Stefan Gradmann,

The Cathedral and the Bazaar, Revisited: Union Catalogs and Federated WWW Information Services

from / *aus*:

Union Catalogs at the Crossroad

Edited by

Andrew Lass and Richard E. Quandt

pp. */ S*. 67-88

Erstellt am 31. März 2005

Impressum

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.ddb.de.

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

This publication is also openly accessible at the publisher's website. Die Deutsche Bibliothek has archived the electronic publication, which is now permanently available on the archive server of Die Deutsche Bibliothek.

Diese Publikation ist auf der Verlagswebsite ebenfalls open access *verfügbar. Die Deutsche Bibliothek hat die Netzpublikation archiviert. Diese ist dauerhaft auf dem Archivserver Der Deutschen Bibliothek verfügbar.*

Available open access / *open acess* verfügbar:

Hamburg University Press / Hamburg University Press
http://hup.rrz.uni-hamburg.de
Die Deutsche Bibliothek archive server / Archivserver Der Deutschen Bibliothek
http://deposit.ddb.de/

ISBN 3-937816-08-9 (print)

Table of Contents

Union Catalogs in a Changing Library World: An Introduction …………….. xi
Andrew Lass and Richard E. Quandt

Part 1 Western Models and Overview

Chapter 1…………………………………………………………………..…. 31
EUCAT: A Pan-European Index of Union Catalogs
Janifer Gatenby and Rein van Charldorp

Chapter 2 …………………………………………………………………….. 51
The Virtual Union Catalog
Karen Coyle

Chapter 3 ……………………………………………………………………… 67
The Cathedral and the Bazaar, Revisited: Union Catalogs and Federated WWW Information Services
Stefan Gradmann

Chapter 4 …………………………………………………………………….. 89
Linking in Union Catalogs
Ole Husby

Chapter 5………………………………………………………………………. 101
Linda: The Union Catalog for Finnish Academic and Research Libraries
Annu Jauhiainen

Chapter 6 ……………………………………..…………………………123
Beyond Technology: Power and Culture in the Establishment of National Union Catalogs
Nadia Caidi

Part 2 Czech and Slovak Union Catalogs

Chapter 7 …………………………………………………………………... 141
The CASLIN Union Catalog
Gabriela Krčmařová and Ilona Trtíková

Chapter 8 ……………………………………………………………………173
LINCA: The Union Catalog of the Czech Academy of Sciences
Martin Lhoták

Chapter 9 ……………………………………………………………………187
CASLIN Uniform Information Gateway
Bohdana Stoklasová and Pavel Krbec

Chapter 10 …………………………………………………………………. 205
The Slovak Union Catalog for Serials
Lýdia Sedláčková and Alojz Androvič

Part 3 Polish Union Catalogs

Chapter 11 …………………………………………………………….…. 227
Are Our Union Catalogs Satisfying Users' Needs?
Błażej Feret

Chapter 12 ………………………………………………………………………. 245
Union Catalogs for Poets
Henryk Hollender

Chapter 13……………………………………………………………………….. 265
Aiming at the Union Catalog of Polish Libraries
Anna Paluszkiewicz and Andrzej Padziński

Chapter 14 ………………………………………………………………………. 281
Implementing KaRo: The Distributed Catalog of Polish Libraries
Tomasz Wolniewicz

Part 4 Hungarian Union Catalogs

Chapter 15 ....………… 297
The Hungarian Shared Cataloging Project: MOKKA
Géza Bakonyi

Chapter 16 ....…………. 305
Subject Cataloging in a Cooperative Cataloging Environment
Klára Koltay

Chapter 17 ....………… 327
Principles of a National Union Catalog: Shared Cataloging in
a Small Country
Erik I. Vajda

Part 5 Baltic Union Catalogs

Chapter 18 ………………………………………………………………………. 341
Using a Shared Cataloging System: The Estonian Approach
Janne Andresoo and Riin Olonen

Part 6 South African Union Catalogs

Chapter 19 …………………………………………………………………… 361
A National Union Catalog for Shared Cataloging and
Resource Sharing by Southern African Libraries
Pierre Malan

Chapter 20 …………………………………………………………………… 381
Regional vs. National Union Database Development:
The GAELIC Perspective
D. L. Man and Lettie Erasmus

Chapter 21 …………………………………………………………………… 407
Why the "Big Bang" Did Not Happen: The CALICO Experience
Amanda Noble and Norma Read

Contributors …………………………………………………………………… 435

Conference Participants …………………………………………………… 441

Chapter 3
The Cathedral and the Bazaar, Revisited: Union Catalogs and Federated WWW Information Services[1]

Stefan Gradmann

1 What This Paper Is Not About ...

In the past 30 years, which have witnessed the advent of library automation, numerous speculations have been published, most of them concerned with either the imminent death of libraries that were seemingly doomed to be replaced by some omnipotent electronic successor, or with "business as usual" proclamations basically stating that libraries—even if electrified to the extreme—would ultimately continue to function the way they had done for centuries.

