Georg Vogeler

Fachspezifische Indexierung von historischen Dokumenten I
Quellen zwischen Zeichenketten und Information – Beispiel Urkunden

aus:

Forschung in der digitalen Welt

Sicherung, Erschließung und Aufbereitung von Wissensbeständen

Herausgegeben von Rainer Hering, Jürgen Sarnowsky, Christoph Schäfer und Udo Schäfer

S. 43–58

Impressum

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
Die Online-Version dieser Publikation ist auf der Verlagswebsite frei verfügbar *(open access)*. Die Deutsche Nationalbibliothek hat die Netzpublikation archiviert. Diese ist dauerhaft auf dem Archivserver der Deutschen Nationalbibliothek verfügbar.

Open access über die folgenden Webseiten:
Hamburg University Press – http://hup.sub.uni-hamburg.de
Archivserver der Deutschen Nationalbibliothek – http://deposit.d-nb.de

ISBN-10 3-937816-27-5 (Printausgabe)
ISBN-13 978-3-937816-27-2 (Printausgabe)
ISSN 0436-6638 (Printausgabe)

Produktion: Elbe-Werkstätten GmbH, Hamburg, Deutschland
http://www.ew-gmbh.de

Bildnachweis: Der Abdruck aller Abbildungen erfolgt mit freundlicher Genehmigung der Autoren bzw. des Autors des jeweiligen Beitrags.

Inhaltsübersicht

Fachspezifische Indexierung von historischen Dokumenten I

Quellen zwischen Zeichenketten und Information – Beispiel Urkunden

Georg Vogeler

1. Einleitung

Die Geschichtswissenschaft digitalisiert sich: Es werden Portale aufgebaut, fachspezifische Metasuchen installiert, Onlinerezensionen und -publikationen verbreiten sich und werden immer häufiger zitiert (wenn auch vorwiegend innerhalb der ‚Gemeinde'). Die ‚Cultural Heritage'-Institutionen wagen sich immer weiter in die digitale Welt, an ihrer Spitze die Bibliotheken, aber dicht gefolgt von Museen und Archiven. Es ist also an der Zeit, sich zu fragen, wie man die von diesen Institutionen bereitgestellten historischen Informationen auch wiederfindet.

Neben den allgemeinen Suchmaschinen wie Google, Yahoo oder Ask gibt es zu diesem Zweck auch fachspezifische Portale und Suchmaschinen. *Chronicon* ist ein solches Angebot, das die Bayerische Staatsbibliothek aufgebaut hat.[1] Es vermittelt den Zugang zu 32 bibliographischen Onlinehilfsmitteln, die mit einer einzigen Suchabfrage gemeinsam abgefragt werden können. Auch das Berliner Portal *Clio Online* bietet eine solche Metasuche an: Hier gibt es neben der Suche in 28 bibliographischen Datenbanken auch ein Metaportal zur Quellenrecherche.[2] Wenn man sich die dahinter stehenden Datenbanken genauer ansieht, stellt man fest, dass es sich um genau

[1] Adresse: http://www.chronicon.de/ (letzte Einsichtnahme am 21.04.2006).

[2] Adresse: http://www.clio-online.de/ (letzte Einsichtnahme am 21.04.2006).

drei handelt: das Archivportal des Landes Nordrhein-Westfalen,[3] die VD17--Datenbank der Herzog-August-Bibliothek in Wolfenbüttel[4] und die Nachlassdatenbank des Bundesarchives.[5]

Ist es etwa so, dass die Grundlage historischer Arbeit, die Quellen, noch gar nicht online suchbar ist? Sicherlich sind im Netz noch nicht so viele Quellenbestände recherchierbar wie mit gedruckten Hilfsmitteln. Die Situation ist jedoch beileibe nicht schlecht. So haben andere Archivverwaltungen auch ähnliche Angebote wie das Land Nordrhein-Westfalen eingerichtet: das Portal des Landesarchivs Baden-Württemberg,[6] das Hessische Archiv-Dokumentations- und -Informationsystem *HADIS*[7] oder *Ariadne*, das *Archive Information and Administration Network* der Archive in Mecklenburg-Vorpommern.[8] In diesen Portalen kann man sich einerseits zu den einzelnen Archiven durchklicken, man kann aber auch in allen Bestandübersichten und Findbüchern, die schon digital vorliegen, suchen – ohne erst das einzelne Archiv anzusteuern. Wir haben es also je mit einer Art ‚Quellensuchmaschine' zu tun.

Neben den Archiven sind auch andere Quellencorpora online zu finden. Ich möchte hier keine umfangreiche Übersicht über solche Corpora und ihre unterschiedlichen Präsentationsformen bieten, sondern kurz zwei Angebote nennen, die sich mittalterlichen Urkunden verschrieben und sie auf besondere Art und Weise durchsuchbar gemacht haben: Da sind die hessischen *Landgrafen-Regesten online*[9] und der *Codice diplomatico della Lombardia medievale*.[10] Das Erstere fällt durch die Möglichkeit auf, in den Volltexten nach lemmatisierten Wörtern zu suchen, das heißt Varianten von Wörtern in Flexionsformen durch Angabe des Grundwortes, das Letztere durch eine umfangreiche Expertensuche, die einen tief in die sachlichen und sprachlichen Strukturen des Textes hineinführt.

[3] Adresse: http://www.archive.nrw.de/ (letzte Einsichtnahme am 21.04.2006).

[4] Adresse: http://www.vd17.de/ (letzte Einsichtnahme am 21.04.2006).

[5] Adresse: http://www.nachlassdatenbank.de/ (letzte Einsichtnahme am 21.04.2006).

[6] Adresse: http://www.landesarchiv-bw.de/ (letzte Einsichtnahme am 21.04.2006).

