Lorcan Dempsey
Eric R. Childress
Carol Jean Godby
Thomas B. Hickey
Andrew Houghton
Diane Vizine-Goetz
Jeff Young
OCLC Research

Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape

Note: This is a revised pre-print version of a paper forthcoming in *LITA guide to e-scholarship* (working title), ed. Debra Shapiro. Date of this version: February 2005. Please cite the published version; a suggested citation appears below.

Abstract

The academic library is not an end in itself. It supports research, learning and scholarship, and it must adapt as research and learning behaviors change in a network environment. This paper briefly considers some of these issues, and takes them as its context, but quickly moves to a very specific emphasis. It considers how such library responses create new metadata management and knowledge organization questions, and it then outlines some of the work in OCLC Research which responds to these issues.

Suggested citation:
Dempsey, Lorcan, Eric Childress, Carol Jean Godby, Thomas B. Hickey, Andrew Houghton, Diane Vizine-Goetz, and Jeff Young. c2004-05. “Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape.” Forthcoming in *LITA guide to e-scholarship* (working title), ed. Debra Shapiro. Available online at: http://www.oclc.org/research/publications/archive/2004/ dempsey-mslitaguide.pdf (PDF:824K/25pp.)

The academic library is not an end in itself. It supports research, learning and scholarship, and it must adapt as research and learning behaviors change in a network environment. The papers in this volume give a good sense of the challenges posed by such developments and the manifold library response.

This paper briefly considers some of these issues, and takes them as its context, but quickly moves to a very specific emphasis. It considers how such library responses create new metadata management and knowledge organization questions, and it then outlines some of the work in OCLC Research which responds to these issues.

OCLC Research is a division of OCLC Online Computer Library Center, Inc. Its primary mission is to expand knowledge in support of OCLC's public purpose of helping libraries serve people. We do this through applied research and development into issues of information management and use.[1]

1 Changing patterns of research and learning

We establish some context through the use of two pictures. This section is short as these issues have been well covered elsewhere in this volume. There is a fuller discussion of these pictures in the recently published OCLC Environmental Scan.[2]

The first is adapted from UKOLN work reported by Liz Lyon.[3]

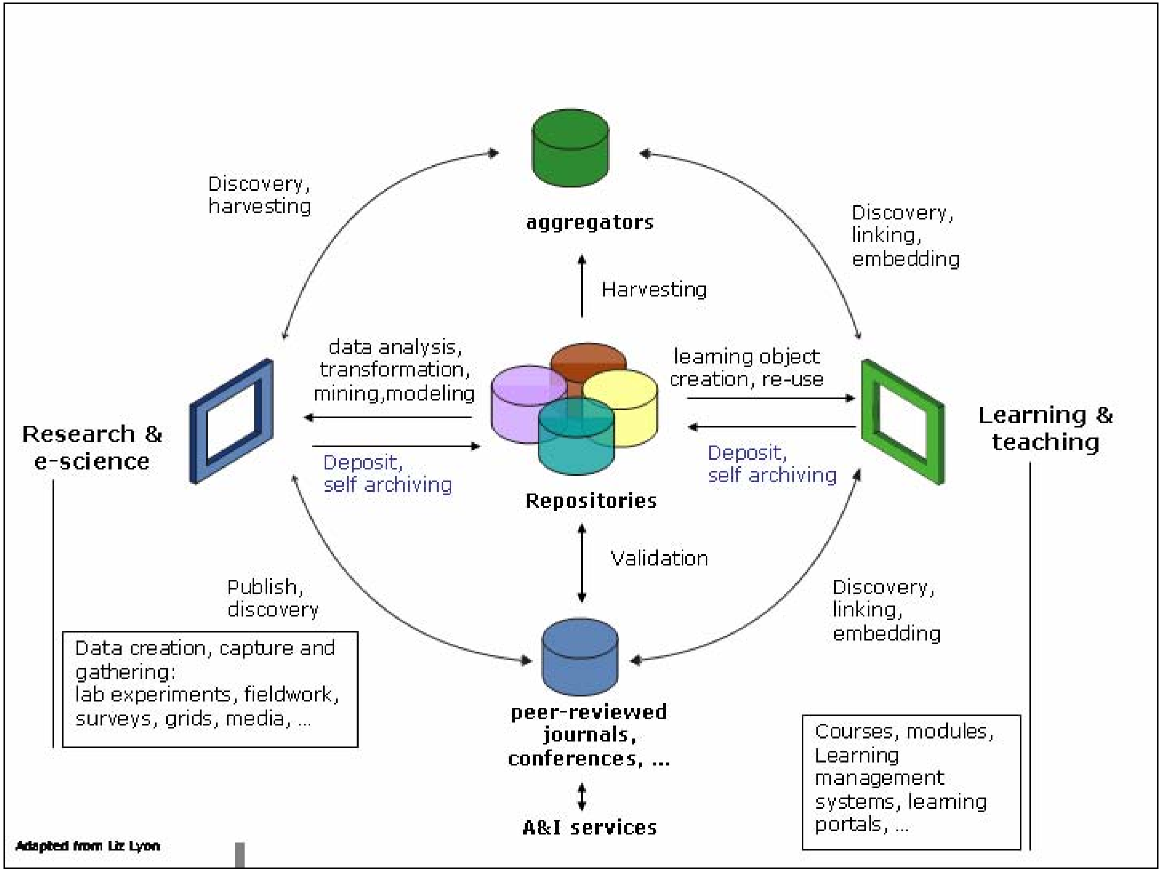


Figure 1: New patterns of scholarly communication

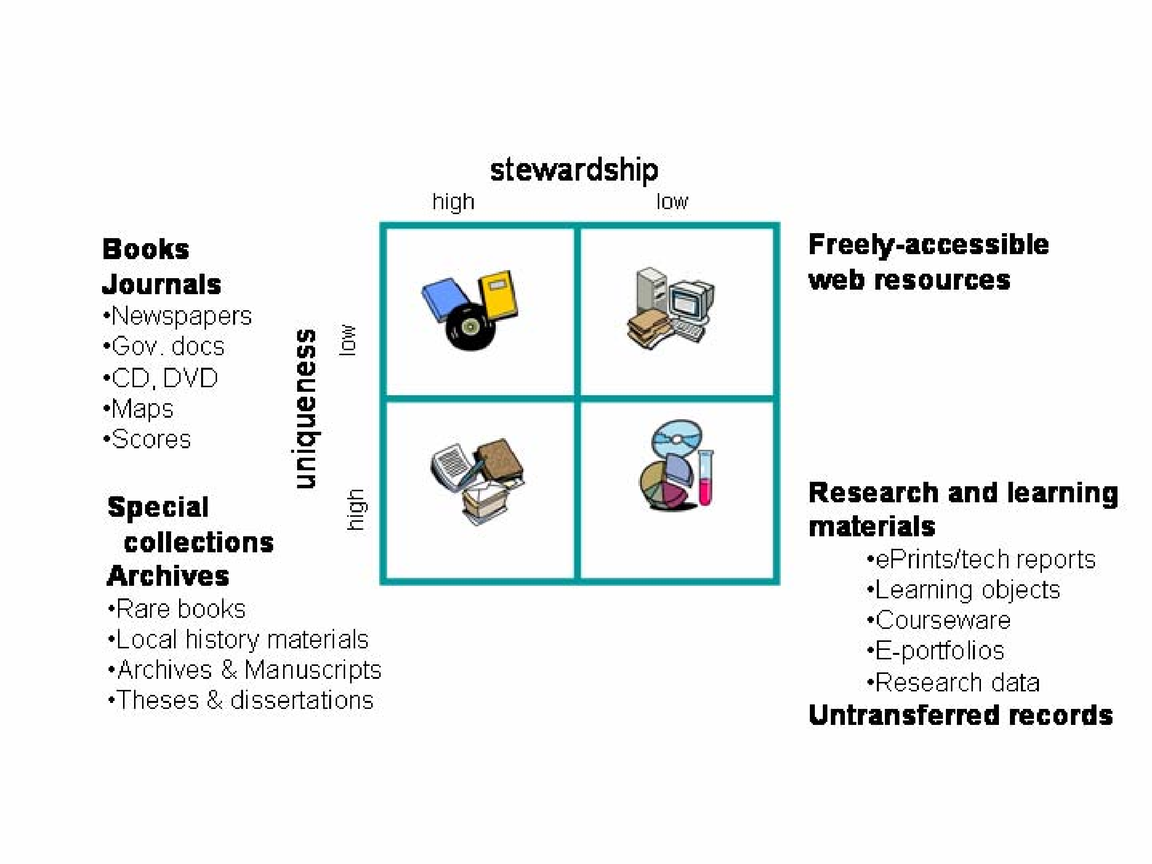
Figure 1 is a simple schematic of how research and learning behaviors are manifesting themselves on the network. In summary, data will potentially flow through a variety of channels; data is manipulated, analyzed and reused; and new roles and responsibilities are being explored. We are now using ‘institutional repository’ as a general summary label for a range of supporting services the library might offer in this environment, working with faculty to care for and disclose a dispersed, complex resource. Community services such as arXiv, Merlot, and CiteSeer have emerged in a short space of time. Multiple repositories of research and learning materials are developing in different organizational and service patterns.

