Final author manuscript (post peer review) version of Sarah L. Shreeves, Joanne S. Kaczmarek, Timothy W. Cole, (2003) "Harvesting cultural heritage metadata using the OAI Protocol", Library Hi Tech, Vol. 21 Iss: 2, pp.159 - 169

HARVESTING CULTURAL HERITAGE METADATA USING THE OAI PROTOCOL

Sarah L. Shreeves
Visiting Assistant Professor of Library Administration

Joanne S. Kaczmarek
Assistant University Archivist & Assistant Professor of Library Administration

Timothy W. Cole
Mathematics Librarian & Associate Professor of Library Administration

University of Illinois at Urbana-Champaign

Abstract

In July of 2001, with funding from the Andrew W. Mellon Foundation, the University of Illinois at Urbana-Champaign undertook a project to test the efficacy of using the Open Archives Initiative Protocol for Metadata Harvesting to construct a search and discovery service focused on information resources in the domain of cultural heritage. To date, the Illinois project has indexed over two million Dublin Core metadata records contributed by 38 metadata repositories in the museum, academic library, and digital library project communities. These records describe a mix of digital and analog primary content. Our analysis of these metadata records demonstrates wide divergence in descriptive metadata practices and the use and interpretation of Dublin Core metadata elements. Differences are particularly notable by community. This article provides an overview of the Illinois project, presents quantitative data about divergent metadata practices and element usage patterns, and details implications for metadata providers and harvesting services.

Keywords: Metadata Harvesting, Digital Libraries, Dublin Core, Open Archives Initiative

Acknowledgements: The University of Illinois Open Archives Initiative Metadata Harvesting Project, including the research described in this article, was funded by a grant from the Andrew W. Mellon Foundation.

Author Information:

Sarah L. Shreeves is Visiting Assistant Professor of Library Administration and Visiting Project Coordinator for the University of Illinois Open Archives Initiative Metadata Harvesting Project. From August 2001 until August 2002 she was the project's graduate research assistant. From 1992 until 2001 she was a member of staff at the Massachusetts Institute of Technology Libraries.

Email: sshreeve@uiuc.edu

Joanne S. Kaczmarek is Assistant University Archivist for Electronic Records Management and Assistant Professor of Library Administration at the University of Illinois at Urbana-Champaign. From September 2001 until August 2002 she served as Visiting Project Coordinator for the University of Illinois Open Archives Initiative Metadata Harvesting Project.

E-mail: jkaczmar@uiuc.edu

Timothy W. Cole is Mathematics Librarian and Associate Professor of Library Administration at the University of Illinois at Urbana-Champaign where he has been a member of the Library faculty since 1989. He has held prior appointments at Illinois as Assistant Engineering Librarian for Information Services and Systems Librarian for Digital Projects. He is Principal Investigator for the University of Illinois Open Archives Initiative Metadata Harvesting Project.

E-mail: t-cole3@uiuc.edu

HARVESTING CULTURAL HERITAGE METADATA USING THE OAI PROTOCOL

The Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH) is designed to facilitate the sharing and discovery of scholarly resources via the World Wide Web. Metadata describing many of these resources are contained in databases, XML documents, or other non-HTML formats and in locations not readily available to current Web search engines. These resources may represent materials that are culturally significant, such as rare books, manuscripts, and personal papers held by library archives, special collections, museums, and historical societies. Some of these resources may be digitized in a variety of formats, for example, digital images and audio or video files, while others may exist in analog format only. Analog-only resources are represented digitally only by metadata, which may be encoded as a MARC record or included as an element in a finding aid digitized in the Encoded Archival Description (EAD) standard. OAI-PMH enables enhanced access to both digital and analog resources. It does this by providing a protocol for the standardized dissemination of the metadata that describes these disparate collections. (Lagoze & Van de Sompel 2001)

One of the ways OAI-PMH–based harvesting services enable effective interoperability between and among content repositories is by facilitating the construction of services that present aggregated metadata to end users through search portals. These portals can be designed with specific communities in mind. The goal of the OAI-PMH project based at the University of Illinois at Urbana-Champaign and funded by the

Andrew W. Mellon Foundation is to test the efficacy of the OAI-PMH model for search and discovery of information resources in the domain of cultural heritage.