[7] Adresse: http://www.hadis.hessen.de/ (letzte Einsichtnahme am 8.04.2006).

[8] Adresse: http://ariadne.uni-greifswald.de/index.html (letzte Einsichtnahme am 21.04.2006).

[9] Adresse: http://online-media.uni-marburg.de/ma_geschichte/lgr/ (letzte Einsichtnahme am 18.04.2006).

[10] Adresse: http://cdlm.unipv.it/ (letzte Einsichtnahme am 21.04.2006).

Ein Konzept für eine fachspezifische Suchmaschine, die dem Historiker dabei hilft, Quellen zu seinen Fragestellungen zu suchen, ja vielleicht sogar nicht nur Quellenreferenzen, sondern auch schon Quellenaussagen, muss sich mit den Fähigkeiten solcher Suchen vergleichen lassen.

Worin liegt denn der Vorteil der beiden Suchen? Sie ermöglichen es dem Historiker, Informationen zu finden, ohne die genaue Zeichenrepräsentation dieser Information zu kennen: Die Suche nach Kaufgeschäften in den hessischen Regesten kann mit einer Suchen nach „kaufen" gestartet werden und muss sich nicht darum kümmern, ob das Phänomen „kaufen" mit einem Partizip „gekauft" oder in Flexionen wie „kauft" ausgedrückt ist.

Die online verfügbaren Editionen des *Codice diplomatico della Lombardia medievale* reichen bis zum Jahr 1150. Ein Historiker, der sich für Anwesenheiten, Aktivitäten und Rezeption der Ottonen in Norditalien interessiert, kann sich zu einer Personensuche vorarbeiten. Dort kann er die 45 Einträge zu „Otto imperator" ebenso aussuchen wie die 3 zu „Otto I., imperatore", „Otto II., imperatore" und zu „Otto III., imperatore". Eine Volltextsuche nach „Otto" (676 Treffer) liefert gänzlich andere Ergebnisse: Ottonis (555 Treffer) wird bei einer einfachen Stichwortsuche übergangen, bei einer rechtstrunkierten Suche dafür auch „Ottobelli" (38 Treffer) mit ausgewiesen. Die ottonischen Kaiser sind also nur als Personen und nur schwierig als Teil des Volltextes identifizierbar.[11]

Die mittelalterlichen Urkunden werden auch weiterhin der Testfall sein, an dem die vorhandenen Suchmaschinentechnologien exemplifiziert werden. Wie weit die für dieses Material sinnvollen Techniken auch für andere Quellengattungen nützlich sind, ist weiteren Studien vorbehalten.

2. Suchmaschinentechnologie

Eine moderne Suchmaschine hat – sehr vergröbert – etwa folgenden Aufbau: Ein Crawler durchsucht das Internet, das heißt er verfolgt alle auf Web seiten vorhandenen Links. Die dabei gefundenen Webseiten übergibt er an Software, die das Dokument in einzelne Wörter zerlegt und daraus einen Index aufbaut. Wenn der Index abgefragt wird, dann wird die gefun-

[11] Es sei noch darauf hingewiesen, dass auch die Personenindexeinträge voneinander abweichen: Otto imperator: 39, Otto imperator augustus: 2, Otto imperatore: 2, Otto, imperator: 2, Otto I, II, III, je ein Eintrag.

dene Menge an Treffern nach Relevanz geordnet. Die Techniken zum Suchen der Dokumente, die Organisation eines schnellen Index und die Methoden zur Bewertung der Relevanz eines Dokumentes für die Suchanfrage sind die drei Bereiche, in denen Suchmaschinenhersteller ihre ganz spezifischen Kompetenzen und Geschäftsgeheimnisse haben. Eine fachspezifische Suchmaschine muss sich eben den selben drei Bereichen widmen. Hier soll es nun vorrangig um den Index gehen, auch wenn einige Überlegungen zur Bewertung mit einfließen.

Aus technischer Sicht ist der Index die Form, wie die Daten so abgelegt werden, dass die Maschine schnell auf sie zugreifen kann. Die einschlägigen Stichwörter lauten ‚B-Baum', ‚Hash', ‚R-Baum' usw. und meinen verschiedene Arten, die einzelnen Wörter so auf der Festplatte abzulegen, dass ein gegebenes Suchwort schnell mit ihnen verglichen werden kann. Die Zugriffsgeschwindigkeit ist dabei die Kernfrage. Welcher Art die Daten sind, interessiert nicht, denn sie sind einfach ein beliebiger Code. Anders formuliert: Die Techniker interessieren sich vorrangig dafür, wie der Computer schnell eine beliebige Zeichenkette in einer Liste findet. Ob diese Liste Personennamen, Archivsignaturen, Berufscodes einer Volkszählung oder Genomsequenzen enthält, interessiert nicht.

Historiker haben da eine andere Perspektive. Sie denken an das Register im Buch, das ihnen hilft, eine bestimmte Stelle im Text zu finden, ein Kapitel, einen Abschnitt, der sich mit Friedrich II. beschäftigt oder die Urkunden, die für den Deutschen Orden ausgestellt worden sind. Sie beginnen die Recherche in einem Quellencorpus im Register, wenn sie den Kontext des Wortes „gubernare" ermitteln wollen oder wenn sie etwas über die Mühlen im 13. Jahrhundert erfahren möchten.

Quellen sind damit für Historiker nicht nur reine Zeichenketten, sondern Informationsträger. Sie brauchen Indices, die an diesen Informationen orientiert sind. Eine Volltextsuche ist dabei nur ein Hilfsmittel bei der Informationsrecherche. Eine fachspezifische Suchmaschine muss also einen Index aufbauen, der möglichst informationsnah gestaltet ist.