‘Explored’ is the keyword here. One is reminded of the William Gibson quote, “the future is here, it is just unevenly distributed.” And, of course, that unevenness is a complicating factor as change does not allow a complete reallocation of resources and attention to the new. As we move forward the library will be engaging with different communities of practice which may have variably well-formed curatorial expectations with regard to their data, variably achieved technical frameworks, and different levels of computational or applications sophistication. For example, there are some well-developed organizational frameworks within which social science data sets are managed.[4] There is a standards framework for geospatial data.[5] There is a growing body of practice and specifications surrounding the management and use of learning materials in automated environments.[6]

The second picture - Figure 2: the ‘collections grid’ - was developed to think a little bit about the different forms of attention and care different collections and their use call forward.[7]

The grid is a schematic way of thinking about categories of collection of interest to libraries. The vertical axis of the grid represents materials’ degree of uniqueness; the horizontal axis represents the degree of stewardship a library provides.

By ‘stewardship’ we wish to suggest the degree of curatorial care that materials receive by libraries. Highly stewarded materials are collected, organized, and preserved. Thus, ‘left of the line’ materials tend to be in library collections, and subject to traditional practices of library management, organization and access. By ‘uniqueness’ we wish to suggest the degree to which materials exist in more than one collection. Thus, ‘below the line’ materials are unique or rare, are currently relatively inaccessible, and are often specific to institutions rather than more widely published materials. This gives us four quadrants and some interesting ways to talk about the materials in them.



Books
Journals
•Newspapers
•Gov. docs
•CD, DVD
•Maps
•Scores

Freely-accessible web resources

Special collections
Archives
•Rare books
•Local history materials
•Archives & Manuscripts
•Theses & dissertations

Research and learning materials
•ePrints/tech reports
•Learning objects
•Courseware
•E-portfolios
•Research data
Untransferred records

Figure 2: Collections grid

The two *below-the-line* categories – research and learning materials, on the one hand, and current special collections on the other – differ in important ways. Research and learning materials are produced by faculty members. This is an area where there is great diversity: courseware, documents, databases, data sets, simulations, and so on. They are hugely variable in approach, in technologies, in scope, in complexity. Research and learning materials are produced throughout a university or college, and the library role is under discussion: responsibility, ownership and interest are organizationally diffused. ‘Special collections’ – as historically construed within the library community – are within the library control and a role is well established.

However, there are also important points of contact. In fact, for many libraries research and learning materials produced by their faculty colleagues may be the special collections of the future: unique local materials, which may require high levels of curatorial care. And the materials in their unique local special collections are increasingly being digitized to support scholarship and learning more widely. There are also many points of contact from a management point of view, especially as they are digitized, which distinguish them from the more traditional library activity:

- They involve digital content management. The library has to develop or acquire digital content management services. It has to put in place workflow and skills to deal with digital assets, which will be increasingly complex, diverse and voluminous.

- Increasing interest in object’s life cycle -- from creation to analysis and re-use. Hitherto, the library did not directly support the creation of resources, nor did it have to provide – except in special cases such as microfilm – additional support beyond reading space for the use of the resource.

- Different types of support may be needed at different places in the life cycle. So, for example, the library may manage and serve up e-prints but have relatively little involvement in their creation. Some resources may be licensed and made accessible, but are not locally managed. In some cases an access apparatus may be built over resources which remain outwith the library management regime. Other resources may actually be managed on behalf of faculty. New advisory and support services will be required. And so on.

- Additional metadata, rights and general interoperability requirements are posed.

- We are seeing a greater ‘archival’ perspective. There is a growing interest in provenance, context and use history. The evidential integrity of the materials is important. The library needs to begin to think about long-term access and management issues if the materials are to be available over time. Appropriate preservation practices need to become part of responsible asset management.[8]

- Multiple, and sometimes new, metadata and content formats continue to arise. This is a major issue as rich, interactive resources are created. Materials may be ‘packaged’ or repurposed in various ways for learning purposes. They may be a variety of derivative or aggregated works. These in turn need to be recombinable with developing research and learning tools.

- We are also seeing a growing interest in harvesting metadata, pulling metadata from different repositories, in fusion and augmentation, and in aggregating in 'union' services. The mechanics of harvesting are becoming routine and well understood, and this is now introducing the interesting challenge of effectively fusing metadata so that a useful retrieval experience can be offered.

In such an environment libraries are potentially looking at:

- Multiple metadata creation and repository environments. The library may have an institutional repository framework, a digital asset management system for digitized special collections, an image database, and so on. They may work with a central shared cataloging system for some materials, but with local metadata creation interfaces for others. Departments, research centers, individual faculty members may have their own environments.

- Multiple metadata formats, some formalized through community agreements or standardization processes, some 'vernacular' where produced by local agreement. Thus they may work with MARC, with Dublin Core, with VRA Core, with GEM, with EAD, and with various other local 'vernacular' or community-specific formats. This presents a need for schema transformation as they move metadata between systems or want to provide metasearch facilities or want to expose metadata for harvesting. This metadata will often not have been created within a framework of consistent practice; approaches to subjects or names will be different for example. It is sometimes suggested that progressive migration to XML will remove many of these issues. Some things may be facilitated, but there will still be issues as we can think of at least three levels at which metadata needs to interoperate:

    - encoding (e.g. XML, ISO 2709),

    - element set or content designation – what statements about a resource can be made and what do they mean (e.g. Dublin Core elements, MARC tags. Is my 'author' the same as your 'creator'?)

    - content/element values – what goes in the fields (e.g. cataloging rules, controlled vocabularies)

- A potential need for multiple controlled vocabularies. Again a variety of materials may present different requirements for control of terms: whether name, place, time period or topic. And again, a variety of general, domain-specific or vernacular approaches may be in place. Support for multiple vocabularies in content management systems is not good.

- A growing interest in promoting metadata from digital resources themselves though automatic means. Automatic categorization and metadata creation is of great interest, and may increasingly be necessary. Traditional labor-intensive practices may not scale up as the volume and variety of digital resources increases, and the economies of shared cataloging are not available in the same way as many of these resources are unique. This becomes increasingly the case as other kinds of metadata are also required: structural or technical for example.

- Managing complex objects. Many digital objects will be complex: bringing together content and metadata in various configurations. We have not yet much experience of this. Research and learning materials will be in diverse formats, sometimes standardized.

This diverse and developing practice means that much of this activity is expensive and existing approaches may not scale. It is expensive in terms of the learning that is required to develop services, in terms of the effort expended to build them, and in terms of the user time to interact with this diversity. Activities have not settled into routine organizational patterns. And certainly, metadata management and knowledge organization approaches are not well understood, let alone routinized.

2 Relevant OCLC Research activity

2.1 Introduction

The question we have asked ourselves is what type of services would be valuable to libraries in this increasingly diverse environment. How can we release the accumulated expertise and value invested in libraries' metadata management and knowledge organization approaches in this new environment? How can we help reduce the cost and improve the quality of services? How can we leverage existing functionality **outside** the systems in which they currently reside, and **inside** new repository and editing applications? How can we unplug and play?

In this context OCLC Research has been looking chiefly at metadata and knowledge organization services that are appropriate to an environment of multiple repositories, creation environments, and discovery mechanisms. We are exploring how you aggregate metadata from different regimes, how you analyze and manipulate it, how you map and transform it so that it becomes more widely usable and useful. We are exploring how and whether the structured knowledge embedded in controlled vocabularies and name authority services can be leveraged as network resources across this diversifying range of need. And we are exploring how you tie these and other services together in the loosely coupled web world.

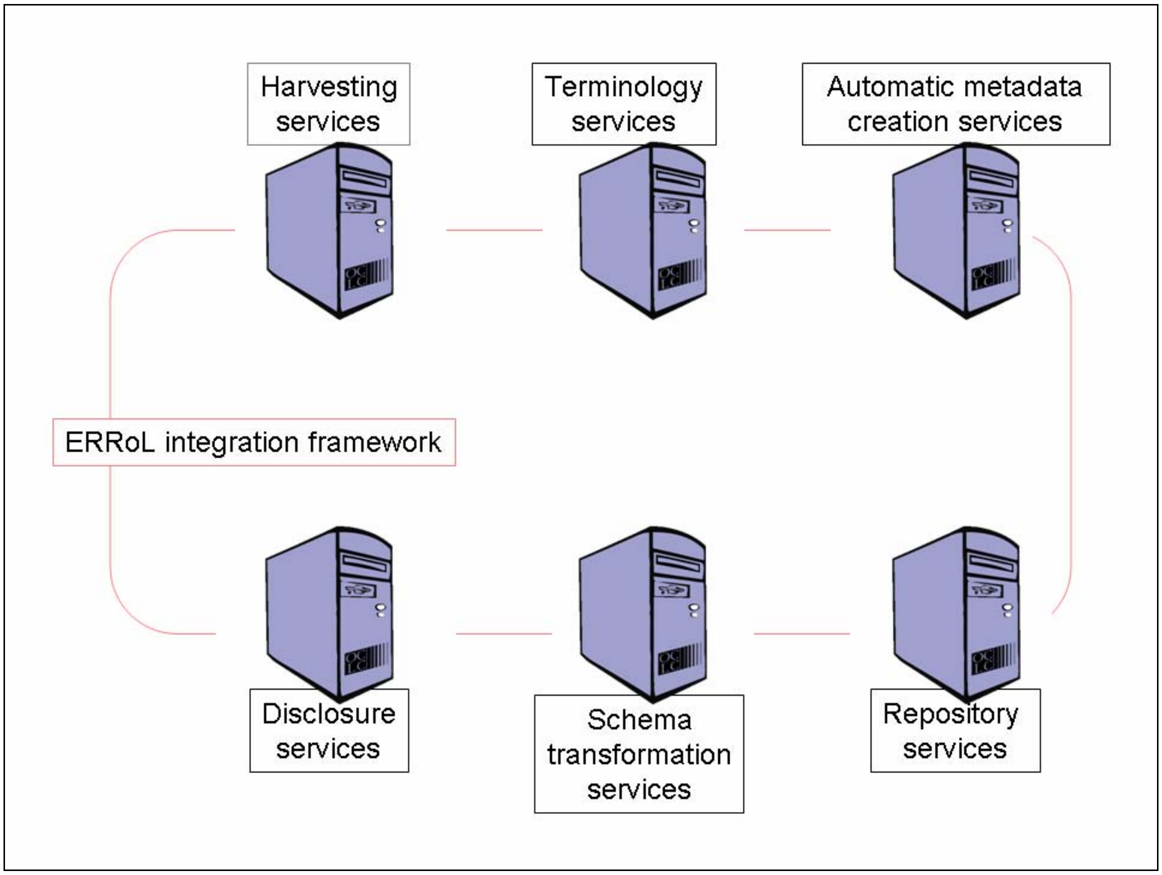

Harvesting services
Terminology services
Automatic metadata creation services
ERRoL integration framework
Disclosure services
Schema transformation services
Repository services

Figure 3: OCLC Research experimental modular services

In this article we will talk about the services shown in Figure 3. Some of this activity was introduced under the general umbrella of the Metadata Switch project.[9] This project looked at constructing experimental modular services that add value to metadata, and at exposing this functionality as 'web services.' We give an example of a name authority web service in the next section to make this concept more real.