The major objectives of the Illinois project were to develop middleware tools for harvesting OAI-PMH–compliant metadata, to build a Web portal through which end users could search harvested metadata to discover cultural heritage resources of interest, to evaluate the potential utility of this approach to providing search and discovery services, and to identify issues that arise when implementing OAI-based services in this domain. (See the University of Illinois Cultural Heritage Repository at http://oai.grainger.uiuc.edu/search/.) The aggregated metadata, encoded in the Dublin Core (DC) metadata schema, has been harvested from a variety of metadata providers (repositories) and describes an array of resources. Much of the metadata has been mapped into DC from other schemas. Because this metadata originates from different communities and describes a heterogeneous collection of resources, a challenge for the Illinois project has been to develop strategies to effectively search and display it.

This paper provides an examination and analysis of a subset of metadata contained in the University of Illinois Cultural Heritage Repository. We provide an overview of the Illinois project including some of our technical findings. We examine the variability of the aggregated metadata both in terms of content and use of the elements in the DC schema. We analyze the metadata of multiple communities for differences and similarities in the use of DC, in particular, that of academic libraries, museums, other cultural and historical knowledge repository organizations, and digital libraries. In particular we explore the use and content of Dublin Core date and coverage metadata elements. Finally, we discuss ongoing and future strategies we plan to use to address the

issues raised by the variability in metadata and better understand the impact of these issues on end-user utility of aggregated metadata.

Overview of the Illinois OAI-PMH Project

The Illinois project, which began in July of 2001, faced two initial technical challenges: 1) to build from scratch middleware tools for implementing OAI-based metadata harvesting services, and 2) to harvest cultural heritage metadata and build a web portal to provide access to these materials.

Building a Harvesting Service

The first phase of the project focused on the construction of middleware tools, specifically a baseline harvesting service. In addition, we refined metadata provider tools developed during the alpha-testing phase of OAI-PMH. These tools aid institutions in making their metadata available in a manner compliant with the OAI-PMH. We have developed VisualBasic (VB) and Java versions of both the harvesting and provider tools and, in an effort to provide widespread access to these tools, we have made both versions available under an Open Source software license. The harvesting and provider tools can be downloaded from http://uilib-oai.sourceforge.net/.

Preliminary testing has shown that harvest times vary according to a few specific parameters. Tests conducted in 2001 and early 2002 demonstrated that for the OAI-PMH-compliant metadata providers, harvesting time was consistently provider- or network-limited rather than harvester-limited, even when relatively modest harvesting hardware was used (e.g., a Pentium III Windows NT workstation). Up to 10 simultaneous harvest jobs can be run from a single workstation without significantly slowing the harvest time of any one job. Harvesting moderate to large blocks of records through the *OAI*

ListRecords command rather than harvesting individual records through the *OAI GetRecord* command greatly reduces the time needed to harvest a collection (by as much as an order of magnitude). Filtering (selectively saving certain records) or normalizing (adding controlled vocabulary terms) during the harvesting process tends to slow harvest times, sometimes by as much as an order of magnitude.

Though harvest times vary due to variations in provider-side performance, typically 150,000 records (the number currently available from the Library of Congress OAI-PMH metadata provider site) can be harvested in as little as two hours. Assuming five simultaneous harvest jobs are running, this suggests that one workstation could harvest 10 million new records daily. This capacity is encouraging and implies fairly aggressive harvesting schedules, even from multiple repositories. It also suggests considerable excess capacity for harvesting the metadata currently available in the cultural heritage domain (a few million records distributed across perhaps 50 repositories). Additional testing is required, but the outlook for harvesting capacity and scalability of the metadata harvesting process itself is optimistic.

We did find that managing an OAI-PMH metadata harvesting service requires ongoing human intervention, at least in the present, still developmental phase in the lifecycle of the protocol. For instance, we estimate 1-3 person days per week would be needed to sustain a service provider of our size (see below). Even when a schedule for routine harvesting of desirable sites is established, scheduled jobs sometimes fail. Although there has been some improvement during the first ten months of active harvesting, failed harvests continue to occur weekly, typically due to the development environment in which we have been working or to instability of providers' baseURLs or

errors in assigning OAI record identifiers. Human intervention is typically required to resolve such issues. The process of identifying relevant metadata provider sites to harvest also takes time and as yet, no automated means exists to do this.