Dabei kann es nicht darum gehen, dass eine Heerschar an Wissenschaftlern alle online verfügbaren Quellenbestände systematisch erschließt. Es muss im Gegenteil darum gehen, Quellenbestände, die ohne Erschließung ins Netz gestellt werden, mit Hilfe von kalkulierbaren Prozeduren automatisch aufzubereiten, so dass sie ihre individuellen Eigenheiten behalten können, im Index der Suchmaschine aber eine Repräsentanz finden, die

mehr über ihren Informationsgehalt verrät als nur eine Liste aller Zeichenketten, die zwischen Worttrennungselementen, das heißt Leerräumen, Satzzeichen usw. stehen.

3. Möglichkeiten zur graphischen und sprachlichen Abstraktion

Eine erste Abstraktionsschicht zwischen Quellentext und Information stellt die Orthographie dar. Wo wir in den jüngsten Diskussionen um ‚alte', ‚neue' und ‚reformierte' Rechtschreibung gelernt haben, dass es sogar in einer orthographisch sehr regelbewussten Zeit Varianten gibt, die gleiche Wörter unterschiedlich graphisch repräsentieren, so sind an Quellen arbeitende Historikerinnen und Historiker sich der um so größeren Vielfalt bis ins 17. Jahrhundert wohl bewusst. Ein erstes Ziel einer automatisierten Abstraktion von den Zeichenketten der digitalen Quellenrepräsentanz ist es also, möglichst viele dieser Varianzen auszuschalten bzw. zu tolerieren.

Die Techniken dazu sind nichts Neues. Sie sind auch schon in komplexe fachspezifische Softwarelösungen eingeflossen.[12] Sie seien hier aber erneut zusammengestellt und auf ihre Nutzbarkeit für eine fachspezifische Suchmaschine geprüft.

Eine grundlegende Technik, graphische oder sprachliche Varianzen in eine Suche einzufügen, sind die ‚regulären Ausdrücke'. Diese in vielen Softwareumgebungen implementierten Mustervergleiche bieten eine Notation an, eine Zeichenkette aus Alternativen bestehen zu lassen. Ein solcher Ausdruck „\<O[dt][.]{0,1}o.*?\>" sucht nach Wörtern, die mit einem „O" anfangen, auf das ein „d" oder ein „t" folgt, dem noch maximal ein zweiter Buchstabe folgen kann, ein „o" folgen muss, bevor das Wort beliebig enden kann. Damit lassen sich zum Beispiel Schreibungsvarianten des Namens „Otto" abbilden: „Oto", „Odo", „Oddo", „Oddonis", „Othonem" entsprechen alle diesem Muster.

[12] Vgl. Manfred Thallers *kleio* (Adresse: http://www.hki.uni-koeln.de/kleio/, letzte Einsichtnahme am 20.04.2006) und darauf aufbauende Anwendungen. Einem dem hier vorgestellten Ansatz vergleichbar ist: Andrea Ernst-Gerlach und Norbert Fuhr: Generating Search Term Variants for Text Collections with Historic Spellings. In: 28th European Conference on Information Retrieval Research (ECIR 2006). Adresse: http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Ernst_Fuhr:06.pdf (letzte Einsichtnahme am 21.04.2006), die jedoch nicht bei der Indexierung, sondern beim Ranking ansetzen.

Einem ähnlichen Prinzip folgen ‚Name Matching'-Algorithmen, von denen es eine größere Menge gibt. Der bekannteste ist das schon 1918 entwickelte ‚Soundex'-Verfahren, das den amerikanischen Einwanderungsbehören half, der Sprachen- und Namensvielfalt der europäischen Einwanderer Herr zu werden. Der klassische Soundex-Algorithmus erhält den ersten Buchstaben eines Namens und fügt bis zu 3 Ziffern an, die für die Konsonanten des Namens stehen und sie in Gruppen ordnen. Aus „Ottonem" wird dann „O355"; aus „Odo" „O300" und aus „Othonis" „O352". Das Verfahren ist relativ stark auf die englische Sprache ausgerichtet, weshalb verschiedene Modifikationen vorgeschlagen worden sind.[13]

Es ist sichtbar, dass auch eine solche Umwandlung immer noch ungenaue Ergebnisse liefert. Ungenauigkeit ist in der Informatik auch aus der Arbeit mit automatischer Texterkennung bekannt: Leichte Abweichungen vom orthographisch eingeführten Zeichenbestand entstehen dabei leicht. Um auch in solchen ‚unsauberen' Texten suchen zu können, sind Algorithmen entwickelt worden, die sich dem Problem des ‚nearest neighbour' widmen. Dabei spielt der Grad der Veränderung eine entscheidene Rolle: Am geläufigsten ist wohl die 1965 vorgestellte ‚Levenshtein-Distanz', die die Anzahl der mindestens notwendigen Operationen beziffert, die es braucht, um aus einer Zeichenkette eine zweite werden zu lassen. Damit kann ein Wert festgelegt werden, bis zu dem zwei Zeichenketten, die nur leicht voneinander abweichen (zum Beispiel in genau einem Zeichen) als gleich gewertet werden.[14]

Aus historischer Perspektive erscheint es unsinnig, beliebige Änderungen gleich zu behandeln: Es ist offensichtlich, dass bestimmte Abweichungen zweier Zeichenketten für einen gegebenen sprachlichen Kontext weniger auffällig sind als andere: Bis ins 17. Jahrhundert kann man davon ausgehen, dass „u" und „v" nur positional und nicht lautlich differenziert verwendet werden. Eine Abweichung von „vnnß" zu „uns" als pure Zeichenkette ist mit einer Levenshtein-Distanz von 3 zu bewerten, obwohl „u" und „v", „ß" und „s" sowie „nn" und „n" für den geübten Leser als iden-

[13] Einen guten Überblick bis zum Stand 1995 gibt: A.J. Lait und B. Randell: An Assessment of Name Matching Algorithms [1995]. Adresse: http://homepages.cs.ncl.ac.uk/brian.randell/ home.-informal/Genealogy/NameMathing.pdf (letzte Einsichtnahme: 31.03.2006), die auch gleich eine weitere Variante (‚Phonex') beisteuern.