These modular services-related research activities address some of the issues identified in the opening paragraph of this section, but by no means have our activities addressed all or even the majority of them. The balance of this article will focus on a select group of issues that OCLC Research is currently pursuing, namely:

Harvesting and disclosing metadata: In the environment sketched above, sharing metadata between sites -- and between sites and aggregators -- has become more appealing. We are exploring issues involved in fusing metadata from different technical and policy regimes, effectively recombining it to support improved service. We are also interested in disclosing such metadata through a variety of interfaces. We want to expose it for searching, for harvesting by other services in turn, and increasingly to the major search engines so that it becomes more visible on the web.

Repository interoperability: We are interested in combining metadata from different repositories, and we are supporting the development of interfaces which facilitate this in open ways.

Metadata schema transformation: We need better ways of managing the variety of metadata schema. Here we are working on services which can automatically convert from one schema to another.

Terminology services: We are working to 'webulate' vocabularies and name authority files so that they can be more flexibly used in network environments. We aim to make vocabularies available as 'pluggable' network resources, which can be accessed and used by other applications. We are also leveraging mappings, transformations, and other services in this context.

Automatic metadata creation: We have done work in this area for some time. We need to reduce the cost of metadata creation by finding ways of promoting metadata from the resources themselves, or by inheriting data from like resources. We want to automatically classify documents to support management.

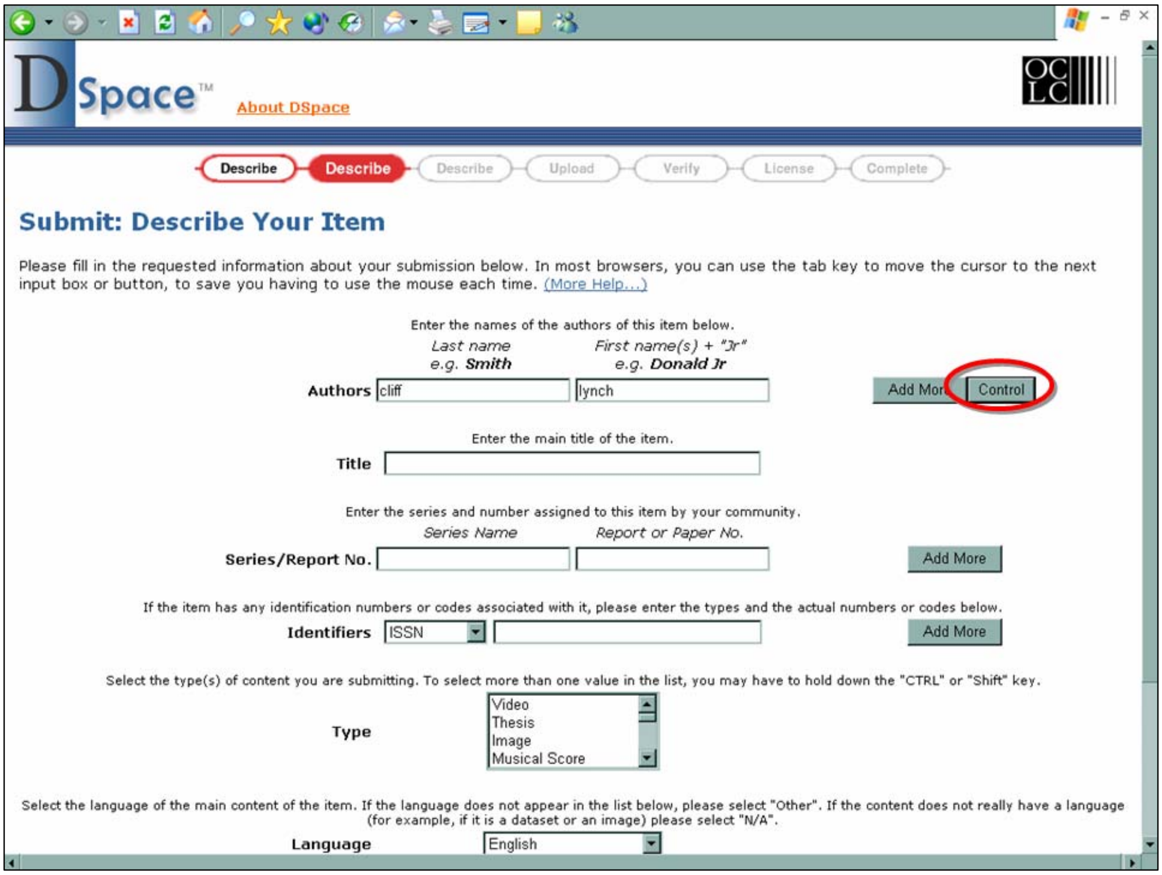
ERRoLs (Extensible Repository Resource Locators): infrastructure to build services on metadata and vocabularies: The services above are made available in various configurations. The ERRoL framework provides an integrated protocol environment within which to explore metadata and vocabulary management issues.[10]

2.2 Web services

Web services are machine-to-machine applications that run over web protocols; they have become very popular as a lightweight way of building distributed applications.[11] From our point of view, they provide an interesting way of proving applications functionality as a network service, as a piece of remote functionality that is 'pluggable' into local applications. Consider Figure 4. This shows an instance of Dspace running at OCLC Research. What we have done is 'plugged' a name look up service into the submission process below (circled 'control' button). When the user clicks this button, the application takes the data in the author entry box and does a lookup of the Library of Congress Name Authority File which we have running on another machine. The user is offered two candidate names, and can select one (Figure 5). Mousing over the name shows data from the authority record which may aid in selection. Clicking on the desired name then populates the field. What this shows is an authority lookup function which is available for plugging into any application that can talk to it. It is a very simple application, and we publish details of how to talk to it.

In this way we are making some functionality available to a variety of applications and repositories. This functionality tended hitherto to be available in a standalone way, or as part of a larger monolithic application (a cataloging system for example). We plan to make various vocabularies available in similar ways and see how they are used.

Of course, there are service issues over and above making data available in this way. For the names service for example, we have the question of what a user does with names that are not in the file. In the longer term, it will be interesting to consider how to articulate current authority processes with institutional directory services, or a more open process of establishment of names.



Submit: Describe Your Item

Please fill in the requested information about your submission below. In most browsers, you can use the tab key to move the cursor to the next input box or button, to save you having to use the mouse each time. (More Help...)

Enter the names of the authors of this item below.

	Last name e.g. Smith	First name(s) + "Jr" e.g. Donald Jr
Authors	cliff	lynch

Enter the main title of the item.

Title

Enter the series and number assigned to this item by your community.

Series/Report No. — Series Name — Report or Paper No.

If the item has any identification numbers or codes associated with it, please enter the types and the actual numbers or codes below.

