traditional and citizen scholars into collaboration through an ongoing editorial process.

The most conservative electronic scholarly editions have used computation chiefly to 'describe and express print-, visual-, and audio-based material in tagged and searchable electronic form' (Schriebman, Siemens & Unsworth 2004: xxv), in many ways mimicking interactive structures more suitable to possibilities of the print medium rather than the digital one; this teleological, codex-based model sees the editor as a single authority, a mediator between the text and the reader, where the editorial entity determines and shapes what is important to the reader, focuses the editorial and analytical lens, and ultimately exerts immense control over how the reader can engage. Social media's interactive online spaces have produced new environments for textual editing, ones which invite us to consider the best way

to move beyond the codex model while preserving qualitative assurance and facilitating community engagement.

2. Materials for The Case Study: The Devonshire MS

If we count all creative textual works – complete poems, verse fragments and excerpts from longer works, anagrams, and other ephemeral jottings – the Devonshire Manuscript consists of 194 items. In addition to 129 courtly verses by Sir Thomas Wyatt, of which sixty six are unique to the manuscript, and eleven poems by Geoffrey Chaucer, the manuscript contains transcriptions of the work of others and original works by prominent figures from the Henrician court including Mary Shelton, Lady Margaret Douglas, Mary (Howard) Fitzroy, Lord Thomas Howard, Henry Howard and, perhaps, Queen Anne Boleyn (Southall 1964: 143). The manuscript itself bears traces of the original contributors' editorial process: besides writing epistolary verse, contributors to the manuscript interacted with one another through editorial annotation. These marginal responses are, at times, quite personal in nature. They include responses that evaluate the quality of certain lines, the crossing out of one word and the insertion of another, or the writing of textual remarks that seem to comment on real-world situations not necessarily noted in the manuscript. Other aspects of the Devonshire Manuscript – its multi-layered and multi-authored composition, its early history and transmission, the ways in which its contents engage with and comment directly on contemporary political and social issues – confirm the volume's function as 'a medium of social intercourse' (Love & Marotti 2002: 63).

3. The Case Study: Fixed and Evolving Editions

The Devonshire MS Editorial Group is modeling the social editing process. We have produced a fixed, authoritative version of the text, which has undergone a thorough review by an international advisory group of Early Modern and Renaissance scholars. This same text, complete with appendices, glosses, commentary, and textual notes is now available in `http://www.bitly.com/DevonMS` , a platform which provides a flexible environment for collaboration, contribution, and discussion by traditional and citizen scholars. The Wikibook, which has already received promising attention from Wikibooks' existing editorial community, is augmented with additional images of the manuscript, witness transcriptions, an extensive bibliography, and xml encoded transcriptions of the Devonshire Manuscript – creating an online version that is both an edition and a research environment. The advisors are currently providing a medial review which is neither as public at the Wikibook discussion pages nor as private as anonymous peer review. They are in conversation with one another over the fixed authoritative edition and the evolving Wikibooks edition in a social media space housed by Iter, a federated site located at the University of Toronto.

After six months online, the Wikibooks edition will undergo a full comparison with the fixed edition. Keeping the best of the Wikibooks additions, the Devonshire MS Editorial Group will incorporate the advisory group's suggestions, and will present a truly socially mediated edition of the Devonshire Manuscript for publication with Iter and Medieval and Renaissance Texts Society. Rather than a terminus, this edition will serve as a snapshot of the living edition, one which represents the perpetual process of scholarly editorial practice.

4. Reflection on Method and the Significance of Anticipated Outcomes

The public editing process for the social edition has been designed to encourage communication across editorial communities, while preserving the peer review process. In addition to producing an edition that allows for multiple editorial perspectives, the Devonshire MS Editorial Group[2] is modeling a social edition methodology. In the interest of refining the process and expounding on its utility for collaborative editors in the Web 2.0 environment, we are using a combination of methods to gather data on the social edition building process. We are conducting qualitative interviews with the members of the advisory group to gather their perspectives not only on the content of our evolving and fixed editions, but also on issues of credit, peer review, and collaborative editing in a social media environment. In addition, we are conducting surveys with self-identified members of the existing Wikibooks editorial community. Finally, we are using analytics to trace the movement of editors through the Wikibooks text, to determine which parts of the edition provoke the most discussion, attract the most attention, and lead to the most community engagement.

Building on existing, expanding, and newly-emerging communities of practice, we can harness the power of specifically social tools, to ensure that the social edition text is fluid, agency is collective, and many editors, rather than a single editor, shape the interpretation of the text. The social edition development process is, by its very nature, under constant review. Relying on dynamic knowledge building, and privileging process, this expansive

method promises to bring together communities of scholarly practice to build a new, social, edition.

References

Love, H., and A. Marotti (2002). Manuscript Transmission and Circulation. In D. Loewenstein and J. Mueller (eds.), *The Cambridge History of Early Modern English Literature*. Cambridge: Cambridge UP, pp. 55-80.

Remley, P. G. (1994). Mary Shelton and Her Tudor Literary Milieu. In P. Herman (ed.), *Rethinking the Henrician Era: Essays on Early Tudor Texts and Contexts*. Urbana: U of Illinois P, pp. 40-77.

Schreibman, S., S. Siemens, and J. Unsworth (2004). The Digital Humanities and Humanities Computing: An Introduction. In S. Schreibman, R. Siemens, and J. Unsworth (eds), *A Companion to Digital Humanities*. Oxford: Blackwell, pp. xxiii-xxvii.

Siemens, R., K. Armstrong, B. Bond, C. Crompton, T. Dickson, J. Paquette, J. Podracky, I. Weber, C. Leitch, M. Chernyk, B. D. Hirsch, D. Powell, A. McLeod, A. Arbuckle, S. Patterson, C. Gaudet, E. Haswell, A. Ciula, D. Starza-Smith, J. Cummings with M. Holmes, G. Newton, J. Gibson, P. Remley, E. Kwakkel, and A. Shirkie (eds). *A Social Edition of the Devonshire MS (BL Ass 17,492)*. `http://www.bitly.com/DevonMS` .

Southall, R. (1964). The Devonshire Manuscript Collection of Early Tudor Poetry, 1532-41. *Review of English Studies: A Quarterly Journal of English Literature and the English Language* 15: 142-150.

Notes

1. In collaboration with The Devonshire MS Editorial Group: Siemens, R., Armstrong, K., Bond, B., Crompton, C., Dickson, T., Paquette, J., Podracky, J., Weber, I., Leitch C., Chernyk M., Hirsch, B. D., Powell, D., McLeod, A., Arbuckle, A., Patterson, S., Gaudet, C., Haswell, E., Ciula, A., Starza-Smith, D., Cummings, J., with Holmes, M., Newton, G., Gibson, J., Remley, P., Kwakkel E. and Shirkie, A.
2. Siemens, R., Armstrong, K., Bond, B., Crompton, C., Dickson, T., Paquette, J., Podracky, J., Weber, I., Leitch C., Chernyk M., Hirsch, B. D., Powell, D., McLeod, A., Arbuckle, A., Patterson, S., Gaudet, C., Haswell, E., Ciula, A., Starza-Smith, D., Cummings, J., with Holmes, M., Newton, G., Gibson, J., Remley, P., Kwakkel E. and Shirkie, A.

Courting 'The World's Wife': Original Digital Humanities Research in the Undergraduate Classroom

Croxall, Brian
b.croxall@gmail.com
Emory University, USA

What would it mean to turn a class of undergraduates – one of the most important emerging digital humanities communities – loose on a text analysis puzzle unintentionally created by England's poet laureate? This poster presentation will report on the process, outcomes, successes, and failures of an original research project for those just learning about the possibilities of humanities computing.

Carol Ann Duffy was named the poet laureate of Britain in May 2009, the first Scot, woman, and openly gay person to hold this position. While this choice signaled a desire for diversity and inclusiveness on the part of the crown, the most important criterion for her selection was her poetry itself. Tackling traditional themes of love along with less conventional ones such as sport, what most sets apart Duffy's poetry from many of her contemporaries is its style, which, on the occasion of her appointment, Sarah Lyall in *The New York Times* described as 'deceptively simple' and which 'produce[s] accessible, often mischievous poems dealing with the darkest turmoil and the lightest minutiae of everyday life' (Lyall 2009).

Perhaps Duffy's style is best exemplified by her 1999 volume, *The World's Wife*, in which she presents short dramatic monologues from the women married to famous men throughout history, mythology, and literature. She presents the stories of 'Mrs. Darwin', 'Mrs. Sisyphus', and 'Mrs. Quasimodo', among others. Clever, humorous, and filled with poems that even rhyme, *The World's Wife* sold tremendously well and began to be used regularly in classrooms. Yet critics felt that the collection was of substantially different quality when compared to her three previous, prize-winning collections. As Jeanette Winterson reported in a profile on Duffy, *The World's Wife* prompted reactions that had critics 'ask[ing] questions about whether Duffy had lost her balance. Had she stopped writing poetry and slipped into verse?' (Winterson).

While Duffy did not comment publicly on such evaluations of her work, undergraduate students in my Spring 2009 poetry class discovered evidence

that she was aware of a difference between *The World's Wife* and those volumes that had preceded it. While exploring her recently acquired papers (http://findingaids.library.emory.edu/documents/duffy834/) in Emory University's Manuscripts, Archives, and Rare Book Library (MARBL), the students found an undated letter from Duffy to the publisher of her previous volumes. She writes that she will be publishing *The World's Wife* with another press: '...this book is not a "normal" poetry collection by me – it's closer to popular entertainment, if you like' (Duffy). So while Duffy later told Winterson that the question of perceived differences in her volumes doesn't concern her (Winterson), her letter makes it clear that she perceives real differences prior to the publication of *The World's Wife*.

This discovery by my students in a 2009 became the seed for the capstone project in my current course, 'Introduction to Digital Humanities' (http://www.briancroxall.net/dh): we are testing Duffy's own words to see whether *The World's Wife* truly *does* differ from her '"normal" poetry collection[s].' To this end, we are reading *The World's Wife* alongside her previous volume, *Mean Time* (1993), which won both the prestigious Forward Poetry Prize and Whitbread Poetry Award. Initially, students wrote essays that employed close reading to make an argument about whether or not there are differences between the two volumes. The class then turned to less traditional modes of analysis. Each student was assigned a number of poems from each volume to transcribe, preparatory for analysis in the suite of Voyant (http://hermeneuti.ca/voyeur), formerly Voyeur (Rockwell 2011), designed by Stéfan Sinclair and Geoffrey Rockwell. Students used different facets of Voyant to explore possible shifts in Duffy's language between the two texts: word frequency, the number of unique words, word collocation, and more. With other tools, we tested the relative line lengths between the volumes and the readability scores of Duffy's language. Students worked in teams on the different modules to understand and interpret their results. The class also turned to the archives and Duffy's notebooks to determine (anecdotally) the rate of changes and corrections between particular poems in the different volumes. Finally, inspired by Stephen Ramsay's arguments that 'Digital Humanities is about building things' and Mark Sample's suggestion that digital humanities is 'about sharing', the students built interpretive arguments about their findings and shared them via the course website.

As mentioned, this poster reports on the processes of designing a digital humanities research project, the process of making those findings public, and the exposure of undergraduates to some of the basic tools and methods of textual analysis. It incorporates not only the results of textual analysis, but also feedback from students about their experience learning new tools and approaches, including a strong emphasis on collaboration – a rarity in most humanities coursework. The presentation continues the trend to discuss the intersection of pedagogy and digital humanities; multiple panels on this subject are being convened at the 2012 Modern Language Convention, for example, and the 2011 Digital Humanities conference at Stanford featured posters by Katherine D. Harris and Beth Bonsignore et al. on the subject (Croxall & Berens 2011; Harris 2011a, 2011b; Bonsignore 2011). This presentation in particular builds on Harris's work on the necessity of uncertainty when designing opportunities for ' play ' by presenting students with a research assignment in which the outcomes are not yet predetermined by the instructor (Harris 2011b).

While reporting on original text analysis research, this presentation simultaneously examines how undergraduates and pedagogy are important facets of the increasingly diverse community of practice that is digital humanities.

References

Bonsignore, B., et al. (2011). The Arcane Gallery of Gadgetry: A Design Case Study of an Alternate Reality Game. *Digital Humanities 2011 Conference Abstracts*, pp. 281-283.

Croxall, B. (2011). *Introduction to Digital Humanities*. http://www.briancroxall.net/dh (accessed 23 March 2012).

Croxall, B., and K. I. Berens (2011). *Building Digital Humanities in the Undergraduate Classroom*. http://www.briancroxall.net/buildingDH/ (accessed 23 March 2012).

Duffy, C. A. (1993). *Mean Time*. London: Anvil Press Poetry.

Duffy, C. A. (1999). *The World's Wife*. London: Picador.

Duffy, C. A. (N.D.). Letter. Box 4, Folder 2. Carol Ann Duffy Papers. Manuscripts, Archives, and Rare Book Library. Emory University, Atlanta.

Emory University (N.D.). Carol Ann Duffy Papers, 1985-2007. http://pid.emory.edu/ark:/25593/8z7md (accessed 23 March 2012).

Harris, K. D. (2011a). Acceptance of Pedagogy & DH MLA 2012. *triproftri*. (14 May 2011). http://triproftri.wordpress.com/2011/05/14/acceptance-of-pedagogy-dh-mla-2012/ (accessed 23 March 2012).

Harris, K. D. (2011b). Pedagogy & Play: Revising Learning through Digital Humanities. *Digital Humanities 2011 Conference Abstracts*, pp. 319-321.

Lyall, S. (2009). After 341 Years, British Poet Laureate Is a Woman. *The New York Times* (2 May 2009). http://www.nytimes.com/2009/05/02/world/europe/02poet.html (accessed 23 March 2012).

Ramsay, S. (2011). Who's In and Who's Out. *Stephen Ramsay* (8 Jan. 2011). http://lenz.unl.edu/papers/2011/01/08/whos-in-and-whos-out.html (accessed 23 March 2012).

Rockwell, G. (2011). Name Change. *Hermeneuti.ca* (19 Oct. 2011). http://hermeneuti.ca/node/212 (accessed 23 March 2012).

Sample, M. (2011). The Digital Humanities is Not About Building, It's About Sharing. *Sample Reality*. (25 May 2011). http://www.samplereality.com/2011/05/25/the-digital-humanities-is-not-about-building-its-about-sharing/ (accessed 23 March 2012).

Winterson, J. (N.D.). Carol Ann Duffy. *Jeanette Winterson*. http://www.jeanettewinterson.com/pages/content/index.asp?PageID=350 (accessed 23 March 2012).

The Academy's Digital Store of Knowledge

Czmiel, Alexander
czmiel@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Jürgens, Marco
juergens@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

1. Introduction

The construction of the Digital Knowledge-Store intends to collect and bundle all digital resources of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)[1]. It extends the existing infrastructure for publishing Digital Humanities knowledge resources in order to increase the visibility of the Academy's research activities in the World Wide Web.

The BBAW is an institution with a long tradition in humanities research. It hosts various humanities long-term projects, which also do research on textual resources in many different languages from all over the world and any historic period. These projects generate research results that are also published digitally and include a variety of different resource types ranging from digital and interactive scholarly editions over collections of databases and images to multimedia content. These digital resources are unique and contain high quality content, for example, digital publications of medieval charters, ancient inscriptions, stained glass studies, manuscripts and much more. The variety of research content and data formats lead to the desire to build a centralized access point to all digital resources of the BBAW.

During the last decade the BBAW was deeply involved in the development of methods, tools, publications and frameworks of digital media and digital resources for scholars in the humanities. With the foundation of the TELOTA[2] (The Electronic Life Of The Academy) working group ten years ago, the Academy became an active part of the ongoing research and development in the field of Digital Humanities. The main tasks of TELOTA are to evaluate and discuss the possibilities that digital technologies offer scholars working at the academy, and to develop tools and digital resources in close collaboration with the Academy's numerous research

projects. Hence TELOTA is also responsible for the development of the Digital Knowledge-Store.

The proposed paper will present the Digital Knowledge-Store of the Berlin-Brandenburg Academy of Sciences and Humanities and show possibilities and methods in the development of an interdisciplinary virtual research environment containing a high diversity of content, languages, data formats and objects.

2. The Digital Knowledge Store of the Academy

The work on the Digital Knowledge-Store was started in 2009 by the TELOTA working group of the Academy. During the current project phase, which started in September 2011 and is funded for three years by the Deutsche Forschungsgemeinschaft (German Research Foundation)[3], it will be improved with extended metadata for better interoperability, a search component based on linguistic methods, and an innovative dialog based user interface. The Knowledge-Store aims to be an infrastructure that offers an easy to use but powerful interface, machine readable APIs and web services for translating queries and analyzing multilingual texts. By connecting the resources semantically with the means of linguistic analysis, text mining and semantic web techniques the infrastructure will support interdisciplinary research in- and outside of the Academy.

2.1. A dialog-based browsing through the Knowledge-Store

The potential of the Knowledge-Store lies in its way of dealing with the process of searching. One can think of the interface as a way rather to interact with the Knowledge-Store than just a query-answer process. The dialog based retrieval interface, the Knowledge-Browser, is conceptually a completely new development and tries to detach from traditional search interfaces by developing entirely new finding strategies. The Knowledge-Browser can be seen as a tool that enables an interactive step-by-step process of computer aided information retrieval.

A possible scenario would be a researcher who has a question about a certain topic but who does not know the BBAW project which could hold the answer. In that case the Knowledge-Store is the tool to solve this problem. The scholar starts the search with a query on a certain subject. In every following step he will receive qualified links to resources which contain the query term or conceptually related terms out of all the Academy's digital resources. With every further step the scholar can browse a related term on the basis of the received results in a different context. As the resources will be linked semantically, every step of the query process offers a new query context. Hence the knowledge browsing is an interactive and multilevel process, which makes implicit connections between different internal and external resources explicit.

Rather than just giving the user the possibility to reduce the search results by filtering, as faceted search interfaces do, it extends the result by offering semantic recommendations for data and documents which were not present in the previous search results. On the basis of these connections the user can search overall in projects and even external project resources to find not only the desired answer but the queried term in different contexts and meanings which will lead to more meaningful connections between the research contexts which were not visible before.

2.2. Challenges and Solutions

The challenges result mainly from the heterogeneous resources which are stored in the various repositories and databases hosted by the BBAW. By defining a metadata scheme for all the different resources in the Academy the Knowledge-Store will provide a basis for the extraction of information by the Knowledge-Browser and a connection to related external projects. This supports interoperability between research data of the BBAW and projects such as Europeana[4], Deutsche Digitale Bibliothek[5], DARIAH[6] or CLARIN[7]. The existing metadata is based on the Metadata Object Description Schema (MODS)[8]. It provides the basic information for search criteria and the description of the resource. For extending the metadata the Europeana Data Model (EDM)[9] will be used to allow sophisticated connections of various heterogeneous resources. With extended metadata the Digital Knowledge-Store is prepared to provide linked open data to be part of the Web of Data[10].

On the level of linguistic analysis an important problem to solve is that the multilingual textual resources in a major part do not follow a normalized orthography or a consistent language space. For the written German language of the 19th century the German Text Archive (DTA)[11] and the Digitales Wörterbuch der Deutschen Sprache (DWDS)[12] provide appropriate tools for an index based solution. For ancient Greek and Latin texts a cooperation with the Max Planck Institute for the History of

Science[13] is established to further the integration of the Donatus / Hopper software package[14]. Donatus is capable of handling texts in Greek and Latin as well as Arabic, German, English, French an others. These services are used for automated lemmatization and stemming as well as text-mining. In combination with Semantic Web technologies such as RDF[15] and RDFa[16] or ontologies, this results in basic semantic retrieval possibilities which enables the researcher to gain new knowledge by a process of searching in different contexts of the resources.

Additionally the Knowledge-Store aims to make all digital textual resources of the Academy accessible as full texts. Therefore it is important to build up an index which is prepared for a search based on linguistic methods, text mining and semantic information retrieval. This index is prepared using the aforementioned services for a search in corpora of different languages and in historic writings as well. The index uses a Lucene[17] / Solr[18] – Framework in combination with XQuery and XSLT-scripts to transform XML-Data and to update the index. Furthermore it is intended to integrate a tool which allows querying the full text without background knowledge of a certain language. This would be possible by sending a query via REST to a translation web service endpoint in order to receive a translated search term. There are several possible translation APIs available. Among those are the Microsoft BING Translation API[19] and the Google Translate API[20] which both provide a translation service that automatically translates text from one language to another.

3. Future prospects

The Knowledge-Store will enable a new way of querying humanities resources. The combination of the Academy's resources, the Knowledge-Browser and the development of information extraction tools will lead to an infrastructure which allows to put the knowledge of the Academy's different projects in different contexts. The Knowledge-Store will be build on open source technologies. On the same time it will realize by these means an interoperable infrastructure which can show new possibilities in the usage of Digital Humanities techniques. It guides the humanities scholar to new ways of thinking about the projects contexts by interconnecting and intercontextualizing the research resources in the BBAW. In developing a semantically interlinked base of resources which is capable of gaining knowledge by querying information in various contexts of the Academy's projects, the Knowledge-Store will serve as a part of the realization of the Web of Data as well as it can be an exemplary project to establish a work flow for the integration of external infrastructure projects into one single endpoint.

A prototype version of the Knowledge-Store which does not include the Knowledge-Browser can be found at `http://www.bbaw.de/en/telota/resources/dkb`.

A much more developed interface will be available at conference time.

References

Bellamy, C. (2010). *What is eResearch in the Arts and Humanities*. `http://www.craigbellamy.net/2010/04/07/what-is-eresearch-in-the-arts-and-humanities/` (accessed 9 March 2012).

Brock, T. (2010). *ReThinking Digital Social Media for Digital Humanities and Community Engagement*. `http://dirt.terrypbrock.com/?p=1116` (accessed 9 March 2012).

Carusi, A., and T. Reimer (2010). *Virtual Research Environment Collaborative Landscape Study*. `http://www.jisc.ac.uk/publications/reports/2010/vrelandscapestudy.aspx` (accessed 9 March 2012).

DINI (2009). *Informations- und Kommunikationsstruktur der Zukunft*. `http://www.dini.de/fileadmin/docs/DINI_thesen.pdf` (accessed 9 March 2012).

Dallmeier-Tiessen, S., et al. (2009). *Positionspapier Forschungsdaten. Arbeitsgruppe Elektronisches Publizieren*. `http://edoc.gfz-potsdam.de/gfz/get/13230/0/a271566a9d9f48030ee78e01ceae7138/13230.pdf` (accessed 9 March 2012).

Davidson, C. N. (2008). The changing profession – Humanities 2.0: Promise, Perils, Predictions. *Publications of the Modern Language Association of America* 123(3): 707-717.

De Virgilio, R. (2010). *Semantic Web information management: a model-based perspective*. Springer.

Heath, T., and C. Bizer (2011). *Linked Data: Evolving the Web into a Global Data Space*. San Rafael, CA: Morgan & Claypool.

Key Perspectives Ltd (2010). *Data Dimensions. Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability*. `http://www.era.lib.ed.ac.uk/bitstream/1842/3364/1/SCARP SYNTHESIS.pdf` (accessed 9 March 2012).

Murray-Rust, P., et al. (2010). *Panton Principles. Principles for Open Data in Science*. Open Knowledge Foundation. `http://pantonprinciples.org/` (accessed 9 March 2012).

Neuroth, H., et al. (2009). Virtuelle Forschungsumgebungen für e-Humanities. Maßnahmen zur optimalen Unterstützung von Forschungsprozessen in den Geisteswissenschaften. *Bibliothek, Forschung und Praxis* 33(2): 161-169.

Rees, J. (2010). *Recommendations for independent scholarly publication of data sets*. Creative Commons Working Paper. San Francisco.

Sierman, B., B. Schmidt B., and J. Ludwig (2009). *Enhanced Publications: Linking Publications and Research Data in Digital Repositories*. http://dare.uva.nl/document/150723 (accessed 9 March 2012).

Notes

1. http://www.bbaw.de/en/ (accessed 9 March 2012).
2. http://www.bbaw.de/en/telota (accessed 9 March 2012).
3. http://www.dfg.de/en/ (accessed 9 March 2012).
4. http://www.europeana.eu/portal/ (accessed 9 March 2012).
5. http://www.deutsche-digitale-bibliothek.de/ (accessed 9 March 2012).
6. http://www.dariah.eu/ (accessed 9 March 2012).
7. http://www.clarin.eu/ (accessed 9 March 2012).
8. http://www.loc.gov/standards/mods/ (accessed 9 March 2012).
9. http://www.europeanaconnect.eu/news.php?area=News&pag=48 (accessed 9 March 2012).
10. http://www.w3.org/standards/semanticweb/data (accessed 9 March 2012).
11. http://www.deutschestextarchiv.de/ (accessed 9 March 2012).
12. http://www.dwds.de/ (accessed 9 March 2012).
13. http://www.mpiwg-berlin.mpg.de/en/ (accessed 9 March 2012).
14. http://archimedes.fas.harvard.edu/cgi-bin/donatus (accessed 9 March 2012).
15. http://www.w3.org/RDF/ (accessed 9 March 2012).
16. http://www.w3.org/TR/xhtml-rdfa-primer/ (accessed 9 March 2012).
17. http://lucene.apache.org/java/docs/index.html (accessed 9 March 2012).
18. http://lucene.apache.org/solr/ (accessed 9 March 2012).
19. http://www.microsofttranslator.com/tools/ (accessed 9 March 2012).
20. http://code.google.com/intl/de-DE/apis/language/translate/overview.html (accessed 9 March 2012).

Building a TEI Archiving, Publishing, and Access Service: The TAPAS Project

Flanders, Julia
Julia_Flanders@brown.edu
Center for Digital Scholarship, Brown University, USA

Hamlin, Scott
hamlin_scott@wheatoncollege.edu
Library and Information Services, Wheaton College, USA

Alvarado, Rafael
ontoligent@gmail.com
Humanities and Arts Network of Technological Initiatives, University of Virginia, USA

Mylonas, Elli
Elli_Mylonas@brown.edu
Center for Digital Scholarship, Brown University, USA

As the language of DH 2011's 'big tent' suggests, in recent years the profile of digital humanities work has expanded to include many scholars and practitioners who draw on a multitude of digital technologies in research and educational contexts, without assuming all the roles required to implement these technologies. This new generation (though some are 'new' only to digital matters) of digital humanists are undertaking intellectually ambitious work with digital methods and tools, but their interest does not necessarily arise from a strong institutional history or infrastructure, or from personal expertise with digital methods. Rather, they are practicing scholars who are increasingly aware of the shifting stakes of technology for the humanities, and who want to explore what may be possible by working in a new way. As a result, their ambitions often outstrip what their own institutions can support: the available infrastructure of digital publishing, archiving, data curation, and repository services may be limited or absent. An individual scholar can gain expertise and achieve interesting results using the TEI Guidelines (http://www.tei-c.org/) or GIS, but it is a slower and more challenging process for a university to develop the institutional infrastructure to support that expertise, in the way that traditional libraries (for instance) support traditional forms of humanities scholarship.

The TEI Archiving, Publication, and Access Service (TAPAS, http://www.tapasproject.org/) is aimed at addressing this gap, by providing repository

and publication services for small TEI projects. TAPAS began with a planning grant from the IMLS (TAPAS 2010), originally proposed by a group of small liberal-arts institutions including Wheaton College, Willamette University, Hamilton College, Vassar College, Mount Holyoke College, and the University of Puget Sound, and later joined by Brown University and the University of Virginia. This planning group conducted an intensive study of the profile of needs, and developed a specification for the TAPAS service. TAPAS is now operating under a two-year IMLS National Leadership Grant to Wheaton College and Brown University which funds the development of the service. TAPAS has also received an NEH Digital Humanities Startup Grant, led by Wheaton College and the University of Virginia, which funds the development of the user interface. Hosted at Brown University, the TAPAS service will provide repository storage, data curation, and simple interfaces for data management and publication. It will also provide an API through which the TEI data can be accessed and remixed. The service thus aims to fill a crucial niche, enabling both a new type of publication and a new model for how scholarly publication is supported. All of these needs are particularly urgent in the liberal-arts community that is the central focus of TAPAS, but they are also strongly evident in the humanities academy more broadly, at a national and international level.

This project takes place within a landscape already well populated with large-scale infrastructural projects (Hedges 2009), such as TextGrid (`http://www.textgrid.de/`), DARIAH (`http://www.dariah.eu/`), CLARIN (`http://www.clarin.eu/external/`), and the Canadian Writers Research Collaboratory (CWRC, `http://www.cwrc.ca/)`). Projects of this kind must all confront a central set of strategic concerns and design challenges, including questions about how much uniformity to impose upon the data, how to accommodate variation, how to create interoperability layers and tools that can operate meaningfully across multiple data sets (DARIAH 2011a, DARIAH 2011b), and how to manage issues of sustainability (of both the data and the service itself). TAPAS is distinctive within this landscape because of its focus on a single form of data (TEI-encoded research materials) and also because of its initial emphasis on serving an underserved constituency (scholars at smaller or under-resourced institutions) rather than on providing an infrastructure that can operate comprehensively. TAPAS is thus able to tackle the questions above in a highly focused way.

The proposed poster will focus on several key areas of the TAPAS project that will be the focus of our attention in the early phases of the project:

1. Architecture and system design. The TAPAS service is built on a Fedora repository, and the user interaction will be managed as a set of modular layers using tools like Drupal. The design of these layers needs to take into account information about how scholars need to interact with the service for activities such as:
   - creating new project records
   - uploading new data files, uploading revised versions of existing data files
   - creating metadata for data files, updating the metadata for existing files
   - configuring options for dissemination, publication, and other modes of access and discovery, such as interface choices, stylesheets, and information to be exposed via APIs.

   The poster will provide a detailed look at the internal architecture and the ways that standards like RDF and METS are used to organize information and enable flexible deployment of repository data.
2. TEI schema development and the challenge of eclecticism. TAPAS plans to accept a broad range of TEI data, but will also need to identify different classes of data that share specific properties, such as genre or the presence of certain encoding features, to determine what kinds of interface tools will or will not be appropriate for a given data set. TAPAS will also use various forms of validation to help TAPAS contributors ensure the consistency and quality of their data prior to upload. The design and use of schemas used within the TAPAS ecology – extending from training, through data creation and management, to long-term data curation is complex and will be an important focus of the project's research. The poster will provide a detailed view of the different roles that schemas of various kinds will play in this ecology, and the principles guiding their design and use.
3. Designing a hosted service. Although TAPAS was prompted by the needs of individual scholars, its implementation as a hosted service means that it also plays an important aggregative role. The TAPAS collection of TEI data has the potential not only to serve as an important corpus of TEI data (of value, for instance, to those interested in the historiography of digital humanities, or in studying how the TEI is used) but also to provide important inter-project connections that may benefit the individual TAPAS contributors and their readers. In addition, designing TAPAS as a hosted service raises a number of issues concerning long-term data curation, rights, and the fiscal sustainability of the service itself. The poster will examine these issues with a particular focus on:

- the membership and sustainability model for the service
- the handling of intellectual property rights
- the design of cross-project tools for searching, exploration, and visualization

4. User interface. Because TAPAS is intended to support scholars who – although they may be expert users of TEI – are not necessarily experts in working with repositories and XML publication, the design of the user interface will be critical in making the TAPAS service approachable. In addition, because some users may be managing very large numbers of files, the user interface will need to provide productive, intuitive ways of visualizing one's data from a management standpoint as well as a publication standpoint.

Funding

This project is made possible by a grant from the U.S. Institute of Museum and Library Services.

References

DARIAH (2011a). Technical Work: Conceptual Modelling. DARIAH Work Package 8. http://www.dariah.eu/index.php?option=com_content&view=article&id=31&Itemid=35

DARIAH (2011b). Technical work: Technical reference architecture. DARIAH Work Package 7. http://www.dariah.eu/index.php?option=com_content&view=article&id=30&Itemid=34 .

TAPAS (2010). Roadmap. http://www.tapasproject.org/roadmap

TextGrid (2010). Roadmap Integration Grid/ Repository. TextGrid, September 2010. http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R121_v1.0.pdf .

Hedges, M. (2009). Grid-enabling Humanities Datasets. Digital Humanities Quarterly 3(2). http://www.digitalhumanities.org/dhq/vol/3/4/000078/000078.html

Author Consolidation across European National Bibliographies

Freire, Nuno
nfreire@gmail.com
The European Library, National Library of the Netherlands, The Netherlands

1. Introduction

The European Library holds several million bibliographic records from the national libraries of Europe. The National Bibliographies are one of the main bibliographic data sources in each country and are key for Digital Humanities scholars.

Their purpose is to list every publication in a country, under the auspices of a national library or other government agency. Depending on the country, all publishers will need to send a copy of every published work to the national legal deposit, or in other countries, a national organisation will need to collect all publications.

Given that the publisher domain is very heterogeneous and that thousands of publishers might exist in a country, National Bibliographies are effectively the single point of reference with which to comprehensively identify all the publications in a country.

In Europe, National Bibliographies are typically created and maintained by national libraries. Whenever a book is published in a country, it is recorded in the corresponding national library catalogue from where the national bibliography is derived.

Currently, *The European Library* holds approximately 75 million bibliographic records in its centralised repository. This number is constantly increasing, as more national libraries' catalogues are included in the centralised repository. By the end of 2012 the total bibliographic universe of *The European Library* is expected to be approximately 200 million records.

The centralisation of European bibliographic data in *The European Library* is creating new possibilities for the exploitation of this data, in order to improve existing services, enable the development of new ones, or provide a good setting for research.

The centralisation of the bibliographic data enables the automatic linkage of the National Bibliographies

across countries, through the use of data mining technologies. Our ongoing work is focusing the linkage of the main entities used to describe bibliographic resources: persons, organizations, locations, historical periods, and subjects.

In this poster, we present the current status of our work on the consolidation of authors across the National Bibliographies of Europe. When complete, it will allow the exploration of an author's work across time and space in the European bibliographic universe.

2. Problem Description

Entity resolution is the process of, given a specific context, determining if two or more references correspond to the same entity. An entity might have multiple different representations, and each representation might match the description of multiple objects (i.e., reference and referent ambiguity). The variations found in the descriptions may have multiple origins, such as misspellings, typing errors, different conventions for abbreviations, naming varying over time, heterogeneous data schemas, etc.

Entity resolution is a common problem to many different research communities, although the term used is not always the same. Common designations include record linkage, record matching, merge-purge, data de-duplication, instance identification, database hardening, name matching, reference reconciliation, reference disambiguation, and object consolidation (Elmagarmid et al. 2007; Dong et al. 2005).

Different communities have proposed several techniques, but most frequently we find applications of algorithms from machine learning, artificial intelligence and data mining (Elmagarmid et al. 2007).

Entity resolution is very dependent on the context. The processes need to be adapted to the data that they are being applied to, in order to achieve acceptable results. Our work focuses on exploring the structural and semantic richness of the bibliographic data, as it exists in the National Bibliographies.

In bibliographic records, references to persons are found as authors or contributors of works, and sometimes as the subject of the work. The value of these references being as complete as possible is recognised by cataloguing rules, which indicate that the references should contain not only the name of the person but also their birth and death years (ALA et al. 2002). However, these dates do not always exist, as they are not always known by the cataloguers. It is also often the case that these dates are approximations of real dates – cataloguing rules comprise conventions for these cases. Although this information is not always fully structured, the common conventions used when encoding the information allow it to be reliably parsed automatically.

3. Approach

Our approach leverages two key aspects of the National Bibliographies and consolidation work on authors carried out by libraries.

The first aspect is that national libraries already individually perform a manual consolidation of authors through their ongoing work to maintain National Bibliographies.

The second aspect is that some European national libraries actively work on the construction of the Virtual International Authority File, or VIAF (Bennett et al. 2006). VIAF is a joint project of several national libraries from all continents. It hosts a consolidated data set containing data that national libraries have gathered for many years about the authors of the bibliographic resources held at the libraries. It is available as open data.

Using VIAF, we can already consolidate the authors across the VIAF participating countries and soon we will exploit this resource to consolidate authors in other countries. By extracting statistics about authors consolidated in VIAF, from the National Bibliographies of VIAF participants, we expect to be able to derive a probabilistic model that will allow us to consolidate the authors from other countries not participating in VIAF.

The author consolidation system is being built as an ETL (Extraction, Transformation and Loading) process, a typical approach for performing consolidation of data in data warehouses.

The process starts with the preparation of data for consolidation processing. This step comprises tasks for selecting the relevant data from the National Bibliographies that will be used to represent an author during the consolidation process. The following data about the authors is gathered at the level of each individual national bibliography:

- Name;
- Variant forms of the name;
- Years of birth and death;
- Known co-authors;
- Known publishers;
- Known titles of publications.

As is typical in ETL processes, the final decision to match two author references is made by reasoning on

the similarity scores obtained by comparing the each of the above data elements.

Based on previous research on this topic (Freire et al. 2008), we use the Jaro-Winkler similarity metric (Jaro 1989) for calculating the similarity between names.

Comparison of remaining data fields is based on the number of common values found in the two records versus the total number of available values. For example, similarity of co-authors is given by the number of common co-authors between two authors, and similar calculations for titles, and publishers.

This information is compared across National Bibliographies to identify, for example, two authors who can actually be considered to be the same.

The solution for reasoning on the outcome of comparisons will be based on supervised machine learned model. It will process the comparison results between two of the above records of the authors and determine the likelihood that the authors are the same entity. Several machine learning techniques for classification will be tested.

As ground truth, for building and testing the machine learned model, we will use a data set extracted from the National Bibliographies of VIAF participants, therefore removing the need for creating a manually annotated collection.

4. Conclusion

This poster presented the current status of our work on the consolidation of authors across the National Bibliographies of Europe. When complete, it will allow the exploration of an author's work across time and space, in a much more comprehensive way then is possible today.

We believe that this consolidation will provide new opportunities for the creation of statistical datasets, resulting from data analysis and mining of bibliographic data. Although no concrete plans are ready at this time, we expect these kinds of data sets to emerge as a future result.

This work will also have impact in intellectual property rights identification processes, such as those of the ARROW project[1] (Accessible Registries of Rights Information and Orphan Works). In such processes, the identification of all publications of a work by an author is essential, and can benefit from this consolidated author's bibliographies.

References

ALA, CLA, CILIP (2002). Anglo-American Cataloguing Rules: 2002 Revision.

Bennett, R., C. Hengel-Dittrich, E. O'Neill, and B. B. Tillett (2006). VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority Files.

Dong, X., A. Halevy, and J. Madhavan (2005). Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data. SIGMOD '05*. ACM: New York, NY, pp. 85-96.

Elmagarmid, A. K., P. G. Ipeirotis, and V. S. Verykios (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on knowledge and data engineering* 19(1): 1-16, DOI: 10.1109/TKDE.2007.250581.

Freire, N., J. Borbinha, and B. Martins (2008). Consolidation of References to Persons in Bibliographic Databases. ICADL 2008. *The 11th International Conference on Asian Digital Libraries, Universal and Ubiquitous Access to Information*.Berlin: Springer-Verlag, pp. 256-265.

Jaro, M. A. (1989). Advances in record linking methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Society* 64: 1183-1210.

Notes

1. 1. Accessible Registries of Rights Information and Orphan Works website: `http://www.arrow-net.eu/`

Historical Events Versus Information Contents – A Preliminary Analysis of the National Geographic Magazine

Fujimoto, Yu
yfujimoto@doshisha-bkj.net
Doshisha University, Japan

1. Introduction

In 2009, Nara University auctioned its full set of 'National Geographic Magazine (NGM)', complete back to its first release in 1888. In response to the event, faculty and student devotees of the magazine launched a small working group in June of 2011. Although in the beginning the working group was satisfied with simply browsing the magazines, as discussion continued, members soon recognised academic importance of the collection.

Since magazines are generally driven by the mass interests of their readerships in order to raise their sales, shifting social conditions would become clear by observing variation in themes and physical information content over time in weekly or monthly magazines. In the case of NGM, its readers are mainly drawn from the intellectual classes, which can be thought to effect on global politics disproportionally, and the magazine covered a vast range of cultural and natural themes for more than 120 years.

Because NGM possesses all these characteristics, the author predicts that social conditions since the end of 19th century would become clear through appropriate analyses. Although the working group's goal is to identify position of 'Japan' in global society from the 1890s to the 1950s as seen through NGM, the author focuses on deepening understanding of the magazine itself by analysing how it has changed overtime. In this paper, the author reports on the shift in the physical amount of information over time as the one of the preliminary analyses employing mathematical methods.

Figure 1: Conceptual schema for NGM research.

2. The research material

NGM has appealed to people all over the world since its first release in 1888 (Hubbard 1888), and Japan has been the focus of an issue's main article more than 90 times since its first appearance in 1894 (Stevens 1894). Some important Japanese historical figures also commented on Japan since then. Additionally, the magazine always took a neutral stance, and even during World War II, conditions in Japan were described objectively; Japan was described as a strong rival to the United States rather than an evil state (Price 1942). Many of photos and illustrations, which are also important characteristics of the magazine, are further significant for Japanology since a huge amount of important materials vanished in 1923 during the great earthquake in Tokyo, and again in 1945 during the American firebombing of Tokyo.

Currently, the working group is planning the 3 phases of research to analyse the 'Japan' described in the magazine. The preliminary phase aims to understand the magazine itself, the resource acquisition phase archives various information concerning the magazine, and the analytical phase attempts to identify position of 'Japan' in global society from the 1890s to the 1950s from new standpoints. Overall, processes should be based on informatics not to put Japanese feelings in historical facts.

In order to understand the position of 'Japan' in global society through studying NGM, first, the characteristics of the magazine itself should be clarified. Therefore, the author designed an entire conceptual schema as a guide line for the project. Figure 1 is the schema that the author draw at the outset of this project. The following sections in this paper reports the results of operations defined on the 'Magazine' class and related classes highlighted in the schema. The author gave much attention to temporal shift in the information content as reflected in ocular information. Analysis of information content used the DVD version of the digital archive, published by

the National Geographic Society, which includes all of the articles since the first issue in 1888 (National Geographic Society, 2010). Many devotees of the magazine enjoy the photographs and 'browse' pages rather than 'read' articles, and the series of analyses in this paper reflect this tendency.

