

The CIC Metadata Portal: A Collaborative Effort in the Area of Digital Libraries

Muriel Foulonneau
Timothy W. Cole
Charles Blair
Peter C. Gorman
Kat Hagedorn
Jenn Riley

ABSTRACT. The CIC consortium includes 12 major Midwestern Universities. Their libraries have decided to share the cost of a joint project (2003-2006) aimed at better understanding the mechanisms by which emerging technologies and standards can facilitate metadata sharing and

Muriel Foulonneau is Visiting Assistant Professor of Library Administration and Project Coordinator for the CIC-OAI Metadata Portal, University of Illinois at Urbana-Champaign, Grainger Library, Urbana, IL 61801 (E-mail: mfoulonn@uiuc.edu). Timothy W. Cole is Mathematics Librarian and Professor of Library Administration, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (E-mail: t-cole3@uiuc.edu). Charles Blair is Co-Director, Digital Library Development Center, University of Chicago Library, Chicago, IL 60637 (E-mail: c-blair@uchicago.edu). Peter C. Gorman is Head, University of Wisconsin Digital Collections Center, University of Wisconsin-Madison, Madison, WI, 53706 (E-mail: pgorman@library.wisc.edu). Kat Hagedorn is Metadata Harvesting Librarian and OAIster Manager, University of Michigan Libraries, Digital Library Production Service, 300E Hatcher North, 920 North University Avenue, Ann Arbor, MI 48109-1205 (E-mail: khage@umich.edu). Jenn Riley is Metadata Librarian, Indiana University Digital Library Program, Wells Library E170, 1320 East 10th Street, Bloomington, IN 47405 (E-mail: jenlrile@indiana.edu).

Science & Technology Libraries, Vol. 26(3/4) 2006
Available online at <http://stl.haworthpress.com>
© 2006 by The Haworth Press, Inc. All rights reserved.
doi:10.1300/J122v26n03_08

the creation of value-added services for their users. The CIC metadata portal project has performed advanced work in the area of Open Archives Initiative Protocol for Metadata Harvesting, collection-level descriptions, metadata transformation and enrichment, and practices and usability of metadata standards. It has provided an opportunity for increased collaboration between CIC academic libraries and a way to highlight the wealth of digital resources held by the participating libraries. This article describes the project and enumerates project accomplishments. The project has helped to better the way in which partner institutions share information about digital content and provide access to digital resources. Four content providers of the project highlight different aspects of the project and the practical benefits they found in the collaboration. doi:10.1300/J122v26n03_08 [Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <<http://www.HaworthPress.com>> © 2006 by The Haworth Press, Inc. All rights reserved.]

KEYWORDS. Collaborative project, metadata sharing, shareable metadata, Open Archives Initiative Protocol for Metadata harvesting, metadata aggregation, thumbnails, collection level description, OAI set descriptions, communication between service providers and data providers, academic libraries collaboration

INTRODUCTION

The Committee on Institutional Cooperation (CIC), established in 1958, is a consortium of twelve major teaching and research universities in the Midwest (the eleven members of the Big Ten Athletic Conference and the University of Chicago). The CIC is committed to advancing academic excellence by promoting and coordinating collaborative activities and sharing resources. The endeavors of the CIC are organized to augment and complement individual institutional programs without supplanting them or reducing their importance. In recent years, the libraries of the CIC member universities have taken a leading role in helping to fulfill the promise of the CIC with initiatives such as the CIC Virtual Electronic Library (initiated in 1992), the establishment of the CIC Center for Library Initiatives (1994), and a new collaboration with CIC member University Presses (announced 2001).

In 2003, to better understand and begin addressing the growing need to manage and showcase the diverse digitized and born-digital collections being developed on the campuses of all CIC member institutions, the CIC Digital Library Initiatives Overview Committee (CIC-DLIOC) undertook a new initiative to investigate metadata sharing and interoperability. This project (<http://cicharvest.grainger.uiuc.edu>) utilizes the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and has so far resulted in the implementation of a collaborative metadata aggregation of more than 500,000 records hosted at the University of Illinois at Urbana-Champaign (UIUC). Some of the specific lessons learned to date about the use of OAI-PMH and the technical issues surrounding harvesting and use of aggregated metadata have been described elsewhere (Foulonneau et al., 2005; Foulonneau and Cole, 2005).