Building the Web Portal

The project has created a searchable database, called the University of Illinois Cultural Heritage Repository (http://oai.grainger.uiuc.edu/search), which contains 1,101,523 original metadata records. The web portal was built using the XPAT indexing and search engine tools developed by the University of Michigan. As of September 2002, we have collected metadata from thirty-nine metadata providers including individual museums and consortia of museums, archives, academic and public libraries, cultural and historical societies, and digital libraries. Table 1 gives a breakdown of metadata providers by institution type. Three of the repositories harvested (CIMI, the Online Archives of California, and the Colorado Digitization Project) are large-scale aggregators of metadata themselves. If we include the number of individual institutions within these aggregators, the Illinois project includes metadata for resources held in approximately 580 institutions. While the number of metadata providers is large it does not offer a full picture of the heterogeneous nature of the collection. In addition to aggregators, many providers have made available distinct and separately maintained collections of metadata using the sets feature of the OAI-PMH.

Table 1—Breakdown of Metadata Providers by Type of Repository

Type of Repository	**Number of repositories**	**Percentage of all repositories**
Museums/cultural and historical societies	7	18%
Academic libraries and archives	16	41%
Public libraries	2	5%
Digital libraries and consortia	14	36%
Total	*39*	*100%*

Of the thirty-nine metadata providers represented in our repository, nineteen have been officially registered with the Open Archives Initiative and certified by the Initiative as "OAI-compliant." We are able to directly harvest their metadata directly. The remaining repositories from which we have gathered metadata have not yet established OAI-compliant metadata provider services. To include metadata from these sites in our repository (desirable in order to better study the diversity and scale of metadata likely to be available via the OAI protocol in the near future), we obtain a data dump, or "snapshot," of metadata (typically using FTP) from each of these sites. After obtaining a snapshot, we mapped the metadata which was not in the DC schema to DC, then made metadata records from the snapshots available via a surrogate OAI metadata provider services running on our servers. These surrogate sites were then harvested using the OAI protocol and harvested metadata included in our repository.

Of the 1,101,523 original metadata records, 339,331 (30%) provide direct access to an online resource (e.g., digitized images, text) via a hyperlink. 53% of the records describe textual materials or sheet music and 27% describe images. Table 2 presents a breakdown of both the metadata and metadata providers by material type.

Table 2—Breakdown of Metadata and Metadata Providers by Type of Material (As of August 20, 2002)

Type of Material	Number of metadata records	Percentage of metadata records	Number of providers	Percentage of all providers
Images (photos, etc.)	305,460	27%	14	36%
Moving images	2,271	.2%	3	8%
Text and sheet music	597,351	53%	12	31%
Audio	934	.1%	1	2%
Physical objects	247,773	22%	4	10%
Websites	685	.1%	4	10%
Archival collections	13,670	1%	15	38%
EAD finding aids	8,730	1%	11	28%
Digital material (any type—image, text, etc.)	339,331	30%	25	64%

There is some overlap between categories. Percentages do not add up to 100%.

As noted in Table 2, 8,730 of the metadata records obtained (all from non-OAI compliant sites) were provided originally as EAD finding aid files (via a snapshot). Each EAD file describes a collection of items (such as personal papers or manuscripts) rather than an individual item. Because such collection-level EAD files do not describe individual items (as do individual DC metadata records), we developed algorithms to derive multiple DC metadata records, each describing an individual item, from EAD collection-level descriptions. (Prom and Habing 2002) The application of current algorithms automatically generates a total of 1,515,595 item-level records from the 8,730 finding aid records listed above. If we include all of these automatically-generated records in our aggregation, our searchable metadata collection exceeds 2.5 million records.