[14] Der Algorithmus wurde veröffentlicht: Vladimir I. Levenshtein: Binary codes capable of correcting deletions, insertions, and reversals. In: Doklady Akademii Nauk SSSR 163 (1965). S. 845–848 (Russisch). Englische Übersetzung in: Soviet Physics Doklady 10 (1966). S. 707–710.

tisch zu bewerten sind. Für das Niederdeutsche hat Jan Strunk einmal daraus die Konsequenz gezogen und einen Algorithmus vorgeschlagen, der diese in der Abstandsberechnung unterschiedlich gewichtet und insbesondere durch Berücksichtigung graphischer Varianzen signifikant bessere Suchergebnisse erzielt.[15]

Im oben angeführten hessischen Beispiel kommt jedoch keine dieser Techniken zum Einsatz. Statt dessen verwendet die Anwendung ein Verfahren, das aus der Computerlinguistik als ‚Stemming' bekannt ist. Dahinter verbirgt sich eine morphologische Lemmatisierung, die Flexionsformen eines Wortes auf ihr Grundwort zurückführt. Auch für das Lateinische existiert Software zur morphologischen Analyse, wie zum Beispiel ‚LemLat'[16] oder das ‚Perseus Word Study Tool'.[17] Diese Ansätze basieren gewöhnlich auf Lexika und Regelwerken zur Flexion, Präfix- und Suffixbildung. Damit verwendet diese Technik nicht nur inhärente Strukturen der Texte, die algorithmisch aufbereitet werden, sondern externes Wissen, das in Thesauri abgelegt ist. Dennoch sind die Verfahren des Stemming keine reinen Wortlisten, die ganze Flexionsparadigmen ihrem Grundwort zuordnen, sondern Wörterbücher der Grundwörter mit formalisierten Angaben zur Morphologie, so dass der Computer jede mögliche Flexionsform des Wortes bilden kann.

Es gibt aber auch Anwendungen, die ganze Wörter in Referenzlisten zusammenführen. Insbesondere prosopographische Arbeit hat Namenslisten erstellt, die leider nur selten öffentlich zugänglich sind: So sind die Register der jüngeren Bände des *Repertorium Germanicum* mit solchen Namenslisten entstanden. Auch die Namensdatenbanken, die im Zusammenhang mit dem *Altdeutschen Namenbuch* und dem Projekt *Nomen et gens* entstanden sind, sind nicht öffentlich verfügbar.[18] Es ist offensichtlich noch nicht wahr-

[15] Jan Strunk: Information retrieval for languages that lack a fixed orthography. Adresse: http://www.linguistics.ruhr-uni-bochum.de/~strunk/LSreport.pdf (letzte Einsichtnahme am 21.04.2006).

[16] Adresse: http://webilc.ilc.cnr.it/~ruffolo/ (letzte Einsichtnahme am 20.04.2006).

[17] Adresse: http://www.perseus.tufts.edu/cgi-bin/morphindex?lang=la (letzte Einsichtnahme am 20.04.2006).

[18] *Nomen et gens*: Adresse: http://www.nomen-et-gens.de/portal.asp (letzte Einsichtnahme am 20.04.2006). Vgl. auch Redmer Alma: A multi-level database. Study on the elite of 15th and 16th century Groningen. In: Prosopography and Computer. Contributions of Medievalists and Modernists on the Use of Computer in Historical Research. Leuven, Apeldoorn 1995. S. 185–194. – Didier F. Isel: Prosopographie des personnages mentionnés dans les textes pour l'époque de

genommen worden, dass diese Materialsammlungen nicht nur zur Vorbereitung onomastischer und prosopographischer Forschung dienen, sondern ein nutzbringendes Hilfsmittel darstellen können, Quellenaussagen aus den Zeichenketten ihrer digitalen Repräsentation zu ermitteln. Eine fachspezifische Suchmaschine muss also den Versuch unternehmen, derartige Datenbestände zu nutzen. Es ist jedoch um vieles leichter, ‚Name Matching'-Algorithmen und approximative Verfahren beim Erstellen eines Index zu verwenden, da hier vorhandene oder einmal zu entwickelnde Algorithmen direkt in die Softwarearchitektur eingebaut werden, statt umfangreiche Zugriffe auf bislang noch nicht einmal öffentliche externe Datensammlungen zu organisieren.

Für eine historisch orientierte Suchmaschine wären demnach Algorithmen zu entwickeln. Das Untersuchungsobjekt ‚mittelalterliche und frühneuzeitliche Urkunden' würde zum Beispiel besonders zwei Verfahren erfordern, eines für die graphischen Varianten des Mittellatein und eine für die des Frühneuhochdeutschen.

Das *DEEDS*-Projekt, das umfangreiche mittellateinische Urkundentexte englischer Klöster digitalisiert hat, verwendet dafür ein existierendes Verfahren, den ‚Double Metaphone'-Algorithmus, der einen echten phonetischen Ausgleich versucht.[19] Erfahrungen damit liegen derzeit noch nicht vor. Auf Grund der Erfahrungen von Jan Strunk, der die Qualität der Suchergebnisse insbesondere durch Verfahren auf graphischer Ebene verbessert sieht, würde ich einen einfachen ‚Graphex'-Algorithmus vorschlagen, der fol-gende Regeln berücksichtigen würde: „h" wird nicht berücksichtigt. „cio" und „tio" werden gleichbehandelt, „e" „ae" „oe" und e-caudata werden gleichbehandelt, ebenso „i", „y" und „ii". Für die Epenthese (zum Beispiel „damnum" = „dampnum") und die Assimiliation (zum Beispiel „auctor" > „autor") werden Ausgleichsklassen gebildet.