The purpose of this article is to illustrate how the collaborative aspects of this project have provided unique opportunities for concrete improvements in metadata quality and in the technical framework by which institutions share content. Each CIC member institution has approached the collaboration from a different context and with varying degrees of familiarity with OAI-PMH. Their experiences, accomplishments, and the experiences of UIUC as the host of the metadata aggregation and portals, are instructive and provide useful insights on some of the benefits of digital library collaborations today. Authors from five of the participating CIC contributed sections for this article, illustrating the wide range of perspectives and experiences embodied in this project.

CIC-OAI PROJECT OVERVIEW

Muriel Foulonneau and Timothy W. Cole
Grainger Library, University of Illinois at Urbana-Champaign

The Choice of the OAI-PMH

While there is some disagreement over the definitions, there are basically two approaches currently in general use for providing unified access via metasearch to distributed resources: Federation and aggregation. In federation, a query is performed through a unique interface, then a system translates and transmits (or broadcasts) the query to a number of data sources (targets), each of which answers with a list of relevant results. The central system merges and sometimes reorganizes the results

sent back by the target systems into a coherent list of results presented to the end user. In the aggregation approach to unifying access to distributed resources, the metadata or other information used to discover and locate resources is collected in a central place. As with federation, the resources themselves remain in their original location. A single centralized interface searches the aggregated metadata and indices centrally and points end users to the resources at their distributed locations.

The infrastructure necessary for aggregation is technically less demanding than for federation. OAI-PMH is a metadata aggregation approach. The University of Michigan and the University of Illinois at Urbana-Champaign were among the institutions involved in the development of the OAI-PMH from the beginning in 2000. Since its introduction, OAI-PMH has become ubiquitous, being used by projects ranging from the American South, American West, OAIster, and the National Science Digital Library (NSDL), to the future and promising Digital Library Federation (DLF) Aquifer projects. To study and experiment with metadata sharing and digital library interoperability, the CIC members thus decided on the use of the OAI protocol. Each institution, not having an active OAI provider agreed to implement OAI. Institutions with existing OAI implementations shared their experience and agreed to further develop and refine their OAI infrastructure. The service provider based at the University of Illinois at Urbana-Champaign harvests the content of all available CIC data providers every three weeks. It reprocesses the data and contributes results to support discussions of the features of both the technical implementation and the metadata by assessing their usability when taken and represented out of their original context (Shreeves et al., 2005).

Starting Point

When the CIC-OAI metadata collaboration began, six institutions had 330,000 metadata records describing resources that fit the project's collection-development policy. These records were available through 12 OAI data provider implementations. The project to date has facilitated both a growth in the quantity of resources available through OAI and improvements in the quality of the metadata and the OAI implementations. The first priority when the collaboration began was the creation of a collection-development policy for the virtual-CIC-metadata collection. This policy includes rules that can be applied postharvest, since multiple factors will affect the evolution (and therefore contents) of each individual OAI repository implementation. As an example, records from the Digital Library of the Commons maintained at Indiana

University (<<http://dlc.dlib.indiana.edu/>>) that are included as part of the Indiana University OAI data provider were excluded from the CIC aggregated collection on the basis that the collection was not specific to the CIC institution. Also, because one of the major objectives of this project is to investigate existing practices and facilitate reuse of metadata in new contexts, it was decided that, at least in the first phase in the development of the portal, the aggregated collection would include descriptions of both digital and analog-only resources.

An original analysis of early CIC-OAI project metadata harvests, done according to a methodology developed by Besiki Stvilia et al. (2004), demonstrated the heterogeneity of the metadata records and the limitations of the metadata in several of the collections. For example, 18% of harvested records had empty metadata elements. This analysis also allowed us to determine which metadata fields could be normalized and to anticipate the difficulties of doing so. Finally, a list of collections already available through OAI and those likely to be made available over the course of the project was created. This listing shown in Table 1 underlines the heterogeneity of the collections, and foreshadowed difficulties and interoperability barriers that would arise.