Analysis of Metadata Variability by Community

One objective of the Illinois project is to explore ways of effectively and meaningfully searching and displaying aggregated metadata. Dublin Core (DC) is a flexible and easily understood metadata schema that can be used for the description of both digital and non-digital resources. All DC elements are optional and repeatable. However the flexibility of the schema raises issues for metadata aggregators. The rather limited generalized guidelines for use are frequently supplemented by locally or community defined rules or application profiles. As a result, metadata authoring practices vary widely. Institutions may use the same DC element in different ways, implement a variety of controlled or local vocabularies, and include different levels of description in metadata records. This has a serious impact on the discoverability and usefulness of the

metadata in an aggregated resource, such as the University of Illinois Cultural Heritage Repository. (Cole et al., 2002)

There have been limited published investigations into how specific communities use DC. Guinchard (2002) conducted a survey on the use in libraries. Her findings showed that most respondents (largely academic libraries) use DC in combination with some other metadata schema. The major challenges faced by those using DC include the paucity of elements and the limited usage guidelines. In suggesting an opportunity for the Dublin Core Metadata Initiative (DCMI) to develop more thorough guidelines, Guinchard notes that "if these [the guidelines] were an integral part of the various application profiles, they might well… foster interoperability among like communities." Perkins (2001) describes the use of the OAI protocols in a museum community and briefly discusses the use of DC by museums. He notes that extensions (or alternatives) to DC are needed to provide the richness of detail the museum community requires. Liu et al (2002), discuss the interactive interface for Arc, a service of the Old Dominion University Digital Library Research Group and an OAI service provider. The Old Dominion researchers examine the variability of metadata harvested from 75 OAI-compliant repositories, but do not break the repositories into communities. They conclude that most repositories tend to use a controlled vocabulary for certain DC fields, but that the type and scope of these vocabularies vary enormously.

Other resources for exploring the varied use of DC are the application profiles and other documents from the DCMI working groups, which are available on the DCMI web site (http://www.dublincore.org). In addition, in 1999 CIMI produced guidelines for use of DC by the museum community. These guidelines provide in-depth descriptions and

suggestions for ways to use DC elements along with sample records for resources commonly found in museums. Although usage guidelines and application profiles exist for specific communities, they do not provide insight into how these communities actually use DC. In order to understand the different metadata-authoring practices across the types of institutions outlined in Table 1, we analyzed a sample set of metadata in our repository.

Methodology

We analyzed metadata originating from twenty-three of the thirty-nine metadata providers. All metadata was formatted in simple DC. Sixteen providers were OAI-PMH–compliant, and their metadata was harvested directly. Seven were from our surrogate provider services, but their metadata was already expressed in DC by the owning institutions. (We excluded from the sample metadata that we had mapped from a native format into DC.)

The metadata was inserted into a SQL database and a total of 613,813 records were analyzed. In order to examine the different authoring practices, we collected information from the SQL database about how each provider used each of the fifteen version 1.0 DC elements. (For definitions, see the Dublin Core Element Set at http://www.dublincore.org/documents/dces/.) We determined the number of records containing at least one instance of an element, as well as the total number of times each element was used. We also extracted the number of unique values for each element. If the number of unique values is low in relation to the number of times an element is used, a controlled vocabulary may be in use. (Liu et al, 2002) We also extracted the values used

for the coverage and date fields. Manual examination allowed us to determine the categories of content that was contained in these elements.

In order to analyze the differences in DC usage among the communities represented, we grouped the metadata providers into three subsets: 1) museums and cultural or historical organizations (6 institutions, 255,800 records); 2) academic libraries, including digital libraries rooted in an academic library (7 institutions, 235,294 records); and 3) autonomous digital libraries (10 institutions, 122,719 records). It should be noted that the results of this analysis are specific to this metadata and may not be fully generalizable.

Results and Discussion

The aggregate analysis of the use of DC elements for all twenty-three repositories is represented in Table 3. Although all of the repositories used date, identifier, and title at least once, only identifier appears in 100% of the records. The least-used elements are source (11% of records), format (32% of records), relation (39% of records), and language (41% of records). Coverage, subject, and type were the most-repeated elements.

Table 3—Use of Dublin Core for Total Sample Set (23 institutions, 613,813 records)

Dublin Core element	Percentage of repositories using element at least once	Number of records containing element	Total times element used	% of total records containing element	Average times used per record
contributor	61%	121,001	228,621	20%	1.89
coverage	61%	335,453	760,884	55%	2.27
creator	96%	395,267	427,077	64%	1.08
date	100%	362,973	408,651	59%	1.13
description	87%	314,857	546,891	51%	1.74
format	78%	199,421	275,597	32%	1.38
identifier	100%	611,553	789,442	100%	1.29
language	52%	249,630	250,276	41%	1.00
publisher	74%	427,195	520,612	70%	1.22
relation	48%	238,122	338,689	39%	1.42
rights	83%	388,551	499,225	63%	1.28
source	39%	66,137	66,455	11%	1.00
subject	96%	369,476	986,998	60%	2.67
title	100%	474,877	630,684	77%	1.33
type	83%	466,628	1,264,294	76%	2.71