Identifiers: ISSN

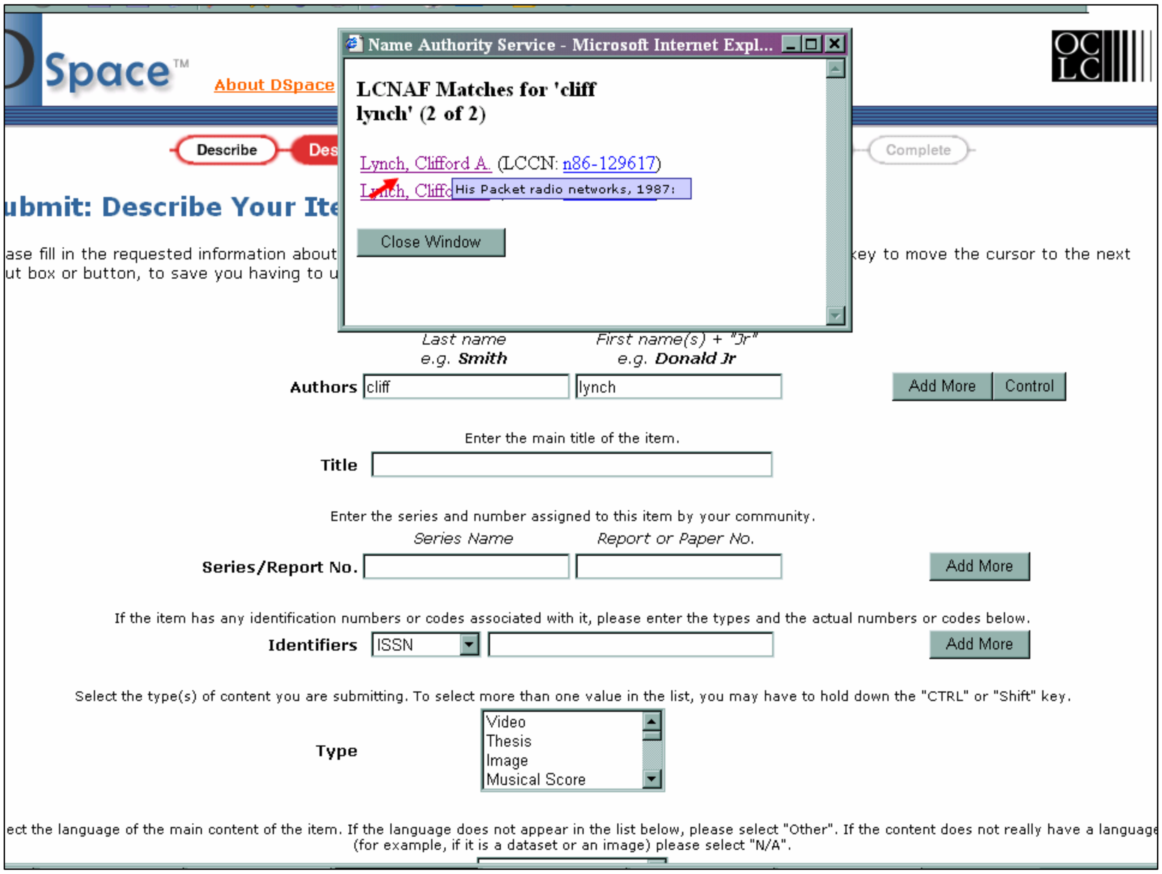
Select the type(s) of content you are submitting. To select more than one value in the list, you may have to hold down the "CTRL" or "Shift" key.

Type: Video, Thesis, Image, Musical Score

Select the language of the main content of the item. If the language does not appear in the list below, please select "Other". If the content does not really have a language (for example, if it is a dataset or an image) please select "N/A".

Language: English

Figure 4: Authorities service in Dspace



LCNAF Matches for 'cliff lynch' (2 of 2)

Lynch, Clifford A. (LCCN: n86-129617)
Lynch, Cliffo His Packet radio networks, 1987:

ubmit: Describe Your Ite

ase fill in the requested information about … key to move the cursor to the next
ut box or button, to save you having to u

Last name e.g. Smith — First name(s) + "Jr" e.g. Donald Jr
Authors: cliff | lynch

Enter the main title of the item.
Title:

Enter the series and number assigned to this item by your community.
Series Name — Report or Paper No.
Series/Report No.:

If the item has any identification numbers or codes associated with it, please enter the types and the actual numbers or codes below.
Identifiers: ISSN

Select the type(s) of content you are submitting. To select more than one value in the list, you may have to hold down the "CTRL" or "Shift" key.
Type: Video, Thesis, Image, Musical Score

ect the language of the main content of the item. If the language does not appear in the list below, please select "Other". If the content does not really have a language
(for example, if it is a dataset or an image) please select "N/A".

Figure 5: Selecting a name from an authority file

2.3 Harvesting and disclosure of metadata

The Open Archives Initiative Protocol for Metadata Harvesting[[12]] has emerged as the leading candidate for doing harvesting and disclosure of metadata in our field. OCLC Research has significant experience harvesting metadata, and has participated in the technical development of OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Our software, has been made available as open source and is used in several applications to provide OAI-PMH functionality, including Dspace and the National Library of Australia.[[13]] Software is available to provide both the harvesting functionality and the disclosing functionality the protocol supports.

As part of our participation in the Networked Digital Library of Theses and Dissertations[[14]], OCLC Research harvests metadata from OAI-PMH repositories of electronic thesis metadata around the world and creates a merged set of this information for subsequent onward harvesting, again using the OAI protocol. Currently, we are harvesting from forty-four institutions, some of which themselves collect theses from several institutions. We have also extracted the metadata for all theses from OCLC WorldCat (the OCLC Online Union Catalog) and made that available. Altogether this amounts to information for more than four million theses, of which something like 100,000 are available electronically. [[15]]

More recently, we have begun to harvest metadata from Dspace repositories. In parallel, our production colleagues have begun harvesting metadata from CONTENTdm servers. CONTENTdm is a digital asset management system distributed by OCLC, which is used in many libraries to manage digitized collections of images, special collections, and so on, the material in the bottom left hand corner of the collections grid.[[16]]

Once metadata has been collected, it is then possible to build services on it. One of the most obvious of these is to make it easy to find the described material. There are several possible approaches here. In the past this would probably involve creating a retrieval system for the metadata. We still do that, however offering the data through a web-based retrieval interface means that it is part of the 'deep' or 'hidden' web. It is becoming more important to get the material noticed by the large search services on the Web, such as Google and Yahoo, and also to be able to share the metadata with other services. We want to explore other forms of 'disclosure,' placing the metadata in contexts where it is more widely available to other services. This leads to the second and third approaches: disclosing under the multi-protocol ERRoL framework OCLC is building, and disclosing to the large web search engines. Currently, these are different approaches, but as we discuss below, they may come closer together over time.

The multi-protocol environment is described below when talking about the ERRoL framework. We are exposing metadata to Google and other search engines for crawling under the framework established in the OCLC OpenWorldCat program.[[17]] In the latter case, for example, we are working with the Dspace Federation to harvest metadata from Dspace repositories and to make it available for crawling by Google.[[18]]

In the past the Web search engines were often criticized for ignoring the 'deep Web,' material that does not show up as normal HTML Web pages, but is contained in databases, such as library catalogs, WorldCat or repositories of images. As database producers recognize the value of visibility on the Web, this is changing and more and more of this material becomes available for indexing.

In the context of the Collections Grid, what this means is that the search engines are becoming interested in collections from all quadrants, and resource operators, libraries among them, are interested in disclosing their collections to the search engines so that they become more visible to their users. For many users, Google and other search engines are increasingly becoming the first and last search resort. In this context, it is interesting to think about so-called 'gray literature,' the materials which were not formally published. These were often of great potential value, but were often difficult to obtain when compared to more formally published materials available through established distribution channels. The Web has turned this inside-out, making 'gray literature' more visible and easier to obtain than some of the formally 'published' materials available through library and other collections. In many cases, the formally published material lies outside the discovery and linking apparatus of the web, but it is precisely this apparatus that makes the web so compelling to users. As protocols for redistribution of metadata become widely used, and union catalogs, such as OAIster[[19]] and WorldCat become visible on the Web, a certain balance may be returning, where the range of materials across the quadrants of the collections grid become more visible on the open web.

This merger of these materials and their disclosure is not without issues, some technical in nature, some more service or business oriented. Consider for example the following:

- *Provenance and trust.* One way of thinking about metadata is as ‘schematized statements about resources.’ As we exchange more metadata outside of existing trust patterns (a shared cataloging environment is an interesting example of a trust pattern enforced by policy) we need to make decisions about which ‘statements’ to trust. Consider, for example, that the search engines do not currently index descriptive metadata embedded in web pages, because they are unwilling to accept the ‘statements about resources’ they find there at face value. What trust and policy patterns for sharing metadata will emerge?

- *Managing identity and difference*. An important research question as we harvest more metadata arises in relation to duplication and identity. We have been using FRBR[[20]] and other techniques to check for duplicate records, to explore what overlap there is between harvested metadata and metadata already in WorldCat, and to think about merging records where relevant. For theses, for example, this overlap is considerable, as many libraries, especially in North America, will already have catalogued these materials. As resources appear in many versions, in derivative works, in aggregate works, and so on, these issues become more interesting.

- *Consistency*. There is value in consistency. Consistency reduces complexity of processing and potentially improves services. However, there is also a cost. The effort expended on creating consistent metadata, or on processing it to achieve better consistency, relates to a value justification. The library community has historically developed approaches to reduce the cost of cataloging and increase its consistency through shared approaches. Facilitating this approach is OCLC’s central activity. However, as we move ‘below the line’ in the collections grid, the uniqueness of the resources changes the shared cataloging value proposition. We believe that there is significant value in consistency and want to reduce the cost of achieving it. Some of the work described below has this focus including automatic extraction of metadata, provision of vocabulary and authority control services in a ‘pluggable’ way. We are also interested in better guidelines, within particular communities of practice, for data entry.

- *Forms of disclosure*. Going beyond this, we are interested in discussions about how best to expose such data to services like Yahoo! and Google, and what the advantages to the library community and others in moving towards some consistency of disclosure are. Again, OAI has potential here. At the moment, data is provided in the form in which the individual search engines expect, which varies between search engines. A provider exports records to flat web pages for Google, whereas for large sites Yahoo expects an XML file of simple records. Google sees the world as web pages, and this accounts for its extraordinary success. Will Google and other search engines want to exploit more of the structure in our records? Can we incentivise different behaviors? Exposing large quantities of metadata poses technical issues which are still being worked through (for example, Google crawled websites up to a maximum number of pages).

- *Harvesting content and complex objects*. We have discussed harvesting metadata above. We harvest digital content into the OCLC Digital Archive (not currently using OAI), and there is growing interest in automating metadata creation as fully as possible. Over time, we will need better programmatic approaches to working with multiple simple and complex content objects.