Auch für das Mittelhochdeutsche wie für das Frühneuhochdeutsche existieren noch keine Ausgleichsalgorithmen. Ein ‚Graphex' dazu müsste die Konsonantenreduplikationen berücksichtigen, „v" und „u" sowie „ß"

Pépin le Bref et de son frère Carloman (741–768) spécialement ceux exerçant une fonction ecclésiastique ou laïque, Quatrième édition (mars 2006): Tableau de concordance entre noms et lemmes (Adresse: http://prosopographie-id.de/Tableau%20de%20concordance.htm).

[19] Vgl. zum Algorithmus: Lawrence Phillips: The Double Metaphone Search Algorithm. In: C/C++ User Journal June 2000. Adresse: http://www.cuj.com/documents/s=8038/cuj0006philips/ (letzte Einsichtnahme am 31.03.2006).

und „s“ gleichbehandeln, „cz“ mit „z“ gleichsetzen und die Schreibweisen „ew“ und „aw“ mit den Diphtongen „eu“ und „au“ identifizieren.[20]

Die Realität dialektal eingefärbter Urkundensprache des 14.–16. Jahrhunderts würde auch ‚Soundex'-ähnliche Ausgleichsmechanismen nahelegen: p=b, mb=b, d=t, g=k=c=q sind geläufige lautliche Verwandschaften. Die deutsche Sprachgeschichtsforschung kennt noch eine größere Menge zeitlich und geographisch begrenzter Abweichungen,[21] kann aber ohne ein gerade erst im Entstehen befindliches Corpus[22] noch keine genaueren Aussagen darüber treffen, wann wo welche Varianten auftreten. Ob also die Berücksichtigung dieser weitergehenden Varianten nutzbringend ist, bedarf genauer Evaluation.

Mit derartigen Ausgleichsmechanismen besteht die begründete Hoffnung, dass die Stichwörter aus dem aktuellen Wortschatz eines Historikers, in die er sein Forschungsproblem übersetzt und die er in eine mit diesem Mechanismen ausgestattete Suchmaschine eingibt, eine Sammlung auch orthographisch und lautlich varianter Quellentexte ermittelt, die Auskunft zu seiner Frage geben, dass er also zum Beispiel Quellendokumente findet, die eine historische Schreibweise des gesuchten Personennames enthalten.

4. Informationstypen

Für historische Arbeit ist darüber hinaus aber noch eine Eigenschaft von Quellentexten zentral, die in einer Suche, die nur auf Zeichenketten orientiert ist, nicht möglich ist: die zeitliche Einordnung. Zeitliche Einordnung beruht nämlich auf numerischen Angaben. Für eine zeitliche Eingrenzung einer Suche auf Quellenmaterial aus der Zeit Ottos III. wären Quellentexte zu filtern, die Datumsangaben aus der Zeit zwischen 996 und 1002 enthal-

[20] Eine Beispielimplementation für den mittellateinischen und den frühneuhochdeutschen Graphex und einen oberdeutschen Soundex in Visual Basic for Applications findet man unter Adresse: http://www.cei.lmu.de/examples/graphex.bas.zip (letzte Einsichtnahme am 01.05.2006)

[21] Vgl. zum Beispiel Frédéric Hartweg und Klaus-Peter Wegera: Frühneuhochdeutsch. Eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit (Germanistische Arbeitshefte 33). 2. überarb. Aufl. Tübingen 2005; Gerhard Philipp: Einführung ins Frühneuhochdeutsche. Sprachgeschichte, Grammatik, Texte. Heidelberg 1980.

[22] Vgl. *Deutsch Diachron Digital*: Adresse: http://www2.hu-berlin.de/ddd/ (letzte Einsichtnahme am 20.04.2006) mit Links auf einzelne Teilprojekte (Adresse: http://www2.hu-berlin.de/ddd/links.php).

ten. Ein Ausdruck „>995 und <1003“ führt in alphanumerischer Ordnung, das heißt also in einer Ordnung rein nach Reihenfolge der einzelnen Zeichen, zu keinem Ergebnis, denn es gibt kein Zeichen, das in alphanumerischer Reihung hinter „9“ und vor „1“ liegt. Ebenso sind Abweichungen, die über approximative Suchverfahren ausgeglichen werden, anders zu beurteilen, indem die Änderung einer Stelle am Beginn einer Zahl eine weit größere Abweichung darstellt als am Ende einer Zahl. Ein fachspezifischer Index muss also zwischen alphanumerischen und numerischen Anfragen unterscheiden.

Zunächst erscheint die Unterscheidung leicht, da für beide Informationstypen bestimmte Zeichen reserviert sind: Ziffern und Buchstaben. Ich möchte das Problem gemischter Formen, lateinischer Zahlen und sprachlicher Ausdrücke für Zahlenwerte beiseite lassen und das Problem ‚Zahl/Wort‘ als allgemeineres Problem von Informationstypen weiterverfolgen.