Community Analysis

Table 4 delineates the use of DC elements by the community subsets. As follows from the aggregate analysis, only identifier is contained in 100% of the records in each subset. Within the museum community, type also is used in 100% of the records, publisher is used in 97% of the records, and subject is used in 93% of the records. Of the elements used by academic libraries, only identifier is used in more than 90% of the records. The next highest elements in use are creator (79% of all records) and title (66%). Of the elements in use in records from the digital library project community, both identifier and title are used in 100% of the records. Other high-use elements are creator (93% of the records) and type (97% of the records).

Table 4—Use of Dublin Core by Community Subsets

	Museums and Cultural/Historical Societies (6 total, 255,800 records)			**Academic Libraries (7 total, 235,294 records)**			**Digital Libraries (10 total, 122,719 records)**		
Element Name	% of repositories using element at least once	% of museum records containing element	Average times used per record	% of repositories using element at least once	% of academic lib. records containing element	Average times used per record	% of repositories using element at least once	% of digital library records containing element	Average times used per record
contributor	67%	45%	1.91	71%	2%	1.93	50%	2%	1.13
coverage	100%	69%	3.41	29%	41%	1.01	60%	51%	1.00
creator	83%	37%	1.02	100%	79%	1.03	100%	93%	1.22
date	100%	64%	1.08	100%	52%	1.06	100%	63%	1.33
description	100%	93%	1.64	71%	13%	2.24	90%	36%	1.88
format	83%	33%	1.77	100%	42%	1.05	60%	14%	1.38
identifier	100%	100%	1.55	100%	100%	1.13	100%	100%	1.06
language	17%	46%	1.00	57%	33%	1.01	70%	44%	1.00
publisher	67%	97%	1.30	86%	45%	1.06	70%	59%	1.19
relation	50%	79%	1.43	43%	11%	1.55	50%	8%	1.03
rights	100%	83%	1.50	86%	48%	1.00	70%	50%	1.07
source	50%	21%	1.00	43%	4%	1.00	30%	4%	1.06
subject	100%	93%	2.75	86%	15%	3.22	100%	78%	2.26
title	100%	77%	1.05	100%	66%	1.93	100%	100%	1.01
type	83%	100%	3.49	86%	39%	1.03	80%	97%	2.34

The least-used elements within records from museum community repositories are source (21% of the records), format (33% of the records), and creator (37% of the records). Within academic library community the least-used elements are contributor (2% of the records), source (4% of the records), relation (11% of the records), description (13% of the records), and subject (15% of the records). Within records from the digital library project community, the least-used elements are contributor (2% of the records), source (4% of the records), relation (8% of the records), and format (14% of the records). Table 5 compares high- and low-use elements and most- and least-repeated elements among the three subsets.

Table 5—Comparison of Use of Dublin Core Elements Across Communities

	Museum and Historical/ Cultural Societies	**Academic Libraries**	**Digital Libraries**
High-use elements (in more than 90% of records)	description identifier publisher subject type	identifier	creator identifier title type
Most-repeated elements (on average used more than twice per record)	subject type	description subject	subject type
Low-use elements (in less than 30% of records)	source	contributor description relation source subject	contributor format relation source
Least-repeated elements (used fewer than 1.1 times per record)	creator language source title	coverage creator date format language publisher rights source	coverage identifier language relation rights source title

The use or non-use of an element can have a significant impact on the discoverability of a specific record or group of records. For example, if an end user were to search only the subject and description elements (which are searched together on the Illinois portal's advanced search interface), the user would unknowingly miss somewhere between 72% and 85% of the records from the academic libraries subset. Cross-community discoverability is also affected by variations in element usage patterns. While the contributor element is used heavily by the museum community (45% of the records), it is used infrequently by the academic library and digital library project community (2% of the records in both communities). Inspection suggests differing interpretations as to the meaning and purpose of the contributor element. Searches for contributor element content will generally only retrieve records from the museum community. In the Illinois portal

contributor and creator element content is searched together to avoid misleading results. Because subject and author information is important for discovery, awareness of how and when these fields are used is key to developing an effective interface. The number of times an element is repeated within a record also could potentially affect a search engine's sorting and ranking of results.