- *Many to many*. As more providers make metadata available in various ways, and as this metadata is indexed in the search engines, there is a 'switch' issue. For example, if a library exposes its own catalog metadata it becomes available to all Google searchers. If a library user searches on Google for a book, he or she is presented with data from many resources. Through its Open WorldCat project, OCLC is interested in providing such a 'switch' service, effectively providing a 'rendezvous' facility between library services and library users. For example, <http://www.worldcatlibraries.org/wcpa/ow/075e71ada9407c10a19afeb4da09e526.html> will take you to a 'rendezvous page for a particular book. Some services are available there, depending on what we know about your privileges based on an IP-checking-based authorization. These include access to the local OPAC, access to interlibrary loan service, access to an openURL resolver, and so on. We plan to enrich the service offering here. This is one example of service issues that will arise. In the future, we might imagine a range of resolution and other services of interest, a range of new service providers, and a range of interesting offerings.

- *Gated resources*. The last paragraph touched on a significant issue: authorization. How to authorize users of particular resources in the open web environment is a major issue moving forward.

Many of these issues revolve around the question of user interface, data context, and data consistency. What will be the user interface of choice for users in the future? Three stages have emerged successively which continue to exist in concert. Data is made available within a particular system and interface, say, for example, a library catalog. More recently we have seen the emergence of user interfaces which sit on top of a federated resource, potentially communicating through machine interfaces with multiple individual resources. Federation may be achieved through harvesting or metasearch. Now we are seeing the web search engines taking the role of user interface. One point to make here is that 'context' becomes much more interesting: where multiple sources of data become available, the services one provides around that data become more important. We are used to the idea of making records available through services; we will become increasingly used to seeing services available through records. Think of some simple examples, pursuing the catalog case again. Say you find a record, and it has links in it to services which will find books like this one, services which allow you to borrow it, services which find books by the same author, and so on. Again, our way of thinking about things may be turned inside out: how do we expose on the web functionality that was part of closed systems on the hidden web?

2.4 Repository interoperability

Some of these records come from repositories managed by software such as CONTENTdm and DSpace. As noted above, we believe that there are advantages where metadata created in institutional repositories is as compatible with other systems, such as library catalogs, as possible. To explore these compatibility issues we have been working with DSpace, contributing to its harvesting, creation, and searching software. To make harvesting work smoothly we have contributed our OAI-Cat software which supports the

OAI-PMH protocol. For searching, we have contributed SRW/SRU [21] support so that DSpace repositories can be searched using a standard protocol. (SRW is Search and Retrieve Web service, and SRU is Search and Retrieve URL service. These protocols recast the functionality of Z39.50 in a form more suitable for the web.) For metadata creation, we have extended DSpace to support different metadata profiles for different collections and have prototyped a connection to a name authority service, which was introduced above.

2.5 Metadata schema transformation

There are cases where differences between metadata records must be reconciled. Incompatible descriptions may hinder effective searching or database management. As noted above, the issues do not simply reduce to ones of encoding. And again, there is a value question. Such work is justified where sufficient value is realized to justify the cost.

The flavor of the problem can be illustrated by the following MARC and Dublin Core record fragments:

100 ‡**a** Shakespeare, William ‡**d** 1564-1616	<dc:creator>Shakespeare, William, 1564-1616</dc:creator>
245 ‡**a** Hamlet	<dc:title>Hamlet</dc:title>
260 ‡**a** New York: b Penguin Books, ‡**c** 2003	<dc:publisher>Penguin Books</dc:publisher> <dc:date>2003</dc:date>

Here it is easy to recognize a correspondence between the MARC 100 *author* field and <dc:creator>; between MARC 245a *title* and <dc:title>, and so on. But the MARC and Dublin Core records are not simply different structural representations of the same information. They also differ in granularity and meaning, some of which is lost when the MARC record is converted to Dublin Core. Thus, in the MARC record, William Shakespeare’s birth and death dates are explicitly coded, but this information is merely implied in the Dublin Core record. And in the MARC record, the 260 field specifies that the work was published in New York by Penguin Books in 2003. But in the Dublin Core record, the date is ambiguous—Is it the publication date? Or the creation date of the metadata record?—and the place of publication is missing altogether.

This example shows that the relationship between metadata standards is semantic and sometimes only approximate. It is usually recorded in a list of mappings, or a so-called *crosswalk*, which asserts equivalences between corresponding fields, such as *author, title* and *publisher* in this example. We have argued that, depending on the complexity of the crosswalk or the user’s need for accuracy and completeness, metadata translation can sometimes be accomplished through relatively lightweight processes. But a more extensive custom-designed solution may be required for high-fidelity translations. We have described this ‘two-path’ approach elsewhere.[22]

The *lightweight translation path* is designed as a publicly available service that uses published metadata standards and crosswalks. It works in the best-case scenario when records are available in XML, differences between the standards involved in the translation are relatively slight, and the conversions can be accomplished with straightforward XSLT code. When these conditions are met, the prospects for promoting re-use and eventual standardization are good.

To encourage this development, we have proposed an application profile for a METS[23] record that associates all of the Web-accessible files required for interpreting and executing a crosswalk. These records are made available to a repository[24] which can be searched and processed by web services, OAI harvesters, and other lightweight software programs that have been developed in the XML community. With this repository, a user can submit the name of the desired target metadata format and a URL to a set of XML records to be converted. The software associated with the repository extracts references to the XML schemas from the data, searches the repository for matching XSLT-encoded crosswalks, and executes the translation.

Our expectation is that the lightweight translation path will encode relatively loose standards for equivalence and round-trip translation, but it remains to be seen whether these standards are adequate for useful levels of interoperability.

A proof-of-concept, an ERRoL-based approach is discussed in section 2.7 and illustrated in Figure 8 below. ERRoL services can be accessed from the OCLC ResearchWorks website. [25]

The *heavyweight translation path* is a custom application that generalizes some of the functionality of OCLC's production systems and makes metadata translation available as a modular, standalone process. It is intended for complex metadata schemas such as MARC, which has standards for accuracy and semantic equivalences with other standards that are not easily represented in XSLT scripts. The heavyweight translation path is designed for high-quality and, hopefully, complete translation of records.

Why two paths? Some time ago, we discussed the issues of metadata translation with Juha Hakala, Director of Information Technology of the National Library of Finland. Like us, he needs effective metadata crosswalks to solve the everyday problem of ensuring consistency in large databases that are built of record streams from multiple sources. As he said, crosswalks (as well as their corresponding XSLT encodings) merely represent a proof of concept. Right now, they need to be augmented with robust systems that handle validation, enhancement, and multiple character encodings and allow human guidance of the translation process. But as standards and the supporting software infrastructure become more mature, the "proof of concept" acquires more and more functionality that can eventually replace major components of a production system. In this view, the lightweight path is our proof of concept and the heavyweight path is our production system. In the future, perhaps only one path will be required.

2.6 Terminology services

A majority of tools and features for accessing the names, subjects, and classification categories assigned to resources are not easily accessed by people or computers. The knowledge organization schemes (KOS) and the features found in cataloging and retrieval systems are often deeply embedded in proprietary formats and software and are rarely linked with other compatible schemes or services.

There are many standards for constructing and encoding knowledge organization schemes[26]. Several large schemes are available for import in the MARC formats, but many others exist only in native or display formats. The Terminology Services project[27] is making selected knowledge organization schemes more accessible for machine and human interaction through a series of processes. Schemes are converted to MARC XML (Authorities or Classification format), SKOS (an RDF Schema for thesauri and related knowledge organization systems) and to Zthes (a Z39.50 profile for thesauri). When applicable, inter-vocabulary mappings are added as part of the processing. During processing, the files are also enhanced with persistent identifiers that include the namespace in which the identifiers are unique. The KOS identifier consists of two elements: scheme and concept. The scheme element identifies the knowledge organization scheme being used, and the concept element identifies a concept within a given scheme.[28]