In the past decade, which has seen the ascent of the Internet, such speculations have been heavily intensified and increasingly focused on aspects of information technology and the information economy exemplified by the information and communication models of the World Wide Web. These speculations have led to sometimes astonishing and radical conclusions and assertions; for example, WWW-based information services such as Google or Yahoo! were supposed to take over library

[1] Although the relation of this paper's title to Eric S. Raymond's essay on "The Cathedral and The Bazaar" is explained in more detail in chapter 4, it should be made clear from the beginning that the title of this paper alludes to this essay.

functions altogether, or librarians were expected to catalog all quality information on the Internet.

None of these radical changes have actually taken place—and yet, a lot has changed. And the speculative striving to make projections and predictions in this field has certainly been fed by the common feeling that something fundamental is happening to our paradigms and techniques of dealing with information, and to our concepts of information themselves. Still, in a period of profound uncertainty, projections that make use of metaphors of the past to predict the shape of future electronic information landscapes do not, in essence, transcend the intellectual qualities of a Star Trek movie, as tempting as they may be.

The present paper tries to avoid bad library science fiction in general, and predictions as mentioned above in particular. Instead, I assume that we can make hardly any valid statements except those concerning the very near future, but that it may be useful instead to describe as precisely as possible what changes and differing approaches can currently be identified in some fields of scientific information technology and economics, and to try to reach an adequate level of abstraction in the description of such changes and differences.[2]

[2] When using the term 'WWW-based information services' in this paper, I am referring to services such as the NASA Astrophysics Data System (ADS) or the NEC Research Institute Research Index, as well as to more generic services such as Google or Yahoo. ADS and the NEC Index are well presented and discussed at length in a very detailed presentation given by Gerry McKiernan at the WilsWorld '02 conference (McKiernan 2002). In the announcement of this presentation on the conference website, the following assertions are made: "In recent years, a number of experimental and operational Web-based information systems and services have emerged that offer advanced and novel features, functionalities, and content. In this presentation, a variety of these innovative services will be profiled, as will their associated technologies. The potential impact of these systems on the development and enhancement of commercial and library information services will also be reviewed and discussed." However, the latter aspect, although announced, is not really discussed in the presentation itself. The present paper therefore can be seen as a complement to McKiernan's work, which is very extensive as far as WWW services are concerned but quite restricted as regards libraries. As a consequence, librarian aspects are stressed to a higher degree in the present paper.

2 ... And What This Paper Does Attempt

This paper is mainly concerned with the differences between the ways in which information is organized; on the one hand in electronic library catalogs and, more specifically, within electronic union catalogs,[3] and, on the other, in genuine WWW-based information services. The main goal here is to identify some of the fundamental differentiating characteristics, whether in terms of the information entities themselves, the way they are conceptualized or the way they are referenced and their identity is established in their respective contexts, or in terms of the actual modes of collaboration within librarian union catalogs and WWW-based information services.

A better understanding of such differences may in turn help us better understand what actually happens within the overlapping zone between both worlds: whenever a union catalog points to information in the WWW domain, or whenever an Internet search engine encounters catalog applications with their index files and librarian metadata, concepts and mechanisms from two different paradigms of information organization are made to coexist and together create a hybrid setting that can be understood better if the originating contexts of the respective mechanisms are kept in mind. The point here is to identify differences and relevant questions (rather than answers) by describing the often complex relation between electronic union catalogs and WWW-based information repositories, in terms of mutual redundancy, competition, and (sometimes and hopefully) convergence.

[3] The term ‘catalog’ is used as a synonym of ‘electronic catalog’ throughout this paper, which is thus implicitly restricted to electronic metadata as part of librarian or WWW-based information infrastructures. The author is aware of the segment of union catalog reality that is thus deliberately excluded from the scope of this paper—on the other hand, a comparison of traditional union card catalogs and WWW-based information services really would not have made much sense.

And if some useful hints can be given at the end of the argument concerning the possible ways for both worlds to coevolve in the near future, this paper will have reached its (modest) objectives.[4]

3 The Risks of Pragmatism: 2½ Examples

In order to illustrate the need for conceptual clarification, one that is of practical interest, it may be useful to consider two concrete examples taken from the authors' daily working context. Both examples are concerned with the coexistence of library catalogs and WWW-based information services.