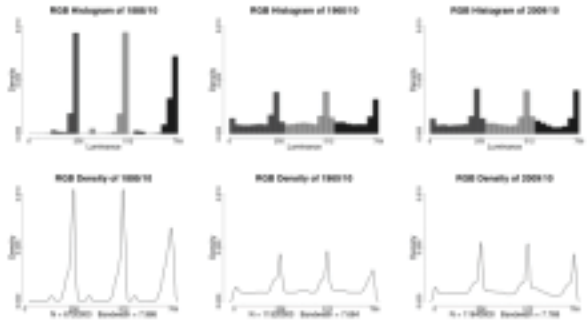


Figure 2: Histogram and density plot of colour use

3. Analyses

Although file size would be an adequate index for the amount of information, as well as the easiest approach, the author chose to also analyse the use of colours and entropy of thumbnails in each volume. The first processes consisted of randomly extracting constant numbers of pixel-values 4,096 times from each page to justify the size of the thumbnails, and in parallel, counting the number of pages in each volumes. After those processes, Normalised Colour Values (NCV) files, which consist of RGB values in each volume, were created. The NCV files were assigned a range of values from 0 to 767, and divided into three frequencies: red values were assigned from 0 to 255, green values were assigned from 256 to 511 and blue values were assigned from 512 to 767.

In this study, NCV files were programmatically created for both full pages and cover pages, and shapes of the colour usage distribution of cover pages in each volume were calculated. Changes in colour were detected using Fourier Descriptors, in other words the coefficient numbers of Fourier Transform defined as below:

$$F_{(x)}=\frac{a_0}{2}+\sum_{n=1}^{N}\left[a_n\cos(nx)+b_n\sin(nx)\right]$$

Any curves are denoted as infinite summations of two trigonometric function having coefficient of *a* and *b* (Fourier Series), and ordered earlier coefficients reflect major changes (lower frequency component), while ordered later coefficients reflect minor changes (higher frequency component) in the curve. Fourier Descriptors are finite version of Fourier Series. In this study, Fourier Descriptors were derived from Fast Fourier Transforms (FFT).

In contrast to the FFT, entropy was used to measure complexity of the magazine (Shannon 1948). Entropy is denoted as below:

$$H_{(x)}=-\sum_{i=1}^{N} p(x_i)\log_b p(x_i)$$

Here, *H(x)* is the average information content, the so-called entropy, of a series of random variable *x*, and *p* denotes the probability mass function of *x*. Therefore, *p(xi)* shows occurrence ratio of each luminance value of RGB. *b* specifies the unit of entropy, and 2 (bit) was used in this study. Using entropy, information content in each volume was described as composite index of pages and colours. Finally, changing points of the shift in entropy were detected by Bayesian Changing Point (BCP) method (Barry & Hartigan 1993).

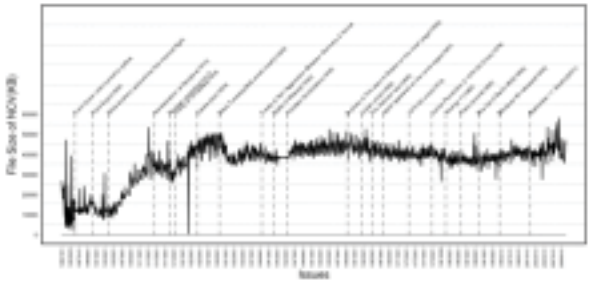


Figure 3: Comparison with information content and historical events

4. Results

The results of these analyses were compared to historical events and tenure of chief editors (Figure 3). First, the shift in file sizes seemed to be related to historical events. The first peak was detected around 1913 after which the curve of information content was decreased until 1919. The curve fit neatly with the duration of World War I. Afterwards, information content rapidly increased until the Wall Street Crash of 1929, known as 'Black Tuesday', after which the curve nosedived. During World War II, the curve had small amplitude, and then the curve increased moderately up to the Cuban Missile Crisis in 1962. The curve was in a gradual decline until the Vietnam War ended. Although a peak was detected around 1979, the year of Iranian Revolution and Oil Crisis, the curve has been stable since then.

Second, the shift of entropy seemed to be related to the history of chief editors. The curve describing entropyover time was unstable until around 1916

(Figure 4). This unstable condition might reflect many challenges of foundational periods. Indeed, another analysis of colour distribution shape for cover pages shows

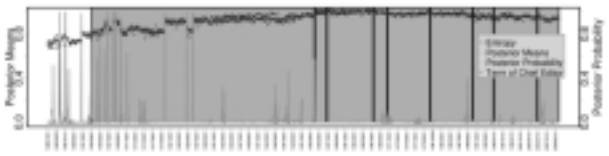


Figure 4: Changing point detection
by Bayesian Change Point(BCP)

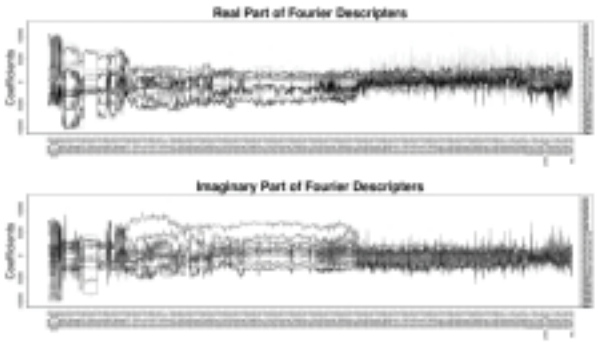


Figure 5: Temporal shift of Fourier Descriptors

similar tendency; although the cover page of NGM had no images at first, the illustrated cover pages had began around 1910 (Figure 5). After 1916, BCP of the shift of entropy was more stable than before, but small peaks appeared near the years in which chief editors changed.

5. Conclusion

The results derived from NCV analysis of the NGM seems to show a relationship to historical occurrences. That is, that NGM represents the conditions of the global society in some way. The results of any analyses should be more clear by overlapping them with these curves. On the other hand, the curves representing physical information content might not fit well with economic curves. Another question is, what does information content actually represent?

The results of these analyses are only a partial implementation of the schema I defined, and I would prefer to refrain from discussing any historical events with these results, under current circumstances. Our working group is currently collecting various kinds of materials, which may concerning NGM. As part of our future research, the author is planning to conduct text mining using the full text of the magazine, and to compare it to the results of this study.

Acknowledgement

This work was supported by Grants-in-Aid for geographic research of Fukutake Science & Culture Foundation (12-GEO-9).

References

Barry, D., and J. A. Hartigan (1993). A Bayesian Analysis for Change Point Problems. *Journal of American Statistical Association* 88(421): 309-319.

Hubbard, G. G. (1888). Introductory Address by the President. *The National Geographic* 1(1): 8-10.

National Geographic Society (2010). [DVD] The Complete National Geographic: National Geographic Society.

Price, W. (1942). Unknown Japan. *The National Geographic* 82(2): 224-252.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379-423, 623-656.

Stevens, D. W. (1894). Japan. *The National Geographic* 4: 192-199.

'Tejiendo la Red HD' – A case study of building a DH network in Mexico

Galina, Isabel
igalina@unam.mx
Universidad Nacional Autónoma de México, Mexico

Priani, Ernesto
epriani@gmail.com
Universidad Nacional Autónoma de México, Mexico

López, José
josemlv@msn.com
Universidad Nacional Autónoma de México, Mexico

Rivera, Eduardo
eduardo.rivera.sumano@gmail.com
Universidad Nacional Autónoma de México, Mexico

Cruz, Alejandro
alejandro_cruz_cervantes@hotmail.com
Universidad Nacional Autónoma de México, Mexico

1. Introduction

As has been noted previously (Borgman 2009; Friedlander 2009; Presner 2009) Digital Humanities (DH) is becoming an established field. However, one of the remaining challenges is to locate and increase collaboration with new emerging DH communities. At the Universidad Nacional Autónoma de México (UNAM) we are working on a) identifying key scholars and projects; b) investigating key local issues in the development of DH projects; c) raising awareness of DH and d) consolidating a community and linking with international DH. Previously we reported (Galina & Priani 2011) on the results from a series of workshops aimed at developing a diagnosis of the DH landscape including the identification of key challenges and issues for the development of projects in order to consolidate and promote DH in the region. An important result of the workshops was the initiative to form a DH community (Red de Humanidades Digitales – RedHD), first in Mexico and then expanding to other Latin American countries. The aim of this poster is to present the multi-faceted approach to DH community building that we are adopting. We will present and discuss the different strategies we have employed and the main challenges encountered. We hope that this model will be useful for other emerging DH communities to build upon.

2. Development of the network

One of the first challenges was moving from a workshop environment to establishing a network of DH practitioners while still remaining a vibrant and participatory group. Although virtual communities can be useful, if not managed appropriately they run the risk of becoming an empty and lifeless virtual space. Although all workshops participants were enthusiastic about promoting DH, in practice we feared that our different disciplines, geographical distance and personal workloads could lead to a weakening of our initiative. Reingold (1993) defines a virtual community as a situation in which 'people carry on public discussions long enough, with *sufficient human feeling*, to form webs of personal relationships'. For this we needed to assure that we had good **communication channels** to **talk**, **organize** and **work** with **relevant people** on particular **topics**.

Therefore, one of the first issues was defining who would form part of the network? Following on from the workshops, where we used a very ample definition of DH, we also wanted to make the community as inclusive as possible. University structures can be inflexible, rigid and hierarchical. Many of our workshop participants were not formal employees of the UNAM and we also had participants from other institutions, as well as non academic organizations. Many DH project participants are not established academic members and RedHD should function in such a way as to incorporate this diversity. The second step was to define specific objectives with concrete goals in order to promote meaningful communication of useful information between us. We needed to develop a web based information system that acted as a hub as well as providing information and communication channels.

From the workshops we also defined the initial topics that needed attention:

1. Information: about digital humanists, digital resources and projects and relevant publications.
2. Formation of human resources: strategies, guidelines, documentation and other activities that promote the formation of human resources required for the development of DH.
3. Project evaluation: necessary actions to generate standards, policies and indicators in order to promote the validity of DH projects and lobbying for the recognition of DH as a valid academic field of enquiry.

Information: Initially we defined three different types of information that we wanted to provide through our website: a database of DH practitioners, a catalogue of DH projects and a bibliography of

published DH results from a particular DH project. The definition of appropriate XML metadata will be described in more detail at a later date, but for the moment it is important to mention that we wanted the databases to work as a system of information with interconnections. For example, for a DH scholar you can also query all DH projects they have worked on as well as the publications they have, and in the same way a DH project will also list the relevant publication and participants and so forth.

Formation: We found a very limited offer of DH related courses and one of RedHD's aims is to build upon this. Initially, as means of promoting DH in general, we are working on short introductory type courses (for example, 'Digital resources and the Humanities', 'Introduction to TEI'). This work is leading up to the design of a longer certificate/ specialization course[1] to be offered at the UNAM beginning of 2012. The aim is that this program will later be developed into a MA course. Human resources can be formed through teaching, but we are also working on the development of documentation, guidelines, publications and other type of support material in Spanish, as there is little available. We published a special issue of Digital Humanities for RDU, a UNAM peer-reviewed, general interest journal (RDU 2011). We are also focusing on the social networking approach for both academic and non-academic audiences. We are also planning a local Digital Humanities conference, the first of its type in the region, in Spring 2012.

Project evaluation: Another key element for the development of DH are the current difficulties involved in the evaluation of research performance of digital humanists as well as the impact and value of DH projects as a research outcome in themselves. One of reasons for this is that evaluation committees do not necessarily have the tools or the know-how to do so. RedHD aims to contribute towards recognition by providing tools and guidelines to help out with this task. This particular type of work requires a more in depth approach and so we have set up a working group that includes face to face meetings. The group is currently revising international available guidelines and will work on developing material not only in Spanish, but also adapted to our own particular institutional and funding structures.

3. Discussion

The consolidation of RedHD has been a huge challenge. Although money is a common problem in Humanities, funding has been a particular ordeal for us. The development of the website, which is the core infrastructure for our network, has consequently been slow. We depend heavily on student work. Long term funding is an issue that we have not yet addressed. In this sense, the registry of DH projects in particular, is a concern. The discontinuation of AHDS, Intute and similar DH resource discovery systems is a negative sign. However, it also presents an opportunity for us to learn from previous initiatives and focus on different approaches to long term sustainability. Initial ideas are linked to our grass root development and working in partnership with libraries (Galina 2011).

In terms of numbers the RedHD has grown little since the workshops. We have not actively sought new members in part due to the delays in the development of our website. The digital humanist registry database is fundamental to our growth as a community and the system must be stable before we encourage new partnerships. However, already there has been a lot of interest and once established we believe that the community will grow rapidly.

We also want to focus on connecting with the international DH community from whom we have had a lot of interest and support. Both DH11 and DH12 have made specific references to emerging digital communities in their call for participation. However, involvement is still low. This could be due to a number of reasons. One is that we are still building momentum. Another factor may be language, as English is the main discourse for publication and participating in DH. It is possible that increased professionalization of the DH field will result in more English speaking members. However, we also we want to focus on Spanish documentation and projects as well. Organizing local, regional and Latin American events will be vital part of our work.

4. Conclusions

There is still a lot of work to do towards consolidating and more importantly maintaining RedHD. Ideally the network will move from being based on enthusiastic individuals to being incorporated into institutional frameworks. Institutional recognition and backing would support long term sustainability. Ideally this network will promote the establishment of DH centres, postgraduate courses, journals, established working groups and other types of academic endeavor. RedHD for the moment works as a proof of concept allowing us to advance towards the next phases.

References

Borgman, C. (2009). The Digital Future is Now: A Call to Action for the Humanities. *DHQ: Digital Humanities Quarterly* 3(4).

Friedlander, A. (2008). Asking Research Questions and Building a Research Agenda for Digital Scholarship. In *Working Together or*

Apart: Promoting the Next Generation of Digital Scholarship http://www.clir.org/pubs/reports/pub145/pub145.pdf

Galina, I. (2011). *El papel de las bibliotecas en las Humanidades Digitales*, 77th IFLA General Conference and Assembly, San Juan, Puerto Rico. 13-18 August 2011 http://conference.ifla.org/past/ifla77/104-russell-en.pdf

Galina, I., and E. Priani (2011*). Is There Anybody Out There? Discovering New DH Practitioners in other Countries. Digital Humanities 2011, Conference abstracts*. Stanford 2011, pp. 135-138

Revista Digital Universitaria-RDU (2011). *Las Humanidades Digitales*, Special Issue 12(7) http://www.revista.unam.mx/vol.12/num7/art68/index.html

Rheingold, H. (1993). *http://www.rheingold.com/vc/book/intro.html The Virtual Community*. Reading, Mass.: Addison-Wesley.

Presner, T., and C. Johanson (2009). *The Promise of Digital Humanities*. UCLA.

Notes

1. Known as *diplomado* this type of academic course is highly specialized and most cover a minimum of 120 hours.

Adaptive Automatic Gesture Stroke Detection

Gebre, Binyam Gebrekidan
binyamgebrekidan.gebre@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Wittenburg, Peter
peter.wittenburg@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

1. Introduction

Many gesture and sign language researchers manually annotate video recordings to systematically categorize, analyze and explain their observations. The number and kinds of annotations are so diverse and unpredictable that any attempt at developing non-adaptive automatic annotation systems is usually less effective. The trend in the literature has been to develop models that work for average users and for average scenarios. This approach has three main disadvantages. First, it is impossible to know beforehand all the patterns that could be of interest to all researchers. Second, it is practically impossible to find enough training examples for all patterns. Third, it is currently impossible to learn a model that is robustly applicable across all video quality-recording variations.

To overcome the three problems and provide practically useful solutions, this paper proposes a case-by-case user-controlled annotation model. The main philosophy for this kind of model is that a model designed to give the best average performance in a variety of scenarios is usually less accurate and less adaptable for a particular problem than a model tailored to the characteristics of that problem. This approach is also grounded in the 'No Free Lunch' theorems, which establish that for any algorithm, any elevated performance over one class of problems is offset by performance over another class (Wolpert & Macready 1997).

We apply our proposed solution to the problem of gesture stroke detection. To be more precise, for gesture stroke detection to be more accurate and more robust, we develop a model that takes intuitive input from the user for a given video and then we apply standard algorithms optimized to the characteristics of the video.

2. Gesture stroke

Gesture stroke is the most important message-carrying phase of the series of body movements that constitute a gesture (or the phrases in a gesture). The body movements usually include hand and face movements. The relevant questions for automatic stroke recognition are: a) what is a gesture? b) where does a gesture start and end? c) what are the phases in the gesture? d) which one is the stroke?

The literature does not give completely consistent answers to the above questions (Kendon 1980, 1972; Kita et al. 1998). However, the most prominent trend is that a gesture unit consists of one or more gesture phrases, each consisting of optional preparation, optional pre-stroke hold, obligatory stroke, optional post-stroke hold and optional retraction (Kendon 1980). Figure 1 shows the different phases in a gesture unit as outlined by Kendon.

Figure 1: Gesture phases (Kendon 1980, 1972)

For the purpose of this paper, a gesture phrase is classified into two: strokes and non-strokes. The non-stroke gesture phases include the absence of movement, the preparation, the hold, the retraction and any other body movements excluding the strokes.

3. Methodology

Our approach to determining gesture strokes involves four stages: a) detect face and hands for every person b) track them c) extract features d) distinguish strokes from non-strokes.

Different algorithms are used to solve each stage. Two features, corners and skin colors, are used to detect faces and hands. These features have been selected because they are usually stable from frame to frame for a given video.

Corners are shown to be good features for tracking (Shi & Tomasi 1993). They have the property that they are different from their surrounding points. A given point in a homogenous image cannot be identified whether or not it has moved in the subsequent frame. Similarly, a given point along an edge cannot be identified whether or not it has moved along that edge. However, the movement of a corner can conveniently be computed and identified, as it is non-ambiguous in its identity. This makes it a good feature for tracking.

For a given application, not all corners in a video frame are equally important. For gesture analysis, the interesting corners are the ones resulting from the body parts, mainly from face and hands. In order to filter out the corners irrelevant to body parts, we mask out corners that do not correspond to the skin color model.

The skin color model is developed with the involvement of the user. The user selects a representative instance of the skin color in one of the frames of the video, usually the first frame. And then the system extracts color information from that instance and finds all points in the frame where the matching of colors with the extracted color is high.

It is important to notice that the on-line selection of the skin region avoids having to design a skin color model for all human races. Off-line skin color model design is as practically difficult as collecting pictures of all human skin colors and its use in the detection for a particular skin color in a given video will be less accurate.

Given the corners from regions of the skin in the video, the tracking is done with the pyramidal implementation of the Lucas-Kanade algorithms (Bouguet 1999; Bradski & Kaehler 2008). Values extracted from the number of corners, clusters and their dynamics across frames (context) are fed into a supervised learning algorithm with class labels 1 for frames inside a stroke and 0 for frames outside a stroke. For the learning algorithm, we used support vector machine with RBF kernel.

4. Experiment data

Various videos have been used to test the detection of face, hands and their movements. Particularly, we experimented with two videos, each of which consists of two people of the same color: black and white skin colors. The resolution of the video with white people is 320x240 and that of the other video with black people is 1280x720. The higher the resolution, the better the detection quality.

For the supervised learning algorithm, we used a stroke and non-stroke annotated video data of 36 seconds long. This is the same video referred to above that has two white people speaking and gesturing. It has 914 frames, 847 of which are annotated. It has 19 strokes each ranging from 4 to 35 frames with mean 13.5 and standard deviation 6.5. It also has 40 non-stroke regions, which include moments of silence, preparation, retraction and holds.

All the videos for the reported experiments in this paper have been taken from the MPI archive `http://corpus1.mpi.nl/ds/imdi_browser/`.

A screen shot of a video with moving corners shown in blue. The rectangle is the region selected by the user as a model for the skin color (white)

Figure 3: A screen shot of a video showing detected skin regions. Very white regions correspond to high probabilities for white skin regions

Figure 4: A screen shot of a video with moving corners shown in blue. The rectangle is the region selected by the user as a model for the skin color (black)

Figure 5: A scrren shot of a video showing detected skin regions. Very white regions correspond to high probabilities for black skin regions

5. Results

The accuracy of results for face and hands detection based on initializations of regions of skin color on one frame in the video and applied to the subsequent frames in the same video shows that this approach can be very effective in identifying the corners belonging to the moving hands/face. Figures 2 through 5 show screen shots of the process and the results for a chosen video frame.

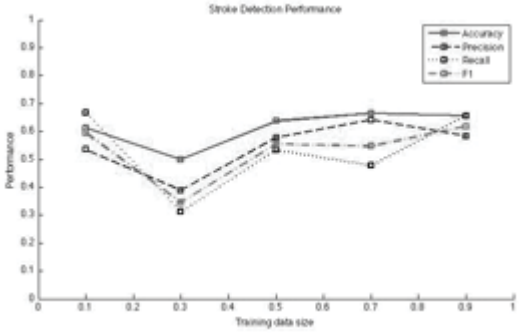


Figure 6: Stroke detection performance varaiation as training data size increases

The stroke/non-stroke classification results are shown in figure 6. The x-axis shows the training data size, which is 75% of the dataset. The y-axis shows the performance of the classification on test data, which is the remaining 25% of the dataset. As we vary the training data size, we get on average a slightly improving performance. The average accuracy, precision and recall achieved are 61.51%, 54.67% and 52.89%. The F1 measure is 53.26%. The average baseline (i.e. classifying every frame as non-stroke) achieves 57.55% accuracy and 0% recall. The performance measures show that there is a lot of room for improvement.

Evaluation for accuracy of frame boundaries for strokes and non-strokes should not be as clear-cut as we assumed in our experiments. One or two frame misses or shifts are tolerable given that even humans do not accurately mark the correct boundary, if any. However, we did not consider that observation in our evaluation results.

6. Conclusion

In this paper, we have put more emphasis on a more adaptive case-by-case annotation model based on the idea that with a little more input from users and facilitated by more user-friendly interfaces, annotation models can be more adaptive, more accurate and more robust (i.e. effectively deal with digital diversity). We have tested our approach on problems of hands/face tracking and automatic stroke detection.

We have noticed that building a skin color model online for all human skin colors will not only make the model more complex but also less accurate when applied on any particular video. However, a model built online for a given video initialized by input from the user achieves higher performance at no more cost than the initialization.

The correct detection of the skin color in combination with feature extraction algorithms can be used to study human gestures. We have shown that unique features (i.e. corners) and their dynamics across frames can be indicative of the presence of strokes, the most meaningful phase of a gesture. In our future research, we will continue to improve the stroke detection performance using more features and learning algorithms. We will also apply our methodology to classify strokes according to their meanings. Success of this methodology will have important impact in human activity recognition from video files.

Funding

The research leading to these results has received funding from the European Commissions 7th Framework Program under grant agreement no 238405 (CLARA). Special thanks go to Saskia van Putten and Rebecca Defina for providing one of the videos used in the experiments reported in this paper.

References

Adelson, E. H., C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden (1984). Pyramid methods in image processing. *RCA engineer* 29(6): 33-41.

Bouguet J. Y. (1999). *Pyramidal implementation of the lucas-kanade feature tracker: description of the algorithm*. Intel Corporation, Microprocessor Research Labs, OpenCV Documents, 3.

Bradski, G., and A. Kaehler (2008). *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media.

Kendon, A. (1972). Some relationships between body motion and speech. *Studies in dyadic communication* 7:177.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication* 25: 207-227.

Kita, S., I. van Gijn, and H. van der Hulst (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. *Gesture and sign language in human-computer interaction*, pp. 23-35.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. U of Chicago P.

Shi, J., and C. Tomasi (1993). *Good features to track. IEEE computer society conference on computer vision and pattern recognition*, pp. 593-600.

Wolpert, D. H., and W. G. Macready (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation* 1(1): 67-82.

Towards a Transnational Multilingual Caribbean Digital Humanities Lab

Gil, Alexander
alexgil@virginia.edu
University of Virginia, USA

The Caribbean, broadly understood, encompasses an archipelago, vast continental regions and a growing global diaspora. For Caribbeanists, the challenge has always been to work in many languages and across a dispersed archive. In recent years, the Digital Library of the Caribbean (DLoC) has lead a multi-institutional, multi-national effort to digitize archives dealing with the region. Prominent Caribbean academic journals have begun their move to digital formats. Many informal communities have begun collaborating and engaging in debates in social networks and personal blogs. In this poster, I outline in more detail the current challenges and opportunities that the new technologies bring to the academic life of the region. In particular, I address the following question: What would a Caribbean digital lab look like? Drawing from my experience as a fellow of the Scholars' Lab and NINES, I argue that the Caribbean Digital Lab of the future must be a combination of both. Despite the fact that the task of digitization and aggregation of vulnerable Caribbean archives remains the number one priority , I argue that the region also needs an R&D agenda, an educational component and an outreach arm in digital humanities.

The poster outlines the four components I see as crucial for a future Caribbean DH Lab.

1. Aggregation

NINES and 18th Century Connect have recently united under the banner of an umbrella organization, ARC, which hopes to welcome many other organizations under its purview. The ARC model provides a set of solutions to three problems in digital scholarship: a) Bringing together specialized digital projects with traditional digital archives under one rubric, one search box; b) Providing peer-review credentials to digital formats; and c) Organizing the digital research of sub-disciplines in the humanities. A Caribbean Digital Humanities Lab (or series of labs) could represent the work of Caribbean Studies at the ARC table. One of the unique contributions of a Caribbean NINES-like aggregator and digital peer-review operation woulb be the fact that the ruling metaphor for the subject of study of Caribbeanists is regional as opposed to temporal. For the region, the construction of a trans-national, polyglot peer-reviewed aggregator would fascilitate the task of discovery and collaboration between Caribbeanists worldwide.

2. Research & Development

Computer programming has the unique distinction of allowing anyone with a computer, access to the internet and time to become proficient at it and contribute new technology to the world. While the resources required to digitize all archives in the Caribbean will prove to be a major challenge for our generation, developing new tools and digital projects appropriate to the needs of the region can be within the range of a trans-national digital humanities lab. Research and Development has become a *sine-qua-non*e of contemporary digital humanities worldwide. In order to be true contributors to the advancement of the humanities worldwide, a Caribbean DH Lab must include an active team of developers and designers working on projects that are both pertinent to the region and of universal application. Given that the archives of the region are made vulnerable for many reasons – political, economic, meteorological, linguistic – I argue that the set of solutions that original digital projects born out of a Caribbean DH Lab can offer could provide alternative perspectives on problems familiar to peers in other regions. In order to argue my point I will briefly talk about the technological challenges of two possible projects: a) A multi-lingual archive of plantation records; and b) A geo-temporal project tracing the development of creole languages in the region.

3. Educational Opportunities

A Caribbean DH Lab, as most of its peers, must be attached to the educational mission of universities in the region. In order to foster a culture of digital humanities that can be sustained across generations, the lab must be able to provide opportunities for emerging scholars to learn and practice digital humanities at the highest order. Training in digital humanities skills can also serve as a way for young scholars in the region to diversify their portfolios and help them face the challenges of the global economies of the 21st century. In order to achieve this goal, I will argue, the Caribbean DH Lab of the future must secure funding for pre-doctoral and post-doctoral fellowships, organize speaker series and seminars and involve emerging scholars in its projects directly.

4. Outreach

Although our networking technologies have made it possible for us to collaborate in ways that were

unimaginable before, many challenges persist for multi-institutional and polyglot collaboration to take place. In this section I will delineate what these challenges are in three categories: funding, language and politics. I offer a set of tentative solutions that all coalesce around the idea of outreach. In order for the Caribbean DH Lab to succeed, I will argue, it must have a strong and well-informed outreach team that can help it get off the ground and sustain itself. Because of the fragility of government funds in the region, our lab must be able to negotiate effectively with private and international sources. Because of the language barriers the outreach team must be able to communicate effectively in the major languages of the region.

NUScholar: Digital Methods for Educating New Humanities Scholars

Graff, Ann-Barbara
annbg@nipissingu.ca
Nipissing University, Canada

Lucas, Kristin
kristinl@nipissingu.ca
Nipissing University, Canada

Blustein, James
jamie@cs.dal.ca
Dalhousie University, Canada

Gibson, Robin
ragibson243@community.nipissingu.ca
Nipissing University, Canada

Woods, Sharon
sewoods397@community.nipissingu.ca
Nipissing University, Canada

NUScholar is an authoritative website and adaptive hypermedia learning environment where undergraduates are introduced to the complex processes of critical scholarly reading.

If Mark Bernstein is right that 'The future of serious writing will lie on the screen' (Bernstein 200): 2), then the future of serious reading is also on the screen. In this SSHRC-funded project called NUScholar, Drs. Ann-Barbara Graff and Kristin Lucas (Nipissing University), Dr. James Blustein (Dalhousie University), and student research assistants Robin Gibson and Sharon Woods (Nipissing University), are asking what digital media can do to support the acquisition of foundational academic skills, specifically active reading, annotating and critical thinking. Our objectives address the problems and practices of scholarly readers.

We are operating from two principles: (1) The act of reading is little understood. (2) Students have no idea what we are asking them to do when we instruct them to 'read this poem'.

The activity of reading is a highly complex cognitive task, involving what Crowder and Wagner (1992: 4) describe as 'three stupendous achievements': the development of spoken language, written language, and literacy. Kneepkens and Zwaan (1994: 126) show that 'In processing text, readers [...] decode letters, assign meaning to words, parse the syntactic structure of the sentence, relate different words and

sentences, construct a theme for the text and may infer the objectives of the author. Readers attempt to construct a coherent mental representation of the text. In this process, they use their linguistic knowledge (knowledge of words, grammar), and their word knowledge (knowledge of what is possible in reality, cultural knowledge, knowledge of the theme).' While empirical studies of the reading of literary texts are in their infancy (de Beaugrande 1992), what is known is that reading is not simply a matter of recall, but a 'complex cognitive and affective transaction involving text-based, reader-based, and situational factors' (Goertz et al. 1993: 35).

For humanities' scholars, the primary task is to determine how meaning can be attributed to texts (Dixon et al. 1993). Literary texts pose particular challenges as they rely on allusion, connotation as well as denotation, verbal and situational irony, metaphor and other figurative tropes, and an awareness of etymology and historical possibility. As Warwick observes, much digital humanities research is involved in 'Digitization projects [which] have revolutionized our access to resources' (Warwick 2004: 375). Beyond resources, there are 3-D graphical maps of *Antony and Cleopatra*, Flash animations like 'What is Print?' etc., which speak to the strengths of new media to visualize text in new ways. Our work differs from and complements such digitization projects because – *avant la lettre digital* – we use new media to enable the teaching and learning of core, traditional literary skills.

NUScholar addresses what digital media can do to support the acquisition of close and active reading skills and their cognate, annotation, in part by modeling a number of strategies and also providing, in the case of annotation, an isomorph to effective techniques used with paper (Blustein et al 2011: 252-259).

As to the second point, we accept that students do not have a reference for the task they are being asked to perform, i.e., actively read a poem or prose text to determine 'how a text means' in John Ciardi's phrase. Critical reading is a complex task that does not have a single entry point nor only one path to success. Because the acquisition of the requisite skills is an open problem (in the cognitive sense), readers need to acquire a wide variety of skills.

Moreover, as the material encountered becomes more complex and as the expectations increase as one advances through a degree, readers must test those skills upon which they have come to rely. NUScholar uses adaptive hypermedia to assist students in their distilling and acquisition of the skills. In brief, this poster shows how we are designing the system to support traditional literary skill acquisition by using an hypermedia platform.

Drs. Graff, Lucas, and Blustein are developing an authoritative website (NUScholar) that offers an adaptive hypermedia learning environment where students will be introduced to the processes of critical reading. The website, NUScholar (http://nuscholar.nipissingu.ca), includes instructional web videos, animations designed to demystify the task of close reading of poetry and interviews with writers (http://ccareads.nipissingu.ca/); it will eventually include an annotated and interactive catalogue of verse, annotation tools, and an editable glossary; it also brings together many resources and potentially points to other tools and projects. As an adaptive hypermedia for education, the design allows for the transfer of the burden to readers/scholars as they proceed through the site. Eventually, scholars/users will be able to import texts in a variety of common formats (.rtf, .pdf, cut-and-paste), to use any of these features.

We seek to harness the power of digital media to demonstrate strategies of critical reading. The intent is similar to the 'writing in plain sight' exercise at Dalhousie University `http://dalgrad.dal.ca/whips/` ; where students watch someone complete a writing assignment "live" in order to demystify the notion that lightning bolts and magic are involved in writing to task or deadline and also Michael Wesch's 'The Machine is Using Us' (final version) `http://youtu.be/NLlGopyXT_g` where students appreciate the power of digital media to provide visualizations of complex tasks.

References

de Beauregard, R. (1992). Readers Responding to Literature: Coming to Grips with Reality. In E. F. Narduccio (ed.), *Reader Response to Literature: The Empirical Dimension*. New York: Mouton de Gruyter, pp. 192-210.

Bernstein, M. (2009). Into the Weeds. In M. Bernstein and D. Greco (eds.), *Reading Hypertext*. Watertown: Eastgate, pp. 1-13.

Blustein, J., et al. (2011). Making Sense in the Margins: A field study of annotation. *International Conference on Theory and Practice of Digital Libraries (TPDL)*. Published September 7, 2011. 10.1007/978-3-642-24469-8_27.

Ciardi, J. (1959). *How Does a Poem Mean?* Boston: Houghton/Mifflin.

Crowder, R. G., and R. K. Wagner (1992). *The Psychology of Reading: An Introduction*, 2nd edition. Oxford: Oxford UP.

Dalhousie University. (2011). Writing in Plain Site. `http://dalgrad.dal.ca/whips` . (accessed 24 January 2011).

Dixon, P., et al. (1993). The Effects of Formal Training on Literary Reception. *Poetics* 23: 471-487.

Goertz, E., et al. (1993). Imagery and Emotional Response. *Poetics* 22: 35-49.

Kneepkens, E. W. E. M., and R. A. Zwaan (1993). Emotions and Literary Text Comprehension. *Poetics* 23: 125-138.

MOMA (2001). What is Print. http://www.moma.org/interactives/projects/2001/whatisaprint/flash.html . (accessed 24 January 2011).

Nipissing University (2012). Canada Council for the Arts Reading Series. http://ccareads.nipissingu.ca/ . (accessed 8 March 2012).

Warwick, C. (2004). Print Scholarship and Digital Resources. In S. Schriebman et al. (eds.), *A Companion to Digital Humanities*. Malden, MA: Blackwell.

Wesch, M. (2007). The Machine is Us/ing Us (Final Version). http://www.youtube.com/watch?v=NLlGopyXT_g . (accessed 24 January 2011).

Latent Semantic Analysis Tools Available for All Digital Humanities Projects in Project Bamboo

Hooper, Wallace Edd
whooper@indiana.edu
Indiana University, Bloomington, Indiana, USA

Cowan, Will
wgcowan@indiana.edu
Indiana University, Bloomington, Indiana, USA

Jiao, David
djaio@indiana.edu
Indiana University, Bloomington, Indiana, USA

Walsh, John A.
jawalsh@indiana.edu
Indiana University, Bloomington, Indiana, USA

The Chymistry of Isaac Newton Project (http://www.chymistry.org/) added a new component to its website that allows researchers and interested readers to use the extensive results of its Latent Semantic Analysis studies of Newton's alchemical manuscripts (NSF STS Project #0620868).

The new LSA web component on the Chymistry website allows researchers to form their own queries to discover correlated passages across the published corpus. The component will return 250-word or 1000-word passages which share significant vocabulary from within a published corpus of 62 manuscripts and over 450,000 words of seventeenth-century English, French, and Latin. At the user's request, the component will also draw network graphs that visualize the structure of semantic relations among the passages in the documents. These network graphs can be examined in *Network Workbench*, which was developed at the Cyberinfrastructure for Network Science Center in Indiana University's School of Library and Information Science.

The new Chymistry LSA component will also allow researchers to use regular expressions to form queries to investigate the relationships between words and draw concept maps as network graphs. Unlike many Bayesian topic analysis techniques in which the words drop out in favor of topics (Blei et al., 2003), latent semantic analysis keeps the vocabulary visible and available for direct text analysis.

The new LSA web component on the Chymistry website currently provides results for the 62

manuscripts that have been released in the Chymistry Project's digital edition. All 119 manuscripts and an expanded LSA component will be released in 2012.

Walsh and Hooper presented initial results of this LSA work with Newton's alchemical papers and with the poetry and literary criticism of Algernon Charles Swinburne at DH 2011. New directions in their collaboration include undertaking further LSA work that combines the King James Version of Bible with Swinburne's corpus to detect patterns of influence and identify Swinburne's extensive borrowing of language from the King James Bible. Similarly, we intend to combine the Geneva bible with the corpus of Newton's theological and alchemical texts to detect patterns of influence, allusion, and borrowing or quoting in Newton's texts.

Since that presentation, there has been considerable interest in their LSA methods from other projects and independent digital humanists.

The Digital Library Program at Indiana University is a partner institution in Project Bamboo the mission of which is to deliver a research environment for humanities scholars and a corresponding infrastructure for librarians and technologists supporting humanities research. Cowan and Jiao of the IU Digital Library Program have been involved in current Project Bamboo activities. They have worked on creating a digital humanities research environment using the HubZero platform, a collaboration and research platform that was originally created to serve research communities in the natural sciences. Cowan and Jiao have built several digital humanities tools such as a Java based page-turning tool and a tool that provides topic modeling analysis of textual contents. Hooper and Walsh from the Chymistry of Isaac Newton project with Cowan and Jiao, from the Digital Library Program, will port the LSA algorithms into a tool on HubZero.

Both the Chymistry Project and the Swinburne Project are partners with the IU Digital Library Program, which provides the technical infrastructure for both projects. Cowan and Jiao raised the possibility with Hooper and Walsh of modifying the code they had developed for those projects and making it available as part of Project Bamboo.

Cowan, Jiao, and Hooper argue that every humanities project will be interpreting and analyzing entirely unique content and their public end-user websites will likewise have their own unique design constraints. What needs to be provided is a framework that will allow each project to bring its data sets to a processing interface and submit it, with the full and reasonable expectation that they will receive standardized outputs that they can process and exploit in their own design contexts.

The Chymistry and Swinburne projects use Perl scripts to extract data streams from TEI/XML-encoded documents, then process the streams in MATLAB on a supercomputer at Indiana to produce CSV data that is subsequently imported into MySQL databases for use with their public websites.

Our collaboration, however, has decided to provide tools that provide tools that accomplish just the numerical processing step in MATLAB, so we ask user projects to submit text data in simple text input structures and we return wimple and well-defined CSV data that can be exploited in any manner desired for end-user web interfaces or for further numerical or algorithmic processing.

The LSA analysis workflow in our MATLAB programs contains several components: first the corpus is read into the system and converted to a term/document frequency matrix. Terms of very low frequency and of high frequency (user-defined stop words) are removed. Secondly a TF/IDF (Term Frequency/ Inverse Document Frequency) matrix is calculated based on the term/document frequency matrix. Then we apply singular value decomposition (SVD) on the TF/IDF matrix and create the matrices that are used to calculate document distances as well as term/document matrices, which are saved to CSV files as the standard outputs and returned to users. Several steps like the SVD calculation and the distance matrix computations are computationally expensive and require high performance computing resources. Therefore, the MATLAB-based workflow is located and executed on a cluster of IBM servers running Red Hat Linux at Indiana University. The process is invoked and monitored by the HubZero submit module. The technical details of the LSA algorithms are thus hidden from the end users. Digital humanities researchers can utilize our LSA algorithm in their research without having to worry about the supercomputing aspect. To use this service, they only need to preprocess their corpus and provide them in the simple format that the tool accepts, and post-process the outputs of the algorithms such as the document/document distances. We expect such a tool would leverage the usage of the LSA algorithm in the digital humanities research, and can serve as a model to expand the application of computing-expensive algorithms in digital humanities.

Machine Learning for Automatic Annotation of References in DH scholarly papers

Kim, Young-Min
youngminn.kim@gmail.com
University of Avignon, France

Bellot, Patrice
patrice.bellot@lsis.org
Aix-Marseille University, France

Faath, Elodie
elodie.faath@revues.org
CLEO, Centre for Open Electronic Publishing, France

Dacos, Marin
marin.dacos@revues.org
CLEO, Centre for Open Electronic Publishing, France

1. Introduction

OpenEdition (`http://www.openedition.org`) is composed of three different platforms dedicated to electronic resources in the humanities and social sciences, Revues.org (journal and book series), Hypotheses.org (research notebooks and scholarly blogs) and Calenda (social sciences calendar). Revues.org is the oldest French platform of academic online journals. It now offers more than 300 journals available in all of the disciplines of the humanities and social sciences. The online publication is made through the conversion of articles into XML TEI (`http://www.tei-c.org`) format and then into XHTML format and allows the viewing of the full text in web browsers. The bibliographical parts of articles on Revues.org are rather diverse and complicated. One main reason of this complexity is the diversity of source disciplines that makes various styles in reference formatting:

- References employ different journal stylesheets or do not follow any stylesheet,
- Labeling should be very precise: performance should be close (or better) those of a 'human labeler'. In case this could not be achieved, the software should provide user with a confidence score. The quality of a digital library depends on the quality of metadata that allow field-based search, citation analysis, cross-linking and so on,
- References can be written in different languages,
- References contain some misspellings,
- Entries may be ambiguous (for example, searching for papers from '*Williams Patrick*' in Google Scholar retrieve several papers: a) '*Colonial discourse* [...]' published in 1996, b) '*Mariage tsigane* [...]' (gipsy wedding) published in 1984, c) '*A randomized trial of group outpatient visits* [...]*: the Cooperative Health Care Clinic*' published in 1997; we can assume that they correspond to different authors with the same name),
- Sometimes, one must refer to other references in order to complete them or else references are grouped together,
- References may be located anywhere in papers: in the body of the text, at the end of the paper and in footnotes.

2. Three new Training Corpora in DH Fields

In order to prepare a data set for the construction of a machine learning model (see below), which automatically estimates reference fields, we adopted a subset of TEI XML tags and annotated manually three different corpora according to three difficulty levels:

1. Corpus level 1: the references are at the end of the article under a heading "References". Manual identification and annotation are relatively simpler than for the two next levels. Considering the diversity, 32 journals are randomly selected and 38 sample articles are taken. Total 715 bibliographic references were identified and annotated.
2. Corpus level 2: the references are in notes and they are less formulaic compared to the ones of the corpus level 1. An important characteristic of the corpus level 2 is that it contains link informations between references. And because of the nature of note data, the corpus contains both bibliographical and non-bibliographical notes. We selected 41 journals from a stratified selection and we extracted 42 articles considering the diversity. We observed 1147 bibliographical references that were manually annotated and 385 non-bibliographical notes, which did not need any manual annotation.
3. Corpus level 3: the references are in the body of articles. The identification and annotation are the most complicated. Even finding the beginning and end of bibliographic parts in a note is difficult. For example, we may accept a phrase citing a book title as reference even if its author is not found near by but in footnote. We selected 42 articles considering different properties of implicit reference in the body of the articles. From these selected articles, we identified and annotated 1043 references.