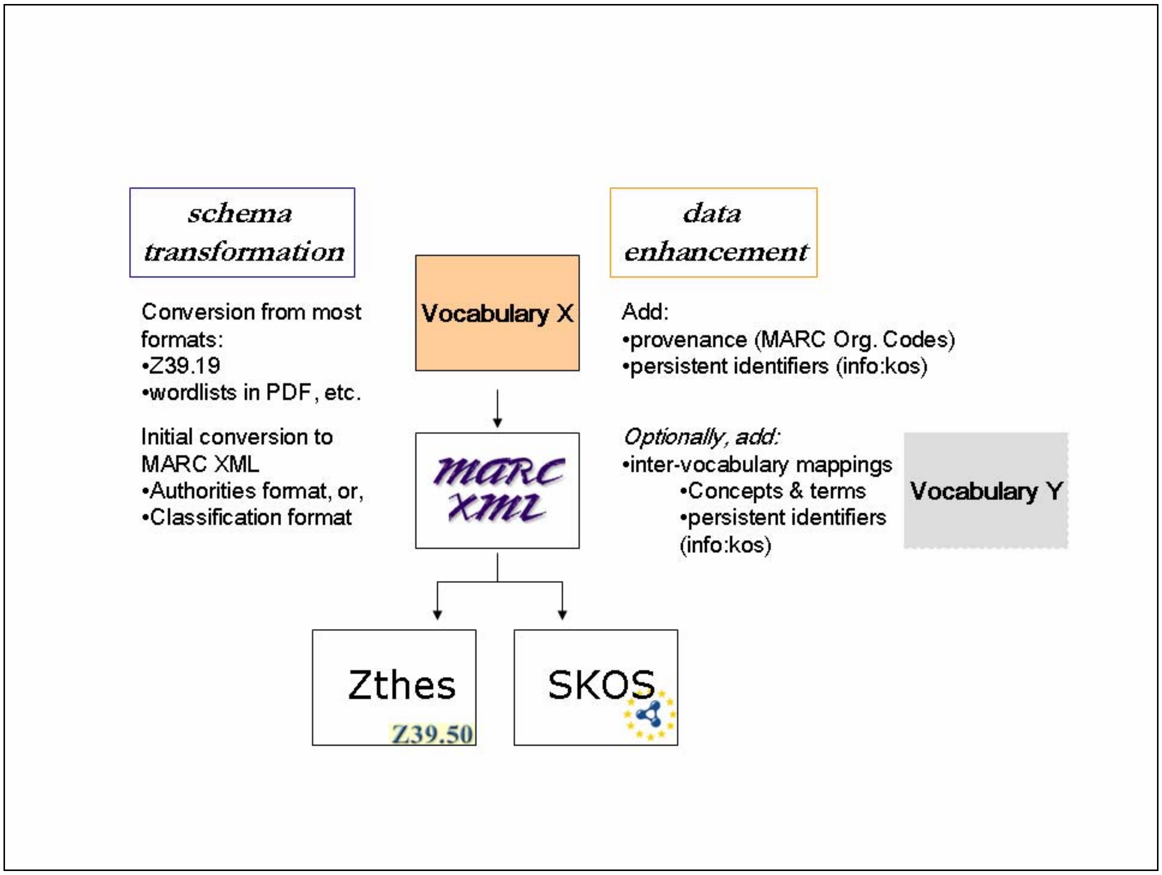


Figure 6: Vocabulary processing

A controlled vocabulary that has been enhanced with mappings and identifiers is now available from OCLC Research. Authority records for *Genre Terms for Fiction and Drama* (GSAFD) were downloaded from Northwestern University Libraries.[29] The file was enhanced with identifiers and mappings to Library of Congress Subject Headings (LCSH) and LC Children's Headings (LCSHac). Versions of the file encoded according to MARC XML, SKOS, and Zthes schemas are available for downloading. The GSAFD vocabulary is also accessible using the OAI Protocol for Metadata Harvesting. Gradually, other terminologies will be brought into this framework, and we will explore how they can be used in different environments.

Modular, accessible vocabularies and classification schemes can be used in a variety of terminology services. During resource discovery, an application might use the full range of term relationships and available mappings to broaden or refine a search. Services invoked during indexing or cataloging might use only mappings produced by a specific organization or for a given scheme. When data is presented or displayed, terminology services can provide translations of terms or category labels that are appropriate for a particular user.

2.7 An integration environment: ERRoLs

The OAI Protocol for Metadata Harvesting gives us a way of identifying, storing, and distributing metadata repositories. The SRU/W protocol gives us a way of searching collections, such as OAI repositories. RSS[30] gives us a way of distributing timely subsets of repositories. OpenURL[31] gives us a way of retrieving items from a repository based on a set of characteristics of the items. OpenURL uses the info:uri resolver to refer to the many formats and protocols.

All of these protocols are ways of building, retrieving and looking at metadata repositories in various ways. For instance, a thesaurus that is put up for access via SRU (which provides retrieval services similar to Z39.50, but in a Web environment) can, with the proper indexes, form the basis of an OAI-PMH service, so that the thesaurus could then be distributed in this way. SRU services can, with the proper style sheets, create HTML pages for interaction and display. An OAI-PMH repository can form the basis of an RSS feed based on the modification dates in the metadata it contains. All these protocols supply one or more pieces of a total set of metadata services that digital libraries need.

What is missing is a way to coordinate all these services, regularize the handling of stylesheets to format the XML responses, and pull them all together. The ERRoL service does this in a very light-weight way by using the tools themselves to do the coordination.[32] The ERRoL server currently depends on an OAI repository[33] maintained at UIUC (University of Illinois at Urbana Champaign) to describe the OAI repositories available. Using information from the UIUC registry it is possible to provide RSS feeds for any of the repositories and views in HTML of many of the records. In addition, if the repositories are built to support other services (such as schema transformation and SRU retrieval), the ERRoL service can detect this through their OAI-identify response and support those services using a set of XML stylesheets. These stylesheets are in turn kept in an OAI database, allowing the full set of ERRoL services to be applied to them.

So, the vocabulary services mentioned above can quite easily support sophisticated retrieval, transformations between metadata schemes, easy syndication via RSS, replication via OAI, and displays and navigation in various formats. In addition, ERRoLs provides a succinct way of creating stable Web identifiers to the thesaurus’s concepts through the use of OAI identifiers, a fundamental feature for references to the concepts to work properly in digital libraries. The scenario under construction is shown in Figure 7. Figure 8 shows the same environment supporting schema transformation services.

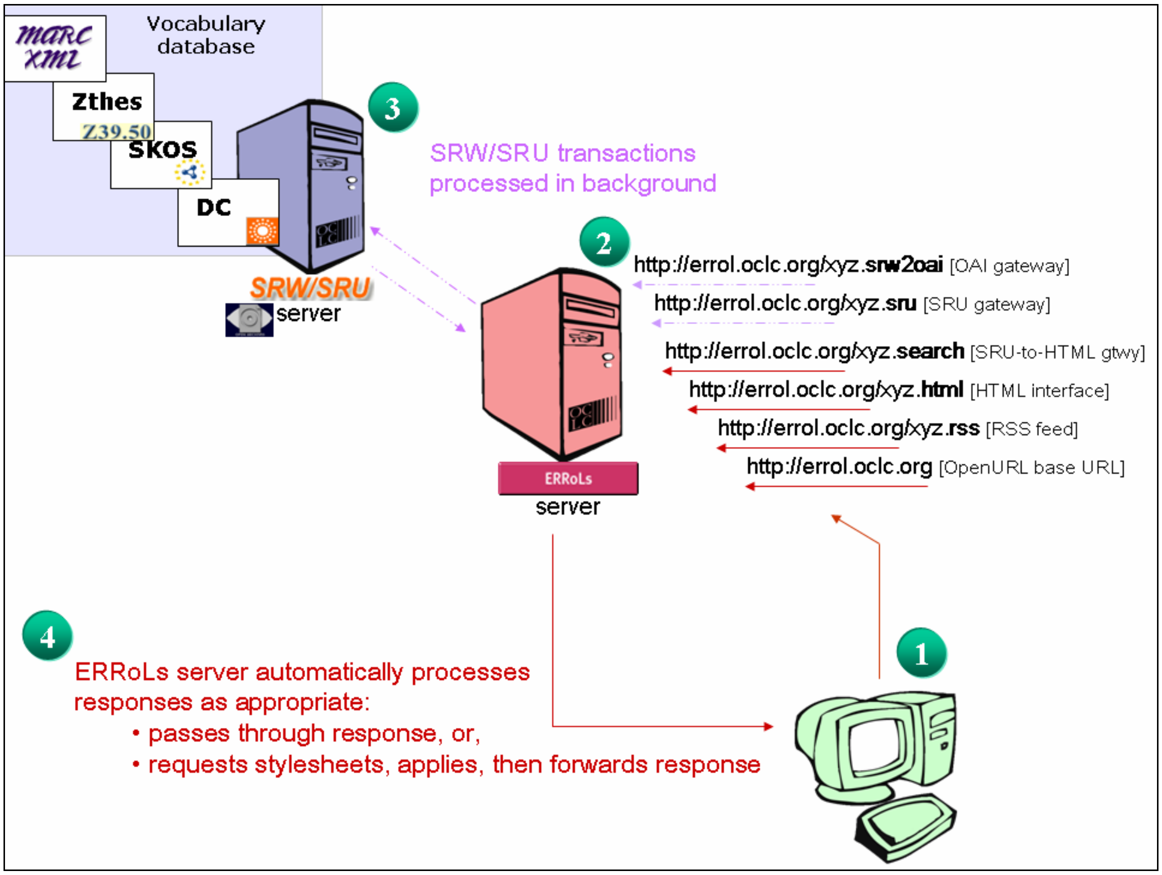

MARC XML
Vocabulary database
Zthes
Z39.50
SKOS
DC
3
SRW/SRU server
SRW/SRU transactions processed in background
2
http://errol.oclc.org/xyz.srw2oai [OAI gateway]
http://errol.oclc.org/xyz.sru [SRU gateway]
http://errol.oclc.org/xyz.search [SRU-to-HTML gtwy]
http://errol.oclc.org/xyz.html [HTML interface]
http://errol.oclc.org/xyz.rss [RSS feed]
http://errol.oclc.org [OpenURL base URL]
ERRoLs server
4
ERRoLs server automatically processes responses as appropriate:
• passes through response, or,
• requests stylesheets, applies, then forwards response
1

Figure 7: Vocabulary services within the ERRoL framework

What does this mean in practice? The ERRoLs server can be used in conjunction with a variety of other services to automatically deliver a versatile array of chained activities. In Figure 8 the ERRoL server is shown as the nexus of a group of SRU/SRW-accessible services that would support:

- the search and retrieval of metadata from a repository,
- the discovery and use of a relevant stylesheet (in this case, a metadata crosswalk) to deliver back to the user the data requested in the user's choice of formats.

(N.B. ERRoL servers also work with OAI servers if the list of available services is diminished.)