Viele der bislang angesprochenen Techniken erscheinen im historischen Alltag überflüssig: Um einen Bibliothekskatalog benutzen zu können, braucht der Forscher eine bibliographisch regelgerechte Referenz, die im 20. Jahrhundert zum Standard wissenschaftlichen Arbeitens geworden ist. Eine Fußnote oder eine wissenschaftliche Bibliographie liefern mir also ausreichend Informationen, eine Angabe gemäß der strikten Regeln des Kataloges zu finden, ja sie sogar für meine Suche nutzbar zu machen, indem klare Ein- und Ausgrenzungen möglich sind. Auch die Suche im Findbuch im Archiv braucht in vielen Fällen keinen Ausgleich historischer Schreibweisen, denn es ist häufig genug von einem Zeitgenossen angelegt worden, der moderner Orthographie folgt – von den Findbüchern des 19. Jahrhunderts sei hier einmal abgesehen. Beide Informationsquellen haben jedoch in der Welt der Bibliotheken und des Internets eines gemeinsam: Sie sind ‚Metadaten‘, das heißt Angaben, die das eigentliche Objekt – das Archival, das Buch – beschreiben. Damit sehen sie grundsätzlich anders aus als der eigentliche Text einer Quelle und sind ‚Informationstypen‘ wie Zahlenangaben oder geographische Angaben mit eigenen Eigenschaften. Weitere geläufige Informationstypen findet man in der Welt der gedruckten Bücher in den Registern: Ortsindex, Personenindex oder Sachindex sind alles Texte, die eine bestimmte Information repräsentieren, die bei einer Suche unterschiedlich ausgewertet werden kann. Die Funktionen dieser Informationstypen sind in der elektronischen Welt über die Anwendungen sichtbar gemacht, in die sie eingebettet sind. So kann zum Beispiel der

Karlsruher Virtuelle Katalog Informationstypen in verschiedenen Datenbeständen zusammenfassen, indem er auf die standardisierten Schnittstellen der internationalen Bibliotheksanwendungen zugreift.[23] Eine fachspezifische Suchmaschine kann das nicht, da die Anwendungen zur Präsentation der histori-schen Dokumente keine solchen Schnittstellen kennen.

5. Strukturen

Dennoch kann sie die Informationstypen unterscheiden. Die technische Entwicklung im WWW entfernt sich von HTML, der ‚HyperText Markup Language', die dem Internet in 1990er Jahren sein äußeres Gesicht gegeben hat. Auch wenn die Diskussion um das ‚Semantische Netz' oder ‚Web 2.0' nicht mehr die Feuilletons beherrscht,[24] XML dominiert die technische Diskussion, wenn es um Publikationen im Internet geht.

Die Phänomene HTML und XML werden in der Informatik unter dem Stichwort ‚Semistrukturierte Daten' diskutiert. Darunter werden Phänomene zusammengefasst, die sich mit relationalen und objektorientierten Datenbanksystemen nicht oder nur ungünstig abbilden lassen, weil unübersichtlich komplexe Datenstrukturen entstehen und fehlende oder wiederholte Informationstypen umfangreiche Leerstellen entstehen lassen. Die Textauszeichnungssprachen HTML und XML sind darüber hinaus so erfolgreich, weil sie nicht nur unregelmäßig strukturierte Daten abbilden können, sondern auch weil sie ‚selbstbeschreibend' sind, das heißt, weil die strukturtragenden Daten Teil der Texte selbst sind.[25]

Ein fachspezifische Suchmaschine wird also auf diesen Eigenschaften der Dokumente im Web aufbauen. Welche Konsequenzen das haben kann,

[23] Über den KVK: Adresse: http://www.ubka.uni-karlsruhe.de/hylib/virtueller_katalog.html (letzte Einsichtnahme 20.04.2006).

[24] Der Einführungsvortrag zur Tagung *.hist2003* widmete sich noch dieser Vision (vgl. Christoph Albrecht: Geschichte und Neue Medien. In: Geschichte und Neue Medien in Forschung, Archiven, Bibliotheken und Museen, Tagungsband *.hist2003*. Hg. von Daniel Burgkhardt und anderen (Historisches Forum 7). Berlin 2005. Adresse: http://edoc.hu-berlin.de/e_hist-for/7_I/PHP/Ueberblicke_7-2005-I.php#001001, letzte Einsichtnahme am 20.04.2006). Auf der Folgetagung *.hist2006* war das Stichwort jedoch aus den Titeln der Vorträge und Sektionen verschwunden (vgl. Adresse: http://www.clio-online.de/hist2006/texts/hist2006_programm.pdf, letzte Einsichtnahme am 20.04.2006).

[25] Vgl. zu den ‚Semistrukturierten Daten': Serge Abiteboul, Peter Buneman und Dan Suciu: Data on the Web. From Relations to Semistructured Data and XML. San Francisco 2000.

möchte ich an den mittelalterlichen und frühneuzeitlichen Urkunden deutlich machen und bei einem Vorhaben beginnen, das zunächst wie das Gegenteil einer Suchmaschine klingt, die automatisch historischen Texten wahrscheinliche sachliche Repräsentanzen zuordnet, ein Vorhaben, das jedoch Konzepte verdeutlicht, die in den Funktionsumfang einer solchen Suchmaschine nutzbringend eingebaut werden können.