"Make the WWW Part of the Catalog"

The first example is concerned with a situation most readers of this paper, at least those from the 'hybrid' library world, will be familiar with: the need to present coexisting printed and electronic manifestations of works to library users in a consistent service model, more specifically in the area of printed and electronic journals.

Until recently, holdings of electronic journals have not been systematically integrated into library union catalogs, even though many participating libraries spend increasing sums of money to enable their users to access such resources via licensing agreements. This has led to a situation where libraries have started to build vast link repositories for electronic journals outside their respective OPAC environments and thus, along with these developments, a very impressive repository of electronic journals metadata and of library 'holdings' (in terms of license agreements) has been built on a national scale in Germany (e.g. the *Elektronische*

[4] It is worth noting that this paper is written from the point of view of a librarian; the author—presently active in the gray area shared by both worlds—has a strong background in the union catalog community, and the present audience are librarians and technicians active in union catalog environments. The paper may thus fail to identify some points that are of specific interest to the W3C community, while it probably overemphasizes issues that may seem completely trivial to those who hold a primarily WWW perspective.

Zeitschriftenbibliothek or EZB).[5] From a user perspective, the major unsatisfying aspect of this situation is the fact that, depending on whether a printed or an electronic resource is to be retrieved, different 'catalog' environments have to be used. There is no way of retrieving both kinds of resources using one single interface. The problem is common to all 'hybrid' library architectures and systematically recurs at all scales—from the context of a single library to the issue of how to relate resources like CORC and WorldCat to each other.

One of the practical responses of the library community to this situation has been to try to integrate as many of the pointers to Internet resources into the library information systems, and thus to make parts of the WWW a part of their catalogs. One of the union catalogs the author of this paper is working with is about to move in that direction. One idea that is currently discussed within this union catalog is to simply add all metadata from EZB (the nationwide repository) to the union database, thus creating holdings data for the participating libraries and thereby ensuring replication of these metadata, together with the 'holdings' information, to the participating libraries' OPAC environments.

However (and quite paradoxically), this creates one specific problem in the case of freely accessible electronic journals such as D-Lib Magazine or First Monday: no license agreements are necessary to access these resources, and as a consequence no library-specific 'holdings' information can automatically be generated for these resources. Here again, a practical solution has been devised: simply add 'holdings' for all libraries participating in the union catalog in the case of such free electronic resources.

The resulting situation is a practical solution to a specific problem that immediately generates numerous new problems of its own. For example, the use of 'holdings' information, which is itself a questionable construct as far as licensed electronic material is concerned, almost completely loses consistency with such an approach. We will come back to this issue as well as to the overall problem of inconsistency later on. At this point it is

[5] "Electronic Journals Library" would be a rough English equivalent. EZB can be accessed via http://rzblx1.uni-regensburg.de/ezeit/

sufficient to highlight the problematic nature of an approach that tries systematically to integrate pointers to WWW resources in library catalogs.

“Make the Catalog Part of the WWW”

The alternative (or possibly complementary) approach is often considered when it becomes apparent that library information resources tend to be ignored within the overall information economy of the WWW. The culprit here is the so called ‘hidden Web’; metadata contained in library catalogs are mostly ignored by the leading search engines, for the simple reason that the application layer used to access these records is not transparent for generic WWW technology, and therefore ‘hides’ the resources it should make accessible.

Solutions to this problem are often discussed in terms of making library catalogs more systematically ‘WWW-transparent’ by making catalogs more generally part of the WWW. The overall aim of such strategies is to ensure the presence of metadata from library environments (OPAC or union catalogs) in result sets generated via WWW-based search engines, and to eventually ensure that these sets receive a high ranking because of their high granularity and the quality of the indexing information they include.

While seemingly logical, the consequences of such a strategy could be far from desirable, especially if such an approach were adopted by all major university and research libraries plus a significant number of union catalogs. The first and most striking effect would be extreme redundancy of information, quickly approaching unwanted levels of information entropy; what user would actually want to be overwhelmed by thousands of metadata records pertaining to James Joyce’s “Ulysses” from libraries all over the world when doing a search for “Ulysses” in Google? Moreover, users would then be confronted with result sets that pointed to information objects in very different ways; while in some cases direct access to an information resource via a URL pointer may be possible, in the case of metadata originating from libraries the user would be faced with differing and various types of mediated access, an effect that would certainly put into question the results of a strategy that reveals library resources.