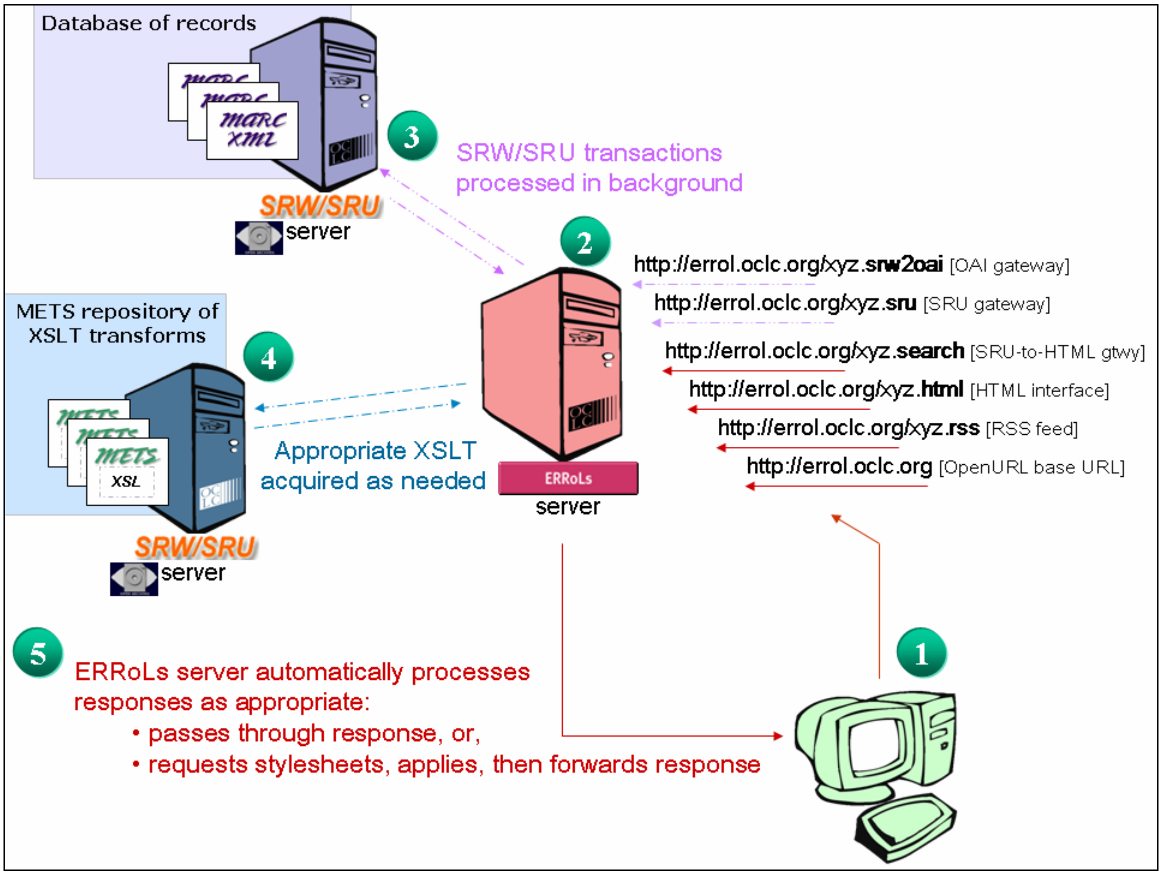

Database of records
MARC
MARC
MARC XML
3
SRW/SRU transactions processed in background
SRW/SRU server
2
http://errol.oclc.org/xyz.srw2oai [OAI gateway]
http://errol.oclc.org/xyz.sru [SRU gateway]
http://errol.oclc.org/xyz.search [SRU-to-HTML gtwy]
http://errol.oclc.org/xyz.html [HTML interface]
http://errol.oclc.org/xyz.rss [RSS feed]
http://errol.oclc.org [OpenURL base URL]
METS repository of XSLT transforms
4
METS
METS
METS
XSL
Appropriate XSLT acquired as needed
ERRoLs server
SRW/SRU server
5
ERRoLs server automatically processes responses as appropriate:
• passes through response, or,
• requests stylesheets, applies, then forwards response
1

Figure 8: ERRoL framework supporting schema transformation

1. The user initiates any one of a number of request scenarios (including OAI, SRU/SRW, RSS feed requests, etc.)
2. The ERRoLs server parses the requests and directs the queries to the appropriate SRU/SRW server (requests already in SRU/SRW format are redirected; non-SRU/SRW formatted requests are converted to SRU/SRW format automatically before sending)
3. The SRU/SRW server responses are processed by the ERRoLs server, and under most scenarios an appropriate stylesheet is applied before passing the results to the user.
4. In the diagram, an optional path to search and retrieve XSLT metadata crosswalks from a METS repository of XSLT metadata crosswalks is shown -- if appropriate for the request, the ERRoLs server would discover and apply the appropriate XSLT metadata crosswalk before passing the requested data back to the user.

Schema transformations can be seen in operation through our OAI viewer, built on the ERRoL framework.[34]

2.8 Automatic classification and metadata extraction

The cost/benefit discussion above points to the need to think seriously about automatic creation of metadata for resources. There is growing interest in this route, especially as the variety of metadata requirements grows in relation to digital materials. Here, for example, technical metadata may be programmatically promoted from the resource itself, documents may be automatically classified for routing, simple records may be created for resources to reduce the manual labor involved. There is also great interest in content analysis. In this section we briefly touch on past work in this area, some of which is now included in OCLC production services, and conclude by describing the ePrints UK project where we hope to leverage some of this work in the context of changing patterns of document and metadata distribution discussed at the beginning of this article.

2.8.1 Automatic classification

One of the first projects undertaken by OCLC Research was to explore options for automatically classifying documents.[[35]] The projects have explored automatic classification (Scorpion[[36]]), and optimized keyword assignment (WordSmith[[37]]).

The Scorpion project initially explored automatic class number assignment of Dewey Decimal Classification (DDC) numbers. Scorpion software automatically parses electronic documents to distill keywords, applies weighting to determine the most important concepts and then matches these terms against a specially-built database to return a ranked group of DDC numbers[[38]]. More recent research has included an exploratory study to determine literary warrant for topics in electronic resources.[[39]] Scorpion has been integrated into OCLC Connexion™[[40]], OCLC’s production cataloging interface, and OCLC WebDewey™[[41]], OCLC’s online version of the DDC. In recent years OCLC has also explored the feasibility of using Scorpion to automatically assign Library of Congress Classification (LCC) numbers[[42]], and whether we can exploit FAST (Faceted Application of Subject Terminology[[43]]) to enhance subject classification[[44]]. Scorpion software is available as open source under the OCLC Research Public License.

WordSmith bears some general similarity to Scorpion, but WordSmith processed target documents to optimize the distillation of key terms that describe a document’s chief concepts[[45]], not with Scorpion’s objective of mapping key terms to an external classifications scheme. WordSmith delivered a ranked set of keywords rather than classification numbers.

2.8.2 Automatic metadata extraction

Another project, Mantis[[46]], delivered a web-based cataloging toolkit that included a metadata extraction feature – users could supply a URL and direct Mantis to parse HTML pages to extract metadata and build a preliminary bibliographic record for a web site. Mantis, Scorpion, and WordSmith were integrated into the OCLC CORC (Cooperative Online Resource Catalog) system[[47]]. The CORC system eventually evolved into OCLC Connexion which includes an automatic DDC assignment feature based on Scorpion (available to WebDewey subscribers), and a metadata extraction feature based on the Mantis concept.

The metadata extraction feature in OCLC Connexion has been significantly improved from Mantis’ original extractor. Mantis employed some simple heuristics to extract metadata. OCLC Connexion's metadata extraction feature by contrast was redesigned and has a growing table of mappings of over one thousand metadata tags from widely-used standards (and selected non-canonical variant tags). Additionally, the OCLC Connexion extractor has significantly greater fault tolerance to handle various common HTTP and HTML shortcomings, and better data correction and enhancement capabilities. The extractor is integrated with Connexion's Dublin Core-MARC crosswalk[48] to improve presentation of the processed data in the Connexion editing environment.

2.8.3 *Automatic classification web services in ePrints UK*

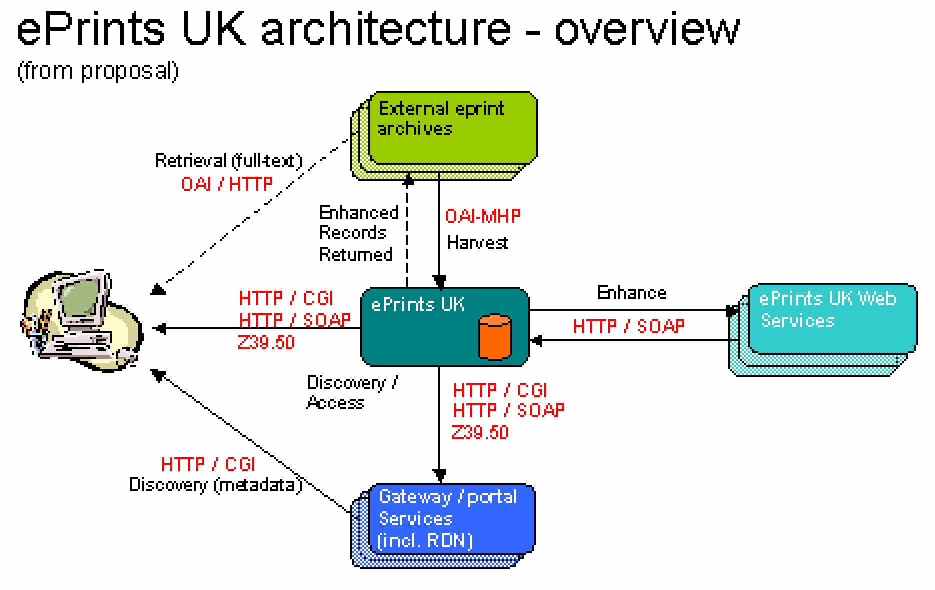


Figure 9: ePrints UK high-level architecture

ePrints UK is a JISC-funded project which is providing aggregation services for UK ePrint repositories.[49] It is harvesting metadata for eprints and eprints themselves from UK university sites. It is then routing the metadata to the national discipline-focused hubs of the Resource Discovery Network.[50] OCLC Research will experiment with an automatic classification service. The experimental service takes an eprint as input and returns relevant classification numbers to facilitate routing. This service is under construction at the moment. We will explore the service in batch mode and as a web service. A component of the project is to evaluate the effectiveness of this approach. In the figure 9, these services are embedded in the 'ePrints UK web services' box.