The following table shows some examples from these three corpora. In total, we annotated 3290 bibliographical references by using the tags presented on Table 1.

Bibliographical entry	entry	<bibl>
	Linked reference	<relatedItem>
Authorship Info.	Author	<author>
	Forename	<forename>
	Name	<surname>
	Organisation	<orgName>
Title info.	Title	<title> t (type a (article) j (journal) m (book) u (scholarly) s (series)
Misc. info.	Date	<date>
	Place of publication	<pubPlace>
	Editor	<publisher>
	Edition	<edition>
	# Pages	<extent>
	Details	<biblScope> t (type vol (volume) pp (pages) issue (issue number)

Table 1: Tags for annotating references

3. Machine Learning for Annotation

Most of the free softwares available on the Internet process pre-identified references against a set of predefined patterns. For example, cb2bib (http://www.molspaces.com/cb2bib/) recognizes reference styles of the publications of the American Chemical Society and of Science Direct but does not work on other styles without a costly adaptation. Some other softwares employ machine learning and numerical approaches by opposite to symbolic ones that require a large set of rules that are very hard to manage and that are not language independent. Day et al. (2005) cite the works of a) C.L. Giles et al. for the CiteSeer system (computer science literature) that achieves a 80% accuracy for author detection and 40% accuracy for page numbers (1997-1999), b) Seymore et al. that employ Hidden Markov Models (HMM) that learn generative models over input sequence and labeled sequence pairs to extract fields for the headers of computer science papers, c) Peng et al. that use Conditional Random Fields (CRF) (Lafferty et al. 2001) for labeling and extracting fields from research paper headers and citations. Other approaches employ discriminatively-trained classifiers (such SVM classifiers). Compared to HMM and SVM, CRF obtained better tagging performance (Peng & McCallum 2006). Some papers propose methods to disambiguate author citations (Han & Wu 2005; Torvik & Smalheiser 2009) or geographical identifiers (Volz et al. 2007). These 'state of the art' approaches seem to achieve good results but they proceed on limited size collections of scientific research papers only and they do not resolve all the difficulties we identified above. We choose Conditional Random Fields (CRFs) as method to tackle our problem of bibliographic references annotation on DH data (Kim et al. 2011). It is a type of machine learning technique applied to the labeling of sequential data. By definition, a discriminative model maximizes the conditional distribution of output given input features. So, any factors dependent only on input are not considered as modeling factors, instead they are treated as constant factors to output (Sutton & McCallum 2011). This aspect derives a key characteristic of CRFs, the ability to include a lot of input features in modeling. The conditional distribution of a linear-chain CRF for a set of label y given an input x is written as follows. The multiplication of all these conditional probabilities for input data is maximized during the iterative learning process.

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_i, y_{i-1}, \mathbf{x}_i)\right),$$

$\mathbf{y} = y_1 \ldots y_T$: a label sequence, $\mathbf{x} = x_1 \ldots x_T$: an input sequence, $\theta = \{\theta_k\} \in R^K$: parameter vector, $\{f_k(y_i, y_{i-1}, \mathbf{x}_i)\}_{k=1}^{K}$ feature functions.

4. Experiments

We separately learn a CRF model for each corpus. Until now, we have conducted various experiments with the first and second corpora for the automatic annotation of reference fields. The standard on the effective local features and appropriate reference field labels are established during the construction of CRF model on the first corpus. This standard is equally applied to the second corpus but in this time, another automated process is necessary for the selection of target notes that include bibliographical information. Another machine learning technique, Support Vector Machine (Joachims 1999) is used for this note classification and we propose a new method, which mixes input, local and global features to well classify notes into bibliographical or non-bibliographical categories. The new feature generation method works well in terms of both classification accuracy and reference annotation accuracy. The features we selected for CRF are listed in Table 2.

Feature name	Description
ALLCAPS	All characters are capital letters
FIRSTCAP	First character is capital letter
ALLSMALL	All characters are lower cased
NONIMPCAP	Capital letters are mixed
ALLNUMBERS	All characters are numbers
NUMBERS	One or more characters are numbers
DASH	One or more dashes are included in numbers
INITIAL	Initialized expression
WEBLINK	Regular expression for web pages
ITALIC	Italic characters
POSSEDITOR	Possible for the abbreviation of editor

Table 2: Features employed for automatic annotation

A remarkable point is that we separately recognize authors and even their surname and forename. In traditional approaches, different authors in a reference are annotated as a field. Compared to the

scientific research reference data used in the work of Peng and McCallum (2006), our corpus level 1 is much more diverse in terms of bibliographical reference formats. However, we have obtained a successful result in overall accuracy (90%) especially on surname, forename and title fields (92%, 90%, and 86% of precision respectively). They are somewhat less than the previous work (95% overall accuracy) but considering the difficulty of our corpus, current result is very encouraging.

The notes in the corpus level 2 are more difficult to treat and had not been studied in the previous works to the best of our knowledge. With our pre-selection process using a Support Vector Machine (SVM) classifier was effective to finally increase the automatic annotation result for corpus level 2. For the comparison, we constructed another CRF model on all notes without classification. Consequently, the CRF model learned with the classified notes using our approach outperforms the other model, especially on recall of three important fields, surname, forename, and title as in the following table. Bold character in the following table means that the corresponding model outperforms in the field with a statistical significance test (see `http://bilbo.hypotheses.org` http://bilbo.hypotheses.org for more information).

Table 3

5. Conclusion

We have constructed three different corpora of bibliographical reference in DH field. Then we have conducted a number of experiments to establish the standards of processing on the corpora. From the basic experiments on the first and second corpora, we continue to refine the automatic annotation via different techniques. Especially we develop an efficient methodology to incorporate incomplete external resources such as proper noun list to a CRF model. According to a trial experiment, the incorporation enhances automatic annotation performance on proper noun fields with corpus level 1. We will try the same method on the second corpus. Given that the annotation accuracies were not so high with the second corpus, we expect a more significance improvement. The integration of the note classification and annotation steps into a process is also expected to improve the efficiency of learning process. One of most important part will be the treatment of corpus level 3. We are now developing several approaches to deal with this corpus.

Acknowledgements

This work was supported by the Google Grant for Digital Humanities in 2011; and the French Agency for Scientific Research under CAAS project [ANR 2010 CORD 001 02].

References

Day, M.-Y., T.-H. Tsai, C.-L. Sung, C.-W. Lee, S. H. Wu, C. S. Ong, and W. L. Hsu (2005). A knowledge-based approach to citation extraction. Information Reuse and Integration. *IRI -2005 IEEE International Conference*, pp. 50-55.

Han, H., W. Xu, H. Zha, and C. Lee Giles (2005). A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations. *Proceedings of SAC'05*, ACM.

Joachims, T. (1999). Making large-scale support vector machine learning practical. Cambridge: MIT Press, pp. 169-184.

Kim, Y.-M., P. Bellot, E. Faath, and M. Dacos (2011). Automatic annotation of bibliographical reference in digital humanities books, articles and blogs. *Proceedings of the CIKM 2011 BooksOnline11 Workshop*.

Lafferty, J., A. McCallum, and F. Pereira (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *International Conf. on Machine Learning* (ICML'01).

Peng, F., and A. McCallum (2006). Information extraction from research papers using conditional random fields. *Information Processing & Management* 42(4): 963-979.

Torvik, V. I., and N. R. Smalheiser (2009). Author Name Disambiguation in MEDLINE. *ACM Transaction on Knowledge Discovering in Data* 3(3).

Volz, R., J. Kleb, and W. Mueller (2007). Towards ontology-based disambiguation of geographical identifiers. *WWW int. conf.*, Canada.

Sutton, C., and A. McCallum (2011). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*. To appear. `http://arxiv.org/abs/1011.4088` (accessed 25 March 2012).

An Ontology-Based Iterative Text Processing Strategy for Detecting and Recognizing Characters in Folktales

Koleva, Nikolina
nikolina.koleva@dfki.de
German Research Center for Artificial Intelligence, DFKI, Germany

Declerck, Thierry
declerck@dfki.de
German Research Center for Artificial Intelligence, DFKI, Germany

Krieger, Hans-Ulrich
krieger@dfki.de
German Research Center for Artificial Intelligence, DFKI, Germany

1. Introduction

Detecting and marking consistently through a folktale the participants that are playing a role in the story can help a lot in assigning in an automatic fashion the typical functions to characters, as those are for example described by (Propp 1968), and we equally expect that the Proppian narrative functions can also be better automatically detected and marked-up in text, if an accurate recognition of the main participants in the story has been performed beforehand. (Lendvai et al. 2010) addresses the issue of semi-automatically assigning Proppian characters and action types to text segments mainly on the base of linguistic analysis.

In this poster/demo article, we describe a complementary approach, which relies first on a knowledge base, in the form of an ontology formalizing family relationships, which is getting populated by iterative applications of the ontology components to a linguistically annotated tale, whereas different natural language expressions referring to an unique character are marked in the iteratively updated knowledge base using the OWL[1] 'sameAs' property[2]. We developed a detailed family ontology[3], which is for the time being embedded in a small folktale ontology that describes the world of the 'Magical Swan Geese' tale[4]. The class hierarchy of this ontology is displayed in Figure 1, in the Appendix.

The current focus on the family ontology is guided by the fact that family relationships play a central role in many tales. Modeling other participants in tales is much more difficult, since their behavior very often do not correspond to the 'normal' entities (so for example a speaking 'river of milk', which acts as a 'helper' in the tale). Nevertheless our approach allows detecting also such entities as characters.

2. Knowledge-based Reference Resolution

A problematic issue in processing folktales is the detection and corresponding annotation of co-referring expressions in text. Folktales are particular in this respect, since people (or characters) are relatively rarely mentioned by name, but are prevalently introduced by their function ("the King"), family status ('the father') or their mere existence ('there lived a woman'). This phenomenon, together with very vague contextual spatio-temporal descriptions in text, makes the recognition of co-referring expressions on the basis of mere linguistic features quite cumbersome. This is a reason why we developed the family ontology, in order to support knowledge-based reference resolution of entities detected in text. On the basis of this semantic resource one can store as a specific individual in the knowledge base each entity of the tales that has been associated with a particular biological or family status.

For the purpose of the knowledge-based reference resolution, we equipped the class hierarchy with a set of inference rules, which are acting in a complementary fashion to the Protégé built-in Pellet reasoner[5]. The rules ensure that instances of the classes **Man** and **Woman** are encoded as instances of the class **Parents,** and therefore are identical to instances of the classes **Father** and **Mother** in case enough evidence about the marital or biological status is given by the text. Different instances of the class **Children** are at the same time instances of the class **Siblings**, using similar heuristics as for the class **Parents**, so that family relationships extracted from text can be completed by the inference rules, and made available for the incremental analysis of the text.

Every class and relation encoded in the ontology is associated with a label in natural language (in four languages: English, German, Russian and Bulgarian)[6]. The labels serve as the interface between the ontology and the text processing system, which is in our case the NooJ platform (Siberztein 2003).

3. A first Iteration of Text Processing: Annotation of Nominal Phrases

We process with NooJ the whole tale and mark especially all nominal phrases (NPs), being simple ('The mother'), coordinated ('a old man and a old woman') or recursive ('a river of milk flowing in banks of pudding') NPs[7].

Our textual analysis is further specifying if an NP is indefinite or definite on the basis of the determiners used ('a woman' – indefinite – vs 'the mother' – definite –), at least for languages using this kind of determiners, like English, German, etc.[8] A specific property of indefinite nominal phrases as introducing discourse referents has been widely discussed in the field of computational semantics and (von Heusinger 2000) is giving a good overview of the past discussions. In the special case of tales (Herman 2000) is providing for examples supporting this view on indefinite nominal phrases, relating them to the introduction of characters of tales. Our actual work with NooJ is implementing some of the views described in the work of Herman. The first step of our iterative approach to text analysis is resulting in the linguistic annotation of the folktale in terms of indefinite and definite NPs.

4. A second Iteration: Storing core Elements of indefinite NPs as candidate Characters in the Knowledge Base

The next iteration is dealing then with the application of the knowledge base to the indefinite NPs in the text. The main elements of the indefinite NPs – the nouns – are extracted and compared with the labels of the classes in the ontology. So the noun 'daughter' within an indefinite NP in the tale is matching the label of the class **Daughter** of the family ontology. As a consequence, this noun is stored in the knowledge base as a potential character of the tale and gets the ID 'ch3' (since before this the program has identified 'man' and 'wife' as the first potential characters occurring in the text), marking it as an individual of the class **Daughter**. This procedure is applied to all indefinite NPs occurring in the tale.

5. A third Iteration: Applying Inference Rules to the stored candidate Characters

We apply then the inference rules described above in Section 2 to the candidate characters stored in the knowledge base. Just to give a simple example: 'ch3' ('daughter') is being automatically encoded in the ontology as an instance of the classes **Girl** and **Sister**, while the relationships to the brother is also automatically inferred. These inferences can be draw also du to the fact that after the first iteration, it appeared that the tale is mentioning only one young female person and only one young person. This iteration offers thus also a kind of consolidation of the results of the preceding ontology population procedure.

6. A fourth Iteration: Merging the stored Characters with the core Elements of definite NPs

In Figure 2 in the Appendix, the reader can see that our approach manage to map the 'ch3' (resulting from the indefinite NP 'their daughter') with occurrences of the string 'girl' and 'sister' occurring in definite NPs elsewhere in the text. This step is for sure benefiting from the results of the application of the inference rules described in Section 4. We apply further a filtering procedure: candidate characters that are mentioned only once in the text (not being matched to the content of definite NPs, for example, or not being involved as agent in an event) are deleted from the knowledge base. On this basis we can eliminate the string 'a handkerchief' from the list of potential characters (as an indefinite NPs), but we can keep the string 'an apple tree' and consolidate the core element 'apple tree' as a character of the tale, since it occurs also in the context of a definite NP, and it is involved in an agentive action (speaking).

7. Conclusion

We demonstrate the potential benefits of the combined use of an ontology, inference rules and textual analysis for identifying characters in the relatively small (and closed) world of a folktale. While first results of our on-going work are promising, we still have to apply the approach to more tales, in other languages, and to evaluate our approach. We plan to use for this purpose the UMIREC Corpus[9].

Acknowledgements

The work reported in this paper has been partly supported by the R&D project 'Monnet', which is co-funded by the European Union under Grant No. 248458.

Appendix

Figure 1: Screen shot of the Ontology

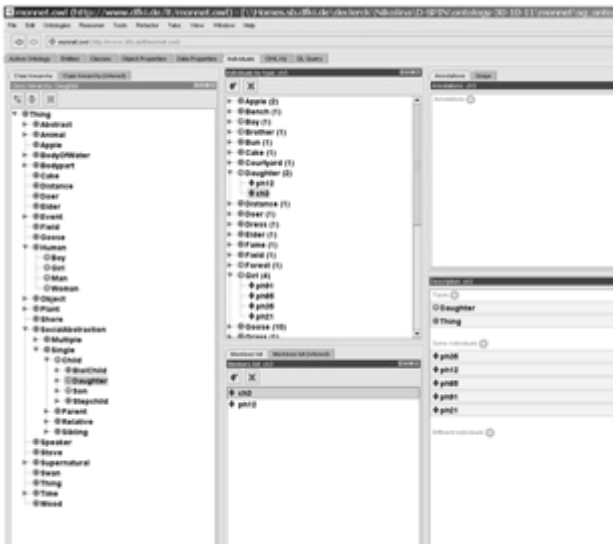


Figure 2: The knowledge base after round4: 'ch3' (Daughter) has been associated (Same individuals) with occurences of 'girl' and 'sister', as those has been identified in the phrases numbered 12, 21, 35, 85 and 91 in ur NooJ XML annotation

References

Geist, L. (2008). *Specificity as referential anchoring: evidence from Russian. Proceedings of SuB12*, Oslo: ILOS 2008, 151-164.

Herman, D. (2000). Pragmatic constraints on narrative processing: Actants and anaphora resolution in a corpus of North Carolina ghost stories. *Journal of Pragmatics* 32(7): 959-1001.

Lendvai, P., T. Váradi, S. Darányi, and T. Declerck (2010). Assignment of Character and Action Types in Folk Tales. *Proceedings of the NooJ 2010 Conference.*

McCrae, J., L. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, and D. Spohr (2012). Interchanging lexical resources on the Semantic Web. *Journal on Language Resources and Evaluation* (in Press).

Propp, V. J. (1968). *Morphology of the folktale*. Austin: U of Texas P.

Silberztein, M. (2003). *Nooj Manual*. `http://www.nooj4nlp.net`.

Thompson, S. (1955). *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends. Revised and enlarged edition*. Bloomington: Indiana UP, 1955-58

Tuffield, M. M., D. E. Millard, and N. R. Shadbolt (2006). Ontological Approaches to

Modelling Narrative. *2nd AKT DTA Symposium*, January 2006, AKT, Aberdeen University.

Uther, H.-J. (2004). *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*. FF Communications no. 284-286. Helsinki: Suomalainen Tiedeakatemia, 2004.

Zöllner-Weber, A. (2008). *Noctua literaria : a computer-aided approach for the formal description of literary characters using an ontology*. Ph.D. Thesis, Bielefeld University.

Notes

1. OWL (Web Ontology Language) is a formal representation language, which is nowadays widely used for describing ontologies. OWL is a W3C standard. See http://www.w3.org/2001/sw/BestPractices/Tutorials for introduction material.
2. The 'owl:sameAs' property used in OWL knowledge bases offers a formal way for stating that various expressions (represented by various URIs) are referring to an identical person (or entity).
3. The ontology has been encoded using the Protégé editor: http://protege.stanford.edu
4. An online English version of this Russian tale is available at: http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/swan-geese.html . We are currently extending the coverage of the ontology, behind its family relationships component, integrating for example elements belonging to the description of characters of narratives, as those are in detail described in (Zöllner-Weber 2008).
5. We use this reasoner since it is incorporated in the Protégé ontology editor, which we selected for designing our ontology. More details about Pellet are given at: http://clarkparsia.com/pellet/
6. The availability of the ontology class labels in 4 languages is reflecting our additional aim in providing for multilingual ontology resources for ontology-based text mining or information extraction procedures applied to folktales in various languages. In this our effort is complementary to two use cases (on the financial domain and eGovernment) defined in the European R&D project 'Monnet' (Multilingual Ontologies for Networked Ontologies, see www.monnet-project.eu (www.monnet-project.eu)).
7. It is important to note that in the actual version of the system only referential expressions are considered. In a next version, the reference resolution work will be extended to all kind of pronominal expressions.
8. For languages not having determiners in their set of categories, like Russian, we are investigating the detection of other linguistic features that can mark indefiniteness, as those are described for example in (Geist 2008).
9. See http://dspace.mit.edu/handle/1721.1/57507.

Integrated multilingual access to diverse Japanese humanities digital archives by dynamically linking data

Kuyama, Takeo
is017080@ed.ritsumei.ac.jp
Ritsumeikan University, Japan

Batjargal, Biligsaikhan
biligsaikhan@gmail.com
Ritsumeikan University, Japan

Kimura, Fuminori
fkimura@is.ritsumei.ac.jp
Ritsumeikan University, Japan

Maeda, Akira
amaeda@is.ritsumei.ac.jp
Ritsumeikan University, Japan

1. Introcuction

This poster provides a summary of our ongoing project for providing integrated access to Japanese multiple digital libraries, archives, and museums. The main goal to construct a federated access system for Japanese humanities databases, which searches multiple databases in parallel and provides on-the-fly integration of the results, has required the system to deal with heterogeneous metadata schemas in various formats. Aggregation and integration of the retrieved results in English and Japanese are complicated if a search needs to be performed from multilingual sources. Ukiyo-e, Japanese traditional woodblock printing, is known worldwide as one of the fine arts of the Edo period (1603-1868). Many museums and organizations in Japan as well as in western countries hold numerous Ukiyo-e prints in their collections. As a result of worldwide digitization over the last decade, many cultural institutions including libraries, archives, and museums started to expose digitized images of Ukiyo-e prints on the Internet. How to find the necessary information effectively from multiple databases is becoming an essential issue for users. In other words, users need an efficient way of searching multiple databases, especially when it is getting more difficult to know which museum has a particular Ukiyo-e print. Thus, federated search of multiple Ukiyo-e databases scattered around the world is a feature expected by humanities researchers of Japanese culture. This poster proposes a method of integrated multilingual

access to heterogeneous Ukiyo-e databases for improving the search efficiency.

2. Related research

In LODAC (Linked Open Data for Academia) Museum, which is a part of Lod.ac project, Kamura et al. (2011) aimed to connect Japanese museums using Linked Open Data (LOD). The LODAC Museum has an online system[1], which can search and browse various data of Japanese museums in Japanese using the concept of Linked Data[2]. In a similar research of the eCultura project, Cornejo et al. (2010) introduced a set of services and applications to access and integrate diverse web-based contents of the cultural domain.

However, we propose a different approach, which has the following three differentiations. First, instead of collecting data from each database, we create links from the search results dynamically. Second, we access not only Japanese databases but multilingual databases from all over the world. Third differentiation is to use authority data for listing related items. We apply our approach to the federated search system for Ukiyo-e databases that we are developing.

3. Proposed approach

3.1. Federated searching system

We are developing a prototype federated searching system for Ukiyo-e prints, which retrieves multiple and heterogeneous back-end databases (Batjargal et al. 2011). In the proposed system, remote databases can be searched and retrieved simultaneously via the federated search protocols such as Search/Retrieve Web service (SRW), Search/Retrieve via URL (SRU), etc. SRU servers return a result set represented in XML, when a search request is received. In addition, web databases are accessible using 'web/screen-scraping' techniques that process a list of search results by reading and extracting data from HTML. At present, our proposed system is capable of retrieving the collections of the Library of Congress (approx. 2,740 Ukiyo-e prints), the National Diet Library (NDL) of Japan (approx. 10,000 Ukiyo-e prints), the British Museum (approx. 15,000 Ukiyo-e prints), the Boston Museum of Fine Arts (approx. 57,000 Ukiyo-e prints), the Victoria and Albert Museum (approx. 38,000 Ukiyo-e prints), and the Ukiyo-e database of the Art Research Center of Ritsumeikan University (approx. 12,500 Ukiyo-e prints). Figure 1 shows the conceptual architecture of the existing federated searching system for Ukiyo-e databases. Figure 2 shows the screenshot of sample search results.

Figure 1: Conceptual architecture of a federated searching system for Ukiyo-e prints

Figure 2: Sample search results

3.2. Utilizing Linked Data

This section discusses our achievements that utilize Linked Data to providing integrated access to multilingual Ukiyo-e prints from all over the world. Linked Data is a paradigm of constructing structured data, which can be read automatically by computers. Linked Data consists of the interlinked data sets, e.g. have relationships between resources. Each resource is identified by a unique URI (Uniform Resource Identifier), and represented in RDF (Resource Description Framework) format. RDF is a data model, in which any resource can be represented by a set of property and value pairs that make statements about that resource. There are three reasons of using Linked Data in our approach. Firstly, Linked Data could be used to link related data in various databases. Taking advantages of a simple model of RDF, we could create various links by combining various RDF data. Secondly, besides linking related data, Linked Data could

enable accessing to diverse databases easily. Prior to Linked Data, users were usually accessing each database one-by-one. However, by using the links of related data, users would be able to access to related data in multiple databases with less trouble. Thirdly, Linked Data could help users to discover some further relationships between individual data.

We utilize the name authorities and subject headings of NDL of Japan, i.e., Web NDL Authorities[3], for obtaining data in RDF. In Web NDL Authorities, name authorities and subject headings can be searched via SPARQL (Simple Protocol and RDF Query Language), which is an RDF query language for searching and manipulating data in RDF format. An authority data for a heading (e.g. personal name, subject, etc.) contains aliases and synonyms of the heading, and information showing the reason for being selected as the heading. In our approach, we use the authority data for identifying a particular person from different representations of the personal name in different databases. The subject heading also distinguishes words that have same representations but different meanings. In an example of authority data as shown in Figure 3, the heading 'Shakespeare, William, 1564-1616' has three different representations in Japanese. An authority data of NDL of Japan contains personal names, family names, corporate names, place names, uniform titles, and general subjects for a certain heading. At present, we utilize only personal names and general subjects in our approach.

representation		heading
シェークスピア	→	Shakespeare, William, 1564-1616
シェイクスピア		
シェークスピヤ		

Figure 3: An example of an authority data

3.3. The proposed approach in action

In this section, we explain about the proposed system that aims to realize integrated multilingual access to Ukiyo-e prints in the world. Outline of the proposed system is shown in Figure 4. At first, in the step 'Search (1)', our system acquires the search results when a user performs a search in a federated searching system. Then in the step 'Select (2)', the system extracts the appropriate authority data for a certain record. In the step (3), the system finds the variants or aliases for that authority data by using SPARQL. After that, the system searches diverse databases by using the variants or aliases as queries. Finally, the system shows the refined retrieved results (4) with the links for further search (5). In general, the user will be able to access to other data easily by using the authority data of the aliases.

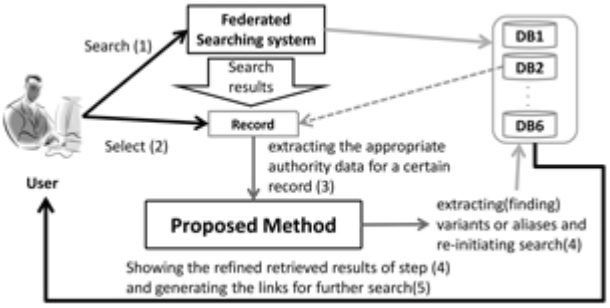


Figure 4: Outline of the proposed system

For instance, as shown in Figure 5, if a user searches using a query 'Hiroshige Utagawa', then our system finds the aliases 'Hiroshige Ando' and 'Hiroshige Ichiryusai' by using an authority data of 'Hiroshige Utagawa'. Using the aliases of 'Hiroshige Utagawa' (e.g. 'Hiroshige Ando' and 'Hiroshige Ichiryusai') our system will show further results.

Figure 5: An example of integrated access

Furthermore, some authority data in Web NDL Authorities has a link to Library of Congress Subject Heading (LCSH). Therefore our system can also provide cross-language access between Japanese and English as shown in Figure 6. Using a link from NDL Subject Headings (NDLSH), the proposed system can generate further links. In this way, users will be able to access to more records with 'Hashirae', 'Pillar Prints' and 'Ukiyo-e' for a given query input 'ukiyo-e' by generating the links for 'Ukiyoe' by using LCSH. Even if the user has no knowledge about the term 'Hashirae', he or she can access them easily and efficiently.

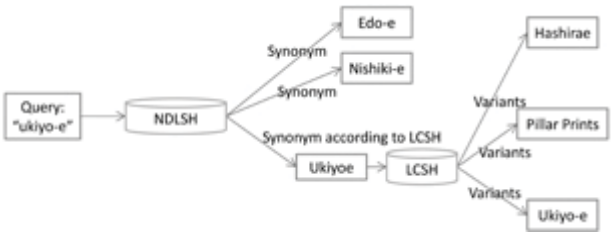


Figure 6: An example of multilingual access using Subject Headings

4. Evaluation of the system

In this section, we discuss our preliminary evaluation of the proposed system. In order to evaluate the usefulness of the system, we calculated the accuracy of the aliases obtained by the proposed system by manually checking whether the records retrieved using the aliases are actually about the same person as the original query (Figure 7). Table 1 shows the results of 6,923 retrieved records for several Ukiyo-e painters. The proposed system achieved the accuracy rate of 99.89% for the aliases of Ukiyo-e painters with full names. The accuracy rate drops to 89.91% for the aliases that consist of either first name or last name of Ukiyo-e painters, because of increased ambiguity in short strings. The overall accuracy rate was 95.45%, which should be enough for most purposes.

Figure 7: Evaluation process of the proposed system

Type of alias	Number of the records	Number of the records of the same person	Accuracy rate
Full name	3,840	3,836	0.9989
Either first name or last name	3,083	2,772	0.8991
Total	6,923	6,608	0.9545

Table 1: The results of the evaluation

5. Summary

In this poster, we proposed a technique using Linked Data in order to realize multilingual integrated access to multiple digital archives. One of the unique features of our proposed approach is that we utilize authority data to deal with the problem of different representation of data in different databases, as well as to generate cross-language links between databases. We believe such a system will help users to find new knowledge, and will also facilitate new directions in humanities research. Our future work include extending our system to other humanities digital archives, and user evaluation of the prototype system.

Funding

This work was supported in part by MEXT Grant-in-Aid for Strategic Formation of Research Infrastructure for Private University 'Sharing of Research Resources by Digitization and Utilization of Art and Cultural Materials' [Grant Number: S0991041].

Reference

Batjargal, B., F. Kimura, and A. Maeda (2011). Metadata-related Challenges for Realizing Federated Searching System for Japanese Humanities Databases. *Proceedings of the 11th International Conference on Dublin Core and Metadata Applications*. The Hague, Netherlands, September 2011, pp.80-85.

Kamura, T., F. Kato, I. Ohmukai, H. Takeda, T. Takahashi, and H. Ueda (2011). Study support and integration of cultural information resources with Linked Data. *Proceedings of the Second International Conference on Culture and Computing*. Kyoto, Japan, October 2011, pp. 177-178.

Cornejo, C. M., I. Ruiz-Rube, and J. M. Dodero (2010). eCultura, a semantically-enriched web-based approach to manage cultural content. *Proceedings of the IEEE International Conference on Information Reuse and Integration, IRI 2010*. Las Vegas, NV, August 2010, pp. 126-131.

Notes

1. http://lod.ac/
2. http://www.w3.org/DesignIssues/LinkedData.html http://www.w3.org/DesignIssues/LinkedData.html
3. http://id.ndl.go.jp/auth/ndla

Linguistic concepts described with Media Query Language for automated annotation

Lenkiewicz, Anna
anna.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Lis, Magdalena
Magdalena@hum.ku.dk
University of Copenhagen, Denmark

Lenkiewicz, Przemyslaw
przemek.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

1. Introduction

Human spoken communication is multimodal, i.e. it encompasses both speech and gesture. Acoustic properties of voice, body movements, facial expression, etc. are an inherent and meaningful part of spoken interaction; they can provide attitudinal, grammatical and semantic information. In the recent years interest in audio-visual corpora has been rising rapidly as they enable investigation of different communicative modalities and provide more holistic view on communication (Kipp et al. 2009). Moreover, for some languages such corpora are the only available resource, as is the case for endangered languages for which no written resources exist.

However, annotation of audio-video corpora is enormously time-consuming. For example, to annotate gestures researchers need to view the video-recordings multiple times frame-by-frame or in slow motion. It can take up to 100 hours to manually annotate one hour of the recording (Auer et al. 2010). This leads to a shortage of large-scale annotated corpora which hampers analysis and makes generalizations impossible. There is a need for automatic annotation tools which will make working with multimodal corpora more efficient and enable communication researchers to focus more on the analysis of data than on the annotation of material.

The Max Planck Institute for Psycholinguistics and two Fraunhofer Institutes are working together on developing advanced audio and video processing algorithms, called recognizers (Lenkiewicz et al. 2011), which are able to detect certain human behavior in a recording and annotate it automatically. However, usage of those recognizers is rather complex and their output is limited to detecting occurrences of predefined events without assigning any semantics. They also work independently from one another delivering annotations according to their specification, like for example hand movement or specific speech characteristics, which later need to be manually analyzed by the researchers in order to find any dependencies between different features.

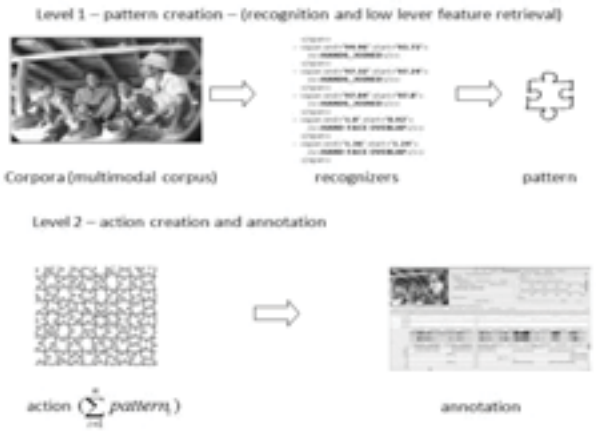


Figure 1: The general structure of the Media Query Language

The core of the work described in this paper is the development of a MQL, which can take advantage of the capabilities of the recognizers and allow expressing complex linguistic concepts in an intuitive manner.

2. Media Query Language

In the current phase of development general structure of the language is defined (Figure 1). The language contains three integrated components: *patterns, actions* and *libraries*.

2.1. Patterns

As a first step in media annotation using MQL human behavior would be decomposed to a single meaningful movement, gesture or speech element and saved for future reuse in form of a pattern as presented in Level 1 of Figure 1. A pattern for the purpose of this work is defined as a template or model, which can be used to save elements of human behavior decoded from media file under a meaningful name. It is the primary element created using MQL and it is the base for future action creation. Pattern provides a sort of architectural outline that may be reused in order to speed up the annotation process by applying search-by-example.

As a first step in pattern creation the corpora will be analyzed by recognizers and by a researcher using

specifically for this purpose designed active device application, which simplifies new pattern creation by marking interesting parts of a recording and including them in the MQL code (Figure 2). All solutions will be integrated with ELAN (Brugman & Russel 2004) a professional tool for the creation of complex annotations on video and audio resources.

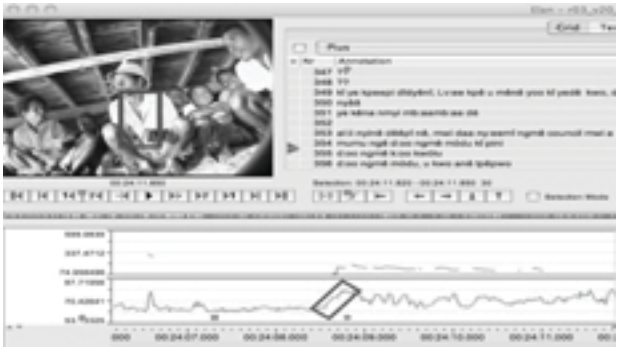


Figure 2: Example of pattern creation using active device as input collector. A to B – start and end point for hand movement. Red square – field determining tolerance in pattern matching in future search

The MQL language will be programmed in order to execute specific recognizers, depending on the nature of the request from the user and using recognizers' specification. The output of the recognizers will be collected and joined in sets of characteristic human behaviors, like for example when, which and how a given hand is moving, including information about direction (left, right, up, down) and speed. At the same time another recognizer can deliver information about speech characterization, like utterance boundaries, word separation or pitch contour. Recognizers will be therefore used to retrieve numerical description of the features mentioned above. The language will serve as a tool, which helps to convert the code received from the recognizers into patterns expressing human behavior in a language similar to natural. An example of a code fragment describing a pattern in MQL is presented in Figure 3.

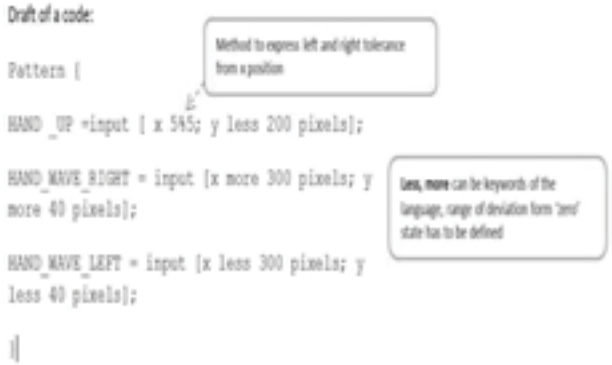


Figure 3: Example of a pattern defined in MQL code

2.2. Libraries

In order to simplify the feature annotation process and assure that already created patterns can be reused, each pattern created by the user needs to be added to a pattern library. Meaningful name and description needs to be given. MQL allows users to name the patterns and actions in a not imposed way. Formulation of the query is also free, as long as it will be in conformity with the language syntax and will include keywords and pattern names. An example of such library can be a set of patterns describing hand movement patterns like hands overlapping, hands rising, one hand movement with specific speed, etc. It is assumed that a pattern library will be dedicated to single actions occurrences. Moreover, it is planned that the system would inform the user whenever a currently created pattern already exists under different name and in which place of the hierarchy. The main goal of this functionality is the creation of well-developed pattern libraries, which in the future can limit the need for new pattern creation. With proper libraries available for the end-users a lot of media annotation work can be carried out using only already existing patterns. Prototype of such a library is shown in Figure 4.

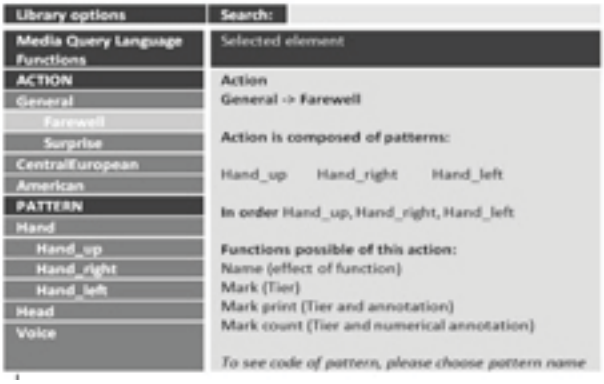


Figure 4: Example of a library defined in MQL code

2.3. Actions

The second level of the language will be dedicated to action creation and file annotation. It is called the executing phase of programming with MQL. On this level the user will be able to combine patterns into actions.

Action in MQL is defined as a set of predefined patterns (human movements, gestures or speech elements) composed together. Advancement of a single pattern to an action is also possible. Actions may be composed of patterns that may require execution of more than one recognizer in order to detect features of interest. An example of a MQL code fragment describing an action is presented in Figure 5.

```
Action (
FAREWELL = [HAND_UP + HAND_WAVE_RIGHT + HAND_WAVE_LEFT]
)
```

Figure 5: Example of an action defined in MQL code

A good example to illustrate the concept of action is annotation of utterance type. In signaling whether an utterance is a statement, a question or a command, different communicative behaviors (e.g. both speech prosodic cues and face expression) can work together. With MQL researchers will be able to aggregate such acoustic and gestural patterns together and save them in the form of an action in an action library according to the same rules as apply to pattern library creation. Thanks to this solution the decision about event classification into specific human behavior and creation of annotation can be taken by researchers rather than an automatic system. Work of the recognizers can be limited to detecting audio and body movement as detailed as possible.

3. Conclusions and future work

Thanks to the proposed MQL solution, together with the recognizers, the major problem of researchers, which is the time needed to manually annotate data, will decrease significantly. The expertise of the researchers will be used better by transferring their focus from highly laborious basic feature annotation to adding meaning to retrieved data and semantics to MQL language components.

Using MQL in the annotation process is going to significantly change the way in which the annotation work is carried out and how it can be reused. Using common ways to describe meaningful features and creating reusable libraries will contribute to creation of the interoperability standards.

The feedback on the current version of MQL received from researchers is encouraging. Flexibility and sufficient expressiveness of complex linguistic concepts is seen as a great advantage.

In the future work the problems of semantic gap coverage and interoperability standards will be tackled thanks to the possibility of expressing complex linguistic concepts using MQL.

Acknowledgment

The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA).

References

Auer, E., P. Wittenburg, H. Sloetjes, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel (2010). Automatic annotation of media field recordings. In C. Sporleder and K. Zervanou (eds.), *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*. Lisbon: University de Lisbon, pp. 31-34.

Brugman, H., and A. Russel (2004). Annotating multi-media / multimodal resources with ELAN *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, pp. 2065-2068.

Kipp, M., J.-C. Martin,P. Paggio, and F. K. J. Heylen, **ed.** (2009). Multimodal Corpora: From Models of Natural Interaction to Systems and Applications. *Lecture Notes in Computer Science 5509*. London: Springer.

Lenkiewicz, P., P. Wittenburg, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel (2011). Application of audio and video processing methods for language research. *Proceedings of the conference Supporting Digital Humanities 2011* (SDH 2011), Copenhagen, Denmark, November 17-18, 2011.

Virtual Reproduction of Gion Festival Yamahoko Parade

Li, Liang
liliang@fc.ritsumei.ac.jp
Ritsumeikan University, Japan

Choi, Woong
wchoi@ice.gunma-ct.ac.jp
Gunma National College of Technology, Japan

Nishiura, Takanobu
nishiura@is.ritsumei.ac.jp
Ritsumeikan University, Japan

Yano, Keiji
yano@lt.ritsumei.ac.jp
Ritsumeikan University, Japan

Hachimura, Kozaburo
hachimura@media.ritsumei.ac.jp
Ritsumeikan University, Japan

1. Introduction

Originated from Heian period (794-1185), Gion Festival, which has been registered in the list of 'Intangible Heritage of Humanity' by UNESCO, is one of the most famous festivals in Japan. Every year on July 17, the festival culminates in a parade of yamahoko, floats known as 'moving museums' because of their elaborate decorations with centuries-old tapestries, and wooden and metal ornaments. The festival is held by the Yasaka Shrine whose parishioners parade 32 floats to represents each self-governing parish. Approximately 150 thousands spectators from all around the world gather to see the parade every year.

With the development of computer graphics and virtual reality technologies, extensive researches have been carried out on digital cultural heritage. For decades, tangible cultural heritage contents including historical crafts, archaeological sites, and historical buildings have been digitally archived. Recently, digital archiving the intangible culture heritage contents, such as traditional festivals and behaviors of participants in cultural events have attracted more and more attention (Gandy et al. 2005; Magnenat-Thalmann et al. 2007; Papagiannakis et al. 2007).

In this research, we try to virtually reproduce Yamahoko Parade in Kyoto Gion Festival. We generated a content that combines motion and acoustics of the floats, crews, and spectators, within a virtual platform of 'Virtual Kyoto' (Yano et al. 2007). In current step, four well-known floats (Fune-hoko, Naginata-hoko, Kanko-hoko, and Kitakannon-yama) out of thirty-two floats were included in this virtual parade. We also reproduced the motion of four types of crews of Fune-hoko (Hikikata, Ondotori, Kurumakata, Hayashikata) using motion capture technique. This work contributes to the research of digital museum and provides a platform that allows the users to virtually experience the atmosphere of Yamahoko Parade in Kyoto Gion Festival.