More Integration Strategies … and the Need for Distinctions

A third prominent integration strategy deserves mentioning here: the systematic use of library systems as gateways to WWW resources.[6] A more generic, and possibly more appealing, variant of such an approach involves all integration strategies that are built around concepts of open and context-sensitive linking as part of library information infrastructures.[7]

Without going into detail at this stage of the argument, it should be said that any over-pragmatic strategy that simply combines library and WWW resources, yet remains unaware of the fundamental differences of the respective information resources, is unlikely to produce satisfying long-term results. This observation does not question the actual need for integration strategies (and we will come back to this point later in this paper), but rather highlights the extent to which strategies need to be built on clearly established distinctions between the information landscapes we ultimately seek to combine.

The following sections of this paper are concerned with such distinctions. For the sake of clarity I will, in what follows, sometimes deliberately ignore 'hybrid' infrastructures. Only after having established the basic, underlying, differences will I reintroduce such hybrid (and mostly secondary) settings.

4 Differing Basic Elements and Concepts: Entities, Pointers, Identities

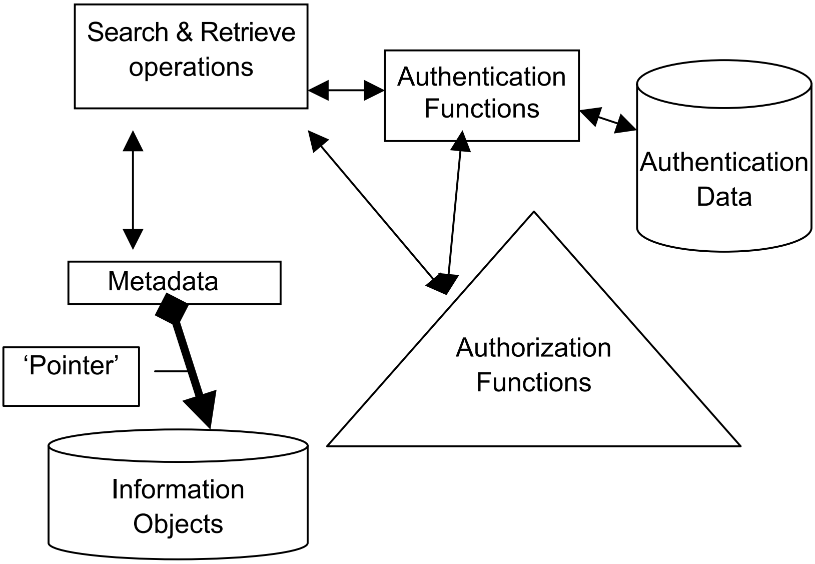
Library and union catalogs on the one hand and WWW-based information resources, such as Yahoo or Google or any repository built on a metadata harvesting protocol [specified, for example, by the Open Archives Initiative

[6] S. Thomas has proposed this, for instance, in her reflections on "The Catalog as Portal to the Internet" (Thomas 2000) that have provoked some interesting discussion (cf. Schottlaender 2000).

[7] Such concepts are presented in detail in the contributions from H. van der Sompel mentioned in this paper's bibliography.

(OAI-PMH)], on the other hand share a number of basic instances and entities as part of their information infrastructure. They mostly contain a distinct metadata layer including pointers to the actual information objects, together with a user interface typically including support for search and retrieval operations. Furthermore, some means of identifying users and information objects must be present somewhere within the respective system: the authentication layer, together with functions that are used to determine what kind of operations a given user (or class of users) may apply to a given information object (or class of objects)—the authorization layer.

From a bird's eye perspective, information systems originating from the library world and from the WWW do indeed have a lot in common. The following diagram visualizes the basic components mentioned above and could be used to describe library information systems and genuine WWW-based systems alike.



However, closer to the ground some basic differences begin to appear. What follows is a closer look at these differences that would be described as ‘distinctive’ (as opposed to variations in detail and granularity).

It may come as a surprise that relatively few of such distinctive/ fundamental oppositions can actually be identified in the areas of search retrieval and of ‘bibliographic’ metadata, or that an assumption is being made here that the main differences reside in the ways information objects themselves are conceived, in the way access to these objects is organized and in the mechanisms of authentication and authorization.

However, search interfaces for electronic library catalogs are a relatively young component of libraries and library cooperation, and from the beginning of their short history have evolved much more in line with features and requirements of generic, non-librarian automation technology than, for example, the books themselves, the nature of which has been shaped over centuries long before the birth of electronic information processing.

As for ‘bibliographic’ metadata, the above assumption may be more controversial, especially within the library community; after all, many librarians still regard the production of metadata (in the sense of cataloging) as the very heart of their business, and it may be hard for them to admit that vital issues may well be defined outside the scope of cataloging principles and practice. The assumption is retained nevertheless: many of the guiding principles of cataloging, that had their origins in the sequential organization of card catalogs and that have initially been preserved in electronic cataloging environments, have either vanished or are at least being seriously reconsidered. And even in those cataloging databases that still contain important layers of data oriented towards card catalog production, the creation of a Dublin Core-like interface is comparatively straightforward. This is much easier, anyway, than converting data the other way round; trying to generate traditional cataloging data from a Dublin Core source would probably turn out to be much more of a challenge, if anyone were even interested in the exercise at all.