Die *Charters Encoding Initiative* (CEI)[26] ist eine Arbeitsgruppe, die sich im Frühjahr 2004 das Ziel gesetzt hat, einen Vorschlag für einen Textauszeichnungsstandard für die Quellengattung ‚Urkunden des Mittelalters und der frühen Neuzeit' zu machen. Ein solches Bestreben beruht auf den Beobachtungen, dass erstens die Historiker, Archivare und Bibliothekare diese Quellengattung in größerem Maß digitalisiert und über das Netz zugänglich gemacht haben und dass zweitens die Urkunden strukturierter Erschließung besonders zugänglich sind.[27]

Die Arbeitsgruppe geht auf einem induktiven und auf einem deduktiven Weg zu einem gemeinsamen Textauszeichnungsstandard. Induktiv sind folgende zwei Maßnahmen: Die Arbeitsgruppe ermittelt erstens Ähnlichkeiten vorhandener Schemata und analysiert zweitens die Arbeitsprozesse bei der Arbeit mit Urkunden. Dabei ist deutlich geworden, dass die an unterschiedlichen Stellen unabhängig von einander entwickelten Datenstrukturen auffällig ähnlich sind. Diese Ähnlichkeit beruht nicht nur darauf, dass natürlich immer das gleiche Objekt beschrieben wird, sondern sogar in den Namen der Strukturelemente finden sich Ähnlichkeiten: Für Abkürzungen verwenden zum Beispiel die den Regeln der TEI angelehnten Editionen der *École nationale des chartes*[28] ebenso <abbr> wie der *Codice diplomatico della Lombardia medievale*.[29]

Die Eigenschaft von XML-Dokumenten, ‚selbstbeschreibend' zu sein, geht also teilweise so weit, dass unabhängig von einander entwickelte Datenstrukturen direkt übertragbar werden. Ein erstes Ziel der Arbeitsgruppe

[26] Adresse: http://www.cei.lmu.de (letzte Einsichtnahme am 20.04.2006).

[27] Vgl. Georg Vogeler und Patrick Sahle (Hg.): Virtual Library Historische Hilfswissenschaften, Sektion Diplomatik. Adresse: http://www.vl-ghw.lmu.de/diplomatik.html (letzte Einsichtnahme am 20.04.2006).

[28] Adresse: http://elec.enc.sorbonne.fr (letzte Einsichtnahme am 21.04.2006). Die XML-Daten dieses Projektes sind über die Webseite nicht einsehbar.

[29] Adresse: http://cdlm.unipv.it (letzte Einsichtnahme am 21.04.2006). XML-Daten können in diesem Projekt über die Suchfunktion (‚Mostra la marcatura XML') auch über die Webseite eingesehen werden.

ist es deshalb, einen gemeinsamen Wortschatz für die Auszeichnung von digitalen Urkundenrepräsentationen bereitzustellen.[30]

Der deduktive Weg geht von einer internationalen Terminologie der Diplomatik aus.[31] Das *Vocabulaire Internationale de Diplomatique* ist nämlich nicht nur ein Wörterbuch, das die in 12 europäischen Sprachen und im Lateinischen üblichen Fachausdrücke zur Beschreibung von Urkunden und ihrer Entstehung zusammenstellt, sondern bildet durch seinen systematischen Aufbau auch eine Struktur diplomatischer Informationen ab.[32]

Eine fachspezifische Suchmaschine wird sich also an diesen Strukturen orientieren. Sie wird die existierenden Strukturen aus den Dokumenten auslesen und sie mit den vorhandenen Strukturen der CEI vergleichen. Obwohl zurzeit noch Handarbeit notwendig ist, erscheint es möglich, mit Hilfe des *Vocabulaire Internationale de Diplomatique* auch XML-Struktu-ren der Ausgangsdokumente automatisch auf einen Standard nach den Vorgaben der CEI abzubilden.

Dabei entstehen im Index der Suchmaschine natürlich sehr ungleiche Strukturen. Welche Bedeutung kann man dann den Strukturen noch zumessen? Zunächst einmal wird aus den existierenden Beispielen deutlich, dass eine fachspezifische Suchmaschine in Datenbeständen sucht, nicht in Webdokumenten. Das heißt die Suchmaschine wird versuchen, das einzelne Dokument aus der Struktur der Datei zu filtern und dafür die Treffergenauigkeit auszuwerten.

Eine ganz neue Perspektive ergibt sich, wenn man sich an Diskussionen im Bereich ‚Information Retrieval' heranwagt. Neben Überlegungen, auch die Semantik eines Textes mit in die Suche einzubeziehen, die neben den oben angeführten Verfahren unter dem Stichwort ‚Textmining' komplexe Konzepte entwerfen, die Bezüge und Bedeutungen von Wörtern formal abzubilden, konzentriert sich ein Strang der Diskussion auf die Strukturen.[33]

---

[30] Vgl. den Vorschlag unter Adresse: http://www.cei.lmu.de/taglibrary.html (letzte Einsichtnahme am 21.04.2006).

[31] Vocabulaire international de la diplomatique. Hg. von Maria Milagros Cárcel Ortí (Collecció Oberta). 2. verbesserte Auflage. València 1997.

[32] Vgl. dazu auch Michele Ansani: Diplomatica (e diplomatisti) nell'arena digitale. In: Archivio storico italiano 158 (2000). S. 349–379, auch in: Scrineum 1 (1999). S. 1–11 (Adresse: http://dobc.unipv.it/scrineum/ansani.htm, letzte Einsichtnahme am 05.03.2000).

[33] Vgl. Norbert Fuhr: Information Retrieval Methods for Literary Texts. In: Jahrbuch für Computerphilologie 5 (2003). S. 147–160 (online unter Adresse: http://computerphilologie.uni-muenchen.de/jg03/fuhr.pdf, letzte Einsichtnahme am 21.04.2006).

Welche Bedeutung die Strukturen für die Bewertung eines Treffers haben können, kann gut am Urkundenbeispiel erläutert werden. Volkssprachliche mittelalterliche Urkunden sind nämlich nicht nur Quellen für Historiker, sondern auch hochgeschätzte Quellen für die Sprachgeschichte, die in ihnen regional und zeitlich klar zugeordnete Textzeugen findet. Das Suchinteresse des Sprachhistorikers ist so in vielen Fällen ein gänzlich anderes: Zwar sucht er auch nach Urkunden, aber er sucht nach Urkundentexten, nicht nach Urkunden als Zeugnissen von Sachverhalten. Ja sogar die oben angeführten Verfahren, variante Darstellungen von Wörtern mit ihren modernen Formen in Übereinstimmung zu bringen, sind für den Sprachistoriker nicht notwendigerweise neue Zugänge zu bislang verschlossenen Informationen, sondern sie können ebenso auch Unterscheidungen überdecken, die zentral sind für die Anfrage.