2. Virtual Yamahoko Parade

We construct the virtual Yamahoko Parade using Vizard software. Vizard enables us to integrate and render CG models, animations, and sounds in a virtual space with real time interaction (Figure 1).

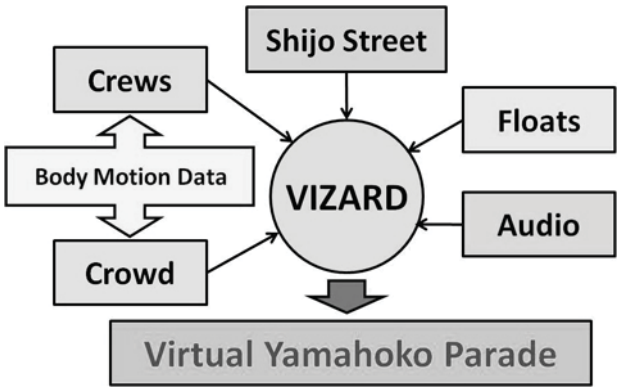


Figure 1: Construction of virtual Yamahoko Parade

Virtual Kyoto

A model of Shijo Street, one of the city's main streets, is reconstructed as the Virtual Kyoto platform for the virtual Yamahoko Parade. Based on MAP CUBE data, the textures of street and building models in Virtual Kyoto are made by capturing photos of the real objects. The model of Shijo Street is transformed into VRML format and imported to Vizard (Figure 2).

Figure 2: Shijo Street in Virtual Kyoto

Crowd Simulation of Spectators

Crowd simulation of spectators is an important element for regenerating the atmosphere of the event. In this work, we arranged about 730 characters on both sides of Shijo Street in Virtual Kyoto (Figure 3).

Figure 3: Crowd simulation of spectators

Three types of character models have been created: high polygon (about 8600 polygons), medium polygon (about 2800 polygons) and low polygon (about 1100 polygons) to build the real time virtual Yamahoko Parade. To make the crowd with realistic behavior, we add basic motions of talking, listening, shouting, and clapping hands to the models.

CG Floats of Yamahoko Parade

Four CG floats of Naginata-hoko, Kanko-hoko, Fune-hoko, and Kitakannon-yama were included in this virtual parade (Figure 4).

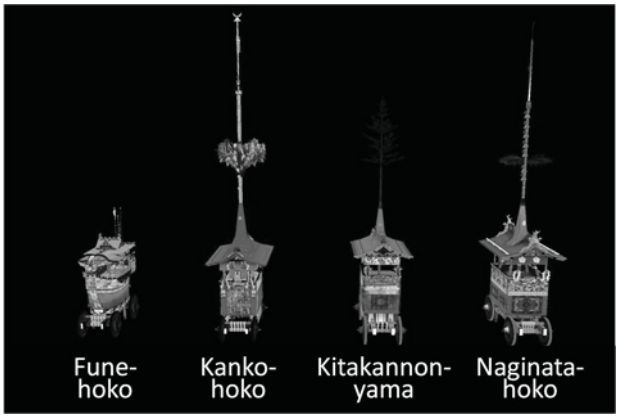


Figure 4: CG models of floats

The models have been created by laser-scanning detailed miniatures of the real floats, as well as surveying the floats' drawings. The textures of the floats are made by capturing photos of the floats during the festival.

Models and Animations of Parade Crews

The Virtual Parade includes four kinds of parade crews: Hikikata who pull the float; Ondotori who lead the parade with Japanese fans; Kurumakata who control the float's directions; and Hayashikata who play instruments on the platform of the float.

We created CG models of the crews using 3ds max and transformed them into Cal3D format to import to Vizard (Figure 5). The textures of their costumes were obtained from the real costumes.

Figure 5: CG models of Fune-hoko crews

Furthermore, since character animation of these crews is crucial for regenerating realistic movements of the parade, we use motion capture technique to reproduce the unique motions of the crews (Figure 6).

Figure 6: Motion capture of Fune-hoko crews

Acoustics of the Parade

We recorded the music of the parade played with the traditional instruments of drum, flute and bell by using multi-point measurement technique (Figure 7). Besides that, we also captured the sounds of ambient noise made by the floats, crews, and crowds.

The collected sounds were integrated and imported to Vizard.

Figure 7: Recording acoustics

Demonstration of Virtual Yamahoko Parade

The virtual Yamahoko Parade can be operated in 3D with high fidelity sounds using an immersive virtual environment system (Figure 8). Users can interactively control the viewing position and angle in the virtual world at real time with a gamepad. Sample screenshots captured under 2D mode are illustrated in Figure 9.

Figure 8: Virtual Yamahoko Parade
in an im-mersive virtual environment

3. Conclusion and Future Works

We virtually reproduced an intangible cultural material of Yamahoko Parade using 3D computer graphics and virtual reality technologies. We received positive feedbacks from the visitors in several exhibitions during which we demonstrated the virtual Yamahoko Parade.

Figure 9: Screenshots of Virtual Yamahoko Parade

Our future works include improving crowd simulation by making more realistic crowd behaviors and creating more Japanese character models; enriching the movements of the floats and the crews; adding more float models to the virtual parade.

Furthermore, we are trying to build a virtual Yamahoko Parade experiencing system by which the users can also experience the vibration as they are riding on the virtual float. We collected the acceleration data of the float with acceleration sensors during the real parade and experimentally reproduced the vibration using a vibration platform. An immersive system which integrates interactive CG animation, high fidelity sound, and realistic vibration is under construction.

Fundings

This research has been partially supported by the Digital Museum Project in the Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors thank the Gion Festival Fune-hoko Preservation Association for the collaboration.

References

Gandy, M., S. Robertson, W. Price, and J. Bailey (2005). The Design of a Performance Simulation System for Virtual Reality. In *Proceedings of Human-Computer Interaction International 2005* . Las Vegas.

Magnenat-Thalmann, N., N. Foni, G. Papagiannakis, and N. Cadi-Yazli (2007). Real Time Animation and Illumination in Ancient Roman Sites. *The International Journal of Virtual Reality* 6(1): 11-24.

Papagiannakis, G., and N. Magnenat-Thalmann (2007). Mobile Augmented Heritage: Enabling Human Life in Ancient Pompeii. *The International Journal of Architectural Computing* 2: 395-415.

Yano, K., T. Nakaya, and Y. Isoda (2007). Virtual Kyoto: Exploring the Past, Present and Future of Kyoto. Kyoto: Nakanishiya.

Complex entity management through EATS: the case of the Gascon Rolls Project

Litta Modignani Picozzi, Eleonora
eleonora.litta@kcl.ac.uk
King's College London, UK

Norrish, Jamie
jamie@artefact.org.nz
King's College London, UK

Monteiro Vieira, Jose Miguel
jose.m.vieira@kcl.ac.uk
King's College London, UK

Managing entities like people, places and subjects across a large corpus of textual documents can be complicated. While the TEI guidelines offer a sound basis for the encoding of a great variety of textual material, there does not seem to be a general agreement on how to manage information that goes beyond the text, like entity information and relationships between entities.

At the Department of Digital Humanities at King's College London, past solutions for entity management have included the implementation of the following:

- Simple XML authority files that contain basic information about all the entities in the corpus;
- EAC files, where a file is created for each entity and linked from the texts using IDs;
- Bespoke databases, that are very tied to a specific project and are not designed to be reusable;
- RDF/OWL ontologies.[1]

But all of these seemed either to be too simplistic and hard to maintain, or too complicated for the requirements we had in the Gascon Rolls project.[2]

In this abstract we will demonstrate how we successfully applied the entity management tool EATS (Entity Authority Tool Set),[3] in the context of this project, to successfully manage entity information.

The Gascon Rolls Project (1317-1468) is an AHRC-funded collaborative venture between the Universities of Oxford and Liverpool, and the Department of Digital Humanities at King's College, London. It began in October 2008, and is due to end in December 2011.

The main aim is to make the unpublished records of the English Government of the Duchy of Aquitaine (1154-1453)[4] available to everyone, both in electronic and printed forms.

The corpus of the Gascon Rolls consists of one hundred and twelve rolls containing up to 67 membranes each containing enrolments by the English royal Chancery of letters, writs, mandates, confirmations, *inspeximuses*, and other documents issued by, and in the name of, the Plantagenet and Lancastrian king-dukes for their Gascon lands and subjects. The rolls contain a large set of information about people and places, in particular, that need to be harvested in order to offer sophisticated indexes and advanced search functions which can give a different research experience to the scholar approaching the online resource. At the time of writing there are 4381 people entities and 2842 place entities in EATS, and there is no issue with having many thousands more. By comparison, the EATS installation at the New Zealand Electronic Text Centre contains almost a hundred thousand entities, drawn from thousands of electronic texts.[5]

The open access online resource offers images of all the unpublished Gascon Rolls (1317-1468), an edition, in calendar (summary) form, indexes of person, places and subjects mentioned, and advanced search features.

The calendar editing framework is based largely on what has been previously developed,[6] [7] for the Fine Rolls of Henry III project.[8] It uses a customised subset of the TEI P5 guidelines;[9] these have been adapted to suit the particular needs of the structural variety presented by the Gascon Rolls.

Calendars are encoded directly in XML. Information is captured about structure (rolls, membranes, entries, openers and closers), dates, people and their offices, places and subjects.

This allows the creation of a number of displays, in both electronic and printed formats, and forms a basis for the construction of indexes and search facilities.

Information about people and their offices, places and subjects are encoded in-line, extracted from the text using a custom-built oXygen XML Editor plugin,[10] and added to EATS, where each entity is given a unique identifier (URI).

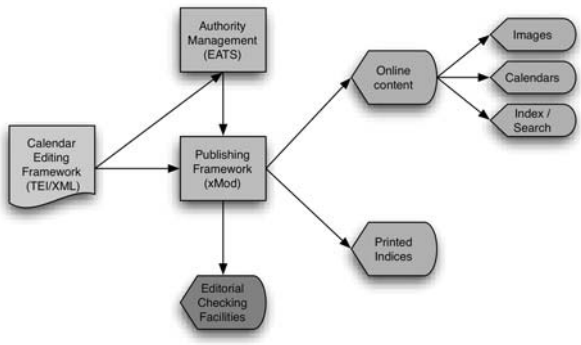


Figure 1: Graphic display of the Gascon Rolls editorial framework

EATS is a web application for recording, editing, using and displaying authority information about entities.[11] It is designed to allow multiple authorities to each maintain their own independent data, while operating on a common base so that information about the same entity is all in one place. It can be accessed by multiple users simultaneously, allowing the research team members to work independently and collaboratively at the same time.

By using a central entity management system to capture information about people, places and subjects mentioned in the rolls, it is possible to track a single entity throughout all the rolls and establish relationships between the entities.

Relationships, along with the variant names of entities, are crucial elements in entity management, since they are not amenable to discovery through simple search interfaces. The Gascon Rolls project harvests relationships from two sources: those explicitly created in EATS by the researchers, and those that can be inferred from TEI markup within the texts. The former encompasses information that is not present in the text, or which is not bounded by time or circumstance; for example, familial relationships. The latter derives from nested name markup, for example specifying that a person is from a particular place.

The publishing framework created at the Department of Digital Humanities at King's College London, xMod,[12] generates the output for the different calendar reading views and for the single entity pages which are the basis for online and printed indexes. The editorial framework and the entity management system together form a flexible structure that can be reapplied for the online and printed publication of other historical sources of the same nature.

One of the advantages of using EATS, as opposed to the other entity management possibilities mentioned above, is the possibility to access EATS directly from the XML editor, by using the plugin, which reduces the time it would take to add or edit information, compared to having to use a separate tool. Another

advantage is that, due to EATS being a database at its core, the information that it stores can easily be exported into other formats as desired, such as RDF/OWL and Topic Maps serialisations.

Further, since the EATS installation is not tightly coupled to the other components of the Gascon Rolls project - that is, it does not reference them, but is only referenced – its information can be easily reused in other projects. Indeed, such projects could use the same installation and add their own information about entities, without causing any changes to the information used by the Gascon Rolls project. Data can be shared without being intermingled.

There are some downsides to the current version of EATS, however. When adding more complex information about an entity, like relationships, the users found it complicated to use at first. It is not easy to add some rich information, like tertiary relationships (Person A related to Person B in Place C), although this was never a problem in our case. It is not possible to define entity type hierarchies; for instance *County* can't be declared as a subtype of *Place*. Such a facility would be extremely useful when trying to classify entities and to create an overview of the total number of entities of an overarching type.

The approach taken to entity management in the Gascon Rolls project has so far served it well, and is well suited to similar undertakings.

Notes

1. Ciula, A., P. Spence, and M. Vieira (2008). Expressing complex associations in the medieval historical documents: the Henry III Fine Rolls Project. *Journal of the Association for Literaty and Linguistic Computing and the Association for Computers and the Humanities* 23(3).
2. http://www.gasconrolls.org/
3. http://www.gasconrolls.org/
4. The Gascon Rolls (Rotuli Vasconie - C61 class in the UK National Archives) relate to the English Government of the Duchy of Aquitaine (1154-1453) and include information on the Hundred Years War, concluding in 1453 with the end of the Anglo-Gascon union.
5. http://www.nzetc.org/ and http://authority.nzetc.org/
6. Ciula, A. (2006). Searching the Fine Rolls: A Demonstration of the Electronic Version. *Paper presented at the International Medieval Congress 2006*, University of Leeds, July 10-13.
7. Spence, P. (2006). The Henry III Fine Rolls Project.*Digital Humanities* 2006. The First ADHO International Conference: Conference Abstracts. Universite Paris-Sorbonne.
8. http://www.finerollshenry3.org.uk/
9. TEI Consortium (2007). *TEI P5 Guidelines*. http://www.tei-c.org/ .
10. http://code.google.com/p/eats/#EATS_oXygen_Plugin
11. Stevenson, A., and J. Norrish (2008). Topic Maps and Entity Authority Records: An Effective Cyber Infrastructure for Digtal Humanities. *Paper presented at the Digital Humanities 2008 conference*. Oulu, Finland, June 25-29, 2008.
12. http://www.cch.kcl.ac.uk/xmod/

TextGrid Repository – Supporting the Data Curation Needs of Humanities Researchers

Lohmeier, Felix
lohmeier@sub.uni-goettingen.de
Goettingen State and University Library, Germany

Veentjer, Ubbo
veentjer@sub.uni-goettingen.de
Goettingen State and University Library, Germany

Smith, Kathleen M.
smith@sub.uni-goettingen.de
Goettingen State and University Library, Germany

Söring, Sibylle
soering@sub.uni-goettingen.de
Goettingen State and University Library, Germany

This poster will show how the TextGrid Repository assists researchers in the curation of their data and how they can make the data available with TextGrid tools to foster scientific re-use. The increasing importance of digitally-aided research methods has caused exponential growth in the creation of research data. New research methods and collaborative ways of using data sets require sophisticated research infrastructures to support researchers in the Digital Humanities and to enable the re-use of existing data. Data curation must be included in the planning stage as a fundamental requirement for all projects dealing with sustainable data. Therefore, this poster will present an overview of the technical infrastructure and the applicability of the TextGrid Repository for humanities researchers.

The TextGrid Virtual Research Environment (VRE), funded by the German Federal Ministry of Education and Research, provides tools, data and services in one integrated interface and supports the long-term archiving and management of research data. It provides a platform for researchers in the Arts and Humanities to curate their data that reflects universally-recognized best practices and standards. TextGrid consists of two main components: the TextGrid Laboratory (TextGridLab), the entry point to the VRE, and the TextGrid Repository (TextGridRep), a long-term humanities data archive. To preserve and maintain research data and ensure its long-term viability, current research practices in all stages of the research lifecycle must be supported. Therefore, the TextGridLab provides common functionalities in a sustainable environment to facilitate the re-use of data, services, and tools, and the TextGridRep enables researchers to publish and share their data in a way that supports long-term availability and re-usability. Rather than acquiring the technical knowledge necessary for data curation themselves, researchers can make use of services and guidelines for long-term data accessibility and sustainability during the initial planning stages of their projects through the TextGrid VRE.

After five years of research and development, TextGrid released a stable, operational version 1.0 in July 2011 and will release a version 2.0 in May 2012.[1] The value of this project is demonstrated by the fact that there are already eight long-term research groups actively using the TextGrid virtual research environment for the creation of scholarly editions, for the analysis of humanities research data, as a basis for the development of project-specific tools for specialized analysis and visualization, and for long-term digital archiving and facilitating world-wide access to research data for the scientific community. (In addition to these established research projects, as of February 2012 there were concrete requests from 18 additional research groups about how TextGrid can be integrated into their projects.) These projects deal with large amounts of humanities data that require specialized tools to reflect individualized requirements. To name a few examples:

- The project Blumenbach-Online (State and University Library, Göttingen) is producing an online resource providing access to the writings and collections of the German physician and anthropologist Johann Friedrich Blumenbach (1752-1840), in addition to secondary literature resources.[2]
- The project 'Hybrid-Edition von Theodor Fontanes Notizbüchern' (University of Göttingen) is creating a critical annotated edition of 67 notebooks from the German writer Theodor Fontane (1819-1898).[3]
- The Virtuelles Skriptorium St. Matthias (University of Trier / City Library Trier / Technical University Darmstadt) is developing a virtual reconstruction of the medieval manuscript collection of St. Matthias.[4]
- The Deutsches Literaturarchiv Marbach is creating an online edition of the letters of Ernst Kantorowicz, a german-jewish historian (1895-1963).[5]

These projects create significant amounts of data during the research process that require curation. This poster will show how the TextGridRep assists researchers in the curation of their data and in ensuring persistent access to data with TextGrid tools to support scientific re-use.

The first section of the poster will give an overview of the technical functionalities and infrastructure of the TextGridRep, which has been fully operational since July 2011. The TextGridRep provides a repository infrastructure based on grid technology. Researchers can decide how and with whom their data will be shared by using the detailed rights management module. Findings and research data can be published directly from the TextGridLab in the repository via a publishing process that guides researchers in preparing the data for long-term accessibility. The middleware consists of various components for handling files in the data grid, rights management in a role-based access control-enabled database, metadata in an XML database, and relations in a Resource Description Framework (RDF) triple store. On a basic level, TextGrid will offer bitstream preservation with redundant grid storage and tape backup for 10 years (as recommended in the guidelines of the German Research Foundation).[6] TextGrid developed its own metadata schema, especially suitable for digital editions, that supports different layers in its object model (item, work, edition, and collection).

When researchers publish their research data via the TextGridLab in the repository, the metadata provided will be automatically validated. The system validates against the TextGrid object model and checks if obligatory metadata fields like rights owner and license are well defined. In the next step, persistent identifiers are allocated by using a reliable handle service that is provided by the center for scientific data processing in Göttingen, GWDG, which is a main developing partner in the European Persistent Identifier Consortium and functions as the computer centre for the Max Planck Society. As part of the publishing process, the data will be frozen and moved to a storage cluster used for long-term preservation. If researchers want to update their data, they can copy it to their workspace, correct or further annotate the data, and publish the data as a new revision that is linked to the old revision. Both revisions will be available but the newer one will be more prominent in search results. The grid storage for the humanities and all connected resources are maintained together with those from the other academic disciplines at the common Grid Resource Centre in Göttingen (which has allotted 275 terabytes for the humanities).

The second section of the poster will show how researchers can make their data available with the TextGrid Repository. There are currently three different ways for research groups to enable access to their data in the repository:

1) All published data is available via the TextGridRep portal, which is already in place. It enables rapid searching with both simple and advanced search capabilities, in addition to the option of browsing repository content, across public research data with fulltext and metadata indexes.[7] Complex editions can be browsed according to the TextGrid object model (see above) and predefined XSLT stylesheets provide HTML representations of TEI documents. Links between texts and images, such as those created in the TextGridLab interface, will be displayed in a synoptical view in a future version of the portal.

2) Research groups who create a digital edition often want to present their data in their own portal with specific graphics, labels, and predefined browse and search options. Therefore, an open REST interface for individual portal solutions is provided so that research groups may provide specific elaborated access to their research collections with common technologies like Javascript, CSS, HTML.

3) Research groups who want to provide complex customized visualizations and complex project-specific search queries for their digital editions often use their own database for their project that is not connected to any long-term archiving solutions. We are developing a straightforward and easy way to sync the data stored in the TextGridRep (for long-term access) with a project-specific XML database (for the project-specific representation of the digital edition). A prototype is already in place that enables users to publish data from the TextGridRep to any eXist database with drag & drop functionality. Users can continuously test the representation of their XML data (e.g., TEI-formatted) in their own environment while they are still working on the digital edition with TextGrid. This allows research projects to annotate their data and develop the representation with XSLT and XQuery scripts at the same time. They can also easily publish new revisions of their data through TextGrid in the TextGridRep as well as in their own environments. TextGrid will provide a TextGrid-specific XQuery-module for the eXist XML database via the newly announced eXist AppRepository.[8] Users who install this module will be able to enhance their eXist environment to interface with the TextGrid metadata schema. To provide an out-of-the-box solution for the representation of a digital edition, TextGrid collaborates with the TELOTA working group at the Berlin-Brandenburg Academy of Science and Humanities to use their publishing framework Scalable Architecture for Digital Editions (SADE).[9]

XML technologies like XQuery and XSLT support the representation of digital editions using a common standard that promotes long-term reusability of and reliable access to research data. Therefore TextGrid facilitates the publication of digital editions in ways that are both easy to use and encourage the use of established best practices.

TextGrid maintains a strong network with other professional associations and eHumanities centres as well as with research infrastructure initatives and projects both nationally (DARIAH-DE,[10] eAqua,[11] D-Spin[12]) and internationally (DARIAH-EU,[13] CLARIN,[14] Bamboo[15]) with the aim to support the establishment of a comprehensive overall research infrastructure for the eHumanities.

Funding

The initial funding phase by the German Federal Ministry of Education and Research (BMBF) lasted from February 2006 to May 2009 (BMBF reference number 07TG01A-H). The second funding phase covered the period from 1 June 2009 to 31 May 2012 (BMBF reference number: 01UG0901A).

Notes

1. TextGrid v1.0: http://www.textgrid.de/1-0.html
2. Blumenbach-Online: http://www.blumenbach-online.de
3. Hybrid-Edition von Theodor Fontanes Notizbüchern: http://www.uni-goettingen.de/de/303691.html
4. Virtuelles Skriptorium St. Matthias: http://kompetenzzentrum.uni-trier.de/projekte/kernprojekte/virtuelles-skriptorium
5. Edition Ernst H. Kantorowicz: http://www.dla-marbach.de/dla/entwicklung/projekte/edition_ernst_h_kantorowicz/index.html
6. Proposals for Safeguarding Good Scientific Practice: http://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/self_regulation_98.pdf
7. TextGridRep Portal: http://www.textgridrep.de
8. eXist AppRepository: http://atomic.exist-db.org/blogs/eXist/AppRepository
9. Scalable Architecture for Digital Editions (SADE): http://www.bbaw.de/telota/sade
10. DARIAH-DE: http://de.dariah.eu
11. eAqua: http://www.eaqua.net
12. D-Spin: http://weblicht.sfs.uni-tuebingen.de
13. DARIAH-EU: http://www.dariah.eu
14. CLARIN: http://www.clarin.eu
15. Bamboo: http://www.projectbamboo.org

RIgeo.net – A Lab for Spatial Exploration of Historical Data

Loos, Lukas
L.Loos@stud.uni-heidelberg.de
University of Heidelberg, Insitute of Geography, GIScience, Germany

Zipf, Alexander
alexander.zipf@geog.uni-heidelberg.de
University of Heidelberg, Insitute of Geography, GIScience, Germany

1. Introduction

Transcultural (historical) research necessitates the integration of both macro and micro levels of research and analysis and emphasizes the dynamic and interactive character of its research objects in their historical and geographical dimension ('How Histories make Geographies' and vice versa). (Appadurai 2011; Döring et al. 2009; White 2010; Knowles et al. 2008; Owens 2007).

The interdisciplinary, interconnected HGIS-projects located in Heidelberg share a common vision: They facilitate research in the humanities by e.g. (semi-)automatic, it-enhanced processing of information, by adding modes of spatial search in heterogeneous data, by supporting visualization of patterns formed by spatio-temporal data or mash up of data from different sources of a different kind. Thereby they ultimately foster the researcher to formulate new questions and theses and eventually find some answers by experimenting in a sort of 'humanities lab'. Only close collaboration of geographers, historians and computer scientists can ensure that new digital tools and methodologies do not stay outside the framework of everyday scientific work in the humanities.

In the RIgeo.net project historians and geographers are working together in order to gain the full potential of the analytical capabilities of (historical) geographic information systems ((H)GIS) for historians. The aim of the project is the development of an – ultimately globally connected – infrastructure for the spatio-temporal analysis of historical sources. The combination of historical and geographical data as a part of a 'GIS-toolbox' provides a kind of 'laboratory' to analyze, recombine and disaggregate (mapped) information encoded in historical evidence, here: the abstracts of the Regesta Imperii[1] (hereinafter: RI) combined with a spatial

perspective and e.g. rearranged with uploaded findings of the RI researchers.

2. Data basis

As a starting point for the development of the project we use the online available and in the European context extremely relevant data base RI and cooperate with the team of researchers of the RI, namely Prof. Dr. Paul-Joachim Heinig. The RI are an inventory of 125,000 mostly German abstracts of documents of all 'German' Emperors from Charlemagne to Maximilian I. The comprehensive data of the RI provides for a continuous evaluation on the technical and content level. With a view to sustainability and utility maximization the implemented procedures are designed to be applicable for other similar data sets.

3. Project Objective

The objective is the development of spatio-temporal thesauri of places and to geocode the places of issue of the documents and the places mentioned in the documents. One objective is hereby is to assist in the research on itinerant kingdoms.

A central question is how to visualize (un-)certainty. One should bear in mind the universal warning that 'all visualizations of information are abstractions, which provide useful approximations of the real world. [...] Visualizations reduces the cognitive weight on the analyst or learner when the quantity of information, both quantitative and qualitative, is great, a problem is complex, and alternative solutions are numerous and surpass the capabilities of human reason' (Owens 2007).

Depending on the source material there are different challenges to be met to allow the user to judge on the degree of confidence of the visualized data (and thus avoid the danger of representing a higher precision than is justified by the historical resource).

The main difficulties are:

- Accuracy and precision of the spatial information: The location isn't just one clearly defined place, but e.g. an diffuse area, a vague offset from a named place ('near Heidelberg') and uncertain due to the credibility of the source itself (e.g. a forged medieval document). (Hill 2006).
- Accuracy and precision of the resolution of the temporal ranges: the 'temporal footprint': the beginning and ending dates are always of a certain fuzziness. The resolution of the dates found in historical sources varies widely, e.g. the RI on when Friedrich II. stayed in a place varies from an exact day (16.3.1217) to the notion of a year (1217) without any further information. (Hill 2006).

4. Implementation

As a first approach we implemented a system that consists of an extract transform and load pipeline (ETL) and a PostgreSQL/PostGIS database with a simple star-schema (see Figure 1). In the dimension tables we store place names, person names, dates and geographic coordinates. Places and dates of issue are being extracted from the XML documents of the RI through Xpath/XQuery queries and stored in the database. In the next step we deploy the GeoTWAIN[2] and Nominatim[3] web services to geocode the place names. From this we calculate an average traveling distance to derive probable whereabouts of unknown places or places that do not exist anymore. This narrows down the area for a manual search and additionally allows to disambiguate and geocode cases where more than one result was received from the web services.

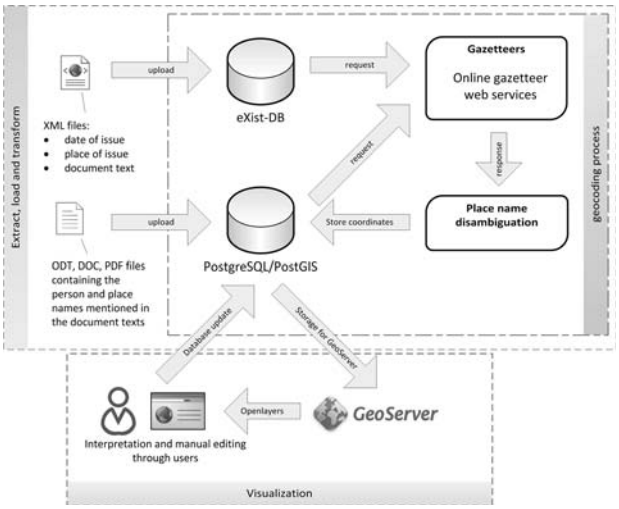


Figure 1: Workflow and system implementation (Source: compiled by the author)

5. Future Work

One focus of future work lies in the field of text mining and Geographic Information Retrieval (Feldman 2007; Leidner 2007). It can be assumed that the language used in the documents correlates highly with the age of the documents. The system allows to automatically annotate the place names and person names mentioned in the documents through the connection of the different datasets of the RI. The annotated documents can be used as training and test data which can be applied for a supervised machine learning approach in order to find regularities in the documents. Due to the sequential properties of the data, a dynamic Bayesian model such as state-of-the-art Hidden

Markov Models (Baum 1966) or Conditional Random Fields (Lafferty 2001) will be considered for learning.

References

Appadurai, A. (2011). How Histories make Geographies. *Transcultural Studies* 1. http://archiv.ub.uni-heidelberg.de/ojs/index.php/transcultural/article/view/6129 (accessed 28.10.2011).

Baum, L. E., and T. Petrie (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics* 37(6): 1554-1563.

Döring, J., and T. Thielmann, eds. (2009). *Spatial Turn: Das Raumparadigma in den Kultur- und Sozialwissenschaften*. 2nd ed. Bielefeld: Transcript.

Feldman, R. (2007). *The text mining handbook – advanced approaches in analyzing unstructured data*. Cambridge: Cambridge UP.

Hill, L. L. (2006). *Georeferencing: The Geographic Associations of Information*. Cambridge, MA: MIT Press, pp. 85-88.

Knowles, A. K., A. Hillier, and R. Balstad (2008). Conclusion: An Agenda for Historical GIS. In: A. K. Knowles and A. Hillier (eds.), *Placing History. How Maps, Spatial Data, and GIS are Changing Historical Scholarship*. Redlands, CA.: ESRI Press, pp. 267-274.

Lafferty, J., A. McCallum, and F. Pereira (2001). Conditional ramdom fields: Probabilistic models for segmentation and labeling sequence data. *ICML, Proceedings of International Conference on Machine Learning,* pp. 282-289.

Leidner, J. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph.D thesis, School of Informatics, University of Edinburgh, Scotland.

Owens, J. B. (2007). Toward a Geographically-Integrated, Connected World History: Employing Geographic Information Systems (GIS). *History Compass* 5(6): 2014-2040.

White, R. (2010). What is Spatial History? *Spatial History Lab: Working paper*. http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29 (accessed 29.12.2011).

Notes

1. http://www.regesta-imperii.de (accessed 20.01.2012).
2. http://geotwain.uni-hd.de (accessed 01.11.2011).
3. http://wiki.openstreetmap.org/wiki/Nominatim (accessed 01.11.2011). (accessed 01.11.2011).

Automatic Topic Hierarchy Generation Using Wordnet

Monteiro Vieira, Jose Miguel
jose.m.vieira@kcl.ac.uk
King's College London, UK

Brey, Gerhard †
dh@brey.org.uk
Independent Researcher, UK

In order to make full use of the rich content of large text collections various finding aids are needed. One very effective way of accessing this kind of collection is via a subject taxonomy or a topic hierarchy. Most subject classification techniques (Sebastiani 2002) are based on supervised methods and need a substantial amount of training data that are used by the various machine-learning algorithms on which they are based. In many cases this constitutes a significant problem if the resources to create these training data are not available.

Unsupervised methods such as clustering algorithms, though not requiring the same resources in data preparation as machine-learning based methods, need considerable attention after the techniques have been applied in order to make the clusters meaningful to users. The use of existing powerful tools such as the semantic tagger developed at Lancaster University[1] avoids these problems, providing semantic tags for each document, but often these semantic tags are very general and therefore not ideal for a user who searches for more concrete subject terms.

The aim of the research described here is the automatic generation of a topic hierarchy, using WordNet (Miller 1995; Fellbaum 1998) as the basis for a faceted browse interface, with a collection of 19th-century periodical texts as the test corpus.

Our research was motivated by the Castanet algorithm, a technique developed by Marti Hearst and Emilia Stoica (Stoica 2004, 2007] to automatically generate metadata topic hierarchies. Castanet was developed and successfully applied to short descriptions of documents. In our research we attempt to adapt and extend the Castanet algorithm so that it can be applied to the text of the actual documents for the many collections for which no abstracts or summaries are available. It should also be a viable alternative to the other techniques mentioned above.

1. Methodology

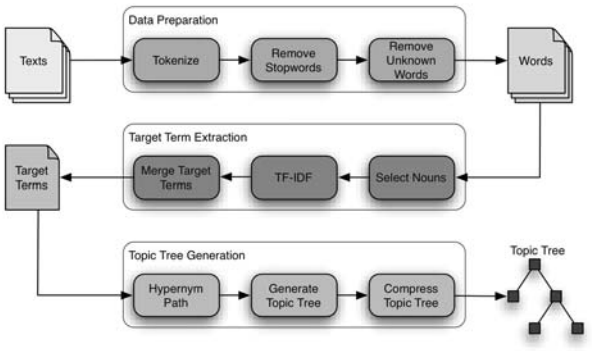


Figure 1: Algorithm Workflow

The algorithm for the automatic generation of the topic hierarchy is implemented using Python with the NLTK,[2] Networkx[3] and PyGraphviz[4] modules. It has three main processes:

1. Data preparation: data needs to be prepared so that the information contained within the texts is more easily accessible (Pyle 1999).
2. Target term extraction: select terms that are considered relevant to classify each text.
3. Topic tree generation: build the tree using the target terms.

2. Data Preparation

The first process in the algorithm, data preparation, is also the most important in any text mining application. Data preparation leads to understanding the data and ensures that the processes that follow will be able to get the most out of the data. The data preparation process has three main steps:

1. Tokenise each of the texts in the corpus.
2. Remove stop words from the list of tokens.
3. Remove unknown tokens from the list by using regular expressions and by doing lookups in WordNet's lemma dictionary. This also has the beneficial effect of removing words that are badly OCRed. Because WordNet is the basis for our topic hierarchy we regard words that not appear in its lexical database as irrelevant.

3. Target Term Extraction

Of all the words left in the texts after the data preparation, the algorithm only uses a subset of the

most relevant terms to create the topic tree. To select the target terms:

1. Select only the ones that are nouns by performing lookups in WordNet.
2. Compute TF.IDF (Term Frequency x Inverse Document Frequency) (Manning 1999) for each one of the nouns. There are many ways to compute term relevance. Initially we tested the algorithms described in Castanet, information gain (Mitchell 1997) and term distribution [Sanderson 1999], but they did not produce good results for our texts as the target terms were not meaningful enough. Therefore, we decided to use TF.IDF to select the target terms as this was a more straightforward approach and the results we were getting were more relevant.
3. Select the fifteen highest scoring terms and merge them all into a unique list of target terms.

4. Topic Tree Generation

The list of target terms constitutes the input for the topic tree generation process:

1. Using WordNet generate a hypernym path for each of the target terms.
2. Using the hypernym paths construct a tree by joining all the paths.
3. And finally, because the hypernym path lengths are varied and to make the tree more usable/readable the tree is compressed. To compress the tree:

1. Starting from the leaves, recursively eliminate a parent that has less than two children, unless the parent is the root node.
2. Eliminate child nodes whose name appears within the parent's node.
3. Eliminate selected top-level nodes, because they denote very general categories and have a very broad meaning.

5. Examples

Consider the following hypernym paths for the nouns:

1. writings: communication.n.02, written_communication.n.01, writing.n.02, sacred_text.n.01, hagiographa.n.01.
2. charter: communication.n.02, written_communication.n.01, writing.n.02, document.n.01, charter.n.01.
3. criticism: communication.n.02, message.n.02, disapproval.n.02, criticism.n.01.

This following image shows the previous hypernym paths combined to create a topic tree. There are many parents with just one child and the height of the tree varies across its different paths.

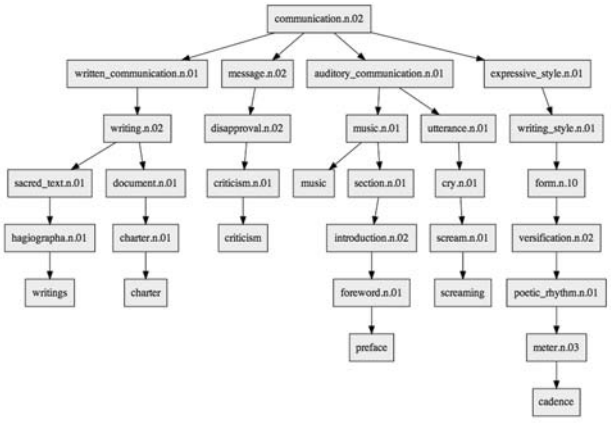


Figure 2: Communication full tree

In the following image we can see how the tree compression works. The grey nodes will be removed because they have fewer than two children, and the black nodes will be removed because their names appear within the parent node.

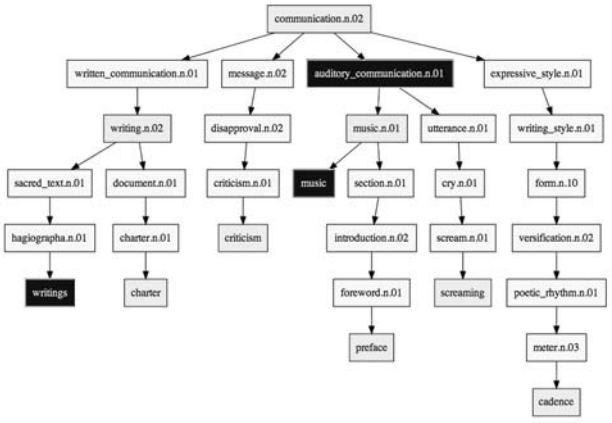


Figure 3: Topic tree compression

After the grey and black nodes are removed we get the compressed version of the tree. This tree is not as deep, therefore making it easier to navigate and use from a faceted browsing point of view.

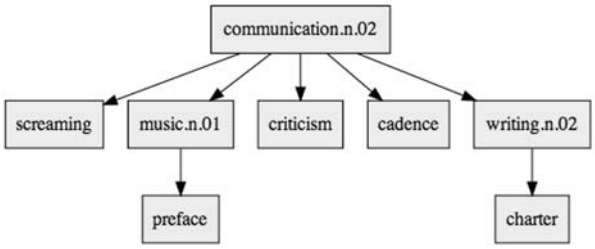


Figure 4: Communication compressed tree

6. Results and Future Work

The text collection used to test the techniques tried in our research is a digital edition of a 19th century periodical, the English Women's Journal (EWJ). The EWJ is one sub-collection of the Nineteenth Century Serials Edition (NCSE),[5] a digital edition of 6 newspapers and periodicals from the nineteenth century. Although the smallest collection within NCSE it was still large and varied enough to serve as a test corpus for our research. EWJ was published monthly between 1858 and 1864. It was edited and written by women and treated mainly literary, political and social contents. The collection is made up of 78 issues containing a total of 7964 articles. The main reason to choose the EWJ collection as our test corpus was that of the 6 periodicals contained in NCSE it is the one with the best OCR quality. Nevertheless the OCR quality of the EWJ was still a problem we had to contend with.

Our test corpus is composed of 1359 texts, with a minimum of 300 characters each, containing about 3.5 million words in total. The list of target terms has 8013 unique terms, when selecting the top fifteen target terms per text. The resulting topic tree has 18234 nodes and after compression it is roughly 50% smaller.

Of the samples we evaluated, over 90% of the topics are relevant, i.e. they clearly illustrate what the articles are about and the topic hierarchy adequately relates to the content of the articles. The results were evaluated by selecting a sample of 20 articles and generating the topic hierarchy for them. After reading each of these articles we compared its contents with the terms in the topic hierarchy, and counted the relevant terms for each one of them.

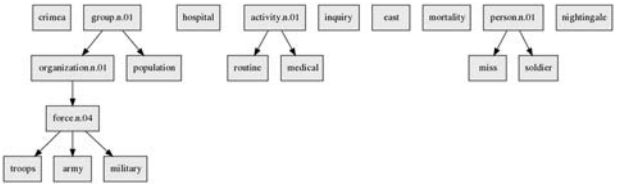


Figure 5: Topic hierarchy for an article about Florence Nightingale and the Crimean war [EWJ 1858]

Even though we consider that the algorithm produces good and promising results, we identified problems that mainly relate to the nature of our corpus. Despite our best efforts to filter out mis-OCRed portions of text, it was unavoidable that some misleading tokens remained. This sometimes led to strange results in the topic tree. For example, the partial text 'Another act in the great European drama' from the original article was OCRed into 'Ano , ther act in the great European drama'. The word 'Ano', which was not filtered out in the data preparation process, is classified in WordNet as the Abu Nidal Organization, which did not exist in the 19th century.

Another source of problems was the erroneous disambiguation of tokens with multiple meanings in WordNet. An example of this is the word 'drama', from the sentence above. It was output in the topic hierarchy as referring to a theatrical play rather than in its figurative sense as a turbulent event.

Future work could explore better ways how to deal with bad OCR. We also plan to enhance topic disambiguation; this could be achieved by analysing the domains of the topics in the hypernym paths before deciding which one to add to the topic tree.

Faceted browsing interfaces based on topic hierarchies are easy and intuitive to navigate (Morville 2006), and as our results demonstrate, topic hierarchies as generated by our approach form an appropriate basis for this type of data navigation. We are confident that our approach can successfully be applied to other corpora and should yield even better results if there are no OCR issues to contend with. And since WordNet is available in several languages,[6] it should also be possible to apply our approach to corpora in other languages.

References

EWJ 1858. *English Women's Journal*, 1 April 1858, pp. 73-79.

Fellbaum, Ch., ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Manning, Ch. D. (1999). Term frequency x inverse document frequency. In Ch. D. Manning, P. Raghavan, and H. *Introduction to Information Retrieval*. Cambridge: Cambridge UP, pp. 116-123.