Furthermore, even the apparently most significant structural differences in the metadata area, such as the ‘holdings’ or ‘copy’ notion of library catalogs that has no real equivalent in WWW-based information services, can be addressed more appropriately as an aspect of pointing and access to information objects (see below.) And so, while I have devoted a good deal of attention in this paper to the topic of metadata, I will continue to maintain that the crucial differences between Web-based information systems and traditional ones do not lie in this area.[8] Instead, some very evident and fundamental differences can be identified in the remaining three component areas. This involves nothing more than recalling some obvious, though often forgotten, truths regarding the relation of library catalogs and WWW-based information services.

Books vs. Digital Information Objects: The Basic Information Entities

The first point to be aware of is the profoundly differing nature of the information objects involved. Library catalogs and automation systems are designed to contain descriptive cataloging records for books and book-like printed information, together with pointers to the actual physical copies of these as present on library shelves. WWW-based information systems are designed to contain identifying (and some basic descriptive) information pertaining to electronic information objects (and most typically hyperlinked objects stored somewhere in the network at any location that can be addressed via HTTP), together with pointers to these objects.

It is worth briefly recalling three of the many consequences that have already received their due of scholarly attention.[9] The first consequence is that paper books and other paper publications are combined presentation and storage media, where the display of information is altogether visual and the content is physically tied to the paper and the pages of the publications. On the other hand, with electronic publications storage and presentation are separate. The second consequence is that additional electronic devices are

[8] This assumption does not contradict assertions made by the present author in an earlier paper (see Gradmann, 1998). The distinctions made there are less concerned with actual bibliographic metadata than with the respective contexts of use and the originating communities of these metadata.

[9] The contributions contained in TEXT-E 2003 are an excellent starting point for entering the relevant scholarly discussion in the area of both semiotics and information technology.

required for access to the content of digital information objects, whereas books can simply be read using our human senses. Finally, the third consequence is that automated operations on content are possible in electronic information objects in a way that is inconceivable for printed material.

The fact that many digital information objects are still modelled upon the example of printed books should not make us forget the fundamental differences between them: digital information objects will evolve from book-like analogies into new forms of information modeling, forms we do not yet have names for, and this fact is about the only excuse for using such terms as 'e-books'.[10]

Shelfmarks vs. Links: The Pointers from Metadata to the Information Objects

The second area where both worlds differ substantially is concerned with the way they organize access to the actual information objects for their respective user communities. To state it simply, library-based information systems are based on the idea of mediated access, whereas the original principle of WWW-based systems is one of direct, instant access. The principal reason for this is the fact that librarian information objects (books and the like) simply are not kept within the information system (the catalog) but on the library's shelves, whereas in the case of WWW information systems the information objects are technically (or at least can be) part of the system.

This seemingly trivial observation has two very important consequences for the respective architecture of these information systems:

In a library information system, the user is interacting with metadata on all levels: not only with 'bibliographic' metadata, but also with a metadata substitute for the real information object within the information system, the copy record, which in turn contains a pointer to some instance outside the

[10] For the very same reason, the term 'digital library' can be considered as intellectually somewhat dubious: an institution either deals with books (and then can be called a library) or with digital information objects (and why should it then be called a library?).

system that will mediate access to the information object for the user. WWW-based information systems have no equivalent of this ‘copy’ or ‘holdings’ layer, because the information objects themselves are a technical part of the system.

As a consequence, the pointers to the actual information objects have fundamentally different functions within the respective systems: the ‘shelfmark’ or ‘lending number’ pointers point to some instance outside the library catalog (a librarian or a lending module) that will interpret it and finally grant access to the information resource in a way the information system has no knowledge about, whereas the URL pointer (or any technical successor in WWW-based information architectures) basically points to the information object itself that is technically kept within the system (not necessarily stored there physically but part of the system’s technical architecture).

These observations account for numerous functional and technical incompatibilities between library and WWW information systems, and it is important to fully understand their implications before combining working principles from both worlds. The ‘copy’ level of a library system is difficult to translate to the WWW world, and the pointers to the actual information resources react to very different functional requirements.