Für ein konkretes Projekt einer fachspezifischen Suchmaschine können diese unterschiedlichen Suchinteressen durch unterschiedliche Anfrageseiten berücksichtigt werden.[34] Es kann aber auch eine andere Erkenntnis daraus verallgemeinert werden: Moderne Suchausdrücke für historische Sachverhalte sind in den Originaltexten seltener und ungenauer zu erwarten als in den Metadaten, die sprachlich normalisiert sind. Eine historische Anfrage an ein digitales Urkundencorpus wird deshalb Treffer in den Kopfregesten höher bewerten als Treffer im Editionstext. Die aus XML-Dokumenten ermittelbaren Datenstrukturen liefern also nicht nur Hinweise auf die Datentypen und die darauf anzuwendenden Werkzeuge graphematischer, phonetischer und lexikalischer Konkordanz, sondern auch Hinweise auf die Relevanz eines Treffers für die Anfrage.

6. Fazit

Zunächst sei festgestellt, dass die angesprochenen Instrumente keineswegs eindeutige, positivistische Sachaussagen aus den Quellentexten werden ermitteln können. Die Frage- und Deutungskompetenz der Historiker stellt eine fachspezifische Suchmaschine nicht in Frage. Sie liefert ihm ein Hilfs-

[34] Vgl. Georg Vogeler: Europäisches Urkundenerbe. Zu Potentialen und Perspektiven eines internationalen Fachinformationssystems digitaler Urkundenpublikationen. In: Elektronische Fachinformationssysteme in der Geschichte, Jahrestagung der AGE, 25.–26.11.2006. Hg. von Franz Götz. München 2006 (im Druck).

mittel an die Hand, das Heterogenitäten der Quellencorpora verringert – und damit die Menge der relevanten Dokumente erhöht (recall) – und die Relevanz der Treffer für eine historische Anfrage besser sichtbar macht – und damit den Anteil an irrelevanten Treffern, die der Suchende verarbeiten muss, verringert (precision).

Eine derartige Suchmaschine muss aber mit zwei Problemen fertig werden:

1. Die oben beschriebenen Werkzeuge entfalten ihre Wirkung erst, wenn sie auf passende Texte angewendet werden: Ein für frühneuhochdeutsche Texte getrimmter ‚Graphex' wird auf altfranzösische Texte angewendet kaum sinnvolle Ergebnisse liefern. Sowohl die Implementierung des im Eingang dieses Textes erwähnten aber nicht weiterverfolgten Crawlers, wie die Analyse der gefundenen Texte, müssen darauf geprüft werden, ob es Instrumente gibt, die Sprache der Texte plausibel zu bestimmen. Im hier verfolgten Testszenario mittelalterlicher und frühneuzeitlicher Urkundencorpora haben wir es noch nicht mit den Materialmengen zu tun, die dazu zwingen, diesen Schritt zu automatisieren. Vorläufig wird die Suchmaschine mit den Corpora von Hand gefüttert, das heißt vor der Indexierung des Corpus kann der Bearbeiter entscheiden, welche Art von Transformationen angewendet werden soll.
2. Die Vielzahl der Techniken ist hier weitgehend theoretisch beschrieben. Es gibt einzelne Evaluationen der vorgestellten Techniken,[35] aber viele sind auch noch nicht evaluiert, ja noch nicht einmal implementiert. Das sind insbesondere die Techniken, die sich mit mittelalterlichen und frühneuhochdeutschen Sprachständen europäischer Volkssprachen und des Lateinischen beschäftigten, mithin also mit genau den Sprachen, die im Testcorpus vorherrschen. Es ist also im Rahmen der Tests ein Beispielcorpus zu bilden, in dem die maximale Trefferpräzision und der maximale Trefferumfang bestimmbar ist, um die mithilfe der existierenden und zu entwickelnden Verfahren mögliche Präzision und Trefferumfang damit vergleichen zu können.

Als vorläufiges Fazit bis zur technischen Vollendung der Anwendung[36] kann man aber Folgendes sagen: Die Qualität einer fachspezifischen Such-

[35] Zum Beispiel Strunk, wie Anm. 15, oder Liar-Randell, wie Anm. 13.

[36] Vgl. dazu den folgenden Beitrag von Markus Heller.

maschine für Historiker beruht nicht so sehr darauf, dass die Anbieter umfangreiche Tiefenerschließung leisten und inhaltliche Normen zum Beispiel in gemeinsamen Thesauri verwenden, sondern zum einen auf einer technischen Entscheidung auf Seiten der Anbieter und zum anderen auf fachspezifischer Softwareentwicklung auch im Bereich der Suchmaschinenindices. Die Anbieter wären darauf zu verpflichten, dass die Datenstrukturen von der Suchmaschine ausgelesen werden können, das heißt dass die Informationen möglichst so in XML-Formaten auf den Servern abgelegt werden, dass die Suchmaschine Zugriff auf die XML-Rohdaten bekommt, sowie dass die Daten mit gemeinsamen XML-Vokabularien beschrieben werden. Die Fachinformatiker tragen zum Erfolg einer fachspezifischen Suchmaschine bei, indem sie die notwendigen ‚Übersetzungstools', wie Lemmatisierung, Graphienausgleich, Lautausgleich, Approximation, wo nötig entwickeln, auf jeden Fall testen und schließlich in die Suche integrieren.