Miller, G. E. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* 38(11): 39-41.

Mitchell, T. 1997. *Machine Learning*. New York: McGraw Hill, pp. 57-60.

Morville, P., and L. Rosenfeld (2006). *Information Architecture for the World Wide Web*. Sebastopol: O'Reilly, pp. 221-227.

Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco: Morgan Kaufman Publishers, chapter 3.

Sanderson, M., and B. Croft (1999). Deriving concept hierarchies from text. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval 1999*, pp. 206-213.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM *Computing Surveys* 34(1): 1-47.

Stoica, E., and A. H. Marti (2004). Nearly-automated metadata hierarchy creation. *Proceedings of HLT-NAACL 2004: Short Papers*, Boston, Mass., pp. 117-120.

Stoica, E., A. H. Marti, and M. Richardson (2007). Automating Creation of Hierarchical Faceted Metadata Structures. *Proceedings of NAACL/HLT 2007*, Rochester, NY, April, pp. 244-251.

Notes

1. http://ucrel.lancs.ac.uk/usas/
2. http://www.nltk.org/
3. http://networkx.lanl.gov/
4. http://networkx.lanl.gov/pygraphviz/
5. http://www.ncse.ac.uk/
6. http://en.wikipedia.org/wiki/WordNet#Other_languages

Hypotheses.org, une infrastructure pour les Digital Humanities

Muscinesi, Frédérique
frederique.muscinesi@revues.org
Cleo, France

Hypotheses.org, at http://hypotheses.org/ , is a French-based platform hosting academic research notebooks, or research blogs, for the use and consultation by researchers, students and research assistants working in the humanities and social sciences. Hypotheses.org now hosts over 330 research blogs, the work of 600 bloggers who have published over 20,000 posts. Created by Cléo, the Centre for Open Electronic Publishing, in 2008, Hypotheses.org is devoted to an unorthodox strand of the digital humanities: academic blogging. It is one of the three platforms included in the OpenEdition program.

At the Digital Humanities 2012 event, we shall be offering a detailed account of how Hypotheses.org works and inviting visitors and interested parties to browse and use the platform. Hypotheses.org is an online platform and is totally adapted to collective and individual presentations. Our data will also be explained in posters.

1. Academic blogging

Objectives

Academic blogging is a very direct means of communication for academic communities. It also provides a focus for collective exchange, collaboration and construction enabling researchers, research teams, librarians, academic communication assistants, etc, to promote their work.

Hypotheses.org connects with the tradition of academic correspondence, offering a new space for academic writing and enabling wider distribution of results and research processes.

Humanities and Social Sciences Research Blogging

While not the sole academic blogging platform available, it does have a major role to play in humanities and social science blogging, as academic blogging is usually confined to the sciences.

In HSS, the blogging field is more limited, so Hypotheses.org has been able to develop in European countries due to the following factors: in France, the http://culturevisuelle.org/ platform

is limited to specialised visual research blogs. In Germany, there is no umbrella platform for HSS blogs. And in Spain, the Madri+d platform hosts research blogs with no distinction between disciplines.

Furthermore, while there are many research bloggers in HSS, their blogs feature on non-specialist platforms and have little visibility, and the numerous existing university platforms offer their services solely to their own researchers.

2. Research, community, technology

Hypotheses.org is an academic tool

The platform has a number of qualities to reinforce its academic credentials. Firstly there is its academic committee, composed of researchers and professionals, which defines the platform's directions, accepts or refuses applications and supports the Hypotheses.org team in its promotion and communication activities. The catalogue offers a documentary classification of blogs with a stable editorial project. Hypotheses.org is also committed to creating long lasting digital research sources and focuses special attention on maintaining long-lasting urls while making research blogs readily accessible to all. The platform features specific technical functions such as bibliographical management, the insertion of footnotes, the integration of a Library Thing library and the importation of word documents, in particular. These characteristics have enabled Hypotheses.org's research blogs to be included in the Ebsco Atoz catalogue in May 2011, and in Isidore in September 2011. Recently they received ISSN classifications from the ISSN France centre. All of which means that Hypotheses.org is a quality academic platform, attracting a specialist readership.

Hypotheses.org is a community

The hypotheses.org bloggers' community has a series of tools at its disposal: training sessions, documentation, a discussion list, *La Maison des Carnets* collating editorial recommendations, technical updates, the latest news, a contact form and a dedicated email address. One person is in charge of running the blog and helping the community. The Hypotheses.org community is an important attraction for independent academic bloggers who wish to take part in discussions.

Hypotheses.org is a technical platform

The platform uses the free WORDPRESS.ORG blogging software. Every new blogger receives a blank blog created on the Hypotheses.org platform. The platform aggregates the various blogs and forms one point of access for visitors, while the blog itself provides another. The blogger has total control over how the blog is presented – on her/his own editorial project and on how contributors are managed, etc. – and can request modifications or technical adjustments from the Hypotheses.org team. New functions are sometimes developed internally; new templates are regularly made available; and the software is frequently updated. Hypotheses.org enables people with no special technical skills to manage a research notebook with innovative functions which is permanently improving.

Hypotheses.org, as an infrastructure, supports research, encourages involvement by academic communities in the digital humanities and contributes to the visibility of research in progress.

3. Development and practices

More readers and bloggers = new practices

Between 2008 and 2011, Hypotheses.org developed impressively in terms of quantity and quality: on the one hand, the number of new blogs, published posts and visits increased by 40% from 2010 to 2011, a sign that `http://hypotheses.org/` hypotheses.org has a huge appeal among readers and bloggers alike; on the other hand, technical developments, which have partially evolved from bloggers' requests, have given rise to, or been accompanied by, new practices.

Case studies

Hypotheses.org can effectively be considered and used as a tool for academic construction, as a support for academic information and as an online discussion space integrated into the broader digital humanities ecosystem. A typology of research notebooks (focussing on seminars, excavations, research programs, etc.) and publication practices (concerned with simplification, publishing sources, research complements, etc.) through case studies and a typology of notebook profiles shows a variety of practices. Furthermore, the tools at the community's disposal create relationships between bloggers, as well as between bloggers and the Hypotheses.org team. The community is jointly responsible for the initiation and development of discussions, new projects, improvements, technical changes and the introduction of new bloggers. The relationship between individual blogs and the collective platform means that a wide variety of projects can be developed and expressed, while the main characteristics of Hypotheses.org, as an academic infrastructure based on a community, also make up its character, the way it functions and the way it grows.

A fast-developing international academic blogging platform

The many practices hosted on Hypotheses.org create a very unique academic space and reveal that academic blogging is genuinely taking root in the French HSS community. The latest trends indicate that Hypotheses.org is outgrowing its linguistic origins and reaching out to German-speaking academic communities (with the creation of http://de.hypotheses.org/), and Spanish-speaking academic communities (with the creation of http://es.hypotheses.org/), as well as Portuguese and English speaking communities, among others. As an academic publication and information space, it has been receiving increasing amounts of visitors and is gradually becoming a relay for online media dealing with general information.

Hypotheses.org is thus an infrastructure exclusively dedicated to research which contributes to the expansion of the Digital Humanities through both creation and visibility. Its positive development shows that academic blogging is an expanding field, a place where digital academic communities can build, develop diverse practices and create tools which respond to the special requirements of their bloggers.

References

OpenEdition sites

Hypotheses.org: http://hypotheses.org

de.hypotheses.org: http://de.hypotheses.org/

es.hypotheses.org: http://es.hypotheses.org/

Hypotheses.org catalogue: http://www.openedition.org/catalogue-notebooks

La maison des carnets: http://maisondescarnets.hypotheses.org

OpenEdition: http://openedition.org

Cléo: http://cleo.cnrs.fr/

Platforms for scientific blogging

Culture visuelle: http://culturevisuelle.org/

Madri+d: http://www.madrimasd.org/blogs/

Academic platforms

Harward University: http://blogs.law.harvard.edu/

Université Paris 5: http://carnets.parisdescartes.fr/

Universidad de Salamanca: http://diarium.usal.es/

References

Blanchard, A. (2008). Ce que le blog apporte à la recherche. *La science, la cité*. http://www.webcitation.org/5iUAxXVIH (accessed 21 March 2012).

Dacos, M., P. Mounier (2009). Sciences et société en interaction sur Internet. *Communication & langages* 159: 123-135.

Dacos, M., P. Mounier (2011). Les carnets de recherches en ligne, espace d'une conversation scientifique décentrée. In C. Jacob (ed.), *Lieux de Savoir, 2. Les mains de l'intellect*, Paris: Albin Michel.

Gunthert, A. (2008). Why blog? *Actualités de la Recherche en histoire visuelle*. http://www.webcitation.org/5iT16xsza (accessed 21 March 2012).

Koenig, M. (2011). Une blogosphère à fort potentiel : petit tour des blogs d'histoire germanophones. *Digital Humanities à l'IHA*. http://dhiha.hypotheses.org/425 .

TXSTEP – an integrated XML-based scripting language for scholarly text data processing

Ott, Wilhelm
wilhelm.ott@uni-tuebingen.de
Universität Tübingen, Germany

Ott, Tobias
ott@hdm-stuttgart.de
Stuttgart Media University, Germany

Gasperlin, Oliver
oliver.gasperlin@pagina-tuebingen.de
pagina GmbH publication technologies, Germany

1. Introduction

With TXSTEP, we present and put up to discussion the prototype of a new, powerful XML-based tool for scholarly research in the text-based humanities. Its architecture is based on more than 40 years of experience in supporting humanities projects at the University of Tübingen and beyond.

The purpose of TXSTEP is not to provide another toolbox containing ready-made solutions for pre-defined problems. Of course, tools like these are adequate for many purposes; but we see no urgency to add a further one to the existing packages of this kind.

In fact, TXSTEP has been designed as a high performing scripting environment for the serious humanities scholar and other professionals in text data processing who face problems not easily solvable by XSLT or other means. TXSTEP gives them complete control over every detail of the data processing part of their projects.

2. Humanities software: basic requirements

Software for serious humanities research has to have certain basic qualities:

- it must be easy to handle, so that the scholar who is an expert in his field, but not in programming or computer science, can use it safely;
- it must be flexible enough to be adapted to the special requirements of each project, be it a philological analysis of a text or the preparation of a critical editon;
- it should support not only single phases of a project, but all its stages and steps, including (for an editorial project) first transcription of the sources and collation of the transcribed texts, evaluation of the variant readings, constitution of the edition text and the critical apparatus, up to (and even beyond) the preparation of the final publication of text, apparatuses, and indexes.

TXSTEP tries to take into account these somewhat contradicting requirements by defining the fundamental operations necessary for the processing of textual data, and by providing a separate program module for each of these basic functions, which can be used without any knowledge of conventional programming or scripting languages.

3. The solution: 1. Modularity

These modules may be combined almost arbitrarily: each module reads from and writes to a single basic file structure. This allows to combine these modules like Unix filters in arbitrary ways.

Where necessary, the single modules can be adapted to special requirements by the user, who may change default parameters (e.g. for providing a sort key for a non-latin alphabet) or provide additional ones (e.g. for the omission of the definite article in the sort key for titles in bibliographic records).

However limited the scope of the single modules may be, the flexibility of their combination can be illustrated by the fact that, for example, there is no dedicated program for generating an alphabetical word list. For this purpose the user has to combine the module for text decomposition (for which he has to provide the parameters defining the single elements and the sort keys), the SORT module, and the module which reduces identical or partially identical records contained in the sorted file to single index entries, and adds – when required – informations like frequency counts and/or references to the source text.

The modules provided by TXSTEP include:

- collation of different versions of a text; output of the differences in a synoptic list (for eye inspection) and for automatic processing in a file showing an appropriate structure;
- text correction and enhancement not only by an interactive editor, but also in batch mode, e.g. by means of correction instructions prepared beforehand (by manual transcription, or by program, e.g. the collation module);
- decomposing texts into elements (e.g. word forms) according to rules provided by the user;

- building logical entities (e.g. bibliographic records) consisting of more than one element or line of text;
- sorting such elements or entities according to the sort keys provided by the preparatory modules, accounting also for non-latin alphabetical rules and other sorting criteria;
- preparing indexes by generating entries from the sorted elements;
- transforming textual data by selecting records or elements, by replacing strings or text parts, by rearranging, complementing or abbreviating text parts;
- integrating external information into a file by means of acronyms;
- updating crossreferences;
- converting textual data from TUSTEP files into file formats used by other systems (e.g. for statistical analysis or for electronic publication) and vice versa.

As the output of any one of these modules may serve as input to any other module, the range of research problems for which this system may be helpful is quite wide.

4. The solution: 2. XML interface to an established text processing and analysis suite

In fact, TUSTEP, the Tübingen System of Text Processig tools, has been developed in the past 40 years along these lines. It has been and still is successfully used for many humanities projects in the German speaking part of the world, as may be detect by visiting www.tustep.org.

But, since TUSTEP's syntax is proprietary, not intuitive and supposed to be difficult to learn, users tend to help themselves with other – often less effective – tools or less specific programming languages.

TXSTEP gives an answer to this situation by providing a user-friendly XML-syntax, allowing beginners and advanced programmers to utilize the whole scope of TUSTEP services in a modern, established scripting environment. The benefits are obvious: support of an open standard, widespread dissemination, programming in every XML-editor, syntax highlighting, code completion and intelligible APIs. Moreover, TXSTEP is aided by the fact that there is no need to change the program's actual core. TUSTEP itself is open source, as TXSTEP is soon going to be as well.

5. The TXSTEP prototype

Development of TXSTEP began in 2009, when Tobias Ott, research associate and lecturer at the 'Stuttgart Media University' and CEO of pagina GmbH (a service provider for publishing houses) first came up with the idea to build an XML interface to the syntax of TUSTEP commands. This would all at once remove most of the barriers usually preventing people from using TUSTEP:

- it would offer an up-to-date established syntax,
- it would allow to draft TUSTEP scripts using the same XML-editor as when writing XSLT or other XML based scripts,
- it would let you enjoy the typical benefits of working with an XML editor, like content completion, highlighting, showing annotations, and, of course, verifying your code,
- it would offer – to a certain degree – a self teaching environment by commenting on the scope of every step,
- it would avoid many syntactical errors, even compared to the original TUSTEP scripting environment.

In the meantime, this idea resulted in a prototype of TXSTEP which we plan to demonstrate in more detail during the poster session. The prototype already contains the most important features of all the modules of TUSTEP listed above.

Not contained in TXSTEP is TUSTEPs typesetting module, which has been designed to meet the ambitious layout demands of publications in humanities research, including those needed by critical editions. The user may however use it in the original TUSTEP environment for publishing in print the results gained by TXSTEP, or he may even include it – in original TUSTEP syntax – into his TXSTEP scripts.

One of the features of TXSTEP is it's capability to process almost all forms of textual data, whether this being XML-data or plain text files. Therefore, even if textual data have to be processed in the first place in order to gain, for example, TEI-data or to enhance the markup of insufficiently tagged XML data, TXSTEP is at it's place.

The proposed demo is based on the mentioned prototype and shows the achieved state of our work in progress. The demonstration of TXSTEPs functionality will include tasks which can not easily be performed by existing XML tools.

Exploring Prosopographical Resources Through Novel Tools and Visualizations: a Preliminary Investigation

Pasin, Michele
michele.pasin@kcl.ac.uk
Department of Digital Humanities, UK

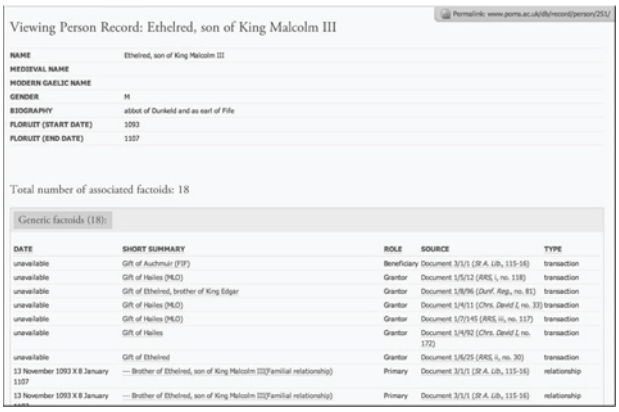
Structured Prosopography provides a formal model for representing prosopography: a branch of historical research that traditionally has focused on the identification of people that appear in historical sources (Verboven et al. 2007). Thanks to computing technologies, structured prosopography has succeeded in providing historians with a mean to enhance their scholarly work and make it available worldwide to a variety of academic and non academic users. Since the 1990s, KCL's Department of Digital Humanities (DDH) has been involved in the development of structured prosopographical databases, and has had direct involvement in Prosopographies of the Byzantine World (PBE and PBW)[1], Anglo-Saxon England (PASE)[2], Medieval Scotland (PoMS)[3] and now more generally northern Britain (BoB).[4]

Pre-digital print prosopographies presented its materials as narrative articles about the individuals it contains. Structured prosopography instead takes a more database-oriented approach as it focuses on isolating information fragments (usually, in textual form) that are relevant to the task of describing the life-events of a particular person. As a result, it is possible to quickly recollect such results in manifold ways using the logical query languages database systems make available.

In particular, DDH has been involved in the development of a general 'factoid-oriented' model of structure that although downplaying or eliminating narratives about people, has to a large extent served the needs of these various projects quite well. The structure formally identifies obvious items of interest: Persons and Sources, and extends to related things like Offices or Places. In our prosopographical model the *Factoid* is a central idea and represents the spot in a primary source where something is said about one or more persons. In other words, it links people to the information about them via spots in primary sources that assert that information (Bradley & Short 2003).

In general, it is fair to say that the issue of representing prosopographical data to the purpose of building large and efficient knowledge bases is no longer a critical problem for digital humanities research to tackle. Thanks to more than twenty years of research in this niche-area, a number of technical approaches such as the *factoid* one just mentioned have been discussed extensively and thus can facilitate enormously the initial design and construction of a structured back-end for a digital prosopographical project.

For that regards instead the visual rendering and final presentation of the contents of a prosopography, the amount of existing research is considerably smaller[5]. In fact it is quite common to present data using a classic database-centric approach: the tabular format. This approach normally boils down to a bibliographical-record-like table containing all the information available about a specific person: his/her recorded appellations and life dates, plus of course a variable number of rows that refer to the excepts that describe that person in the primary or secondary sources examined. We can see an example of this classic visualization approach in Fig. 1 (the example can be found online at `http://www.poms.ac.uk/db/record/person/251/`).


Viewing Person Record: Ethelred, son of King Malcolm III

NAME	Ethelred, son of King Malcolm III
MEDIEVAL NAME	
MODERN GAELIC NAME	
GENDER	M
BIOGRAPHY	abbot of Dunkeld and as earl of Fife
FLORUIT (START DATE)	1093
FLORUIT (END DATE)	1107

Total number of associated factoids: 18

Generic factoids (18):

DATE	SHORT SUMMARY	ROLE	SOURCE	TYPE
unavailable	Gift of Auchmuir (FIF)	Beneficiary	Document 3/1/1 (*SR A. Lib.*, 115-16)	transaction
unavailable	Gift of Hailes (MLO)	Grantor	Document 1/5/12 (*ARIS*, i, no. 118)	transaction
unavailable	Gift of Ethelred, brother of King Edgar	Grantor	Document 1/8/96 (*Dunf. Reg.*, no. 81)	transaction
unavailable	Gift of Hailes (MLO)	Grantor	Document 1/4/11 (*Chrs. David I*, no. 33)	transaction
unavailable	Gift of Hailes (MLO)	Grantor	Document 1/7/145 (*ARIS*, iii, no. 117)	transaction
unavailable	Gift of Hailes	Grantor	Document 1/4/92 (*Chrs. David I*, no. 172)	transaction
unavailable	Gift of Ethelred	Grantor	Document 1/6/25 (*ARIS*, ii, no. 30)	transaction
13 November 1093 X 8 January 1107	— Brother of Ethelred, son of King Malcolm III(Familial relationship)	Primary	Document 3/1/1 (*SR A. Lib.*, 115-16)	relationship
13 November 1093 X 8 January	— Brother of Ethelred, son of King Malcolm III(Familial relationship)	Primary	Document 3/1/1 (*SR A. Lib.*, 115-16)	relationship

Figure 1: A typical prosopographical record in tabular format

The tabular format has the advantage of offering a wealth of information in a clean and well-organized interface, thus simplifying the task of finding what we are looking for during a search. However, by combining all the information in a single view, this approach also hides some of the key dimensions used by historians in order to *make sense* of the materials at hand. For example, such dimensions could be deriving from a spatio-historical, genealogical or socio-political consideration of the data.

In other words, we acknowledged that although the tabular format succeeds in creating a comprehensive and condensed version of the information relevant for a search, it would also be interesting to examine if we could present the same data in a more piecemeal

fashion, according to predefined *pathways* or *views* on the dataset that aim at making explicit some of the *coherence principles* of the historical discourse.

We believe that this kind of approach could be desirable for both non-expert users (e.g., learners) – who could simultaneously access the data and get a feeling for the meaningful relations among them – and for experts alike (e.g., academic scholars) – who could be facilitated in the process of analyzing data within predefined dimensions, so to highlight patterns of interest that would be otherwise hard to spot.

With these ideas in mind we started to investigate the creation of innovative methods for presenting prosopographical data to users. For the moment these experiments have been developed in the context of a single prosopography, the 'Paradox of Medieval Scotland', but we reckon that they could be easily generalizable to other projects too, due to the intrinsic similarity of the approaches we used.

In particular, we have classified these exploratory tools into three broad categories:

1. **Visualizations that focus on specific aspects of the dataset.** In this group we have for example visualizations of people that focus on the historical dimension (such as timelines, or views of people's longevity – as in Fig. 2 below) or the geographical one (e.g., through maps). Also, we can have visualizations that aim at representing the *importance* of a person in the context of the database i.e., in terms of how often that person is cited. Other visualization approaches such as tag clouds seem to be very relevant in this context too, as they can be customized so to contain visual cues (e.g., through size or color) that give users a quick overview of the data (Bateman et al. 2008).
2. **Digital storytelling systems**. With this term we refer to the class of interactive applications aiming at letting users explore the database contents in a more incremental manner, following a story-inspired approach that reflects the key *coherence principles* of a discipline (Mullholland et al. 2004; Lawrence et al. 2010). For example, the 'friend-of-a-friend' exploratory tool (Fig. 3) lets you inspect the available connections between one person and the other people he or she was involved with (obviously, based on the information available in the database). This procedure may lead to dead-ends - for we are looking at a less-known person, e.g., a peasant – or it may take you to the discovery of unanticipated relations among people. This approach can be further elaborated so to pivot around specific social relationships (e.g., *witnesses* in medieval *transactions*), to the point that the path a user takes become analogue to the unfolding of a story only partly commanded by the computer (Murray 1997).
3. **Game-like interactive systems**. This group of interactive applications are a direct evolution of the digital storytelling systems just introduced (point b), with the difference that a series of user-feedback mechanisms are put into place with the aim of pushing users to engage more actively with the system (e.g., by rewarding certain decisions they make). These approaches are normally more relevant in the context of a classroom, where it is important to provoke students with challenging and mind-stimulating activities (Wolff et al. 2004).

The experiments can be accessed online at `http://www.poms.ac.uk/db/labs/` ; we invite the reader to try them out so to better develop a critical understanding of their potential usefulness.

Although we are still in a very early stage in the development of such exploratory interfaces, we already had a number of enthusiastic early reviews from project partners and work colleagues. As a result of this initial response, and from the evolution of our thinking around these issues, we are currently refining some of these tools and also developing other and more specialized ones. For example, an interesting aspect we would like to explore further concerns the potential applicability of the approaches herewith presented within other prosopographical scenarios, such as the ones focusing on ancient and modern history. Furthermore, in order to gain more empirical evidence on the potential usefulness of these tools for historians, a more formal user-evaluation session involving both high-school students and academic scholars is being planned.

At the conference we intend to present a preliminary analysis of these results, together with a practical demonstration of the most successful prosopographical exploratory tools we developed.

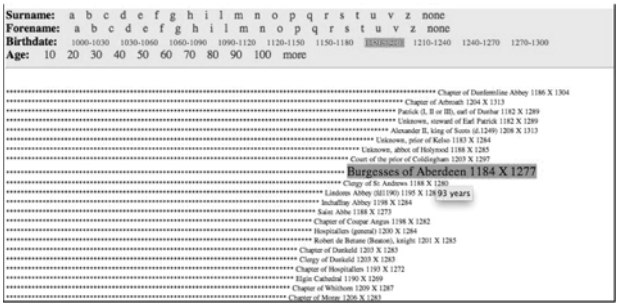


Figure 2: Explorative tool based on people's life-lengths. Each dot represents a year in the lifeline of a person.

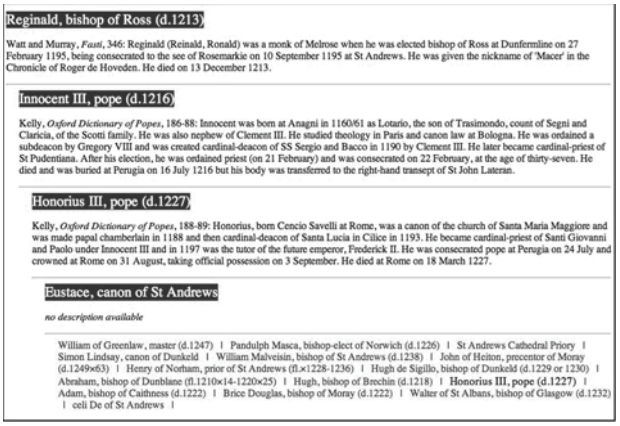


Figure 3. Explorative tool based on the social connections among people. Each indented level (top to bottom) represents a step in the 'friend-of-a-friend' chain.

References

Bateman, S., C. Gutwin, and M. Nacenta (2008). Seeing things in the clouds: the effect of visual features on tag cloud selections. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*. New York, NY, pp. 193-202.

Bradley, J., and H. Short (2003). Texts into databases: The Evolving Field of New-style Prosopography. *ACH/ALLC conference, Athens, Georgia*.

Lawrence, K. F., et al. (2010). Scanning Between the Lines: The Search for the Semantic Story. Panel Session in *Digital Humanities 2010*.

Mullholland, P., T. Collins, and Z. Zdrahal (2004). Story Fountain: intelligent support for story research and exploration. In *9th International Conference on Intelligent User Interface*.

Murray, J. H. (1997). *Hamlet on the Holodeck*. MIT Press.

Verboven, K., M. Carlier, and J. Dumolyn (2007). A Short Manual to the Art of Prosopography. In K. S. B. Keats-Rohan (ed.), *Prosopography Approaches and Applications A Handbook*. University of Oxford, Linacre College Unit for Prosopographical Research.

Wolff, A., P. Mullholland, and Z. Zdrahal (2004). Scene-driver: a narrative-driven game architecture reusing broadcast animation content. In *ACM SIGCHI International Conference on Advances in computer entertainment technology*.

Notes

1. http://www.pbw.kcl.ac.uk/
2. http://www.pase.ac.uk/index.html
3. http://www.poms.ac.uk/
4. http://www.breakingofbritain.ac.uk/
5. The Berkeley Prosopography Services toolkit (http://berkeleyprosopography.org/) is one of the very few enterprises that aim explicitly at the investigation of alternative visualization mechanisms in digital prosopography. However at the time of writing the results of this research have not been made available yet.

Heterogeneity and Multilingualism vs. Usability – Challenges of the Database User Interface 'Archiv-Editor'

Plutte, Christoph
plutte@bbaw.de
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

1. Starting point

The Archiv-Editor[1] is a multilingual desktop program created for working with prosopographical data and the Personal Data Repository (PDR), a server-client database system developed since 2009 by the DFG-Project Personal Data Repository[2] at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). Further development of the Archiv-Editor has become part of the DARIAH-DE initiative for digital infrastructure in 2011. The aim of the Person Data Repository is to provide a decentralized software system for research institutions, universities, archives, and libraries that allows combined access to biographic information from different data pools. It is especially designed to meet the requirements at the BBAW where divers research projects in Humanities collect personal data throughout their work. This data gathered from different research perspectives according to different research questions is usually stored in completely different formats ranging from text documents to relational databases and can hardly be maintained after the research project has finished. Since 2009, the Person Data Repository has provided a productive solution for storing, searching and exchanging person data including long term maintenance and tools such as the Archiv-Editor for working on the data and collecting new data.

The development of the PDR and the Archiv-Editor are based on the experiences that have been made at the BBAW since 2007 with a first version of the Archiv-Editor originally developed for the historical research project Preußen als Kulturstaat investigating archival materials (Czmiel & Holtz 2007). Concepts and approaches that had proved to be useful were generalized and implemented to be thoroughly customizable and extendable.

2. Challenges

The requirements of the starting point produce three main challenges for the data model and the architecture of the PDR and the Archiv-Editor. However, the Archiv-Editor's interface design must to ensure usability over a complex data model because on the user side no special software knowledge can be presupposed. At the beginning of the PDR project evaluations of repository software and prosopographical databases were made (Körner, Plutte, Roeder & Walkowski 2010) that led to the decision to construct a new system and a new data entry interface rather than extending any existing software. Although this was an expensive approach the experiences with the first version of the Archiv-Editor were very promising and a redevelopment of the software could lead to a perfectly well fitting program, and this was proven to be true.

Enfolding the Heterogeneity of Data

Personal data from different research and editing projects at the BBAW are not only provided in different formats and have to be transformed into the PDR data model (Körner 2010), but come from divergent research perspectives and are semantically heterogeneous: The PDR integrates by means of transformation and the Archiv-Editor data from historians, musicologists, linguists, and philologists. The scientific cultures and disciplines at the BBAW and the kind of personal data and perspectives for which it is gathered differ very much. The PDR has therefore chosen to use a very open data model based on XML that does not narrow the type of information stored but allows the retention of any kind of statement about persons which can then be classified according to customizable semantic, time and spatial dimensions (Walkowski 2009). The approach does not define a person as single data record, but rather as compilation of all statements concerning that person. Thus, it is possible to display complementing as well as contradicting statements in parallel, which meets one of the basic challenges of biographic research. To enable a very deep classification of proper names and notions, words can be marked with a TEI[3] compatible and fully customizable mark-up. Thus a very high level of atomization and complexity is implemented.

To ensure interoperability with other data stores transformation scripts are developed that allow the export of personal data into other standards such as TEI person description.

Multilingualism

Already in 2009 when the PDR started to develpe the repository software it began a cooperation with the Rom based German-French-Italian musicological project MUSICI[4] (Roeder & Plutte 2010a). Therefore

the Archiv-Editor was not only required to be translated into Italian, French and English but the classification schemas itself had to be designed to be language independent. Researchers in Rom want to classify their data in Italian but the data should be semantically compatible with those personal data collected at the BBAW in Berlin.

Usability

The complex data model of the PDR is diametrically opposed to the usability of the Archiv-Editor. The first step to usability was to encapsulate the XML – users do not see or edit XML. Since early 2011 when productive work with the Archiv-Editor began it has become clear that it is not enough to encapsulate XML: Research projects demanded to narrow the data entry fields and to enhance user guidance. In contrast to the – greatly appreciated – complexity and openness of the PDR data model a familiar database interface such as a formula with certain fields (Name, Profession etc.) was desired – as desktop program for offline work as well as an online version for web access.

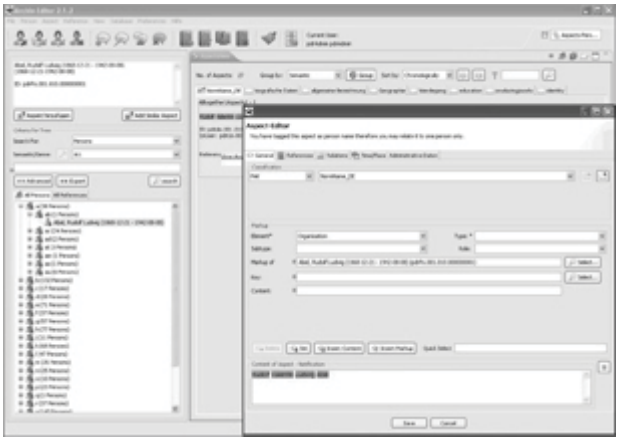
3. Approaches to these challenges

Modularity

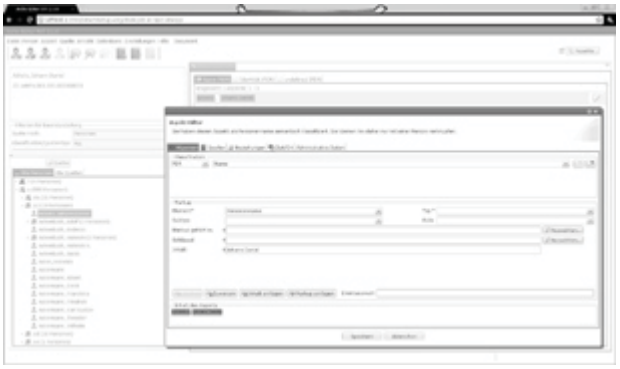
To combine theses divergent requirements the Archiv-Editor is designed modular. The Eclipse Rich Client Platform[5] was chosen because of it's extensibility and because it allows individual plugins for individual project requirements. While the PDR provides a solution to combine heterogeneous data und different research perspectives through a complex data model, the Archiv-Editor provides several general solutions plus individual adaption without cutting back the interoperability and exchangeability of the data.

One Archiv-Editor = 3 different Editors

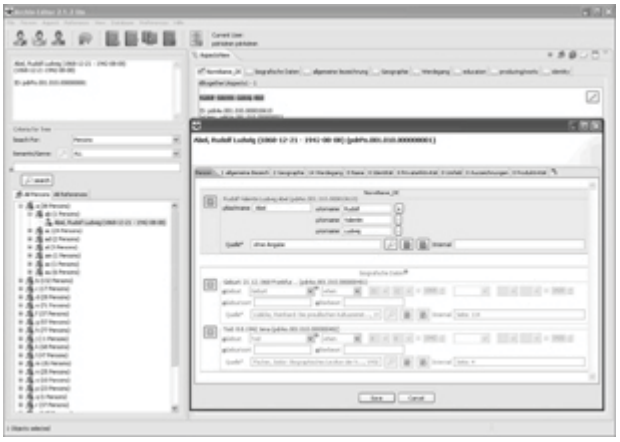
Currently two desktop versions of the Archiv-Editor are provided and an online version will be published in 2012 porting both versions to the web.



The lite version of the Archiv-Editor simplifies the appearance of the editor and provides only the very necessary functionality to make it easy for researchers to get started with the editor. The complex data model is represented in fully customizable formulas with user guidance in order to make the editor look as familiar as a simple database, although the openness of the data model is always in close reach through one click.



The advanced version of the Archiv-Editor contains all the functionality of the lite version and provides editors to fully edit all aspects of the data model. Besides the simple search different levels of advanced search are available as well as options for grouping and filtering statements. Furthermore, tools for datacleaning and automated PID (e.g. PND[6], VIAF[7], LCCN[8]) search are included.



Archiv-Editor RAP is the online version of the Archiv-Editor and includes both the advanced version and the lite version. It uses the Eclipse Rich Ajax Platform[9] Architecture to port the RCP Application to the web and hold most of the desktop functionality for working from anywhere and without prior installation. The Archiv-Editor RAP will be published in summer 2012.

Language independent classification schemes

Semantic classifications and ontologies are hardly language independent and notions and proper names

do not retain exactly the same definition, connotation and discriminatory power when translated into another language. Furthermore, the mapping of classification systems is very problematic (van Ossenbruggen, Hildebrand & de Boer 2011). In order to ensure interoperability between classifications in different languages used in different projects (e.g. in Rom and Berlin) without having to map categories afterwards the Archiv-Editor and the PDR use a hierarchical classification system. Superordinate concepts and ordinate concepts such as 'personName', 'placeName', 'profession', 'intellectual profession', 'musical profession' are predefined using unique English terms and are translated for presentation in all interfaces to project specific languages. Subordinate concepts such as 'violinist', 'violinista' or 'cantate' can be added according to project specific requirements without prior standardization. Thus classifications in different languages can be compared on high and mid-level terms without any mapping efforts. Although the classification system is thus not thoroughly standardized and internationalized it combines both the need for interoperability and the openness to easily manageable project specific extensions. Even if subordinate concepts such as 'violinist' and 'violinista' can not directly be mapped they can be identified as subcategories of the same language independent superordinate concept 'musical profession'.

References

Czmiel, A., and B. Holtz (2007). Quellenarbeit im Projekt 'Preußen als Kulturstaat' – Strukturierte Informationserfassung mit dem 'Archiv-Editor'. *Beiträge der Tagung .hist 2006 – Geschichte im Netz: Praxis, Chancen, Visionen*. vol 10, part vol. I.

Körner, F. (2010). Datenarchäologie und Datenaufbereitung, digiversity, September 30. http://digiversity.net/2010/datenarchaeologie/ .

Körner, F., C. Plutte, T. Roeder, and N.-O. Walkowski (2010). Software-Evaluation für ein Personendaten-Repositorium. Research Paper, Berlin. <urn:nbn:de:kobv:b4-opus-15111>

Neumann, G., F. Körner, T. Roeder, and N.-O. Walkowski (2010). Personendaten-Repositorium (PDR): *Berlin-Brandenburgische Akademie der Wissenschaften. Jahrbuch 2010*, Berlin, pp. 320-326.

Plutte, C. (2011). Archiv-Editor – Software for Personal Data. In S. Gradmann, F. Borri, C. Meghini, and H. Schuldt (2011). *Research and Advanced Technology for Digital Libraries*, Berlin: Springer, pp. 446-448.

Roeder, T., and C. Plutte (2010a). Un repositorio per i musici stranieri nell'Italia dal 1650 al 1750. Concezione del database del progetto ANR/DFG: Deutsches Historisches Institut in Rom, January 27.

Roeder, T., and C. Plutte (2010b). Die MUSICI-Datenbank. Personendaten-Repositorium und Archiv-Editor: *Jahrestagung der Gesellschaft für Musikforschung, Deutsches Historisches Institut in Rom und École Française de Rome, November 04–05.*

Roeder, T. (2009). Kooperationsmöglichkeiten mit dem Personendaten-Repositorium der BBAW: <urn:nbn:de:kobv:b4-opus-9231>

Schattkowsky, M., and F. Metasch (2008). *Biografische Lexika im Internet*. Thelem, Dresden.

Van Ossenbruggen, J., M. Hildebrand, and V. de Boer (2011). Interactiv Vocabulary Alignment. In S. Gradmann, F. Borri, C. Meghini, and H. Schuldt (2011). *Research and Advanced Technology for Digital Libraries*, Berlin: Springer, pp. 296-307.

Walkowski, N.-O. (2009). Zur Problematik der Strukturierung und Abbildung von Personendaten in digitalen Systemen: <urn:nbn:de:kobv:b4-opus-9221>

Walkowski, N.-O. (2011). Das Konzept einer polysemischen Datenbank und seine Konkretisierung im Personendaten-Repositorium der BBAW. In G. Braungart, P. Gendolla, and F. Jannidis (2011). *Jahrbuch für Computerphilologie – online*.

Notes

1. http://pdr.bbaw.de/software
2. http://pdr.bbaw.de/english
3. http://www.tei-c.org
4. http://www.musici.eu
5. http://wiki.eclipse.org/Rich_Client_Platform
6. http://www.d-nb.de/standardisierung/normdateien/pnd.htm
7. http://viaf.org/
8. http://www.loc.gov/marc/lccn_structure.html
9. http://www.eclipse.org/rap

Medievalists' Use of Digital Resources, 2002 and 2012

Porter, Dot
dot.porter@gmail.com
Indiana University, USA

1. Introduction

The field of medieval studies has a long and established history of scholarship assisted by technology. The first person who we now call a digital humanist was a medievalist, Fr. Roberto Busa, who in the 1940s first conceived of the *Index Thomisticus*, a complete lemmatization of the works of Saint Thomas Aquinas, which was developed through the 1970s with the collaboration of IBM. Today there are dozens of digital resources aimed at medievalists: online collections of digitized manuscript images, full- text databases, online scholarly editions, and secondary sources such as books and journals. Much attention is paid to medievalists who actively contribute to the design and implementation of digital resources but relatively little attention is paid to 'regular' or 'traditional' medievalists who use these resources for their own research (for example, the Digital Medievalist Community of Practice is a community for medievalists who build resources, not medievalists who use resources). On the other hand there has been a fair amount of work done on the use of digital resources by other scholars in the humanities (see for example the series of studies by Diane Harley on the use of digital resources by humanities scholars) but very little specifically about medievalists. The only study of the use of digital resources by medievalists that I have found thus far in my literature search is my own master's paper, cited below, which was undertaken in 2002 and is thus quite bit out of date.

2. Background

In 2002 as part of my MSLS degree work I undertook a research project entitled 'Medievalists' Use of Electronic Resources,' which surveyed a selected group of medievalists on faculty at universities across the US on their use of and attitudes towards electronic resources (`http://ils.unc.edu/MSpapers/2807.pdf`). The survey asked faculty how they preferred to access various different types of resources (journals, translations, facsimiles, etc.) and attempted to gauge preferences for using digital resources across the community (junior faculty vs. chaired faculty, preferences in different fields, etc.).

Much has changed since 2002. Google Books was first introduced in 2004. ACLS released its final report on Cyberinfrastructure, 'Our Cultural Commonwealth' in 2006. The NEH, which had long supported digital projects, awarded the first Digital Humanities Start-Up Grants in 2007, and founded the Office of Digital Humanities in 2008. In medieval studies, more and more digital resources are released every year, there is an established community of 'Digital Medievalists', and the International Congress on Medieval Studies has had a steady growth in the number and profile of presentations and events on issues relating to digitization. Digital projects in the medieval studies are ubiquitous. Are they being used?

Now is the perfect time to return to this project, to undertake the survey again, and to see just how (or if) attitudes and use have changed over the past nine years. Looking forward, this research can help resource developers design projects to best serve their audiences and can help librarians understand how scholars assess and use digital resources.

3. The Study

In September 2011 I distributed two parallel surveys, one 'open' and one 'closed'. Both surveys include the same questions; the surveys are very similar but not identical to the 2002 survey. Updates made include asking about use of electronic books and clarifying some of the other questions with input from librarians and both digital and non-digital medievalist colleagues. The closed survey was sent to a selected group of 100 faculty drawn from institutions listed in the CARA database at the Medieval Academy of America (this is the same approach taken for the 2002 survey and will allow for a close study between the findings between the two groups). The open survey was distributed publicly, and announcements sent out to listservs, to my friends on Facebook, and posted to my network on Twitter. This project will compare findings from the open survey both with both the survey from 2002 and the closed survey from this year.