The latter difference in particular needs to receive additional attention. The ‘shelfmark’ string in the library system may contain almost any information that can be interpreted by humans, from the actual shelfmark (“X 1989/1234” or the like) to information like “go to room 202 and ask there,” or even simply “go and ask the librarian”. And should the copy or call number be erroneous, the lending system module will not recognize it—but ultimately some librarian will be there to help with the matter; the pointer goes outside the system, and the responsibility for resolving the pointing information is outside the system as well. This is the reason why our union catalogs and library OPACs containing such an amazing quantity of incorrect shelfmark information nevertheless continue to function.

The situation is radically different with URL pointers within WWW-based information systems; one character missing in a URL will simply generate code 404 and not reveal any information beyond this error message. Mostly, no external instance can be called upon to correct the pointing information; the correctness and reliability of the pointer are a vital constituent of the information system. This is why the protocols for constructing and resolving HTTP pointers are relatively strict and elaborate (even though insufficient: there will be successors to URLs as we know them today!) whereas shelfmarks and copy numbers are variable string values with almost no restrictions at all.

Of course, notions of direct access to resources have been added to library-based systems in the recent past, and access control mechanisms and restrictions have been implemented in various ways in WWW-based systems—but still the original governing principles of mediated vs. direct access have been at the origin of the respective systems' design and of the pointing mechanisms used. This is an important fact to remember when one tries to understand what happens to Internet pointers in library systems.

Identity and Credentials: Authentication and Authorization

Instances that are taken for granted in one information environment may cause near-metaphysical problems in another.[11] This fact can be illustrated with one simple yet striking example (considering the way persons and information objects are identified in both worlds and the way authorization to use a given resource is determined).

In the 'real' world, when trying to establish the identity of a library user, one simple and effective way would be to ask for a passport or ID card. A certain number of additional checks can then be performed; if the ID-document bears the same name the user claims to have and the photograph therein bears at least some resemblance to the owner, and, furthermore, the document has been issued by a trustworthy authority, the librarian may decide that the identity of that user has been established to a sufficient degree. And if that user wanted to borrow a book reserved, for instance, for local residents, a simple check of the address in the user's ID document would quickly solve the issue. Authentication and authorization can thus be established to a sufficient degree using simple and robust techniques.

[11] A very sound introduction to the issues of authenticity and integrity is given in Lynch (2000).

However, one of the key factors for the efficiency of this approach is indicated by the words “to a sufficient degree”: the user’s identity is never established with 100% certainty, and there is no need to do so, since a complex set of context information is combined to dynamically evaluate the level of trust required and the degree of certainty needed as a consequence.

The situation gets far more complex once we look at digital authentication scenarios: in this context; identification and authentication information must often be established 100% or simply fails to be established at all. In binary logic, identity is either established or not, and no such notion as “to a sufficient degree” can ease the task. As a consequence it has to be established to a degree that is almost never required in ‘real life’ environments. Or, as Clifford Lynch puts it:

> In the digital environment […] computer code is operationalizing and codifying ideas and principles that, historically, have been fuzzy or subjective, or that have been based on situational legal or social constructs. Authenticity and integrity are two of the key arenas where computational technology connects with philosophy and social constructs. (Lynch, 2000)

And the annoying fact is that this holds not only for persons operating in digital information environments, but for digital information objects as well: the identity and integrity of a printed book is far easier to determine than the identity and integrity of its digital equivalent.

Moreover, while such information is far more difficult to establish in digital environments, ambiguous authentication and identification information can completely block a digital information system, while some flexible strategy of dealing with this lack of information in conventional information environments can almost always be devised.

As a consequence, tremendous efforts have to be made in digital information environments in order to determine what kinds of operations a given user may perform upon a given object, and this places constraints upon the way such environments function, a situation that is almost unknown in ‘conventional’ library contexts.

5 Differing Modes of Collaboration: The Cathedral and the Bazaar

In addition to the differences in the two types of information systems mentioned above, important differences can also be located (and must be accounted for) in the way the respective communities cooperate; library union catalogs and federated information environments on the WWW have very different traditions of organizing and experiencing cooperation.

The first striking, and almost trivial, difference concerns the types of cooperating partners: libraries—as different as they may perceive themselves to be—are a more homogeneous group of organizations by far, both in terms of decision making and in terms of user requirements, than the heterogeneous groupings of companies, individual scientists and more or less formally organized parts of the academic community that typically make up the user/production base of federated information services on the WWW.

This basic difference leads to an important secondary observation: rules and guiding principles, as well as common policies for information management, can be imposed much more effectively in a relatively uniform and close user group such as the library sector, while the typical setting within the Internet can never be prescriptive to such a high degree.

The difference is also similar, to some extent, to those described by E. Raymond in his essay on "The Cathedral and The Bazaar" between different modes of collaboration and differing modes of communication when comparing the traditional community of software engineers, for whom the 'cathedral'-building metaphor is used, and the open source development community, to whom the 'bazaar' metaphor is applied.[12] What follows is a brief outline of some of the directions that a closer analysis of this issue should pursue.