The proceeding results and analysis on a per record basis tends to confirm Guinchard's repository-level survey, which reports that the most frequently used DC elements by repository were creator (97% of repositories responding) and title (93% of respondents). Identifier was used by 86% of respondents. (Guinchard 2002) Our results were similar: 100% of the academic library repositories included the creator, title, and identifier elements at least once. Like Guinchard's results the least used elements by repository are source (43% of respondents) and relation (43% of respondents).

Analysis of Coverage and Date Elements

In the previous analysis, we focused on how museum, academic library, and digital library communities use DC elements when structuring metadata. We also analyzed differences in the content within DC elements. In particular, we examined the *coverage* and *date* elements in the sample set. These elements are interesting because they often provide overlapping temporal information about a resource. By manual examination we determined that there were eight categories of information contained in the date and coverage elements:

- date of creation/publication/copyright (appeared variously in date and coverage)
- date of digitization (in date only)
- date of collection (in date only)
- date of metadata creation (in date only)
- era or range of years to which the resource belongs (temporal coverage) (in date and coverage)

- geographic area to which the resource belongs (spatial coverage) (in coverage only)
- subject of the resource (subject coverage) (in coverage only)
- type of resource (genre coverage) (in coverage only)

Table 6 provides a breakdown of the use of the coverage and date elements.

Table 6—Content of Dublin Core's* Coverage *and* Date *Elements by Institution

Content of element	**Repositories using *date***	**Repositories using *coverage***
Any temporal information	96%	26%
Created/published/ copyrighted	87%	4%
Digitized	13%	0%
Collected	4%	0%
Metadata created	9%	0%
Temporal coverage	9%	22%
Spatial coverage	0%	43%
Subject coverage	0%	13%
Genre coverage	0%	9%

The analysis showed a range of temporal information within the date element and overlapping uses of the date and coverage elements. The date of creation, publication, or copyright as well as temporal coverage were found in both elements. In addition to the range of categories of content, dates are displayed in a variety of formats, including standard ISO 8061 (e.g., 1900-12-31); range (e.g., 1940-c.1960); and general term (e.g. 19th century).

Vocabulary used for spatial coverage also varies from standard Library of Congress geographic headings to specific coordinates. These variations compound the challenges to discoverability already made difficult by differing structural uses of DC.

Strategies to Enhance Discoverability: Ongoing and Future work

The variability of the metadata as presented above raises a number of issues about the ability to collocate and effectively search the University of Illinois metadata aggregation. Use and non-use of elements that are particularly primarily used for resource discovery (such as creator, title, subject, and description) are particularly important for a service provider to understand when building a portal to aggregated metadata. We continue to investigate and implement a number of strategies to enhance the discoverability of our records.

Metadata Normalization

Normalization can enhance the discoverability of metadata records in a cross-collection repository. We investigated normalizing the type and the temporal aspect of the date and coverage elements and found the normalization process beneficial for these elements.

The goal of normalizing metadata is to enable users to get consistent and predictable results when searching across a heterogeneous collection of resources. It is likely that there will be some disagreement among metadata providers regarding the appropriate use of particular DC elements, even among members of a single community. Therefore, to effectively normalize metadata, it's important to understand how the element was initially interpreted by the metadata providers.

Once the use of a particular element is understood, we identified which, if any, vocabularies are used in it. If the majority of metadata providers use an existing controlled vocabulary, it could potentially be mapped and applied across the repository. It also may be possible to build a local controlled vocabulary to apply to records within our

repository. We were able to create a controlled vocabulary that applies to the type element and to the temporal aspects of the date and coverage elements. We believe that this enables end users to narrow their searches by type of material (such as image or text) or by range of years. The other obvious candidate for normalization is subject. However, the variability of vocabulary and values within this element is great, and the task of building such a vocabulary depends largely on human effort. (Liu et al, 2002)

Once the local controlled vocabulary is built and content is mapped to it, a programmer can write scripts that automatically augment records with the additional controlled vocabulary terms added in appropriate elements. A drawback of normalization is that as metadata is added from new repositories, some manual examination must occur in order to provide appropriate mappings. The value of normalization for end users remains to be tested in a structured manner.