I will be taking research leave in December 2011 and January 2012 in order to undertake analysis of the survey data and to draft initial research findings for presentation and publication. I will be publishing several articles on this research, aimed at several distinct audiences, including medievalists, digital medievalists, academic librarians, the digital libraries community, and he digital humanities community. By July 2012 I will have completed all data analysis and I will have a least one article submitted. The poster session would give me the opportunity to present the findings to a digital humanities audience.

References

Harley, D., et al. Final Report: Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines. *CSHE 1.10* (January 2010) (URL: http://cshe.berkeley.edu/publications/publications.php?id=351)

Porter, D. C. Medievalists' Use of Electronic Resources: The Results of a National Survey of Faculty Members in Medieval Studies. UNC Chapel Hill (December 2002) (URL: http://ils.unc.edu/MSpapers/2807.pdf)

Cross-cultural Approaches to Digital Humanities – Funding and Implementation

Rhody, Jason
jrhody@neh.gov
National Endowment for the Humanities, USA

Kümmel, Christoph
christoph.kuemmel@dfg.de
Deutsche Forschungsgemeinschaft, Germany

Effinger, Maria
Effinger@ub.uni-heidelberg.de
Heidelberg University Library, Germany

Freedman, Richard
rfreedma@haverford.edu
Haverford College, USA

Magier, David
dmagier@princeton.edu
Princeton University, USA

Förtsch, Reinhard
foertsch@uni-koeln.de
Universität zu Köln, Cologne Digital Archaeology Laboratory, Germany

---

Beginning in 2008, DFG and NEH provided joint-funding through a bilateral program in digital humanities, a program that was based in part on the previous year's similar opportunity offered by the NEH and the Joint Information Systems Committee (JISC) in the United Kingdom. Such collaboration between US and European humanities funding organizations was inspired, in part, by the Report on Cyberinfrastructure in the Humanities and Social Sciences, commissioned by ACLS, which charged that such entities 'should work together to promote collaboration and skills development – through conferences, workshops, and/or grant programs – for the creation, management, preservation, and presentation of reusable digital collections, objects, and products' (5). Equally influential was the assertion that 'defining and building cyberinfrastructure should be a collaborative undertaking [...] designed to foster and support collaboration across disciplinary and geographical boundaries [...] to bring new perspectives to bear on the exploration of the cultural record' (28). Within the spirit of these charges, funding organizations increasingly believed that a call for increased international collaboration should be supported by efforts to collaborate with each other; in short, they

should not demand a model that they were unable to emulate.

This poster session will highlight work from digital humanities projects funded through these international collaboration efforts. Staff from US and German funding agencies will be present to discuss the challenges and opportunities to be found in international collaboration, with input solicited from the projects themselves. The discussion will cover the organizational and financial side of international cooperation as well as issues of compatibility with regards to content. What is good practice in starting new projects in an international environment? What challenges have to be met getting different, established infrastructures to work together smoothly? How can different approaches be combined in a fruitful manner? How can funding organizations allow for flexible collaboration between institutions (including their own) often bound by complex administrative requirements? Why is it important to build infrastructure within an international context? What are successful strategies in combining infrastructure development and humanities research? Finally, what are the next steps for generating successful – even ad-hoc, fluid – international collaborations as part of the natural workflow in digital humanities production?

Brett Bobley, Jason Rhody (both of the National Endowment for the Humanities in the United States) and Christoph Kümmel (Deutsche Forschungsgemeinschaft e.V., DFG, in Germany) will be available to discuss strategies, surprises, and challenges in developing the collaborative framework for the bilateral grant programs. Four projects funded through this process will be emphasized, with each representing a different type of digital humanities project that have unique institutional and disciplinary approaches to bilateral cooperative efforts.

The German Sales Project: Art Works, Art Markets, and Cultural Policy (Dr. Maria Effinger, Universitätsbibliothek Heidelberg/ Heidelberg University Library) allows conversation about opportunities for shared expertise across cultural boundaries. A partnership with the Kunstbibliothek – Staatliche Museen zu Berlin, the Universitätsbibliothek Heidelberg, the Forschungsstelle 'Entartete Kunst' at the Universität Hamburg, and the Getty Research Institute, the project team is locating auction catalogs that record art transactions in Germany, Austria and Switzerland between 1930 and 1945. By documenting and digitizing thousands of catalogs in Europe and the United States and producing a database of these records that will form part of the Getty Provenance Index, they will ultimately allow scholars to explore wider issues concerning the German art market beyond traditional provenance research.

The Music Encoding Initiative (MEI) reveals the challenges and opportunities in developing and implementing standards while participating within a coalition of users for international projects. MEI, supported by two grants through the DFG/NEH Bilateral Digital Humanities program, developed an extensive XML DTD that allows for rich tagging of notated scores, which are mostly commonly available only as image files. By encoding music scores, scholars can carry out the same kinds of operations that are commonly performed on electronic textual sources, such as compiling musical corpora, data interchange, and comparative analysis. This project began as a collaboration between the University of Virginia and the University Paderborn, and has transitioned into a prototype gathering exemplar projects involving dozens of scholars, librarians, and universities crossing interdisciplinary and national boundaries. Dr. Freedman co-directs the NEH-supported 'Les Livres De Chansons Nouvelles De Nicolas Duchemin,' a digital facsimile project centered on a dozen sets of Renaissance music books, which also serves as an exemplar project for MEI; Dr. Freedman also is a member of the MEI advisory board.

The Yemen Manuscript Digitization Initiative (YMDI) is a collaborative project between Princeton University Library and the Freie Universität, Berlin. David Magier (Princeton University) reports on international cooperation in recovering cultural heritage materials and their subsequent incorporation into research libraries. The private manuscript libraries of Yemen comprise one of the world's largest and most important collections of Arabic manuscripts. Collectively, these 6,000 private libraries possess some 50,000 codices, many of which are unique, recording a rich cultural legacy of Arabic and Islamic literature from the eighth century to the present. Because Yemen is relatively remote from the central lands of Islam, it has preserved many extremely rare sources, including some of the earliest extant Qur'an fragments and theological tracts, and works of great importance for the study of classical Islam, Arabic literature, science, and history. In recent years, however, Yemen's private libraries have suffered great losses, in part due to sectarian extremists; over 10,000 manuscripts, including several entire libraries, have been destroyed. YMDI is working with a local foundation in Yemen in order to address this critical situation by devoting itself to the digital reproduction of these private collections.

The Hellespont Project, a collaboration between the German Archaeological Institute (DAI) and Tufts University, combined and expanded the collections

of two of the oldest and most established digital projects in Classical Studies – Arachne and Perseus – in order to work toward a single comprehensive digital library about the ancient world. Dr. Reinhard Förtsch of the Cologne Digital Archaeology Laboratory offers commentary on the challenges and cooperative models in building interoperable resources across international boundaries.

In addition to these topics (sharing expertise; building standards and garnering users; international cooperation in recovering cultural heritage; and interoperability), Bobley, Rhody and Kümmel will be available for broader discussion about the future of international infrastructure for digital humanities, data-driven approaches, the rich entanglement of research methods with infrastructure development, and the rationale of bilateral and multilateral approaches in funding and research.

The following participants will be available during the poster session:

- Brett Bobley, Director, Office of Digital Humanities, National Endowment for the Humanities
- Christoph Kümmel, Program Officer, Scientific Library Services and Information Systems, Deutsche Forschungsgemeinschaft
- Jason Rhody, Senior Program Officer, Office of Digital Humanities, National Endowment for the Humanities
- The following participants contributed to the poster session:
- Dr. Maria Effinger, Universitätsbibliothek Heidelberg (Heidelberg University Library), German Sales Project
- Professor Dr. Reinhard Förtsch, Universität zu Köln, Cologne Digital Archaeology Laboratory, Hellespont Project
- Richard Freedman, Haverford College, Music Encoding Initiative & Les Livres De Chansons Nouvelles De Nicolas Duchemin
- David Magier, Associate University Librarian for Collection Development, Princeton University,Yemen Manuscript Digitization Initiative

CWRC-Writer: An In-Browser XML Editor

Rockwell, Geoffrey
geoffrey.rockwell@ualberta.ca
Department of Philosophy and Humanities Computing, University of Alberta, Canada

Brown, Susan
sibrown@ualberta.ca
University of Alberta and University of Guelph, Canada

Chartrand, James
jc.chartrand@gmail.com
Open Sky Solutions, Canada

Hesemeier, Susan
s.hesemeier@ualberta.ca
University of Alberta, Canada

1. Introduction

The Canadian Writing Research Collaboratory (CWRC) has developed an in-browser text markup editor called CWRC-Writer for use by collaborative scholarly editing projects. This poster will demonstrate the editor and discuss the named entity annotation features that use stand-off RDF for text annotation. The combination of the poster and demonstration will:

- Introduce CWRC-Writer so that attendees can try it on their own,
- Show the hybrid markup model that combines in-text XML and stand-off RDF, and
- Explain the agile development process followed and recruit testers.

We deliberately propose this as a poster for two reasons. First, we want to provide an opportunity for attendees to try CWRC-Writer. Second, we want to recruit a larger circle of individuals and projects willing to test it with real editing needs.[1]

2. CWRC-Writer

CWRC-Writer is an open-source scholarly editor that is undergoing extensive testing and real-world use in scholarly editing projects. CWRC-writer has or will have the following features:

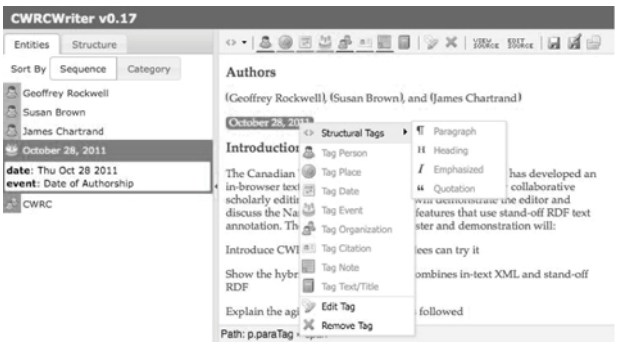


Figure 1: Screen shot of CWRC-Writer showing the tagging options

- Close-to-WYSIWYG editing and enrichment of scholarly texts with meaningful visual representations of markup
- Ability to add named entity annotations to texts
- Ability to combine TEI markup for the text and stand-off RDF for named entities
- Ability to export using 'weavers' that recombine the plain text, the TEI, and the RDF into different forms (including an embedded TEI-compliant XML)
- Documented code to allow editorial projects to incorporate CWRC-Writer into their environments

3. Background to CWRC-Writer

The Mashing Texts project, which was funded by the Social Science and Humanities Research Council of Canada, developed a prototype text collection and editing environment (called JiTR) that included a Java Web Start XML editor built on Eclipse by Open Sky Solutions.2 The idea was to test the viability of an easy-to-use editor that could be launched from a web-based collections manager with the text that the user wanted to edit. This meant that users would not have to buy and install a complicated editor. Simultaneously, Open Sky Solutions had been developing a similar XML editor for the Russell Letters project led by Dr. Nicholas Griffin at McMaster University.[2] While these editors worked, the time they took to load via the Internet was too long, and developments in browser-based JavaScript editors made it feasible to re-implement the editor as something easy to use in the browser itself.

CWRC, which is funded by the Canada Foundation for Innovation to develop a collaborative editing and publishing environment, is therefore developing an in-browser editor dubbed CWRC-Writer.[3] CWRC has built the first usable version of the editor and is working on a second version for the Spring of 2012. The development is led by James Chartrand of Open Sky Solutions. CWRC-Writer is based on TinyMCE, a javascript editor using jQuery to extend functionality.[4]

4. Agile Development Process

This project uses an agile development model to develop the editor in close consultation with CWRC partner projects and member projects. As part of the JiTR project, we developed personas and usage scenarios for those personas. CWRC, once funded, then developed specific use case scenarios for the XML editor with wireframes showing how it might be launched (from an editorial environment) and how it might look. Now we are developing this editor in iterations with input from partner projects that use it in their editing or born-digital writing. With the partners we follow an agile process that involves:

1. Presenting prototypes to the partners with suggestions of what we want tested and where we need suggestions. Susan Hesemeier manages this process.
2. Summarizing the feedback and prioritizing the next features to be developed. Dr. Rockwell and Dr. Brown do the prioritizing in consultation with the developer.
3. Responding to queries as Open Sky Solutions develops the next iteration of the prototype.
4. Initial testing of the prototype by a researcher to address any obvious bugs so as not to waste partner time.
5. Presenting it back again to the partner participants to be used with their texts. Back to 1.

Each iteration takes about a month and we have completed three. Partner projects are committed to iterative development and have research assistants to help with testing in context.

5. Partner Projects

The partner projects for the first iterations include:

Orlando Project: This ongoing collaborative experiment in digital women's literary history has since 1995 involved more than 100 people, many junior scholars, in using a custom semantic tagset based structurally on the TEI but specific to literary history. Orlando's flagship publication appeared in 2006: Orlando: Women's Writing in the British Isles is an on-line cultural history generated from the lives and works of over 1200 writers. Orlando continues to produce new materials.

Wilfred Watson and Sheila Watson Projects: The international Editing Modernism in Canada project is producing scholarly print and digital editions of texts by modernist Canadian authors. Through

partnerships with several university libraries, University of Alberta Press and CWRC, the EMiC group at the University of Alberta, led by Dr. Paul Hjartarson, is producing digital and print editions of the literary manuscripts of Wilfred and Sheila Watson, who rank among the best late modernist writers in Canada.

Russell Letters Project: Philosopher and social critic Bertrand Russell was one of the twentieth century's great letter writers and a highly prolific one. His letters are a hugely important resource for philosophers and historians and anyone interested in twentieth-century culture and politics. The Collected Letters of Bertrand Russell project, led by Dr. Nicholas Griffin, is digitizing, transcribing, annotating and indexing more than 40,000 letters from the Russell Archives to create an on-line electronic edition.

Canada's Early Women Writers Project: Despite the prominence of star authors like Margaret Atwood, little is known of most of Canada's earlier women writers. This project updates and expands a bio-bibliographical database of 470 Canadian women writers housed at Simon Fraser University. The enlarged semantically-tagged version (of well over 1000 names) will include all notable English-language writers active before 1950 who lived in or wrote about Canada.

6. Named Entity Annotation

One of the issues flagged early on was that many CWRC partner projects wanted sophisticated annotation for names, places, titles, organization names, dates, citations, and events, as well as great freedom for personal annotation, including overlapping annotations. There is also a desire for interoperability across projects, the coordination of authority lists for these entities, and the ability to harvest some annotations.

The solution is to provide an editing tool that uses a custom in-memory javascript data structure, that can export structural markup that conforms to schemas such as the TEI and Orlando ones, along with Open Annotation Collaboration (OAC) RDF for deeper semantic annotation (using controlled authority lists and vocabulary) of references to people, places, events, organizations, bibliographic material, and dates. CWRC-Writer thus supports creation of structural XML and the RDF in a WYSIWYG environment (WYSIWYG in that the annotations are displayed in the editor as they might appear in the public interface). The formatting of the text within the editor relieves the user of the complexity of interacting directly with RDF formats and even with angle brackets, though both can be viewed on demand and edited by those comfortable with code. There is also provision for exporting enhanced text either as structural markup combined with representations of entities as (potentially overlapping) RDF or, for those who opt for hierarchical markup, entirely as nested tags. The poster will present this in detail with examples.

7. Authority List Management and Lookup

CWRC-Writer provides forms to look up or construct annotations for people, places, events, organizations, dates, and bibliographic references. The annotations are applied much like formatting is applied in a WYSIWYG editor: the end user highlights the text to be annotated, then clicks on an icon to trigger the annotation lookup or edit form. We are now developing an API so that CWRC-Writer can retrieve a list of recommended entities to present to the user for selection and nominate new entities for inclusion. This is a first step towards a system that can manage people, places, organizations and other entities centrally across multiple projects. In the first instance, this will work with entity data from several CWRC pilot projects and be developed collaboratively with the Watsons project, but we are working towards use of Cool URIs so that CWRC entities can be exposed as and interact with other linked open data.

CWRC's decision to design an editor that can be used without a full understanding of markup or RDF will undoubtedly be controversial, but we feel such an editor is needed by projects that bring on collaborators for specific tasks who are uninterested in the deeper technology, just as the accessibility of a web-based editor will be useful to many digital projects for light editing, correction, enhancement, and annotation of dynamic collections. We welcome the opportunity to engage in this debate by demonstrating the CWRC-Writer to interested members of the DH community.

References

CKEditor: http://ckeditor.com/

Collected Letters of Bertrand Russell project: http://russell.mcmaster.ca/brletters.htm

Cool URIs: http://www.w3.org/TR/cooluris/

CWRC: Canadian Writing Research Collaboratory: http://www.cwrc.ca/

CWRC-Writer: http://www.cwrc.ca/cwrcwriter

JiTR (Mashing Texts) project: http://tada.mcmaster.ca/Main/MashTexts

jQuery: http://jquery.com

Open Annotation Collaboration: http://www.openannotation.org/

Open Sky Solutions: http://www.openskysolutions.ca/

ORE, a component of the Open Archives Initiative: http://www.openarchives.org/ore/

Orlando Project: http://www.ualberta.ca/orlando and http://orlando.cambridge.org

TEI Lite: http://www.tei-c.org/Guidelines/Customization/Lite/

TinyMCE: http://www.tinymce.com/

Notes

1. For current information and to contact us visit the current CWRC-Writer site: http://www.cwrc.ca/cwrcwriter
2. See http://russell.mcmaster.ca/brletters.htm
3. See http://cwrc.ca
4. For TinyMCE see http://www.tinymce.com/ . For jQuery see http://jquery.com.

The Musici Database

Roeder, Torsten
roeder@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

Plutte, Christoph
plutte@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

1. About the Project

The Musici Database is a research platform that focuses on European musicians who migrated to Italy in the 17th and 18th century. It supports the work of the international research group *Musici*,[1] whose main interest is to compare the cultural impacts and the social roles of foreign musicians in the cities of Rome, Naples and Venice. As the sources on that historical period are dispersed over a number of archives, the researchers decided to share their data on a common platform. For the duration of the project, the Musici Database serves as a research infrastructure, which is intended to be published at the end of the project. The database will contain information on approximately 1,000 persons.

Database Architecture

The database uses an existing repository architecture that provides a broad support of typical research workflows, which can be configured individually. This architecture is known as the Person Data Repository (PDR).[2] The aim is to create a musicological database that is fed with data by a number of researchers, and is also capable of interoperating with other data sources (see Roeder 2009). As the quantity of accessible historical data can be critical for the results of the project, it is a main interest to find collaboration partners whose interest is to share research data, helping to enrich the information available to the researchers. The work is facilitated by the client software Archive Editor, which is a substantial element of the PDR.

2. An Approach to Digital Musicology

The configuration of the Musici Database revealed some data constellations that are typical for the field of digital musicology. Further specialties were accounted for by the focus on migrant biographies. An approach to this was presented by Roeder and Plutte (2010a).

Notation and Performance

The focus on music requires some special information types to be taken into consideration. Usually, music needs to be represented on two different levels: the musical performance, which can be described as the event when one or more musicians produce music; and the composition, usually written in musical notation, which makes it easier to remember and reproduce musical ideas of higher complexity.

Both these forms of musical representation are of relevance for the study of musical biographies. While the written representation is well-preserved in the archives, the performative representation did not exist before the 1860s and can only be reconstructed by coeval descriptions and historical performance practice. As the analysis of musical performance can serve as a key to the social role of music, while the analysis of written music offers insights into creation processes, musical ideas and inspirations, the Musici Database records descriptions of musical events and instruments as well as the various creation levels of compositions from draft to print.

Musical Biographies

Since the project takes biographies of musicians and composers into consideration and compares personal motivations and ethnic backgrounds to social roles they had abroad, a basic question is how musicians integrated their personal background and their expertise in a foreign structure. Minority cultures existed in most Italian cities, thus giving musicians an opportunity to adapt to a familiar local structure and to the individual musical scene of the city. Many musicians did not only work in the field of music, and in some cases, the migrant musicians' language skills supported second careers in other areas.

The Musici database was prepared to record personal background data and personal relationships as well as itineraries, residences and various occupations; the vocabulary was configured with a broad range of musical professions.

Quantity of Traces

A major challenge of the historical period of 1650-1750 and of the research subject of the Musici project is the relatively little quantity of sources, compared to the 19th and 20th century. Most musicians did not leave many traces on their itineraries, and the available information is more fragmentary. When a person has been identified as musician, sometimes it is not possible to find out his or her name. The Musici Database supports the combination of research data, which helps to identify identical persons by logical comparison of the available information.

Collated Names

Another essential problem is the lack of officially accepted proper names. The German musician Heinrich Schütz was known as Enrico Sagittario in Italy; the settlement of Aachen was known as Aquisgrana; the region of Sachsen was known as Sassonia. Much depended on the language and the cultural background of the writer. The Musici Database collocates various noun forms of the same entity and maintains the original form.

This also meets the requirements of the multilingual research group, where Italian, German and French are the principal languages, and of the final publication of the database, which is also intended to be internationally accessible.

3. The Musici Data Repository

The PDR architecture, which is used for the Musici Database, provides an environment that allows researchers to store, edit and share information on historical persons. It serves as a platform for both research and publication, and it functions as an infrastructure that is able to interoperate with other data sources.

A Data Model for Historical Persons

The PDR data model is centered on information snippets, which can be connected to corresponding persons and sources. An information snippet contains information on one or more persons, usually one descriptive phrase or a quotation from a source (described in MODS). All information can be marked-up with proper names, keywords and dates.[3] It is possible to configure the mark-up vocabulary in different languages, while original text and a language-independent standard form can be maintained.

Adapting the Data Model for Musicians

The PDR data model has proven its capability to meet most of the requirements for the musicological research, as it is well suited to describe both events, seen as information that is connected to all involved persons, and compositions, seen as source objects (see Roeder & Plutte 2010b). However, for the future it might be of interest to extend the data model by other types than persons, in order to describe also places, events and relics in more detail.

4. Project Workflow

Client Software

Archive Editor, the client software, is available for different platforms and facilitates the process of data ingestion.[4] It is based on Eclipse RCP, which allows extensibility and customization through plug-ins.[5] For the multilingual Musici project, localization packages were developed for German, English, French and Italian (presented by Roeder & Plutte 2010b).

Adapted Vocabulary for Musicological Research

To meet the requirements of musicological research, a vast vocabulary of some hundred semantic tags from musical and social history was developed before the researchers started the data ingestion. This project vocabulary was developed by the researchers with Archive Editor. In addition, a number of special source types were defined by the researchers, facilitating the source recording of archive documents like diaries, letters, contracts, bills, drafts, compositions, and etcetera.

Linking to other Resources

Archive Editor allows connecting and uniting data from different projects. The PDR data model utilizes standard identifiers defined by national libraries (PND, ICCU, BNF, LCCN). Data sets bearing the same identifiers can be united by choice. This improves the interrelation of directly and semantically connected personal data inside the Musici Database as well as among other person data repositories or other data sources using the same identifiers and standards.

Data Conversion

During the first phase of the project, a small quantity of data (about 150 musicians) that had already been collected had to be converted to the PDR data format. The project vocabulary was used to enrich the semantic layer of the data. Additionally, some services provided by PDR, like date and place identification tools, were applied during the automatic conversion.[6]

Database

Standard visualization methods for the repository data are available through the PDR basic infrastructure, although other visualizations need to be developed to meet the interests of the project. These are chronological and geographical views of itineraries and settlements of musicians, as well as social network views of musicians and other social groups, such as clerics or merchants. This could also be achieved collaboratively with other musicological projects.

Interoperation

It is desired to connect the Musici Data Repository to other repositories and databases to exchange data. Contacts have been maintained with French and German musical database projects. The repository can exchange data through RDF and is prepared to be integrated in the Linked Open Data network.

5. Outcome

The results of the Musici Database are twofold. Firstly, an infrastructure for musicological research was created by configuring the PDR system. It became clear that international projects require some special services which probably would not have appeared in a monolingual project. Secondly, a semantically rich database was created, which by its flexible vocabulary allows for structured searches and various forms of visualizations, configurable to correspond directly to specific research questions of the project. It is planned to publish all data together with the research results at the end of the project in December 2012.

Figure 1: The Musici Database

References

Körner, F. (2010). Datenarchäologie und Datenaufbereitung. *digiversity*, September 30, 2010, http://digiversity.net/2010/datenarchaeologie/.

Plutte, Ch. (2011). Archiv-Editor – Software for Personal Data, TPDL Conference, September 26-28, 2011.

Roeder, T. (2009). Kooperationsmöglichkeiten mit dem Personendaten-Repositorium der BBAW. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, urn:nbn:de:kobv:b4-opus-9231 (urn:nbn:de:kobv:b4-opus-9231)

Roeder, T., and Ch. Plutte (2010a). Un repositorio per i musici stranieri nell'Italia dal 1650 al 1750. Concezione del database del progetto ANR/DFG. Deutsches Historisches Institut in Rom, January 27, 2010.

Roeder, T., and Ch. Plutte (2010b). Die MUSICI-Datenbank. Personendaten-Repositorium und Archiv-Editor. Jahrestagung der Gesellschaft für Musikforschung, Deutsches Historisches Institut in Rom und École Française de Rome, November 04–05, 2010, http://www.gfm-dhi-rom2010.de/programm/hauptsymposien/hauptsymposium-iii/#AbstractPlutte

Walkowski, N.-O. (2009). Zur Problematik der Strukturierung und Abbildung von Personendaten in digitalen Systemen. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, 2009, urn:nbn:de:kobv:b4-opus-9221

zur Nieden, G. (2010). Musici. Poster-Präsentation auf dem Workshop 'Personen – Daten – Repositorien'. Berlin, September 27-29, 2009, http://pdr.bbaw.de/workshop/poster/musici

Notes

1. Musici. Europäische Musiker in Venedig, Rom und Neapel (1650-1750). Musik, nationale Identität und kultureller Austausch. http://www.dhi-roma.it/musici.html
2. Personendaten-Repositorium, developed by the Telota initiative of the Berlin-Brandenburg Academy of sciences and Humanities. http://pdr.bbaw.de
3. The approach of this data model is precisely described in Walkowski (2009)
4. Archive Editor is available for Windows, Mac OS and Linux. http://pdr.bbaw.de/software http://pdr.bbaw.de/software
5. The software includes a complex help system and a software update mechanism (Plutte 2011). It was extended by a search interface that connects directly to various online research services. Scenarios to integrate the Archive Editor into other research environments through plug-ins are currently being examined.
6. The methodolocigal approach for data conversion was described by Körner (2010).

The TEICHI Framework: Bringing TEI Lite to Drupal

Schöch, Christof
christof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Achler, Stefan
stefan.achler@gmx.de
University of Kassel, Germany

This contribution presents the TEICHI Framework, a light-weight set of modules for the Drupal content management system serving to display, search and download electronic documents encoded according to the Text Encoding Initiative's Guidelines for TEI Lite (Burnard & Sperberg-McQueen 2006). This publishing framework was developed jointly by humanities scholars at the University of Würzburg and computer scientists at the University of Kassel in Germany. To showcase the various features and concrete usability of the framework, we present a poster as well as a tool demo.

In comparison to other available solutions for the publication of electronic scholarly editions, such as the combination of an eXist database with a Fedora Commons repository or the Scalable Architecture for Digital Editions (SADE), the focus of the TEICHI Framework is clearly to be a lightweight, low-resource, and easy-to-use solution. Instead of competing directly with more comprehensive but also more demanding solutions, the TEICHI Framework focuses on low technical barriers and optimal usability (see Pape, Schöch & Wegner 2012). The overall feature set of the TEICHI Framework focuses on the requirements of humanities scholars working on straightforward, small-scale digital edition projects and the needs of teachers of electronic textual editing. It allows both these target groups to set up a complete web-based digital scholarly edition with minimal resources and technical support and see the results of their or their students' work online in a timely manner.

Currently, the TEICHI framework consists of four modules, one each for displaying documents, for searching them, for downloading them, and for browsing associated digital facsimiles. The modular approach means that not all of the modules need to be installed for the others to function, and additional modules may be added in the future. In addition, any digital edition published with the TEICHI Framework can take advantage of the full website infrastructure and community support Drupal provides, including useful features such as

revision history for each page or fine-grained users' rights management and optional extensions such as blog or forum modules.

At the core of the framework is the TEI Content module, which lets editors upload TEI documents via the Drupal GUI, order them in a hierarchical structure, and modify them online. The module stores XML/TEI documents in the Drupal database and displays them through an XSL transformation and a CSS stylesheet using Drupal's content filter mechanism. The default stylesheet included with the module provides specific support for many of the features that are part of the TEI Lite (P5) set of elements: author and editor notes are displayed in the right sidebar and scribal corrections and editorial emendations are catered for. A toggle mechanism allows for the display of two alternative readings via the 'choice' mechanism. There is currently no support for the use of encoded names, dates and places in an index.

There are several additional modules for the framework. The TEI Download module allows downloading various renderings of TEI-encoded documents. The editor/administrator can choose which of the downloading options should be available to the user for each document: EPUB files for ebook readers, plain text files for use with basic analysis tools, and XML/TEI files for more advanced analysis. There are various additional user-side options for downloading: either one of the alternative readings can be chosen, editor notes can be included or excluded, and several pages can be downloaded as one large file or as separate files. The TEI Image Viewer module allows to associate the displayed texts with their corresponding digital facsimiles. This module is based on Seadragon Ajax, a JavaScript library for interactively viewing high resolution images. It lets the user call up any given digital facsimile, scroll through the available facsimiles, and zoom in and move around in each of them. The TEI Search module, which is still experimental, adds some basic TEI-aware search functions to the framework, allowing the users of the edition to limit their search in ways, for instance, to exclude results from editor notes, to only include results from quotes, or to restrict the search to one of two alternative transcriptions distinguished by the 'choice' mechanism. A basic TEI Editor module is in development.

All modules are fully integrated into Drupal's administrative graphical user interface, allowing the editor or administrator to adjust various settings online. For instance, many settings of the TEI Content module, such as color schemes and button texts can be modified with a few clicks; also, the TEI Download module lets the administrator select metadata fields and a cover image for the EPUB files made available for download; finally, the TEI Image Viewer's control panel is defined as a block which can be moved to any block region on the website.

The poster will indicate the features each of the modules provides to the overall publishing framework and visualize how each of them is embedded into the Drupal environment from a technical point of view. The tool demo will showcase both the frontend and the backend of the framework: the ways in which readers/users can interact with an electronic scholarly edition supported by the framework will be shown based on one project using the TEICHI framework, the digital edition of Bérardier de Bataut's *Essai sur le récit* (Bérardier de Bataut 2010). A local installation will be able to show how easy it is to perform some of the typical tasks of editors/administrators of an electronic edition using the TEICHI framework, such as installing the modules in a Drupal environment, adding, organizing and modifying XML/TEI files, changing some basic settings and associating digital facsimiles to the transcriptions.

References

Bérardier de Bataut, F.-J. (2010). *Essai sur le récit, ou entretiens sur la manière de raconter* [1776], digital edition ed. by Ch. Schöch `http://www.berardier.org`.

Burnard, L., and M. Sperberg-McQueen, eds. (2006). *TEI Lite: Encoding for Interchange*. Text Encoding `http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite`.

Pape, S., Ch. Schöch, and L. Wegner (2012). TEICHI and the Tools Paradox. *Journal of the Text Encoding Initiative 2* `http://jtei.revues.org/432` ; DOI: 10.4000/jtei.432.

Software

Drupal. `http://www.drupal.org`.

eXist-db. `http://exist.sourceforge.net`.

Fedora Commons Repository. `http://www.fedora-commons.org`.

Scalable Architecture for Digital Editions (SADE). `http://www.bbaw.de/telota/projekte/digitale-editionen/sade/sade-1.`.

The TEICHI Framework. `http://www.teichi.org`.

Seadragon Ajax. `http://gallery.expression.microsoft.com/SeadragonAjax`.

What Has Digital Curation Got to Do With Digital Humanities?

Schreibman, Susan
susan.schreibman@gmail.com
Trinity College Dublin, Ireland

McCadden, Katiet Theresa
katietmccadden@gmail.com
Trinity College Dublin, Ireland

Coyle, Barry
barrypcoyle@gmail.com
Trinity College Dublin, Ireland

The preservation and curation of digital objects begins at the time of the creation. Life-cycle management involves a wide variety of tasks and skills, from applying appropriate metadata to issues of access and reuse, transformation and migration. (What is Digital Curation?)

For the vast majority of humanities scholars, curation has been the preserve of librarians, archivists, and museum curators. But the increasing involvement of humanities scholars in the creation of digital objects, from the creation of texts encoded in TEI, to the development of large datasets suitable for datamining, to the creation of audio/video corpora, puts a new onus on humanities scholars to create durable data.

In a survey of a majority of the MA degrees awarded in digital humanities,[1] none offered a course/module that specifically focused on curation/preservation. More than one of University College London's courses/modules, as well as one at the National University of Ireland, Maynooth ('Creating Digital Humanities Artefacts') and Trinity College Dublin ('Theory and Practice of Digital Humanities') contains elements of digital curation within modules with a wider pedagogic focus.

Equally, much shorter-term training in the digital humanities often overlooks digital curation as an area of core curriculum. Yet, as digital humanists increasingly find themselves working in positions from alt-ac roles within a wide variety of memory institutions, to the creators of digital objects as part of a more traditional scholarly profile, we might do well to consider integrating digital curation more centrally in the curriculum. This gap in educational/training courses is being investigated by a pan-European project, Digital Curator Vocational Education Europe (DigCurV). DigCurV is a project funded by the European Commission's Leonardo da Vinci Lifelong Learning programme to establish a curriculum framework for vocational training in digital curation, `http://www.digcur-education.org/` ; with participants in Ireland, the UK, Germany, Italy, and Lithuania.

In order to inform development of the framework, DigCurV has been carrying out a variety of activities in the library, museum, archive and cultural heritage sector. Two surveys have been conducted, the first focused on training or competence centres to gather information about training opportunities, while the second examined the training needs of practitioners. Focus groups took place in Autumn 2011 in Germany, Ireland, Italy, Lithuania and the United Kingdom in order to learn more about the skills and competences required of those working (in the broadest terms) in digital curation.

Through research carried out by DigCurV, it has become clear that many of the skills and core competences required of people working in digital curation/preservation map well onto areas traditionally covered in digital humanities degree/training courses. Preliminary DigCurV findings enrich and expand our thinking about what competences are considered core to a digital humanities curriculum. For example, training on how to promote websites using social media and digital marketing strategies emerged as an important part of the digital lifecycle management process. Many of the technical skills required, such as proficiency in XML and TEI and a knowledge and understanding of standards, best practice and the latest technologies, also proved to be a shared competences. Project management skills play an important role in this field as people dealing with digital objects often find themselves influencing attitudes about technology and encouraging digital activities to be seen as core institutional activities.

DigCurV found that many of the skills and core competences requiring training are not strictly in the realm of 'digital curation' (ie object creation and preservation), but also deal with the management of objects within an institution, social media savvy, ability to manage organisational change, and cross-disciplinary communication skills. These findings should also expand our ideas of what digital humanities training/education is.

Thus the curriculum framework being developed by DigCurV might also be useful within a digital humanities context. Due to the high degree of interdisciplinarity between digital curation and digital humanities, a case can be made for more emphasis on digital curation in digital humanities curricula. This poster will explore recent findings of the DigCurV network in light of the DH curricula.

References

Bermès, E., and L. Fauduet (2011). The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France. *The International Journal of Digital Curation* 6(1) , http://ijdc.net/index.php/ijdc/article/view/175

Fino-Radin, B. (2011). Keeping it Online, http://rhizome.org/editorial/2011/aug/5/keeping-it-online/ .

Fulton, B., P. Botticelli, and J. Bradley (2011). DigIn: A Hands-on Approach to a Digital Curation Curriculum for Professional Development. *Journal of Education for Library and Information Science* 52(2).

Hank, C., and J. Davidson (2009). International Data Curation Education Action (IDEA) Working Group: A Report from the Second Workshop of the IDEA. *D-Lib Magazine*. 15(3/4) http://dlib.org/dlib/march09/hank/03hank.html ,

Kraus, K. (2011). When Data Disappears. *New York Times Sunday Review*, http://www.nytimes.com/2011/08/07/opinion/sunday/when-data-disappears.html?_r=3&ref=opinion

Lee, Ch. A, H. R. Tibbo, and J. C. Schaefer (2007). Defining What Digital Curations Do and What They Need to Know: The Digcurr Project. *ACM Digital Library*. Proceedings of the 7th ACM/IEE-CS Joint Conference on Digital Libraries.

Long, C. (2011). Developing and Implementing a Master of Archival Studies Program: A Collaborative Effort of a State University, a State Archive, and the National Archives and Records Administration. *Journal of Education for Library and Information Science* 52(2).

Spiro, L. (2011). Knowing and Doing: Understanding the Digital Humanities Curriculum. *Paper presented at the DH 2011*. See blogpost for further information: http://digitalscholarship.wordpress.com/2011/06/20/making-sense-of-134-dh-syllabi-dh-2011-presentation/

Stanton, J. M. Y. Kim, M. Oakleaf, R. D. Lankes, P. Gandel, D. Cogburn, and E. D. Liddy (2011). Education for eScience Professionals: Job Analysis, Curriculum Guidance, and Program Considerations. *Journal of Education for Library and Information Science* 52(2).

Tibbo, H. R., C. Hank, Ch. A. Lee, and R. Clemens, eds. (2009). Digital Curation: Practice, Promise & Prospects. *Proceedings for DigCurr2 (2009)*, http://www.lulu.com/product/paperback/proceedings-of-digccurr2009-digital-curation-practice-promise-and-prospects/4994819

'What is Digital Curation?' Digital Curation Centre. http://www.dcc.ac.uk/digital-curation/what-digital-curation

Notes

1. King's College London, University College London, Trinity College Dublin, National University of Ireland, Maynooth, University of Alberta, Loyola University.

Orbis Latinus Online (OLO)

Schultes, Kilian Peter
kilian.schultes@zegk.uni-heidelberg.de
Heidelberg University, Germany

Geissler, Stefan
stefan.geissler@zegk.uni-heidelberg.de
Heidelberg University, Germany

Transcultural (historical) research necessitates the integration of both macro and micro levels of research and analysis and emphasizes the dynamic and interactive character of its research objects in their historical and geographical dimension ('How Histories make Geographies' and vice versa).[1]The interdisciplinary, interconnected HGIS-projects located in Heidelberg share a common vision: They facilitate research in the humanities by e.g. (semi-)automatic, it-enhanced processing of information, by adding modes of spatial search in heterogeneous data, by supporting visualization of patterns formed by spatio-temporal data or a mash up of data from different sources of a different kind. Thereby they ultimately foster the researcher to formulate new questions and theses and eventually find some answers by experimenting in a sort of 'humanities lab'. Only close collaboration of geographers, historians and computer scientists can ensure that new digital tools and methodologies do not stay outside the framework of everyday scientific work in the humanities.

The online gazetteer 'Orbis Latinus Online' (OLO) is in its contents based on the three-volume, 1972 edition of the Latin-German-dictionary of the same name ('Großausgabe')[2] and published as a data base in cooperation with the Bayerische Staatsbibliothek Munich (BSB).[3]The project contributes to the ongoing interdisciplinary Heidelberg HGIS-projects RIgeo.net and uses the geocoder GeoTWAIN, which is part of the Heidelberg Research Architecture (HRA, Cluster of Excellence 'Asia-Europe in a Global Context'), in addition to further funding mentioned above. The OLO has currently reached beta-status and will be made accessible to the public in 2012.

The OLO provides coordinates to every Latin place name listed in the Orbis Latinus[4], additional contextual information missing in the print version as well as links between the entries and different cross-references. The user can, for instance, search for latin place names or their modern equivalents by using a multilanguage webfrontend (php/mySQL). The place names are stored in a PostgreSQL db to use the advanced GIS-functionality of PostGIS. The original print version provided by the Münchener Digitalisierungszentrum (MDZ) is linked and the place is shown on Google Maps (later: Open Street Map) to give the user the opportunity to verify the entry. A web service based on XML-RPC is currently tested. Due to copyright restrictions, this involves as by now just the BSB and RIgeo.net. The basic gazetteer services are accompanied by a kind of 'geo-wiki' which provides commentaries and alternatives to the shown locations or the assignment of the Latin place names to the modern ones (or vice versa). The backend of the OLO consists of several parts such as a tool for parsing the ocr-results and 'cutting them up' according to the needs of the relational data base, a user management as well as a versioning system.

At first sight, it seems to be a quite simple and one-off task to provide the coordinates, but the design of the database proved not to be without challenges, – a discussion of which will be the main topic of our poster. The first semi-automated georeferencing of the places added from the print version was made via Google Maps. Most geocoders like Google Maps provide satisfactory results merely for the main settlements, since Google Maps provides just the coordinates of the 'most important' place. This can but must not necessarily be the searched location (e.g. a town like 'Neustadt' exists more than 50 times in the world). Therefore besides establishing a connection to additional gazetteers and georeferencing-services with advanced structured information, a collaborative multi-user-environment with rights-management to add, correct and comment on the entries was needed.

Registered users can now provide alternative modern day place names and/or coordinates and comment on their choice, the level of probable historical correctness of their entries/of their entry, the time span for which their equivalent is valid and on the sources they have used. The search option for alternative/additional coordinates uses the functionality of a web based tool of the HRA called GeoTWAIN.[5] Among other options, it offers a quick visual search for coordinates including plausibility checks. During its interview-like search, a range of all coordinates is presented to the user, all of them fitting to given place names worldwide in all languages based on the search in different gazetteers (UTF-8), especially including the Getty Thesaurus of Geographical Names (TGN, which was licensed by the HRA).

To prevent vandalism and obvious incorrect data, new entries will be reviewed by the editors. The new coordinates or place names submitted are presented parallel to the older ones, open for rates and comments by fellow users. The entries on the results page are arranged according to their rating.

Contributors are shown by full name (if they want to), so that credits can be assigned to individual researches (and maybe the context of research they used the OLO for).