3 Conclusion

It is clear that the library will become more engaged with research and learning practices. The library will need to consider carefully how best to add value as it develops a role contributing to the management of research and learning materials, and as it makes its special collections available in ways which facilitate greater use by faculty and students. Librarians will work to understand what incentives and expectations influence faculty behavior. They will need to understand diverse information management regimes, and the research, teaching and learning contexts of resource creation and use. At the same time, organizational models and institutional patterns are still taking shape.

Supporting the creation, use and reuse of diverse complex resources, the creation of repository frameworks, the creation and merger of metadata and knowledge organization services, the provision of distributed disclosure, discovery and deposit services – these will all be difficult and interesting. Combining and recombining them in the emerging loosely coupled world of web services will be challenging.

We have described some initiatives here which we believe will make a contribution to understanding how we should move forward. We have a pragmatic focus, aiming to build resources which are immediately useful and advance our understanding. We welcome experimentation with, and discussion of, our results and outputs.

References

Note: a growing proportion of OCLC Research's published output (stretching back 25 years) is available from the Research Publications Repository, which uses some of the technologies described above. The repository is available at: http://www.oclc.org/research/publications/

[1] Further information about OCLC Research is available at: http://www.oclc.org/research/about/.

[2] De Rosa, Cathy Lorcan Dempsey, and Alane Wilson, *The 2003 OCLC Environmental Scan: Pattern Recognition*. (Dublin, Ohio: OCLC, 2004). This report is available online at: http://www.oclc.org/membership/escan/ The scholarly communication picture is discussed in the Research and learning landscape chapter. The collections grid is discussed in an appendix.

[3] Lyon, Liz, "eBank UK: building the links between research data, scholarly communication and learning," *Ariadne*, 36, 30-July-2003 http://www.ariadne.ac.uk/issue36/lyon/

[4] See for example the Inter-University Consortium for Political and Social Research (http://www.icpsr.umich.edu/) or the UK Data Archive (http://www.data-archive.ac.uk/). See also the Data Documentation Initiative, a documentation specification, at: http://www.icpsr.umich.edu/DDI/

[5] See for example the work of the Federal Geographic Data Committee: http://www.fgdc.gov/

[6] See for example the range of specifications produced by the IMS Global Learning Consortium, Inc.: http://www.imsproject.org/

[7] The collections grid was developed by Lorcan Dempsey and Eric Childress. It is described in *2003 OCLC Environmental Scan: Pattern Recognition*. http://www.oclc.org/membership/escan/

[8] Lavoie, Brian and Dempsey, Lorcan, "Thirteen ways of looking at …. digital preservation," *D-Lib Magazine*, 10, nos. 7/8, July/August 2004. http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

[9] OCLC Metadata Switch project web site: http://www.oclc.org/research/projects/mswitch/

[10] ERRoLs (Extensible Repository Resource Locators) are described at: http://www.oclc.org/research/projects/oairesolver/

[11] Gardner, Tracy, "An Introduction to Web Services," *Ariadne* 29, 02-October-2001. http://www.ariadne.ac.uk/issue29/gardner/ ; See also the Wikipedia definition: http://en.wikipedia.org/wiki/Web_service

[12] OAI web site: http://www.openarchives.org/

[13] The software is available from OCLC Research's Open Source Software pages: http://www.oclc.org/research/software/ Some applications which use this software are described in: Storey, Tom, "University repositories: an extension of the cooperative," *OCLC Newsletter*, 261, July 2003, pp. 7-11. Available in PDF from: http://www.oclc.org/news/publications/newsletters/oclc/2003/261/n261.pdf

[14] NDLTD web site: http://www.ndltd.org/

[15] For further details see the ETD project website: http://www.oclc.org/research/projects/etd/

[16] CONTENTdm web site: http://contentdm.com/
[17] For details about the Open WorldCat initiative see: http://www.oclc.org/worldcat/open/
[18] For details on the Dspace-OCLC-Google collaboration, see: http://www.oclc.org/research/projects/dspace
[19] OAIster is an OAI aggregation service provided by the University of Michigan; web site: http://oaister.umdl.umich.edu/o/oaister/
[20] Functional Requirements of Bibliographic Records. See Hickey, Thomas B., Edward T. O'Neill, and Jenny Toves, "Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR)," *D-Lib Magazine*, 8, no. 9, September 2002. http://www.dlib.org/dlib/september02/hickey/09hickey.html

[21] OCLC SRU/SRW web page: http://www.oclc.org/research/projects/webservices/
[22] Godby, Carol Jean, Devon Smith, and Eric Childress,"Two Paths to Interoperable MetadataIn *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop, September 28-October 2, 2003*, Seattle, Washington,USA Available online at: www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf
[23] Metadata Encoding and Transmission Standard (METS) maintained by the Library of Congress; web site: http://www.loc.gov/standards/mets/
[24] Godby, Carol Jean, Jeffrey A. Young, and Eric Childress,."A Repository of Metadata Crosswalks," *D-Lib Magazine*, 10, no. 12, December 2004. http://www.dlib.org/dlib/december04/godby/12godby.html
[25] OCLC ResearchWorks: http://www.oclc.org/research/researchworks/
[26] Koch, Traugott, "Activities to advance the powerful use of vocabularies in the digital environment - Structured overview." A presentation at given 30-September-2003 at the 2003 Dublin Core Conference, DC-2003: Supporting Communities of Discourse and Practice—Metadata Research & Applications, September 28-October 2, in Seattle, Washington (USA). http://www.lub.lu.se/~traugott/drafts/seattlespec-vocab.html
[27] Vizine-Goetz, Diane, "Terminology services: Making knowledge organization schemes more accessible to people and computers," *OCLC Newsletter*, 266, Oct.-Dec. 2004, p. 19. See also the OCLC Terminology Services project web site: http://www.oclc.org/research/projects/termservices/
[28] The info:kos scheme is described at: http://www.oclc.org/research/projects/termservices/resources/info-uri.htm
[29] GSAFD resources (files and services) are available at: http://www.oclc.org/research/projects/termservices/resources/gsafd.htm
[30] Really Simple Syndication, a protocol for serially distributing data as a "feed"; see Wikipedia entry: http://en.wikipedia.org/wiki/Rss
[31] OpenURL, a protocol for providing citation-like references in the body of a URL; See: Van de Sompel, and Oren Beit-Aire, "Open Linking in the Scholarly Information Environment Using the OpenURL Framework, "*D-Lib Magazine*, 7 no. 3, March 2001. http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html
[32] For a fuller description of ERRoLs, see the project website: http://www.oclc.org/research/projects/oairesolver/
[33] Experimental OAI registry at UIUC: http://gita.grainger.uiuc.edu/registry/
[34] See the OAI viewer at: http://errol.oclc.org/
[35] Automatic classification project web page: http://www.oclc.org/research/projects/auto_class
[36] Scorpion project web page: http://www.oclc.org/research/software/scorpion/
[37] WordSmith project web page (archived): http://orc.rsch.oclc.org:5061/
[38] Thompson, Roger; Shafer, Keith, and Vizine-Goetz, Diane, "Evaluating Dewey Concepts as a Knowledge Base for Automatic Subject Assignment." Paper presented at the first ACM Digital Libraries Workshop, January 1997.
[39] Vizine-Goetz, Diane, and Julianne Beal, "Using literary warrant to define a version of the DDC for automated classification services." In *Knowledge Organization and the Global Information Society; Proceedings of the Eighth International ISKO Conference, 13-16 July 2004, London, UK*, ed. Ia C. McIlwaine. (Vol. 9 in the *Advances in Knowledge Organization* series; ISBN 3-89913-357-9.) Würzburg (Germany): Ergon Verlag. Available online at: http://www.oclc.org/research/publications/archive/2004/vizine-goetz-beall.pdf
[40] OCLC Connexion web page: http://www.oclc.org/connexion/

[41] See OCLC Dewey Services web page: http://www.oclc.org/dewey/

[42] Godby, Jean, and Jay Stuler, "The Library of Congress Classification as a knowledge base for automatic subject categorization." In *Subject Retreival in a Networked Environment; Proceedings of the IFLA Satellite Meeting held in Dublin OH, 14-16 August 2001*, ed. Ia C. McIlwaine. (Vol. 25 in the *UBCIM Publications -- New Series*; ISBN 3-598-11634-9.) München (Germany): K.G. Saur. pp.163-169.

[43] FAST project web page: http://www.oclc.org/research/projects/fast

[44] FAST as a knowledge base for automatic classification project web page: http://www.oclc.org/research/projects/fastac

[45] Godby, Carol Jean and Ray R. Reighart, "The WordSmith Indexing System," *Annual Review of OCLC Research 1998*. http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003487

[46] Mantis project web page (archived): http://orc.rsch.oclc.org:6464/

[47] Hickey, Thomas B., "CORC--Cooperative Online Resource Catalog," *Annual Review of OCLC Research 1998*. http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003480

[48] Childress, Eric, "Crosswalking Metadata in the OCLC CORC Service," *Journal of Internet Cataloging* 4, nos.1/2, 2001, pp.81-88.

[49] See the project website for more details: http://www.rdn.ac.uk/projects/eprints-uk/

[50]Resource Discovery Network (RDN) web site: http://www.rdn.ac.uk