If one examined the respective ways in which a WWW development and library staff are collaborating, one would immediately find that the

[12] Raymond then goes further than I want to go here: he proclaims the bazaar model to be more powerful than the cathedral model, whereas I have no intention of transposing that conclusion to the context of this paper. This is where the reference to Raymond's paper has its clear limits.

librarian collaboration model is almost entirely obsessed with rules, whereas such rules hardly play an important role in the WWW environment, where their structural position is taken over by protocols. Likewise, library environments tend to be highly prescriptive as compared to the rather experiment-oriented WWW environments. And finally, library settings seem to have a strong tendency to establish pre-coordinating frameworks, whereas WWW environments tend to assemble collaborative resources first and then post-coordinate their actual collaborative use.

In the field of communication modes, similar observations can be made. Whereas library communities tend towards hierarchical communication models, WWW communities have a rather flat information culture. The channeled vs. broadband perceptions of the communication lines seems to be another relevant distinguishing factor. And one could also argue that the way of organizing communication in libraries is very much oriented towards aggregation of information, whereas the WWW communication paradigm seems to be heavily oriented towards distribution of information, the two worlds thus focusing on two very different aspects of communication practice.

One could even speculate on the differing modes of perception and of mental organization of information units that seem to be at the roots of the respective communities, and might then end up reflecting on the community difference in terms of identity vs. difference, but I will leave such philosophical ruminations for another occasion.

The point now is to create an awareness of the ways in which respective communities differ ‘culturally,’ in their modes of communication and of collaboration. This, together with the insights made in an earlier section, provides sufficient basis for a discussion of possible scenarios for the future relations of these two cultures.

6 Modes of Coexistence: Future Choices and Bridging Concepts

Coexistence? Coexistence!

It should be clear by now how the recognition of the fundamental differences between the two information paradigms helps us to understand better the often unexpected effects produced when transposing objects and methods from one world to another. While such combinations of objects and methods stemming from very different contexts cannot be avoided altogether and, in order to be sure, must be accounted for systematically in 'hybrid library' settings, it is still useful to keep in mind the side-effects that are produced with such an approach.

The recognition of these differences can also help conceptualise the possible future relation between library catalogs and WWW-based information services, without falling back into the bad habit of excessive and fruitless prediction-making mentioned in the beginning of this paper.

In this attempt to take a modest look ahead, I make two assumptions. First, that libraries with their catalogs and WWW-based information architectures will coexist for quite some time, and even though one paradigm of information organization may eventually gain the upper hand, such a possible future situation is far beyond the scope of this paper. The second assumption is less evident: it is that real choices can actually be made in organizing this coexistence and that the coevolution of both paradigms is not governed by some obscure cybernetic natural law that causes fatal things to happen. The end of this paper is devoted to actual choices we could, and should, make in this area.

Redundancy, Competition, Convergence, Integration

The possible relations of present and future coexistence can be described using (at least) four different concepts. To begin with, two of these are rather unproductive and ultimately inappropriate. Redundancy may be the least desirable one: modeling the same information objects redundantly in two contexts is expensive, inefficient and carries a high risk of long term overall inconsistency. This is true for all approaches resulting in redundancy, be they based on parallel, unconnected activities in both environments that are not acting in concert in any way, or on data replication scenarios. Competition is not an appropriate characteristic either, even though it may appear inevitably in many political contexts where both paradigms are competing for the same resources (usually money) and therefore are perceived as functionally and technically competing, although they serve fundamentally different needs.

Two other characteristics could be more fruitful and may help to establish productive and realistic objectives. Provided the fundamental conceptual differences between both paradigms are well understood, their relation could evolve either in terms of convergence or of integration. Convergence in this context would mean that both worlds move towards the same objectives, getting continuously closer to each other and possibly creating more and more overlapping areas without, however, blending both paradigms altogether. Catalogs and WWW-based information systems remain clearly discernible worlds in this approach. Integration, in contrast, would mean that both worlds are actually blended into something new, embracing both paradigms and serving the needs of their respective communities in one common approach of information modeling.

Examples of all four characteristics on organizing coexistence can be found in our present professional experience, and most readers of this paper will be able to quickly identify examples of redundant, competing, converging, or integrating scenarios in their own working context. The author of this paper is convinced that (at least) these four scenarios of coexistence will remain valid options in the short term, and that it is up to the stakeholders of both worlds to make their choices among them. Such choices will be triggered by many factors: money, politics, economic interest, to name just a few powerful ones outside the scope of what readers of this paper will typically be able to influence. There are, however, two concepts in the area of information architecture that may help to orient this coevolution in the direction of convergence or integration, and the promotion of these two concepts would be a very useful contribution of the union catalog community to the shaping of future cooperative scenarios.