Together with GeoTWAIN and RIgeo.net the OLO is one of the various interconnected services of a virtual research environment, the 'gis-toolbox', envisioned and realized by the Heidelberg HGIS-research which is being developed with the aim to provide researchers with a set of databases and tools, e.g. to facilitate the analysis/mapping/combining of itineraries of medieval nobility and thereby define the spatial dimension of power and control beyond simplified linear boundary lines.

Project coordination

Dr. Kilian Schultes, Stefan Geißler, Jun Zhu: History Department, Heidelberg University in cooperation with Prof. Dr. Bernd Schneidmüller

Project members

Konrad Berner, Arina Chithavong, Matthäus Feigk, Philipp Franck, Anuschka Gäng, Peter Gietz, Timo Holste, Lukas Loos and Thomas Seitz

Project financing

Mostly a private initiative, Heidelberg University, Cluster Asia-Europe in a Global Context; Prof. Dr. Bernd Schneidmüller, History Department, Heidelberg University & RIgeo.net, "Frontier"-project, Heidelberg University (Prof. Dr. Alexander Zipf/Dr. Kilian Schultes in cooperation with Prof. Dr. Paul-Joachim Heinig, Akademie der Wissenschaften Mainz, Deutsche Kommission für die Bearbeitung der Regesta Imperii)

Project cooperations

Heidelberg Research Architecture (HRA) Cluster Asia-Europe in a Global Context, History Department - Universität Heidelberg

Project presented by Stefan Geißler/Dr. Kilian Schultes

Figure 1

Figure 2

References

Appadurai, A. (2011). How Histories make Geographies. *Transcultural Studies* 1.

Available from: http://archiv.ub.uni-heidelberg.de/ojs/index.php/transcultural/article/view/6129 . [Accessed 22 Jan 2011].

Döring, J., and T. Thielmann, eds. (2009). *Spatial Turn: Das Raumparadigma in den Kultur- und Sozialwissenschaften*. 2nd ed. Bielefeld: Transcript.

Graesse, J. G. T., F. Benedict, and H. Plechl, eds. (1972). *Lexikon lateinischer geographischer Namen des Mittelalters und der Neuzeit: Großausgabe, bearbeitet und herausgegeben von Helmut Plechl unter Mitarbeit von Sophie-Charlotte Plechl*, 3 vols. Braunschweig: Klinkhardt & Biermann.

Knowles, A. K., A. Hillier, and R. Balstad (2008). Conclusion: An Agenda for Historical GIS. In A. K. Knowles and A. Hillier (eds.), *Placing History. How Maps, Spatial Data, and GIS are Changing Historical Scholarship*. Redlands, CA: ESRI Press, pp. 267-274.

Owens, J. B. (2007). Toward a Geographically-Integrated, Connected World History: Employing Geographic Information Systems (GIS). *History Compass* 5(6): 2014-2040.

White, R. (2010). What is Spatial History? *Spatial History Lab: Working paper* [online] (Feb.). Available from: http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29 [Accessed 29 Dec 2010].

Notes

1. Döring et al. 2009; White 2010; Appadurai 2011; see also Knowles et al. 2008 and Owens 2007
2. Graesse et al. 1972. The less extensive one-volume edition from 1909 is online since 2000 as a text-only-html-website and has attracted some 370.000 visits: http://www.columbia.edu/acis/ets/Graesse/contents.html [Accessed 12 Feb 2012].
3. The BSB is the owner of the online publication rights. The pdf-version of the 1972-Orbis Latinus is published here: http://daten.digitale-sammlungen.de/~db/0005/bsb00050912/images/index.html [Accessed 12 Jan 2011]
4. Currently 140.000 Latin place names and 40.000 modern place names.
5. http://geotwain.uni-hd.de [Accessed 24.3.2012].

Semantically connecting text fragments – Text-Text-Link-Editor

Selig, Thomas
selig@ztt.fh-worms.de
Fachhochschule Worms, Germany

Küster, Marc W.
kuester@fh-worms.de
Fachhochschule Worms, Germany

Conner, Eric Sean
conner@fh-worms.de
Fachhochschule Worms, Germany

Text-Text-Link-Editor (TTLE) is a tool designed to allow researchers to link arbitrary text fragments across document boundaries. The tool's architecture was developed with the goal of creating a generic, easy to use tool that can support various disciplines of text research, e.g. for annotating texts or to present an original text together with its translation.

TTLE is developed as part of the TextGrid project and will be integrated in the TextGridLab. Being part of TextGridLab allows TTLE to benefit from the advantages TextGrid offers like user management, storage, search tool and metadata.

The authors are currently not aware of existing tools with a similar set of functionalities as those planned for TTLE. They do, however, closely collaborate in the development with Centre for e-Research (CeRch) at King's College London, to maximize synergy. They would also be interested in learning about possible other work on similar tools currently ongoing.

Due to scholars usually having different access privileges to various documents they want to work with, TTLE will offer three ways to select text fragments, depending on the accessibility of the source document. If the source document is writable for the user, corresponding tags to mark start and end of a selection will be inserted into the document. If a document is immutable for all users, a document-offset for start and end of the selection will be calculated, this also allows the selection of any text fragment.

Figure 1

Documents that are read only to the TTLE-user but are writable for other users are more difficult to handle. If tags in such documents have unique identifiers, those can be targeted as selection for TTLE. If no unique identifiers are available or do not match the selector's needs a copy of the document can be created for the TTLE-user. This selection mechanism still requires TTLE to validate all links on a regular basis, as any writable document can be changed and so can be linked text fragments. TTLE will therefore keep hash sums of all text fragments that are referenced by links and inform the user of any changes.

Any number of text fragments can be linked together. These links can be either free form or can be assigned specific types. These link types are pre-defined and belong to a specific person, who can share these types with selected projects. The personalization of link types automatically hides all link types created by other people. This will prevent the user from being confused by hundreds of similar link types. Also, this concept enables working groups to use the same link types without requiring everyone to create their own set. Text can be added to every link. As free form link types do not define any specifications, only a single text block can be added to each link of type free form. Pre-defined link types instead can carry an XML schema fragment, which then has to validate against the text attached to any link of that specific type. Due to the nature of TextGrid as a distributed, collaborative system all link types are stored in a centralized triple-store and are exposed to the web through open interfaces. Additionally, this online storage makes it very easy to assigned groups of link types to specific projects the owner wants to share them with. For future stages of the project an enhancement for handling link types is foreseen. Plans are: grouping link types together, handing over link type groups to other people and making link type groups publicly available to form a text linking community. These future plans will be kept in mind while implementing the current stage.

As mentioned earlier, any text fragments can be linked together. But additionally other targets can be included in a link as well. One possible target is an external URI. This allows to reference sources not directly accessible to TTLE. But additionally this allows to include non text objects into links, for example references to persons, e.g. using their FOAF (`http://www.foaf-project.org/`) or dbpedia (`http://dbpedia.org/`) identifiers. Identical URIs are treated as identical objects by TTLE. Another possible target for a link is another link. This allows grouping links and commenting link groups. For example you can create multiple links in multiple versions of a document showing that part a is followed by part b and then create a link referring all the links of all the documents showing that part a always comes before part b except in document version x. Links to a single target are also possible, offering an easy way to comment text passages when using the free form link type.

All links created will be stored in a specific link document. This document is a TEI compliant XML file. Each link target will be represented there together with a locator specifying the selected text depending on the chosen selection method (see above). Also information about the link type and the additional data required by the link type will be stored in this file. To be more useful to working groups, information about the user who created a link and the user who last changed a link will also be stored here. To ensure easy access to linked data, the links stored in a TTLE-file will be sortable by link types and target documents.

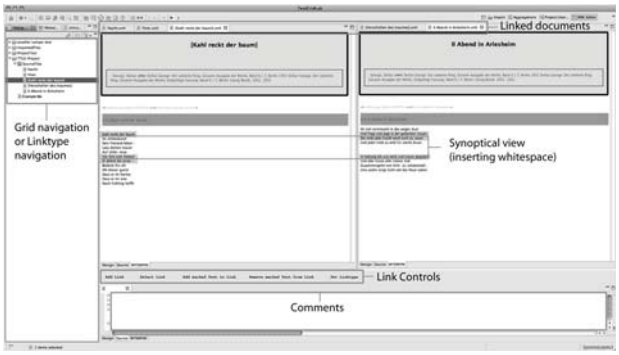


Figure 2

To work with linked texts, TTLE will offer different views, always directly comparing two documents in two viewports. If more than two documents form a link, each of these documents can easily be selected and made visible in one of the two viewports. At a later stage of the project plans are to use a customly defined XSL-transformation to even more adapt the displayed data to the needs of the researcher.

TTLE is developed with open interfaces in mind. One such interface enables other applications to propose links for specific documents. To demonstrate this, a separate webservice will also be implemented, which scans specified documents for similar text fragments and returns a link proposal to TTLE offering to link the text passages found to be similar.

Several prototypic implementations have been created to aid the selection of the best suited technology for this project. A web based solution has been found to be best fit. This web application will be integrated into TextGridLab using the Eclipse browser component, combining the utilization of all existing TextGrid tools, with the rendering capabilities of modern browsers and ease of maintenance of a web deployable application. Basic functionality is currently implemented and a working prototype of the application is expected for May 2012.

Plans to enhance TTLE even further after May 2012 have already been made. The following features were identified as particularly important and, subject to funding, are foreseen to be implemented starting June 2012:

1. Scientists often present the results of their research on their own web portal. To easily integrate the TTLE-results stored in TextGrid into these portals, an API which allows external applications to directly access links and the linked text fragments would be useful. This can either be done by export functionality within the TextGridLab or by offering a web service interface to access the data stored in the Grid.
2. Map like visualization of all linked documents in a project can give a good overview of the whole project, while displaying all documents of a specific link type can offer a very good view of a document's development. Both types of visualization would be important to be supported in TTLE.
3. Predefined link types are an essential element of TTLE. Having a comfortable editor for managing these link types would be helpful. Important functionality would be: grouping link types together, offering link types and type groups to other people, making link types and type groups publicly available, allowing modifications to link types which are in use and transforming specific link types into others.

The Melesina Trench Project: Markup Vocabularies, Poetics, and Undergraduate Pedagogy

Singer, Kate
ksinger@mtholyoke.edu
Mount Holyoke College, USA

If we take seriously Stephen Ramsey's call to arms at the 2010 MLA that Digital Humanities means making something, then it behooves us to teach coding – or other 'back end' digital skills – early and often. While many practitioners are focused on creating new courses teaching coding and theory, another important pedagogical approach is to consider how digital 'poesis', or making, might change the standard classroom experience. This poster presents some experiments with teaching TEI encoding to upper-level undergraduates at a liberal arts college, specifically in a seminar on women's poetics in the eighteenth- and nineteenth-centuries. As part of an upper-level undergraduate seminar at Mount Holyoke College, my students set about building a TEI digital edition of Melesina Trench's long poem 'Laura's Dream; or the Moonlanders,' perhaps the first science fiction text written by a woman in Britain. Since so much of the republication and dissemination of women's texts from this period occurred through electronic editions and digital scholarship, it seems only right to discuss the 'radiant textuality' of these materials.[1] One can only properly historicize the last thirty years of research in this field by discussing the ways in which digital encoding has aided and abetted our current thinking about the history and materiality of women's poetics. While many digital sites were originally built for scholars, it seemed time to consider what kinds of coding and interfaces might work best for students. Most importantly, in a course designed to think carefully about poetic literary terms, TEI seemed like a wonderful way to teach hands-on close reading. Perhaps one of the most interesting and important skills TEI can teach is to help our students become reflective about the various kinds of vocabularies they use to describe texts.

After spending the first few weeks discussing the benefits and problems of several digital editions of women writers, students were taught a series of exercises to learn the basics of encoding. During a two-hour workshop, they were introduced to TEI, especially the difference between html as a formatting language and XML as a descriptive or

analytical one. Rather than starting students with the document's header, they were given a guided tour through it, but spent their time learning the grammar and syntax of line groups and lines, persons and places, figurative strings, and the elements of CSS stylesheets. Students placed in small groups then had a month to tag fifty lines, with the help of a student mentor during lab hours. They were encouraged to use creativity by picking out three or four tropes (figures that repeated throughout the passage). They coded these using seg@ana – the segment element for isolating an arbitrary segment of test of their choice and the @ana type to signal an analytical interpretation of their choice. Finally, students color-coded their mark-up using a CSS stylesheet to visualize their own encoding trends. Two groups of students doubled up on each section of the poem, so that the class could compare two different methods of markup.

In a reflection submitted along with their XML documents, they were asked to describe any especially noteworthy analytical experiences from tagging, any limitations (or frustrations) encountered, and asked to propose possible additions to our tag set. While students readily understood tagging hierarchies, they had a difficult time distinguishing between those proper nouns reserved for <persName> and <placeName> and more generic or figurative people and places, including personifications (e.g., Terror or the Moon). This became an excellent topic for discussion, particularly of Romantic-era poems where real and illusive places tend to intermix. The ambiguity of the seg@ana tagging was likewise initially nebulous to students, but it eventually encouraged them to define what types of repeated figurative language spoke to them, labeling the @ana field (<seg ana="??">) with a signifier of their choosing. For example, one group summarized their process this way: 'The first theme we chose was "decay," but we had difficulty deciding if that also meant "dirt-related" imagery. We were unable to include as many words as we had initially wanted, so we revised our choices for tags. "Decay" became "morbidity" [...] "dirty" became its own tag, and evolved first to "earthy" and then to "earth" so we could include all organic imagery.' There were also some interesting pleas for more refined tags. One group desired an element for inanimate objects, to parallel the tags for places and people, but almost all groups wanted more granular tags for implied meanings of metaphors and for controversial and debatable topics.

At least three interlocking outcomes suggest how helpful teaching TEI to upper-level undergraduates might be – despite the time-intensive nature of including such a technology in an English seminar. Students not only produced multi-layered, associated observations based on multiple tropes color-coded through a single passage, but they were also quite reflective about the relationship between various elements and figurative trends. In one group's reflection, for example, the trope of 'religion is always within a line that is also tagged for emotion or male dominance.' In larger discussion of this observation, two groups of students had a particularly important debate about the gender differences between 'spirituality' and 'religion' within a lunar voyage poem, where the main female character Laura dreams that she goes to the moon and falls in love with one of its inhabitants. (See Figures 1 and 2 below.)

Second, and perhaps more important, TEI enabled students to undermine and find vocabularies for formal elements employed by women writers other than the classical rhetoric of figures of speech and figures of thought that still dominate poetic vocabulary in the academy today. Since even the most virtuosic women poets were autodidacts and not classically trained in English public schools, it remains to be seen how some of their more complex formal experimentation can be satisfactorily categorized by such rhetorical figures and forms. TEI – with its ambiguous and flexible vocabulary for poetry – actually allowed students to find larger, more amorphous terms for more organic formal choices in Trench's poem. By codifying their own vocabularies to isolate and name tropes, students began to consider terminology as descriptive and case-based rather than prescriptive and universal. For example, students decided that the long, multi-lined blocks Trench used in her poems were neither stichic or strophic in quality, but either both or neither. They tended either to decide that the term 'line group' gave them more leeway to describe how her units of verse operated or to suggest they needed a new term. This kind of interpretive creativity would not have occurred to them without the problem of trying to figure out TEI's standards.

Third, the limitations and possibilities of both classical terminology and TEI raised questions about the politics or the ideological traces in both types of categorical vocabularies. Having 'rediscovered' women writers in the past twenty years, many Romanticists and Victorianists are now beginning to more closely assess women's formal experimentation, considering how genre and forms carry with them histories and ideologies.[2] While TEI provides both the allure of extensibility, it also provoked particular kinds of observations based on its categories. For example, when discussing the <persName> and <placeName> elements, students felt pressured to locate nameable people and places within the poem and within Trench's prose writings that we read. They wondered whether the emphasis

on person- and place-ographies place a premium on women's connections to other famous figures and places (such as London or Bath), occluding Trench's poetic tactics of omitting concrete places while forging intersubjectivities between the 'moonlanders' and Laura. They worried that the emphasis on people and places – even imaginary – already framed their ways of looking at a poem for its connections to notable, concrete, and often masculine realities.

These are just a few of the ways in which the TEI terms, culled from many different disciplines, aided a reflective study of poetics and categorical language for younger learners. It likewise provided them with the opportunity to think about creating a scholarly edition by students and for students based on their reading and interpretive inquiries.

Figures 1 and 2: Two groups' TEI-encoded markups of the same two stanzas, with figurative tropes and place names

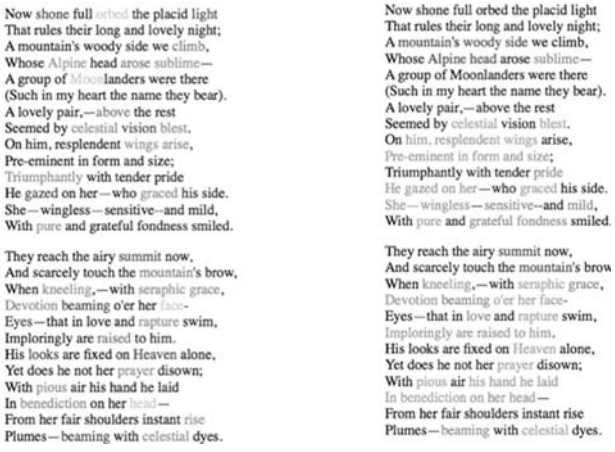


Figure 3: The browser view of figures 1 (left) and 2 (right), using CSS stylesheets. In the left-hand column, green=place; mint green=place with a proper name (<placeName>); red="ascension"; pink="spherical"; light blue="spiritual"; purple="vision." In the right-hand column, the color-coding is as follows: green=place; yellow="male dominance"; lavender="religion"; plum="feeling"

References

Blackwell, Ch., and Th. R. Martin (2009). Technology, Collaboration, and Undergraduate Research. *Digital Humanities Quarterly* 3(1). Accessed February 27, 2011. http://digitalhumanities.org/dhq/vol/3/1/000024/000024.html.

Buzzetti, D., and J. McGann (2006). Critical Editing in a Digital Horizon. In L. Burnard, K. O'Brien O'Keeffe, and J. Unsworth (2006). *Electronic Textual Editing*. New York: The Modern Language Association of America, pp. 53-73.

Curran, St. (2010). Women Readers, Women Writers. *The Cambridge Companion to British Romanticism*. 2nd ed. New York: Cambridge UP.

Digital Humanities Questions and Answers. 'Can you do TEI with students, for close reading?' Association for Computing in the Humanities. Accessed February 27, 2011. http://digitalhumanities.org/answers/topic/tei-encoding-for-close-reading.

Fairer, D. (2009). *Organising Poetry: The Coleridge Circle, 1790-1798*. New York: Oxford UP.

Flanders, J. (2006). The Women Writers Project: A Digital Anthology. In L. Burnard, K. O'Brien O'Keeffe, and J. Unsworth. *Electronic Textual Editing*. New York: The Modern Language Association of America, pp. 138-149.

Fraistat, N., and St. Jones (2006). The Poem and the Network: Editing Poetry Electronically. In L. Burnard, K. O'Brien O'Keeffe, and J. Unsworth (2006). *Electronic Textual Editing*. New York: The Modern Language Association of America, pp. 105-121.

McGann, J. (2004). Marking Texts in Many Dimensions. *A Companion to Digital Humanities*. Eds. R. Siemens, S. Schreibman, and J. Unsworth. Malden, MA: Blackwell. Accessed February 28, 2011. http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-4&toc.depth=1&toc.id=ss1-3-4&brand=default.

Piez, W. (2010). Towards Hermeneutic Markup. *Digital Humanities 2010 Conference*, King's College London. Accessed February 28, 2011. http://www.piez.org/wendell/papers/dh2010/index.html.

Project Tango. University of Virginia. Accessed February 26, 2011. http://uvatango.wordpress.com/2010/08/28/introducing-project-tango-2/.

Rudy, J. (2009). *Electric Meters: Victorian Physiological Poetics*. Athens, Ohio: Ohio State UP.

Notes

1. See, for example, Jerome McGann's *Radiant Textuality* and Julia Flander's comments on assembling an archive of women's writing in 'The Women Writers Project: A Digital Anthology.'
2. This so-called 'new formalism' can be seen in monographs on poetry including, to name just a few, Jason Rudy's *Electric Measures*, David Fairer's *Organising Poetry* or the recent British Women's Writing Conference panel on 'Formal Curiosities.'

Digital Edition of Carl Maria von Weber's Collected Works

Stadler, Peter
stadler@weber-gesamtausgabe.de
Universität Paderborn, Germany

Carl Maria von Weber (1786-1826) was a german composer who is nowadays mainly known for his best selling opera *Der Freischütz* although he has composed more than 150 works (ranging from operas to piano pieces) in total and has been of substantial influence on the history of musical composition in the whole 19th century. Next to his musical Œuvre he wrote several articles for music journals and even sketched a novel which unfortunately remained unfinished. Additionally to these works intended for the public the Carl-Maria-von-Weber-Gesamtausgabe (WeGA)[1] aims at publishing Weber's private correspondence and diaries as well until Weber's 200th anniversary in 2026. The WeGA was genuinely a 'traditional' scholarly project funded by the German Academy of Sciences and Humanities but has turned during the last years more and more into a 'modern' digital editions' venture both in the field of musical and text edition. The proposed poster should illustrate the current state of development of the text edition and it's framework and some envisioned future perspectives:

1. TEI based markup of every type of text

Every primary (e.g. diary entry, letter, article) as well as every secondary textual object (e.g. prosopographies, commentaries, descriptions of works) are encoded in conformance with the current TEI P5 guidelines (additionally the encoding of metadata for musical works follows the guidelines of the Music Encoding Initative[2]). Schemata have been developed and documented for every type of text making use of TEI's meta language ODD. The overall encoding focus – besides the textual features – lies on the markup of persons, places, works and roles (e.g. in an opera or play).

These files (currently > 15.000) are stored in a subversion system and then exported to an eXist production database for online presentation and to another eXist development database for internal work.

2. Web application

The website of the WeGA which presents the edition to the user is driven by the native XML database eXist.The TEI files are the raw data that are transformed on-the-fly to XHTML for convenient reading while, for a maximum of transparency, they are still accessible in raw XML format through a dedicated tab. A third view on the text is facilitated by the digital facsimile which is presented if not restricted by copyright issues.

The extensive markup (and identification) of persons, places, works and roles in the TEI data allows for a dense linkage between these objects in the web application; additionally, for every such link a preview tooltip with the most relevant information is generated to avoid unnecessary page exits.

Further incorporated website standards are a search function and faceted filtering.

One of our main concerns is to present not only the textual data but also to provide the documentation of encoding practice and the necessary schemata, along with examples, to keep record of the making of the edition and to provide stimuli for envisaged correspondence editions, especially in the field of musicology.

3. Relations

The web application mimics the look of current social network sites and centers around persons and their social networks. These networks (the contacts of a given person) are computed through the available correspondence and are presented in a separate frame on every person's 'homepage'. (Two persons A and B are connected if there is some correspondence material in our corpus that was sent from Person A to B or vice versa.) It is a future goal to elaborate on relations of arbitrary objects and to compute and visualize their distance (e.g. by which nodes are Carl Maria von Weber and Mozart's *Zauberflöte* connected?)

4. Linked data and web services

The WeGA tries to develop tools and views not only for a 'reader' (who is close reading the texts) but also for 'clients' who want to access the (raw) data by automatable tasks. Furthermore do we try to incorporate authority files to allow for automated linkage with other repositories. An already established example of *Linked Data* is the PND beacon[3] that connects prosopographic information about a given person from different sources. The information from these sources (e.g. Wikipedia or the German Biographic Dictionaries) are then directly embedded in our web application via according tabs or simply listed as web links.

A future goal is to provide a well documented API to allow for machine interaction with our data.

Notes

1. http://www.weber-gesamtausgabe.de
2. http://www.music-encoding.org
3. http://de.wikipedia.org/wiki/Wikipedia:PND/BEACON

Data sharing, virtual collaboration, and textual analysis: Working on 'Women Writers In History'

van Dijk, Suzan
suzan.van.dijk@huygens.knaw.nl
Huygens ING - Royal Dutch Academy of Arts and Sciences, The Netherlands

Hoogenboom, Hilde
Hilde.Hoogenboom@asu.edu
Arizona State University, USA

Sanz, Amelia
amsanz@filol.ucm.es
Complutense University Madrid, Spain

Bergenmar, Jenny
jenny.bergenmar@lit.gu.se
University of Göteborg, Sweden

Olsson, Leif-Jöran
leif-joran.olsson@svenska.gu.se
University of Göteborg, Sweden

1. Introduction

The European COST Action IS 0901 ('Women Writers In History'; 2009-2013; http://www.womenwriters.nl) will present, through this poster, different aspects of its current activities. These aim at the constitution of a Europe wide (and beyond…) network of researchers, who plan to collectively rewrite European literary history: to give women their due place in transnational literary historiography before 1900, from the Medieval period up to the early 20th century.

Current ways of studying women's writing, by country or even individual author, do not do justice to women's contribution to the literary field. Recent publications show that women played important parts as authors, readers, and intermediaries, particularly on the international level: many were allowed to learn languages, and turned their expertise into the profession of translator.

This active female participation needs to be studied more broadly, without treating each woman as an exceptional case. Therefore the project has developed a digital research infrastructure, the *WomenWriters* database (http://www.databasewomenwriters.nl), in which the project members have the possibility of sharing their research data (inevitably relevant to each of them), collaborating together in the same tool, as individual researchers *and* as interconnected projects, creating means for textual and comparative analysis of texts (primary and secondary texts) written in different languages.

Our poster will present:

1. the *WomenWriters* database, which is currently being developed into a broader Research Environment (plans for development will be included),

2. an example of the types of sources we use to develop large-scale approaches that encourage the serendipitous finding of connections between women authors, works, and countries through readers

as well as two examples of the projects, which will be interconnected to the database (in its next version):

3. a project concerning one individual author: the Swedish Selma Lagerlöf, and

4. an annotating project presently developed and tested for future use in establishing connections between texts, and formulating the significant relationships existing between them.

2. Adapting the Structure to the Research Questions (Suzan van Dijk)

The *WomenWriters* database allows researchers to generate *new knowledge* about women's impact on their contemporary or early readers: we proceed by large-scale data entry about their works' reception, without focusing on particular authors. This is how serendipitous findings are provoked: by including information either about authors we did not know as yet, or about reception we would not have been aware of or looking at. The tool currently contains over 20.000 manually processed reception data, which prepare an important empirical basis for research in European women's literary history. We will gradually complement (or even: replace?) manual processing by technological solutions, and collaboration with other online projects.

One of the challenges is in the way of handling the activities of these authors starting from the *reception* side. Characteristic of the *WomenWriters* database (present form) is that it is *not* just composed of the two 'classical' entities: *Author* and *Work*. A third one has been added: the *Reader*. This structure materializes Todorov's scheme, putting in between writer and reader the imaginary world created by the one and to be 'constructed' in the other's head (Todorov 1980). From our point of view a book/text only gets historical relevance by *having been read* – this is why we added, next to the Author-Work, the Work-Reader connection.

This symmetrical structure however fails to do justice to the real *dynamics* of ongoing literary communication: *receivers* can become *senders*, even without profiling themselves as intermediaries or translators. Which leads to a view of literary communication as being a series of texts generating each other through the intervention of a *reader* turned *writer*. A new data model has been presented doing justice to this, allowing at the same time interconnectivity with related projects.

3. Data Sharing (Hilde Hoogenboom)

In this project, we use different kinds of large-scale sources that contain information about these authors' early reception, and provide the data we work on. They include bio-*bibliographic compilations* of European women writers, compiled after the example of Boccaccio's *Famous Women* (1361-75) as the databases of their time. This genre spread across Europe, to become an on-going transnational phenomenon that exists today as bio-bibliographic dictionaries and databases. Recording the traces of women and works, some of whom seem to have left no other mark in literary history, and are part of what Margaret Cohen refers to as the 'great unread' (1999), they have taken many forms over the centuries. Compilers rely on, disagree with, and often simply borrow their predecessors' work. Together with other sources, this material allows us to understand and describe the presence of women authors (as a category, or taken individually) over the centuries both using numbers and studying the tropes by which these women are represented. Some of these compilations are available online, which facilitates their use for this kind of analysis.

WomenWriters is the only transnational database of women writers that brings together the whole spectrum of sources for literary markets and literary fields of small and large nations, and can be considered as the ultimate bio-bibliographical compilation.

4. Comparative Textual Analysis (Jenny Bergenmar, Leif-Jöran Olsson)

Arguments for investing resources in digital scholarly editions are often based upon the cultural dignity of the authorship (Shillingsburg 2004). For Selma Lagerlöf (first woman to win the Nobel Prize for Literature), this is above any doubt. But scholarly editions that focus on single, canonized authorships represent the author in splendid isolation. On the other hand, for *WomenWriters* it is important to include as many data as possible on the production and reception of those authors who may have been role models for writers all over Europe – in particular as these data already exist in digital form and are interconnectible.

The Selma Lagerlöf Archive aims to create such a broader (gender) historical context that includes the intensive contacts within networks of European translators and writers that were important for Lagerlöf's success in Europe. These can in particular be traced thanks to our participation in the COST-WWIH Action, which has facilitated, and will continue to do so, the discovery of unknown connections between European women and Lagerlöf. Standardizing the different entities data will allow, by way of micro services currently developed, exchanging information on these levels.

Here, we benefit from work carried out in COST Action 'Interedition' (Joris van Zundert et al.), where the SLA also participated, and where *Juxta* and *CollateX* have been developed, to explore patterns in the reception of women writers by critics, and whether they correspond to specific themes in the literary texts. The same method may be used to collate texts by different female authors as a shortcut to discovering common topics. The advantage of this method is, again, that researchers don't have to know beforehand what to search for: unexpected similarities may appear, that researchers might not have imagined.

5. Sharing Annotations (Amelia Sanz)

Within this COST-WWIH Action, we are proceeding toward a next version of the *WomenWriters* database, to make it available for other applications and usable outside its original application in such a way that different communities of scholars can ask questions from the same sets of data. Complutense University teams are now testing *@Note*, an annotation tool capable of supporting collaborative work on texts, using women's European reception as a case study.

They are developing an innovative annotation method, which will be used in our project for specifying and formulating significant relationships between the texts we are studying. They can be annotated using a shared annotation schema; adding new terms and relationships to the schema as they are required during the annotation process; making it possible for authorized users to edit and to restructure the schema.

As a concrete example, we will present the outcome of an experience of collaborative annotation of the Spanish translation of Mme de Graffigny's famous novel *Lettres d'une péruvienne* (1747): *Cartas de*

una peruana (1792, 1823) by María Romero, both translations available in *Google Books* and *Hathi Trus*t, and included in the `http://neww.huygens.knaw.nl/receptions?fromreceptionsearch=1&sort=upper%28authors_works.name%29&page=1&searchtoggle=on&workauthor=Graffigny&worktitle=&receptionauthor=&receptiontitle=&receptionyear=&references=&per_page=20&x=17&y=24` .

References

Cohen, M. (1999). *The sentimental education of the novel*. Princeton: Princeton UP.

Shillingsburg, P. (2004). Hagiolatry, Cultural Engineering, Monument Building, and Other Functions of Scholarly Editing. In R. Mondiano, L. F. Searle, and P. Shillingsburg (eds.), *Voice, Text, Hypertext. Emerging Practices in Textual Studies*. Seattle, Washington: U of Washington P.

Todorov, T. (1980). La lecture comme construction. In *Poétique de la prose*. Paris: Seuil.

Storage Infrastructure of the Virtual Scriptorium St. Matthias

Vanscheidt, Philipp
pvanscheidt@uni-trier.de
University of Trier, Germany

Rapp, Andrea
rapp@linglit.tu-darmstadt.de
Darmstadt University of Technology, Germany

Tonne, Danah
danah.tonne@kit.edu
Karlsruhe Institute of Technology (KIT), Germany

1. DARIAH

DARIAH aims to support and enhance digitally-enabled research across the arts and humanities and builds and maintains a research infrastructure for the wider digital humanities community. The project is based on national contributions whereat this paper emerges from the German part DARIAH-DE (DARIAH-EU 2012; DARIAH-DE 2012). In order to develop adequate strategies it is necessary to orientate on existing projects in the humanities. For this reason contact was established with the 'Virtual Scriptorium St. Matthias'. Interdisciplinary collaboration like this can work both ways: genuine questions of research within humanities regarding duration and change of cultural and medial identities can get new impulses and possibilities through the representation of analog artifacts in the digital medium and projects in the humanities can learn to define their technical requirements in a more precise way to get fit and proper solutions.

2. Virtual Scriptorium St. Matthias

The 'Virtual Scriptorium St. Matthias' will be an online edition with images of more than 450 medieval codices, mostly written between the eighth and sixteenth century, and a database with information from several manuscript catalogues. Although the supply of the library was dislocated in the time of secularisation most codices remained in the Public Library of Trier and in the library of the Episcopalian Seminary Trier (Becker 1996: 101-103). The latest catalogue of the codices used for the reconstruction can be found in Becker (1996: 66-71, 105-234). The project wants to enable the user to analyse a codex from any place at any time but also

to present the codices as an ensemble of medieval writing and reading culture and as an institution of scholarship and knowledge (Embach et al. 2001: 492).

The project is coordinated by the Public Library and Archive of Trier and the Center for Digital Humanities at the University of Trier. It is supported by the German Research Foundation and realised in cooperation with many institutions all over the world that now possess manuscripts from the library and scriptorium of the Benedictine abbey St. Eucharius / St. Matthias in Trier. The digitised content will not only be shown on the project homepage but also on the portals of the TextGrid Repository and Manuscripta mediaevalia (Virtuelles Skriptorium St. Matthias 2011-2012; TextGrid Repository n.d.; Manuscripta Mediaevalia n.d.)

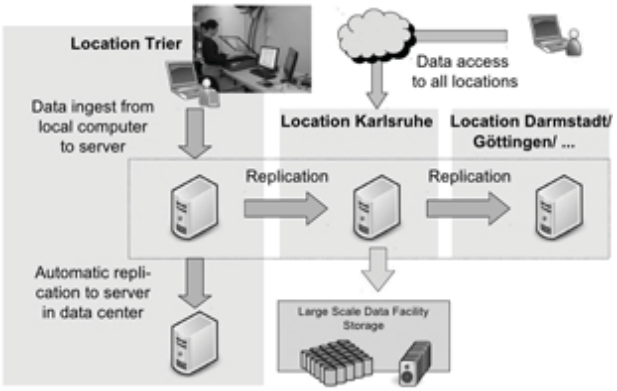


Figure 1: Proposed architecture

3. Supporting Storage Infrastructure

Requirements for a Storage Infrastructure

Analysing the current digitisation and access process in Trier several requirements can be formulated to design an architecture fitting to the 'Virtual Scriptorium St. Matthias'. Due to the distributed manuscripts data ingest and access to the storage resource should be possible worldwide. The images ingested should be replicated and their integrity should be checked regularly to ensure a reliable, long-term storage.

Proposed Architecture

To support the 'Virtual Scriptorium' an architecture with different locations is proposed as shown in figure 1. The images are produced and ingested into the system in Trier. An automatic replication places a copy in the local data center to ensure the web view currently offered. Another replication process transfers the data to Karlsruhe, where it is securely stored inside the Large Scale Data Facility (Stotzka et al. 2011). Karlsruhe is the coordinating data center and is responsible for the further replications to all remaining locations. With these procedures the actual number of copies depends on the number of attached locations but at least three copies are provided at every time. Although the ingest process described is handled locally, data can be ingested worldwide using the web. For a huge amount of data it is possible to take advantage of a local storage server and connect it to the storage and replication system. Similar procedures can be taken into account for the access as the data is either accessed via web or via a local copy of the data which was replicated from Karlsruhe to the specific location. The replication process itself and the data integrity checks have to be provided by a software for storage virtualisation which is needed because several servers have to be dealt with and which offers the opportunity to extend the system as needed by the users of the 'Virtual Scriptorium' by attaching additional servers.

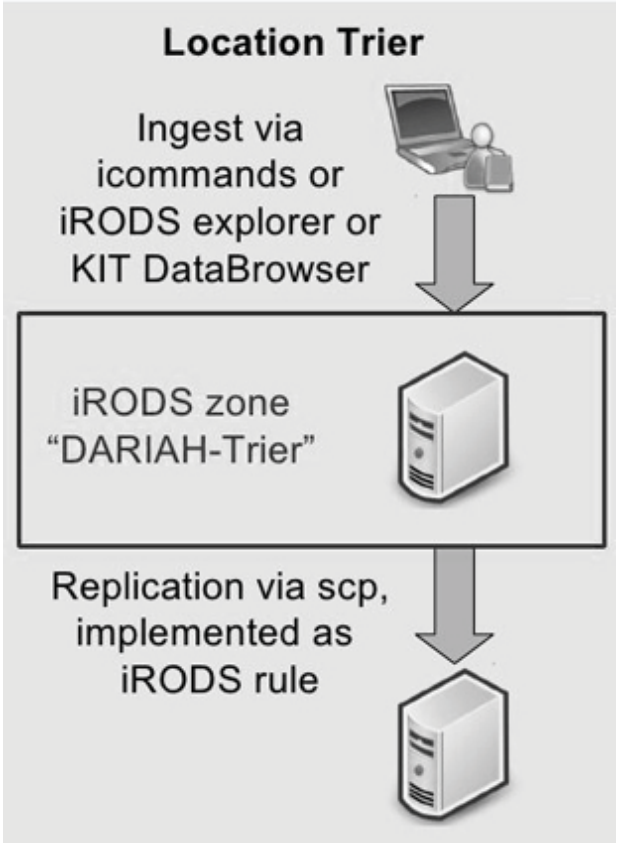


Figure 2: Implementation of the first iRods zone

Implementation

Although different software solutions are theoretically possible to realise the architecture, iRODS with its overall flexible structure seems to provide the most convenient mechanisms to deal with the humanistic research data of the 'Virtual Scriptorium St. Matthias' (iRods 2012). In this implementation two iRODS zones are used to realise a distributed, reliable storage resource.

The responsibilities for data inside a zone and for the replications needed are clearly assigned to this specific zone. Additionally the overall performance is improved by using two data bases.

The data is ingested into the 'DARIAH-Trier' zone (figure 2) using the iRODS Explorer or the command line tool iCommands with the iRODS specific transfer protocol. The replication to the local data center in Trier resp. to Karlsruhe, which are located outside 'DARIAH-Trier', is realised with iRODS rules which are triggered by a data ingest and then activate shell scripts to replicate via *scp* resp. iCommands.

The replication inside the second zone 'DARIAH-MUSE' (figure 3) is also realised with iRODS rules, which copy new data to every server of the zone and synchronises the servers attached. Mean transfer rates of 9 MB/s (30 MB/s, 11 MB/s) could be achieved in the first transfers between Karlsruhe and Trier (Trier and local data center Trier, Trier and Karlsruhe). Access to both zones is possible using various iRODS clients. Additionally to the mechanisms described bit preservation services are executed. An MD5 checksum is computed while uploading a file, stored inside the iRODS system and checked while downloading a file. In Karlsruhe the checksum is recomputed once a month and a corrupted file is replaced with a valid copy if the values mismatch.

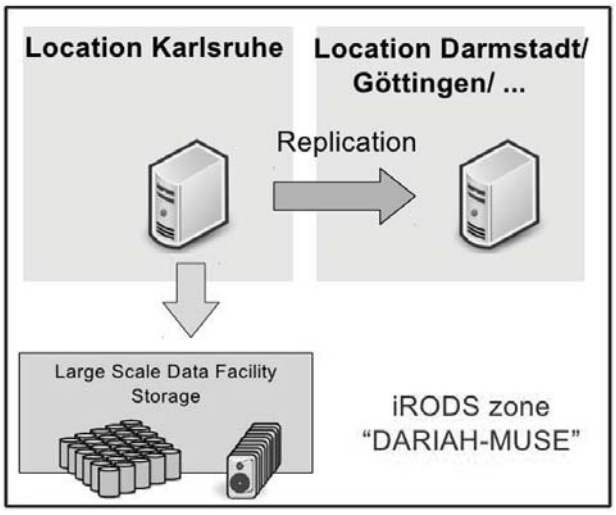


Figure 3: Implementation of the second iRODS zone

4. Discussion

An initial implementation was realised with the locations Trier and Karlsruhe and will be extended in the near future. Up to now several codices were successfully ingested and replicated. Because of the humanistic context of this implementation several challenges had to be dealt with. Most of the features are highly automated to be as easy to use as possible for the humanistic researcher. The 'Virtual Scriptorium St. Matthias' provides a large amount of files which have to be stored. This issue is faced by using a scalable data management system and by building a distributed system to increase efficiency and reliability. As humanistic researchers rather think in centuries than in years, the system provided has to be longterm available and sustainable. By integration of the implementation to the DARIAH infrastructure this demand is approached but a comprehensive solution is still under research.

Additionally, the architecture has a generic design to be able to provide a storage resource for other humanistic projects as well. This feature is necessary for the infrastructure the DARIAH project aims to build. The implementation described is therefore a basic but nevertheless fundamental component as it ensures preservation on bit level and is a basis for further on-going developments.

Fundings

This work has been supported by DARIAH-DE which is partially funded by the German Federal Ministry of Education and Research (BMBF) under the D-Grid initiative by agreement 01UG1110A-M.

The work has also been supported by the KIT startup budget for the 'Build-up of an Experimental Research Data Repository (e-Repos)'.

References

Becker, P. (1996). *Die Benediktinerabtei St. Eucharius – St. Matthias vor Trier*. Berlin, New York: de Gruyter.

DARIAH-DE (2012). Available from: `http://www.de.dariah.eu` (Accessed 16 March 2012).

DARIAH-EU (2012). Available from: `http://www.dariah.eu` (Accessed 16 March 2012).

Embach, M., C. Moulin, and A. Rapp (2011). Die mittelalterliche Bibliothek als digitaler Wissensraum. Zur virtuellen Rekonstruktion der Abteibibliothek von Trier-St. Matthias. In R. Plate, and M. Schubert (eds,), *Mittelhochdeutsch. Beiträge zur Überlieferung, Sprache und Literatur. Festschrift für Kurt Gärtner zum 75. Geburtstag*. Berlin, Boston: de Gruyter, pp. 486-497.

iRods (2012). Available from: `http://https://www.irods.org/` (Accessed 16 March 2012).

Manuscripta mediaevalia (n.d.) Available from: `http://www.manuscripta-mediaevalia.de/` . (Accessed 16 March 2012).