Implications for Metadata Providers and Harvesters

A problem with the normalization of type was that, due to the manual effort involved, the process was applied only to terms that appeared at least 100 times in our repository. Some metadata providers used type rarely or not at all. While evaluating ways to restructure the indexes on XPAT to improve performance, we chose to base the index groups on type of material, as grouped top-down by repository or set (rather than by record). We believe that this approach is both more efficient and more global, as it covers those metadata providers who do not use type. Whether end users will find this an effective means of grouping records will be evaluated as an ongoing activity.

As a result of our experience with the diversity of metadata we harvested, we suggest that metadata providers give priority to dividing their metadata into sets. While there are

any number of logical groupings, we have found the most useful divisions to be by subject area, sub-collection, and/or type of material. Metadata aggregators may also want to use sets to indicate which institutions are included in their collection. Since the OAI-PMH allows for one record to belong to more than one set, it is possible for one collection to be divided into multiple sets. However, this geographic or institutional division may be less useful for end users and harvesting services.

An approach by which harvesting services can deal with differing uses of related elements is to index (or present for search) similar elements grouped together. For example, the Illinois portal searches together the contents of subject and description elements as well as searching together the creator and contributor elements. Our rationale is that the disctinctions originally assigned to these elements varies from repository to repository and in any event is not likely apparent to a diverse and broad spectrum of users. Since our goal is to provide a high-level discovery tool for a general and diverse user community, we believe this approach effective. In addition the initial simple search is a keyword search in every element of every record. In this way we attempt to provide a Google like search tool and avoid some of the issues raised by different interpretations of specific elements like subject or creator.

Data Mining

We have begun to explore the use of text-oriented data mining tools developed by NCSA at the University of Illinois at Urbana-Champaign. These tools apply systematic algorithms to data sets, identify document clusters of potential interest, and provide visual displays of these clusters and document similarities. We hope these tools will supplement gross manual-based groupings and sub-aggregations of metadata and enable the

automated co-location and clustering of similar resources. In particular, we hope data mining will be a useful tool to analyze similarities in and relationships between the subject and description elements, since manual analysis of the contents of these fields is prohibitively time and labor intensive.

While there is significant analysis yet to be done on the metadata in the metadata aggregation developed for this initial OAI-based project, preliminary findings support the belief that there is potential for effective search and discovery services built using the OAI protocol.

References

Cole, T.W., Kaczmarek, J., Marty, P.F., Prom, C.J., Sandore, B., & Shreeves, S.L. (2002), "Now that we've found the 'hidden web' what can we do with it? The Illinois Open Archives Initiative Metadata Harvesting experience" in D.Bearman & J.Trant (Eds), *Museums and the Web 2002: selected papers from an international conference p. 63-72*. http://www.archimuse.com/mw2002/papers/cole/cole.html, Accessed: 12 September 2002.

Dublin Core Metadata Initiative. (1999), "Dublin Core Element Set, Version 1.1, Reference Description", http://www.dublincore.org/documents/dces, Accessed: 17 September 2002.

Guinchard, C. (2002), "Dublin Core use in libraries: a survey", *OCLC systems & services*. v.18, no.1, pp 40-50.

Lagoze, C. & Van de Sompel, H. (2001), "The open archives initiative: building a low-barrier interoperability framework", in *JCDL 2001: Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries 2001 p.54-62.*

Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M.L., Knudson, F. & Holtkamp, I. (2002), "Federated searching interface techniques for heterogeneous OAI repositories", Journal of Digital Information, v. 2, iss.4. http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/, Accessed: 17 September 2002.

Perkins, J. (2001), "A new way of making cultural information resources visible on the web: museums and the Open Archives Initiative", in D.Bearman & J.Trant (Eds), *Museums and the Web 2001: selected papers from an international conference.*

http://www.archimuse.com/mw2001/papers/perkins/perkins.html, Accessed: 17 September 2002.

Prom, C.J. and Habing T.G. (2002), "Using the Open Archives Initiative Protocols with EAD", in G. Marchionini & W. Hersch (Eds), *JCDL 2002: Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries, July 14-18 2002 p.171-180*