Bridging Concepts: FRBR and the Semantic Web

Two important bridging concepts in that sense might well be the metadata layering model expressed in IFLA’s “Functional Requirements of Bibliographic Records” (FRBR) and concepts currently taking shape in the “Semantic Web” approach.[13] The general reason is that both concepts raise the level of abstraction concerning information entities that are present in

[13] This assumption is by no means meant to be exhaustive: there are certainly more examples of bridging concepts, and the author merely tried to identify two prominent ones.

both information paradigms sufficiently high in order to potentially embrace both worlds, and thus may play an effective bridging role.

Semantic Web technology and, more specifically, methods based on semantic Web ontologies are likely to make new and productive use of the fine-grained semantic metadata that libraries have been traditionally producing, thus enhancing the taxonomies of semantic Web ontologies. Assertions based on the use of classifications and indexing schemes could easily be transposed into taxonomic elements that, in turn, greatly broaden the basis to which inference rules can be applied. This results in a much richer taxonomic base for ontological operations, and could well generate an ongoing process of library work being fed into semantic Web ontologies.

Likewise, the integration of semantic Web techniques in library catalogs, not only for search and retrieval operations but also, for instance, to generate proposals for classification attributes using inference rules, may well help a lot in everyday library work: a rule of the type “If a work by a given author has a given classification element associated with it and if the publication year of another work by an author with the same name is adjacent, the same classification element is likely to apply to this item” would probably yield useful and time-saving classification proposals for newly cataloged items.

It is assumed here that semantic Web-based approaches will primarily contribute to the dynamics of convergence.

The FRBR model that results in a layered metadata architecture has the strategically important advantage of making possible a combination of metadata architectures typical of library union catalogs (and as discussed above in section 3) and of the ‘flat’ metadata models that are typical for WWW information architectures. As a consequence, applying FRBR-based approaches to the development of their catalogs, librarians could substantially decrease the annoying effects that were described above and that today contribute to keeping library metadata resources within the ‘hidden Web’.

Establishing coherent unified concepts of what semantic entities, expressions/manifestations and item derivates actually are and relating these in one model that makes ‘hybrid’ information settings appropriately conceivable is one of the major advantages of FRBR. Clearly, approaches based on the FRBR model probably have a very high integrative potential.

To conclude, while it does not seem very wise to predict future developments too emphatically, library and WWW communities would probably be well advised to invest concerted efforts in semantic Web technology and in hybrid information models based on the FRBR-approach.

References

Gradmann, Stefan. “Cataloging vs. Metadata: Old Wine in New Bottles?” In *64th IFLA General Conference August 16–August 21, 1998, Proceedings*. Online at http://www.ifla.org/IV/ifla64/007-126e.htm. Also in *International Cataloging and Bibliographic Control* 28.4.1998: 88–90.

Lynch, Clifford A. “Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust.” *Authenticity in a Digital Environment*. Washington: Council on Library and Information Resources, 2000: 32-50. Online at http://www.clir.org/pubs/reports/pub92/lynch.html.

McKiernan, Gerry. “eProfiles. Innovative Information Systems and Services.” Online at http://www.wils.wisc.edu/events/wworld02/present/eProfiles.ppt.

Raymond, Eric S. *The Cathedral & the Bazaar*. Beijing [etc.]: O’Reilly, 1999.

Schottlaender, Brian E. C. *Commentary on “The Catalog as Portal to the Internet” by Sarah E. Thomas*. Washington: Library of Congress, 2000. Online at http://www.loc.gov/catdir/bibcontrol/schottlaender_paper.html.

Text-e: *Le texte à l’heure de l’Internet*. Sous la direction de Gloria Origgi et Noga Arikha. Paris. BPI, 2003. Online at: http://www.text-e.org/.

Thomas, Sarah E. “The Catalog as Portal to the Internet.” Washington: Library of Congress, 2000. Online at http://lcweb.loc.gov/catdir/bibcontrol/thomas_paper.html.

Van de Sompel, Herbert, and Oren Beit-Arie. “Open Linking in the Scholarly Information Environment Using the OpenURL Framework.” *D-Lib Magazine* 7:3 (March 2001). Online at http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html.

Van de Sompel, Herbert, and Oren Beit-Arie, “Generalizing the OpenURL Framework beyond References to Scholarly Works: the Bison-Futé Model.” *D-Lib Magazine* 7:7/8 (July/August 2001). Online at http://www.dlib.org/dlib/july01/vandesompel/07vandesompel.html.