Stotzka, R., V. Hartmann, T. Jejkal, M. Sutter, J. van Wezel, M. Hardt, A. Garcia, R. Kupsch, and S. Bourov (2011). Perspective

of the Large Scale Data Facility (LSDF) Supporting Nuclear Fusion Applications. *Proceedings of the 19th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP)*, Ayia Napa: IEEE Press, pp. 373-379.

TextGrid Repository (n.d.) Available from: http://www.textgridrep.de/ . (Accessed 16 March 2012).

Virtuelles Skriptorium St. Matthias 2011-2012**.** Available from: http://www.stmatthias.uni-trier.de (Accessed 16 March 2012).

Digital Emblematics – Enabling Humanities Research of a Popular Early Modern Genre

Wade, Mara R.
mwade@illinois.edu
Society for Emblem Studies/University of Illinois, USA

Stäcker, Thomas
staecker@hab.de
Herzog August Bibliothek, Wolfenbüttel, Germany

Stein, Regine
r.stein@fotomarburg.de
Bildarchiv Foto Marburg, Germany

Brandhorst, Hans
jpjbrand@xs4all.nl
arkyves, The Netherlands

Graham, David
david.graham@concordia.ca
Concordia University, Montreal, Canada

Emblematica Online offers a rewarding site for exploration of the seismic shifts in digital humanities from 1.0 to 2.0,[1] while providing perspectives on how Emblems 3.0 can also contribute to Digital Humanities 3.0. Emblem studies are large enough to offer a significant corpus for study, yet small enough to present a finite overview. As a historical field of endeavor – emblems are a Renaissance text-image genre that informed all aspects of early modern literary and material culture – it offers a unique perspective on key questions of digital humanities at 2.0.

The international scholars who cooperate as the OpenEmblem research group have created a vibrant model for humanistic collaborative research, culminating in the international project 'Emblematica Online,'[2] including the OpenEmblem portal for emblem studies.[3] The digital emblem projects at Glasgow,[4] Utrecht,[5] Urbana (Illinois),[6] Wolfenbüttel,[7] Munich,[8] La Coruña (Spain),[9] and elsewhere serve the needs of an international community of scholars. The long-term success of working with a consortium has ensured that their research is not preserved in a digital silo, but accessible to a greater range of future projects.

Emblematica Online and its OpenEmblem Portal allows for new levels of research surpassing

that which was previously available in emblem scholarship:

1. Quantity – access to more than 700 fully digitized rare works
2. Quantity – access to more than 70,000 individual emblems
3. Quality – extremely rich metadata for approximately 15,000 emblems
4. federated searching across multiple projects at geographically widespread locations at the book level
5. federated searching across multiple projects at geographically widespread locations at the level of the individual emblem
6. emblem motto database
7. Iconclass notations with associated hierarchies
8. Iconclass notation with multilingual thesaurus
9. An emblem exchange format, or emblem schema, taking advantage of namespaces such as TEI and SKOS
10. High quality digitization suitable for rigorous scholarly research needs
11. Multiple access routes to the book and to the emblem
12. Sustainable links at the sub-object level
13. Multiple levels of granularity
14. Worldwide unique emblem identifiers at emblem level
15. Linking book identifiers and emblem identifiers
16. a prototype emblem registry providing persistent identifiers by means of a handle service

The OpenEmblem portal has moved from generic digitization to customized, sophisticated digitization of a complex Renaissance genre consistent with high scholarly demands. The design and scope of Emblems 2.0 makes these Renaissance resources more interesting to a wide range of scholars who study the Renaissance from diverse and divergent perspectives.

While the current researchers for Emblematica Online have gained experience in creating standard formats for automated metadata exchange,[10] digital philology,[11] and portal creation,[12] through the use of authorities and controlled vocabularies of their metadata, they have indicated relationships among resources, making this work extensible in the future. The Iconclass notations and labels reflect a hierarchy, while the OpenEmblem Portal simultaneously anchors the very rich metadata to the original repositories. We can search at different levels of granularity across integrated corpora and visualize and present whole and part relationships in a meaningful way.

The panelists, all of whom have committed to the DH 2012 conference in Hamburg, will present the following aspects of humanities based research in the digital medium:

Mara R. Wade traces the culmination of the research through Emblems 2.0 and introduces as its key feature unique emblem identifiers. In addition to helping us present the whole-part relationship, the unique identifier reflects the forward vision of the emblem community to anticipate a time when researchers can include annotation functions such as 'identical with' or 'similar to' in textual and pictorial research. It creates an authoritative basis for expanding the portal for book emblems to the material culture of early modern Europe, enabling the future study of art forms where emblems feature prominently. The unique identifier is a significant scholarly step in emblem scholarship and shows how community driven research serves larger scholarly communities as well.

Thomas Stäcker demonstrates how the project sets digital emblematics to work, focusing on practical issues of sharing, and the distribution and mining of emblem metadata and data. Future digital emblem research, in particular, and digital humanities, in general, must achieve more integrated systems grouping resources together, making them available through uniformly designed graphical interfaces and allowing searches on standardized fields and vocabulary. Emblematica Online established a common format for indexing digital emblem material. Staecker emphasizes 1) the development of a particular XML-schema with a separate emblem namespace based on standards such as TEI, METS MODS, and SKOS; 2) how data originating from different international projects are shared and distributed, e.g. via OAI; and 3) how emblems and theirs parts may be reliably identified and quoted by persistent identifier by means of a central handle service at the University of Illinois.

Two papers focus on image indexing and its broad potential for the digital humanities.

Regine Stein demonstrates how the issuing of a unique identifier at the level of the individual emblem opens emblem studies up for cross-domain research and discusses the Linked Data approach. She argues that the CIDOC Conceptual Reference Model (ISO 21127, CIDOC-CRM) provides the 'semantic glue' necessary to mediate between different sources of cultural heritage information and thus provides the mechanism to open emblem metadata to cross-domain research from both the text and the image

research communities, including material culture, such as prints, paintings, illuminated manuscripts, and even architectural ornament. Based on the emblem schema, it is worth exploring a Linked Data implementation using the CIDOC-CRM. Because we aim also for maximum compliance to widely used authorities and controlled vocabularies such as Iconclass, Emblematica Online can offer a best practice case in terms of linking different sources instead of building data silos.

Hans Brandhorst presents the advantages of Iconclass as a shared vocabulary tool demonstrating that it has a vital role to play in making visual information accessible and retrievable. 'Meaning' is not an intrinsic quality of a picture, and cannot be detected with image recognition techniques (alone). Non-trivial subject retrieval of visual sources is impossible without rich textual metadata. A classification as a tool for the production of metadata can only survive in the distributed environment of the internet if it functions as a flexible, central webservice, and if the user community itself can correct, edit and expand the vocabulary. He discusses Iconclass 1) as a data production tool; 2) a shared vocabulary tool, open to community editing; 3) as an information retrieval tool; and 4) as Linked Open Data in SKOS/RDF and JSON representations. He ends with a discussion of *arkyves* as a mixed business model to sustain Iconclass development.

David Graham outlines Emblems 3.0. It can be argued that our earliest efforts were primarily aimed at gaining individual access to digital data that minimally replicates the books comprising our corpus, and in particular at creating ways to digitize, store, and display visual content. The second phase was about making collaborative access possible by ensuring interoperability and independence from particular vendors and platforms. It now seems clear that the third phase will have as its primary focus the goal 1) of massively, pervasively, and permanently *interconnecting* huge amounts of textual and visual data and metadata in multiple forms, contexts, and purposes, 2) of *integrating* materials from a wide diversity of sources, and 3) of enabling *collaboration and interaction* – both among scholars and between the scholarly and lay communities – on an unprecedented scale.

In the spirit of the emblem itself with its visual and textual components, Emblem 3.0 will be interactive and collaborative; it will extend to emblems in material culture. Emblem 3.0 will open up heretofore unanticipated qualitative research questions based on well curated and designed quantitative digital resources.

Notes

1. The Digital Humanities Manifesto (2009) maintains 'Digital Humanities is not a unified field but an array of convergent practices that explore a universe in which: a) print is no longer the exclusive or the normative medium in which knowledge is produced and/or disseminated; instead, print finds itself absorbed into new, multimedia configurations; and b) digital tools, techniques, and media have altered the production and dissemination of knowledge in the arts, human and social sciences.' http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20/
2. This is a bilaterally funded initiative of the University of Illinois and the Herzog August Bibliothek (HAB), Wolfenbüttel. http://emblematica.grainger.illinois.edu
3. http://www.germanic.illinois.edu/news/emblem
4. http://www.emblems.arts.gla.ac.uk
5. http://emblems.let.uu.nl/index.html
6. http://emblematica.grainger.illinois.edu/
7. http://www.hab.de/forschung/projekte/emblematica.htm
8. http://mdz1.bib-bvb.de/~emblem/
9. http://rosalia.dc.fi.udc.es/emblematica/
10. http://www.hab.de/bibliothek/wdb/emblematica/regelwerk.htm
11. See the link to the Spanish project (fn 9) for the link to DEBOW, 'Digitized Emblem Books on the Web,' a regularly updated list of digital emblem books.
12. The Portal will be launched in April 2012. Presently book level data can be searched here: http://emblematica.grainger.illinois.edu/Browse/Books/DigitizedBooksByTitle

DTAQ – Quality Assurance in a Large Corpus of Historical Texts

Wiegand, Frank
wiegand@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Germany

1. Project

The DFG-funded project *Deutsches Textarchiv* (German Text Archive, DTA[1]) started in 2007 and is located at the *Berlin-Brandenburgische Akademie der Wissenschaften* (Berlin-Brandenburg Academy of Sciences and Humanities, BBAW[2]). Its goal is to digitize a large cross-section of German texts from 1650 to 1900. The DTA presents almost exclusively the first editions of the respective works. Currently there are more than 700 texts available (i. e. 500 million characters), most of them transcribed by non-native speakers using the double keying method.

The DTA provides linguistic applications for its corpus, i. e. serialization of tokens, lemmatization, lemma-based and phonetic search, and rewrite rules for historic spelling variation.

Each text in the DTA is encoded using the XML/TEI-P5 format. The markup describes text structures (headlines, paragraphs, speakers, poem lines, index items, etc.), as well as the physical layout of the text down to the position of each character on a page.

2. Problem Statement

Even though our corpus of historic text exhibits very good quality, many errors still occur in the transcription, in the markup, or even on the level of presentation. Due to the heterogeneity of the corpus, e. g. in terms of text genres (novels, prose, scientific essays, linguistic reference works, cookbooks, etc.) there is a strong demand for a collaborative, easy to use quality assurance environment.

As of October 2011, the corpus consists of more than 260,000 pages (half a billion characters), several gigabytes of XML. Even though our digitization providers assure an accuracy rate of 99.95 %, many errors remain undetected, not to mention problems in the presentation layer of the DTA or workflow mistakes.

There are many kinds of possible errors in our transcribed texts: transcription errors (e. g. due to illegible text or text written in foreign scripts like hebrew, greek, runic, etc.) sometimes require specialized background knowledge, so we created various assorted tools to aid users in finding potentially problematic spots in our texts, and to help transcribers to obtain better and faster results.

In addition, DTA provides an interface DTAE[3] for the integration of external text transcriptions along with the corresponding images and metadata. These transcriptions should be encoded in XML/TEI using the DTA 'base format'[4]

Quality assurance (QA) also has to take into account other levels of error prone representations and tasks, namely metadata, XML/TEI annotation, HTML presentation (and other media), and the robustness of workflow. DTAQ is our QA system dealing with all these potential errors: They need to be reported, stored and fixed.

3. DTA Quality Assurance

DTAQ[5] is a browser-based tool to find, categorize, and correct various kinds of errors. Using a simple authentication system combined with a fine-grained access control system, new users can easily be added to our QA system. The GUI of our tool is highly customizable, so we can offer diverse views of our source images, transcriptions, and presentations.

Figure 1: DTAQ: parallel view of image and rendered transcription

Our linguistic tools (*CAB,* [6] PoS-tagging) are integrated into this environment, not only to check their performance for errors, but also to provide alternate views of our texts. To avoid unnecessary repetitions in proofreading, users can mark pages as proofread. Using this technique, we are also able to provide several quality levels of our pages or books. All tickets and proofread pages are stored in a database, thus DTAQ provides in-depth analysis and visualisation of the accuracy of the DTA corpus.

The backend of DTAQ is built upon many open source packages, using Perl as a glue language. The system runs on *Catalyst* [7], connects to a *PostgreSQL* database via the *DBIx::Class* ORM and builds its web pages with *Template Toolkit*. The frontend makes heavy use of *jQuery* and *Highcharts JS* to create a very interactive and responsive user interface.

Our XML/TEI files are automatically split up into individual pages and stored in a *git* [8] repository. The development of DTAQ itself also occurs within a distributed *git* repository.

4. Poster and tool demonstration

Our poster will show the DTAQ workflow patterns, along with a live demonstration showing the various views, tools, and powerful features of the quality assurance platform.

References

Geyken, A., et al. (2011). Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In S. Schomburg, C. Leggewie, H. Lobin, and C. Puschmann (eds.), *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 20./21. September 2010. Beiträge der Tagung*. 2., ergänzte Fassung. Köln: HBZ, pp. 157-161.

Jurish, B. (2010). More than Words: Using Token Context to Improve Canonicalization of Historical German. *Journal for Language Technology and Computational Linguistics (JLCL)* 25(1).

Jurish, B. (2012). *Finite-state Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam 2012 (urn:nbn:de:kobv:517-opus-55789).

Unsworth, J. (2011). Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI. *Journal of the Text Encoding Initiative* 1 (`http://jtei.revues.org/215` , 2011-08-29).

Notes

1. Deutsches Textarchiv (German Text Archive): `http://www.deutschestextarchiv.de/` .
2. Berlin-Brandenburgische Akademie der Wissenschaften: `http://www.bbaw.de/` .
3. DTAE (DTA Extensions): `http://www.deutschestextarchiv.de/dtae` .
4. DTAB (DTA "base format"): `http://www.deutschestextarchiv.de/doku/basisformat` .
5. DTAQ (DTA Quality Assurance): `http://www.deutschestextarchiv.de/dtaq` .
6. For each text, DTAQ provides a transformation to a normalised modern spelling form using the Cascaded Analysis Broker CAB, cf. Jurish:2010 + Jurish:2012.
7. Catalyst MVC Framework, written in Perl: `http://www.catalystframework.org/` .
8. git, a distributed version control system written by Linus Torvalds et al.: `http://git-scm.com/`

The Digital Averroes Research Environment – Semantic Relations in the Editorial Sciences

Willems, Florian
f.willems@uni-koeln.de
University of Cologne, Germany

Gärtner, Mattias
gaertner@informatik.uni-koeln.de
University of Cologne, Germany

The Digital Averroes Research Environment aims to provide structured access to the complete oevre of Abū l-Walīd Muhammad Ibn Ahmad Ibn Rušd, better known under his latin name Averroes, including scans of thousands of manuscripts, chunked full texts as interim editions of all his works and a comprehensive bibliography of primary and secondary sources. The primary data format is XML as TEI P5, but the complex tradition of Averroes's body of works necessitates a higher level of abstraction than TEI is able to provide. Consequently, this high level abstraction should pave the way for crosswalks to lower level formats, as FRBRoo and such like will become more and more common.

Averroes himself wrote mostly in Arabic, but nearly all of his works were translated into Hebrew and Latin. As was the case with many medieval manusccripts, scribes often edited the texts they were copying, and, in the case of Averroes, numerous translations were crafted. Sometimes texts were translated from the Arabic original into for example Hebrew, and then, after generations translated back into Arabic. This creates a very inconsistent but very challenging tradition of the works of Averroes. Many of Averroes's texts also have quite a few witnesses, often numbering in the hundreds, where the TEI doesn't provide a satisfactory mean to manage that many manuscripts. Semantic interconnection of XML/TEI data to knowledge bases encoded in semantic web technologies may be a solution to that problem.

To this end, the connections between manuscripts, translations, abstract works and traditions have to be modeled as semantic data. Compatibility considerations led to an abstraction of the Europeana Data Model EDM, done in OWL-DL. A possible implementation of the DARE semantic model in CIDOC-CRM is also in the making with the integration of WissKI into our environment. Much work and thought went into the user interfaces, supported by experiments with different types of front ends and different types of users – how should a philologist be able to interface with the semantic data providing the glue between the documents? How should inference engines be used on the data? How will an editor use the manuscripts?

Performance is also in issue in highly interconnected data like that. We will show our backend solution for storage and maintenance of semantic data as well as the implementation of the inference engines. A completely new XML back end, XELETOR, was programmed to serve the sheer amount of XML data in reasonable time. User management and the history of a given triplet is also an issue.

Another challenge is posed by the inherent writing direction of Arabic and Hebrew script. Inputting XML in right-to-left script is tricky, but what about RDF triplets? The on-screen input methods have to be carefully chosen to facilitate bidirectional input. The problem of displaying said information is solved by the DARE, because intelligent usage of CSS- and XSLT properties provides easy display of bidirectional texts.

By using a given standard and giving much thought to interfaces and usability, DARE will have one of the first semantic representations of the complex tradition of one of medieval philosophy's foremost commentator.

These central questions concerning especially the usability of the interface on the semantic data shall be presented as a demo. Two computers will allow interested parties to try DARE and its semantic technologies out.

All our technology is available as free and open source code.

Funding

This work is supported by the Deutsche Forschungsgesellschaft.

References

Hissette, R. (2010). *Averroes Latinus. Commentum Medium Super Librum Praedicamentorum Aristotelis*. Translation Wilhelmo De Luna Adscripta, ed. by R. Hissette, Arabic-Latin apparatus with notes by A. Bertolacci, glossaries by Hissette & Bertolacci and with † L.J. Bataillon (Averrois Opera. Series B. Averroes latinus 11). Leuven 2010.

Wirmer, D. (2008). *Averroes, Über den Intellekt*. Auszüge aus seinen drei Kommentaren zu Aristoteles' De anima. Arabisch – Lateinisch – Deutsch, herausgegeben, übersetzt, eingeleitet

und mit Anmerkungen versehen von David Wirmer (Herders Bibliothek der Philosophie des Mittelalters 15). Freiburg 2008.

http://dare.uni-koeln.de/

http://wiss-ki.eu/

http://developer.berlios.de/projects/xeletor/

AV Processing in eHumanities – a paradigm shift

Wittenburg, Peter
peter.wittenburg@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Lenkiewicz, Przemyslaw
przemek.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Auer, Erik
erik.auer@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Lenkiewicz, Anna
anna.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Gebre, Binyam Gebrekidan
binyamgebrekidan.gebre@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Drude, Sebastian
sebastian.drude@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

1. Introduction

Speech research saw a dramatic change in paradigm in the 90-ies. While earlier the discussion was dominated by a phoneticians' approach who knew about phenomena in the speech signal, the situation completely changed after stochastic machinery such as Hidden Markov Models [1] and Artificial Neural Networks [2] had been introduced. Speech processing was now dominated by a purely mathematic approach that basically ignored all existing knowledge about the speech production process and the perception mechanisms. The key was now to construct a large enough training set that would allow identifying the many free parameters of such stochastic engines. In case that the training set is representative and the annotations of the training sets are widely 'correct' we could assume to get a satisfyingly functioning recognizer. While the success of knowledge-based systems such as Hearsay II [3] was limited, the statistically based approach led

to great improvements in recognition rates and to industrial applications.

However, most humanities and social science research does not deal with proper signals that allow to apply neither the purely statistical nor the rule-based approach. Speech is spoken in natural situations embedded in noise, it is widely spontaneous, often one has to deal with much variation for which we lack advanced models and the existing corpora are small. Thus there is no chance to apply holistic speech recognizers that take a speech signal and would produce a useful transcription.

The situation for moving image processing (video) is even worse, since we do not have an accepted target – researchers target annotations are mostly semantic functions that are associated with gestures, mimics and other body motions, which are very much dependent on cultures, situations and other parameters. Only for sign languages we can consider a situation comparable to oral speech, since we can correlate between a stream of video observations and a target transcription. However, sign languages use various information channels, which are easy to be comprehended by the human eye, but difficult to process by machinery.

This situation is not satisfying since we see that the gap between the amount of material that has been recorded by researchers and the amount of material that is available for research purposes gets larger since the available time for creating annotations did not change substantially despite new and more efficient annotation software such as ELAN [4]. Thus in psycholinguistics and in many other humanities disciplines dealing with natural scenes a new 'paradigm shift' was required.

2. Interactive Recognition Paradigm

This new paradigm is based on four equally important pillars:

- training many statistic recognizers on small phenomena and improve robustness, i.e. widely reducing the complexity of the phenomena to be recognized
- including interactive learning methods
- providing a highly efficient usability framework that brings researchers back into an active role
- make the existing and partly complex audio/video recognition technology available to the researchers

For data streams in the humanities we cannot assume that there is one recognizer that does it all. What researchers want is to have the possibility to train a recognizer to detect a specific hand movement or a specific intonation contour for example. These recognizers will then create probabilistic annotations at a specific tier. All these detectors are adding annotations ending in a complex lattice. The type of approach to realize a recognizer depends very much on the type of pattern to be detected. A cascaded recognizer, a recognizer that makes use of annotations of earlier ones, could be rule-based or based on statistics.

Interactive learning methods need to be implemented to allow the researcher to quickly improve the representation of a specific pattern including its variation. First, a single sample might be sufficient. Supervised learning techniques might help to improve the annotation accuracy to acceptable values after a few iterations.

Figure 1: Three layers of annotation representing three steps of the automated analysis. First layer is the result of the uniform document segmentation, second layer is the division into speech/no-speech parts, third layer recognizes different speakers in the recording. All the steps have been performed automatically

All recognizers will create erroneous annotations, i.e. we need a framework that allows users to quickly scan annotation patterns and correct them. Here we can build on the ELAN annotation tool and TROVA search tool that have already been optimized over the years.

One of the basic assumptions of such an approach is that there are many recognizers available to the researcher. Although there exists a lot of partly complex technology in the specialist labs, it is not accessible yet. A change of culture is required to make such technology accessible, as is currently being worked out by CLARIN[1]. A standardized mechanism for invocation is required to allow starting such recognizers and to interact with them. The positive WebLicht[2] experience motivates us to continue extending the Service Oriented Architecture to audio/video services.

This new paradigm is shifting complexity to the annotation lattice. From many discussions with researchers we can expect that the method can work, since researchers are mostly not interested in 'complete' annotations, but they are interested in certain specific phenomena. For these kind of

selected annotations the interactive paradigm seems to be appropriate. In addition, we have seen in first experiments that the researchers will become engaged again, since they are not confronted with a trained machine where they do not know what the values stored in the many parameters mean. Now they can use the patterns to be looked at and the annotation created as part of their theorization process.

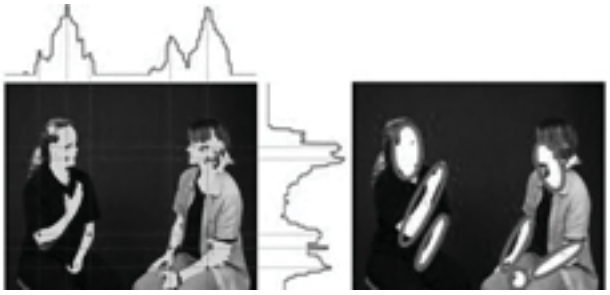


Figure 2: Left – a video frame with skin color pixels marked and histograms of those in two dimensions, X and Y. Right – example of ellipses approximating the skin areas in given image

3. The AVATecH Approach

The AVATecH project [5], started in 2009 as a joint work between MPI, IAIS[3] and HHI[4] experts, is working along the Interactive Recognition Paradigm. A number of audio/video recognizers have been implemented and integrated, the usability framework ELAN has been improved and a method for remote invocation has been established that can be extended to a Web Services scenario.

Audio Detectors

One of the recognizers provides a fine-granular segmentation of the audio stream into homogeneous segments allowing the user to control the granularity of segmentation. Another recognizer is able to label audio segments containing human speech, regardless of the language of the recording.

A language-independent speaker clustering recognizer is able to find segments spoken by the same person within a given recording (Figure 1).

A pitch contour detector can allow researchers to graphically specify pitch contours and search for similar patterns. The detector can tag segments in audio recordings and annotate with pitch and intensity properties such as for example minimum, maximum, initial or final fo frequency, or volume.

Video Detectors

A shot and sub-shot recognizer is able to detect shots of similar video content and label them. All further algorithms rely on the results of this shot/cut detection.

Accurate motion analysis allows distinguishing between different types of video content and it can be used to segment a video in order to select only the parts, which are relevant for the researchers. For each frame in the video a motion vector map is computed using the Hybrid Recursive Matching (HRM) algorithm [6].

A skin-color detection algorithm [7] can be used to identify seed points where the hands and heads regions most likely occur. The resulting regions, where heads and hands are identified, are approximated by an ellipse for each video frame (Figure 2). After the hands and head have been labeled, the recognizer can detect strokes, gestures and relation between hands and head (Figure 3).

Figure 3: Results of the Hands and Head Tracking recognizer. On the video file the positions of hands and head are marked for every frame of the video. The annotations created are fully automated and include the time ranges in which left and right hand movement occurs, when the hands join and when there is an overlap of hand and face

User interaction

In close collaboration with the experimenting researchers the ELAN tool has been adapted to become a powerful framework to invoke the various detectors via an API described in XML and to search for annotation patterns and to manipulate them.

4. Conclusions

Automatic audio and video processing of natural scenes is a tough task and the project team from MPI, IAIS and HHI worked hard on the 4 pillars mentioned beforehand. A first number of about 10 recognizers can be used; the ELAN tool has been extended to a comfortable research environment and a mechanism supporting the required complex interaction between the ELAN (and finally the user) and the recognizers has been developed. The first experiments with researchers working on real scenes

has shown efficiency gains of about 70% only for the segmentation case which is very promising for speeding up the annotation work. In the realm of CLARIN we will collaborate with more of the technology providers in the specialist labs to tune their algorithms so that they can be integrated in this new interactive paradigm.

References

[1] **Rabiner, L., and B. Juang** (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE* 3(1): 4-16.

[2] **Hopfield, J. J.** (1988). Artificial neural networks. *Circuits and Devices Magazine, IEEE* 4(5): 3-10.

[3] **Erman, L. D., et al.** (1980). The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *ACM Comput. Surv.* 12(2): 213-253.

[4] **Wittenburg, P., et al.** (2006). Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference (LREC* 2006.)

[5] **Auer, E., P. Wittenburg, H. Sloetjes, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel** (2010). Automatic annotation of media field recordings. In *ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 2010: University of Lisbon, Portugal, pp. 31-34.

[6] **Atzpadin, N., P. Kauff, and O. Schreer** (2004). Stereo analysis by hybrid recursive matching for real-time immersive video conferencing. *Circuits and Systems for Video Technology, IEEE Transactions* 14(3): 321-334.

[7] **Terrillon, J. C., et al.** (2000). Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference.*

Notes

1. www.clarin.eu (`www.clarin.eu`)
2. weblicht.sfs.uni-tuebingen.de (`weblicht.sfs.uni-tuebingen.de`)
3. www.iais.fraunhofer.de (`www.iais.fraunhofer.de`)
4. www.hhi.fraunhofer.de (`www.hhi.fraunhofer.de`)

Index of Authors

Achler, Stefan 514
Agugiaro, Giorgio 365
Akdag Salah, Alkim Almila 79
Albrezzi, Francesca 342
Alessio, Piccioli 221
Alexander, Marc 82, 427
Almas, Bridget 105
Alonso Garcia, Nuria 84
Alvarado, Rafael 448
Andresen, Jens 22
Andrews, Tara Lee 85
Antonio, Lamarra 221
Arazy, Ofer 88
Armaselu, Florentina 91
Armstrong, Helen 148
Arnold, Kerstin 70
Arnold, Matthias 429
Auer, Erik 538
Auvil, Loretta 158, 397
Baca, Murtha 342
Baechler, Micheal 94
Bärenfänger, Maja 97
Barker, Elton 99
Barner-Rasmussen, Michael 102
Batjargal, Biligsaikhan 473
Bauman, Syd 33, 52
Bański, Piotr 52
Beaulieu, Marie-Claire 105
Beavan, David 107
Beer, Ruth 109
Beißwenger, Michael 259
Bellamy, Craig 111
Bellot, Patrice 467
Bennett, Bradford 413
Benzon, Bill 268
Beretta, Francesco 431
Bergenmar, Jenny 113, 527
Berry, David M. 151
Bestmann, Oliver 288
Binder, Frank 97
Bingenheimer, Marcus 115, 230
Blake, Catherine 273
Blandford, Ann 35
Blanke, Tobias 117
Blustein, James 463
Bod, Rens 57
Bode, Katherine 119
Bodenhamer, David 41
Boggs, Jeremy 299, 433
Bohl, Benjamin W. 435
Bol, Peter Kees 43
Boot, Peter 121
Bradley, Adam James 123
Bradley, John 124
Brandhorst, Hans 532
Brey, Gerhard † 491
Brock, Anne 126
Brockmann, Christian 15
Broeder, Daan 169
Brown, Susan 35, 324, 508
Brown, Travis 64
Brughmans, Tom 129
Brüning, Gerrit 131
Brunner, Annelen 135
Bryant, Michael 117
Büchler, Marco 137
Bunke, Horst 94
Burch, Thomas 375
Burnard, Lou 52
Burrows, Toby Nicolas 139
Buzzetti, Dino 142
Capitanu, Boris 158, 397
Caplan, Alison 84
Carusi, Annamaria 221
Chan, Peter 208
Chartrand, James 508
Chaturvedi, Manish 148
Cheesman, Tom 151
Chen, Howard 154
Choi, Woong 480
Chong, Dazhi 155
Chun, Su 88
Clark, Ashley M. 438
Clavaud, Florence 72
Clement, Tanya 72, 158

Cocciolo, Anthony 322
Conner, Eric Sean 520
Coppage, Samuel 155
Coulter, Kimberly 206
Cowan, Will 465
Coyle, Barry 516
Cristina, Marras 221
Crombez, Thomas 249
Crompton, Constance 441
Croxall, Brian 443
Cruz, Alejandro 456
Crymble, Adam H. 162
Cummings, James 52
Czmiel, Alexander 445
Dacos, Marin 467
Daelemans, Walter 249
Darányi, Sándor 163
Day, Shawn 22
De Wilde, Max 28
Decker, Eric 429
Declerck, Thierry 163, 470
Della Costa, Dave 190
Dershowitz, Nachum 264
Dobrin, Lise M. 167
Dobson, Teresa 35
Douglass, Jeremy 79
Drude, Sebastian 169, 538
Duke-Williams, Oliver William 173
Dunn, Stuart 176
Earhart, Amy 179
Eckart, Thomas 137
Eder, Maciej 16, 181
Effinger, Maria 506
Ehrmann, Alison 151
Eide, Øyvind 18, 185
Ell, Basil 359
Ell, Paul 41
Elmaz, Orhan 188
Ermakova, Maria 259
Ermolaev, Natalia 392
Faath, Elodie 467
Faisal, Sarah 35
Fendt, Kurt E. 190
Fiorentino, Carlos 35
Fiormonte, Domenico 72
Fischer, Andreas 94
Flanagan, Kevin 151
Flanders, Julia 448
Fong, Grace 43
Förtsch, Reinhard 506
Fraistat, Neil 331
Franchi, Stefano 192
Freedman, Richard 506
Freire, Nuno 450
Frizzera, Luciano 35, 324
Fujimoto, Yu 453
Físseni, Bernhard 57
Gaiffe, Bertrand 52
Galina, Isabel 456
Gan, Yu 375
Gannod, Gerald 148
Gärtner, Mattias 537
Gasperlin, Oliver 497
Gayoso-Cabada, Joaquin 195
Gebre, Binyam Gebrekidan 458, 538
Gedzelman, Séverine 198
Geissler, Stefan 518
Geyken, Alexander 54, 259
Giacometti, Alejandro 35, 88
Gibson, Robin 463
Gil, Alexander 433, 462
Girardi, Gabrio 365
Gius, Evelyn 24
Gloning, Thomas 54
Glorieux, Frédéric 202
Goel, Ankita 158
Goicoechea-de-Jorge, Maria 195
Good, Jeff 280
Gooding, Paul Matthew 204
Graf von Hardenberg, Wilko 206
Graff, Ann-Barbara 463
Graham, David 532
Graham, Wayne 299
Grassi, Marco 303
Gray, Steven 348
Gregory, Ian 41
Gu, Xiangyi 155
Hachimura, Kozaburo 480

Hallam, Julia 41
Hamlin, Scott 448
Hammarström, Harald 296
Hangal, Sudheendra 208
Harris, Trevor 41
Hedges, Mark 176
Heer, Jeffrey 208
Heller, Brooke 35
Henrich, Verena 126
Henzel, Katrin 131
Herzog, Rainer 288
Hesemeier, Susan 508
Hettel, Jacqueline 308
Heydel, Magda 212
Hinrichs, Erhard 20, 126
Hirvonen, Ville 215
Hodgson, Eric 148
Holloway, Steven W. 438
Hoogenboom, Hilde 527
Hooper, Wallace Edd 465
Hoover, David 33, 218
Hope, Jonathan 151
Hours, Bernard 431
Hrachovec, Herbert 221
Hsiang, Jieh 43, 224
Hsieh, Cheng-en 115
Hu, Jiajia 228
Hudson Smith, Andrew 348
Huentelmann, Raphael 221
Hung, Jen-Jou 115, 230
Hyvönen, Eero 22
Ilovan, Mihaela 35
Ingold, Rolf 94
INKE Research Group 324
Innocenti, Perla 232
Isaksen, Leif 22, 99, 236
Ivanovs, Aleksandrs 405
Iwata, Yoshimi 239
Jankowski, Nicholas 394
Jiao, David 465
Jockers, Matthew 242
Johnson, Eric 433
Jolivet, Vincent 202
Jordanous, Anna 176
Juola, Patrick 245
Jürgens, Marco 445
Katelnikoff, Joel 247
Kautonen, Heli Johanna 215
Kay, Christian 427
Kelley, Wyn 190
Kestemont, Mike 249
Ketzan, Erik 252
Kim, Young-Min 467
Kimura, Fuminori 473
Klein, Lauren Frederica 254
Knab, Cornelia 429
Koleva, Nikolina 470
Kong, Jung-Wei 224
Krieger, Hans-Ulrich 470
Kristel, Conny 117
Kruse, Sebastian 137
Kümmel, Christoph 506
Küster, Marc W. 375, 520
Kuyama, Takeo 473
Kwok, Jieli 230
Lam, Monica S. 208
Laramee, Robert S. 151
Lawrence, Faith 176
Lay, Marie Hélène 256
Lee, Sangmi 313
Lemnitzer, Lothar 259
Lendvai, Piroska 163
Lenkiewicz, Anna 477, 538
Lenkiewicz, Przemyslaw 477, 538
Lestock, Brooke 433
Levy, Noga 264
León, Carlos 57
Li, Liang 480
Lis, Magdalena 477
Litta Modignani Picozzi, Eleonora 483
Littauer, Richard 268
Little, Hannah 268
Liu, Jyi-Shane 270
Lohmeier, Felix 486
Longaker, Rachel 342
Loos, Lukas 488
Lou, Burnard 221
Löwe, Benedikt 57

Lucas, Kristin 463
Lucic, Ana 273
Luyckx, Kim 249
López, José 456
MacEachern, Alan 162
Macé, Caroline 85
Maeda, Akira 473
Magier, David 506
Mäkelä, Eetu 22
Maly, Kurt 155
Mandell, Laura 148
Manovich, Lev 79
Marse, Alex 313
Marttila, Ville 396
Mavillard, Antonio Jiménez 306
McCadden, Katiet Theresa 516
McClure, David 299
Meister, Jan Christoph 24
Michura, Piotr 35
Middell, Gregor 418
Miller, Ben 313
Mimno, David 64
Monroe, Megan 158
Monteiro Vieira, Jose Miguel 483, 491
Moran, Steven 276
Morbidoni, Christian 303
Moulin, Claudine 9, 375
Muller, A. Charles 61
Muscinesi, Frédérique 494
Myers, Tom 280, 283
Mylonas, Elli 448
Měchura, Michal Boleslav 278
Nagasaki, Kiyonori 61
Naji, Nada 94
Nakhimovsky, Alexander 280, 283
Nathan, David John 286
Nelson, Brent 35
Nelson, Robert K. 64
Neumann, Bernd 288
Neuroth, Heike 20
Nishiura, Takanobu 480
Noecker Jr., John 292
Nordhoff, Sebastian 296
Norrish, Jamie 483
Nowviskie, Bethany 299, 433
Nucci, Michele 303
Nuñez, Camelia Gianina 306
Olsson, Leif-Jöran 113, 527
Opas-Hänninen, Lisa Lena 308
Ore, Christian-Emil 18
Ott, Tobias 497
Ott, Wilhelm 497
O'Halloran, Kieran Anthony 310
Pablo-Nuñez, Luis 195
Park, Tae Hong 313
Pasin, Michele 499
Peng, Sheng-Yang 270
Petris, Marco 24
Piazza, Francesco 303
Pichler, Alois 221, 318
Pierazzo, Elena 72
Piez, Wendell 33
Pigney, Stephen 124
Pitti, Daniel 70
Pleyer, Michael 268
Plutte, Christoph 502, 511
Porter, Dot 505
Pravida, Dietmar 131, 418
Praxis Program Collaborators 433
Priani, Ernesto 456
Prokic, Jelena 276
Pytlik-Zillig, Brian 327
Rabina, Debbie L. 322
Radzikowska, Milena 35, 324
Rahtz, Sebastian 18, 52
Ramsay, Stephen 327
Rapp, Andrea 375, 529
Rees, Gethin Powell 329
Rehbein, Malte 72
Remondino, Fabio 365
Reside, Doug 331, 333
Rhody, Jason 506
Richards, John 232
Richards-Rissetto, Heather 365
Ridge, Mia 25, 335
Ritsema van Eck, Marianne Petra 336
Rittberger, Marc 359
Rivera, Eduardo 456

Roberts, Sean 268
Rochester, Eric 299
Rockwell, Geoffrey.....
..... 26, 35, 72, 324, 339, 508
Rodriguez, Omar 88
Rodríguez, Nuria 342
Roe, Glenn H. 345
Roeder, Torsten 511
Romary, Laurent 52
Ross, Claire Stephanie 348
Roueche, Charlotte 176
Roued-Cunliffe, Henriette 351
Röwenstrunk, Daniel 435
Rudman, Joseph 353
Ruecker, Stan 35, 88, 324
Ruiz, Cesar 195
Ryan, Michael 292
Rybicki, Jan 16, 212
Sahle, Patrick 72
Salah, Albert Ali 79
Sancho-Caparrini, Fernando 385
Sanz, Amelia 527
Sanz-Cabrerizo, Amelia 195
Sarasa-Cabezuelo, Antonio 195
Savoy, Jacques 94
Sayers, Jentery 357
Scharnhorst, Andrea 394
Scheel, Julian 288
Schindler, Christoph 359
Schnöpf, Markus 362
Schöch, Christof 514
Schomaker, Lambert 336
Schreibman, Susan 72, 516
Schüch, Lena 24
Schultes, Kilian Peter 518
Schwartz, Robert 41
Schwarz, Roland 375
Schwerin, Jennifer von 365
Selig, Thomas 520
Sellers, Jordan 397
Seppänen, Tapio 308
Sharma, Narayan P. 368
Shaw, Ryan Benjamin 370
Shimoda, Masahiro 11
Shrestha, Ayush 313
Siemens, Lynne 373
Siemens, Raymond 373, 441
Sierra-Rodriguez, Jose-Luis 195
Sievers, Martin 375
Simon, Agnès 70
Simon, Rainer 99
Sinclair, Stéfan 26, 35, 72, 339
Singer, Kate 522
Smith, Kathleen M. 486
Smithies, James Dakin 380
Solth, Arved 288
Sondheim, Daniel 35
Söring, Sibylle 486
Speck, Reto 117
Stäcker, Thomas 54, 532
Stadler, Peter 525
Stalnaker, Rommie L. 413
Stein, Regine 532
Stokes, Peter 264, 382
Storrer, Angelika 259
Storti, Sarah 433
Suhr, Carla Maria 396
Sung, Allan 224
Suárez, Juan-Luis 385
Swafford, Joanna 433
Tabata, Tomoji 388
Takahashi, Koichi 61
Tasovac, Toma 392
Tatum, Clifford 394
Tcheng, David 158
Terras, Melissa 204, 348
Thaller, Manfred 27, 72, 142
The ARTFL Project 345
The Devonshire MS Editorial Group ... 441
Thiel, Stephan 151
Toljamo, Tuomo 308
Tomabechi, Toru 61
Tonne, Danah 529
Trilsbeek, Paul 169
Tupman, Charlotte 176
Turkel, William J. 162
Turner, Jonathan 313
Tyrkkö, Jukka Jyrki Juhani 396

Underwood, Ted 397
van Dalen-Oskam, Karina ... 33, 400, 402
van Dijk, Suzan 527
van Hooland, Seth 28
van Zundert, Joris Job 331, 400, 402
Vanscheidt, Philipp 529
Varfolomeyev, Aleksey 405
Veentjer, Ubbo 486
Verborgh, Ruben 28
Vernus, Pierre 431
Vershbow, Ben 331
Versley, Yannick 126
Viglianti, Raffaele 435
Vitali, Stefano 70
Wade, Mara R. 532
Wakelnig, Elvira 176
Wallman, Jeff 61
Walsh, John A. 465
Wang, Ning 228
Wangchuk, Dorji 15
Warwick, Claire 35, 204, 348
Welger-Barboza, Corinne 407
Wenzel, Claudia 411
Wieber, Sabine 232
Wiegand, Frank 535
Wiesner, Susan L. 413
Willems, Florian 537
Williams, David-Antoine 415
Windsor, Jennifer 35
Winters, James 268
Wissenbach, Moritz 418
Wittenburg, Peter 20, 458, 538
Wittern, Christian 422
Wolf, Lior 264
Woods, Sharon 463
Wu, Harris 155
Yano, Keiji 480
Zancarini, Jean-Claude 198
Zhang, Jia 190
Zhang, Lu 88
Zhao, Geng 151
Zipf, Alexander 488
Zöllner-Weber, Amélie 318
Zubair, Mohammad 155

www.dh2012.uni-hamburg.de
ISBN 978-3-937816-99-9