Accomplishments to Date

In the two years of the project so far, multiple views of the CIC metadata aggregation have been implemented and six additional OAI metadata providers have been implemented by CIC members. The number of identified collections in existing metadata providers grew from 89 to 179. The set and collection descriptions provide context to individual results. While the CIC metadata aggregation remains at this stage experimental (demand for an exclusively CIC aggregation remains uncertain), a number of useful findings have emerged. Several of these are related to the importance of set and collection-level descriptions that preserve valuable context for item-level metadata and allow making resource items more comprehensible to and discoverable by end users. Over the course of the project, we have also seen new use of richer metadata formats, such as MODS, and the implementation of new metadata providers (the current providers and harvested metadata formats are described in Table 1). Three institutions not previously using OAI have implemented OAI data providers. The project framework guaranteed technical support to some of the partners, although generally the main difficulty consisted in assessing the usability of the metadata exposed in the context of the CIC metadata portal. In a majority of cases,

TABLE 1. Configurations of CIC Metadata Repositories

Repository Name	Metadata Prefix Supported	Granularity	Deleted Records Support	Compression Method	# Sets	Set Description	Record Count Support	Resumption Token
Indiana University Digital Library Program	Oai_dc, mods	YYYY-MM-DD	Transient	gzip	7	No	Yes	Yes
Indiana University Bio Software Archive	Oai_dc	YYYY-MM-DD	Persistent	-	61	No	No	Yes
Michigan State University Digital and Multimedia Center	Oai_dc	YYYY-MM-DD	No	Gzip/deflate	3	Yes	No	Yes
PSU CONTENTdm Repository	Oai_dc	YYYY-MM-DD	Transient	-	5	No	No	Yes
Penn State Electronic Thesis and Dissertation Collection	Oai_dc, Oai_marc, Oai_rfc1807, Oai_etdms	YYYY-MM-DD	No	-	1	No	No	Yes
The KnowledgeBank at OSU	Oai_dc	YYYY-MM-DDThh:mm:ssZ	Persistent	Gzip/deflate	42	No	No	Yes
OhioLINK Electronic Thesis and Dissertation Center	Oai_dc, Oai_etdms	YYYY-MM-DD	No	-	15 (1 harvested)	No	No	Yes
UIUC Aerial Photographs	Oai_dc	YYYY-MM-DD	Transient	-	17	No	No	Yes
The University of Chicago Library Metadata Repository	Oai_dc, dc_qual, mods	YYYY-MM-DDThh:mm:ssZ	Transient	-	8	Yes	Yes	Yes
UoI at Chicago	Oai_dc	YYYY-MM-DD	No	-	0	No	Yes	Yes
Firstmonday repository	Oai_dc	YYYY-MM-DD	No	-	0	No	No	Yes
UIUC Engineering Documents Center Collection	Oai_dc	YYYY-MM-DD	No	-	0	No	No	Yes
UIUC Archival Finding Aids	Oai_dc	YYYY-MM-DD	Transient	-	2	No	No	Yes
UIUC Sheet Music	Oai_dc	YYYY-MM-DD	Transient	-	2	No	No	Yes
University of Michigan Library Digital Library Production Service	Oai_dc	YYYY-MM-DD	No	-	113	Yes	Yes	Yes
UIUC Digital Imaging And Media Technology Initiative	Oai_dc	YYYY-MM-DD	Transient	-	7	No	No	Yes
University of Minnesota IMAGES	Oai_dc	YYYY-MM-DD	No	-	0	No	No	Yes
University of Wisconsin-Madison Scout Archives	Oai_dc, nsdl_dc	YYYY-MM-DD	No	-	0	No	No	Yes
University of Wisconsin-Madison Library	Oai_dc, dc_qual	YYYY-MM-DD	No	Gzip/deflate	8	Yes	No	Yes

the technical improvement was less due to proper technical support than to the open dialogue that was built into the framework of the project.

A specific interface was created to make use of collection-level descriptions together with item-level descriptions in order to encourage the data providers to expose collection description thus improving the discoverability and interpretability of their data. Originally, no CIC data provider exposed collection-level descriptions. Currently four OAI repositories expose collection descriptions for 130 collections. These collection descriptions are available to any service provider willing to take advantage of them.

Success to date suggests a clear benefit of the collaboration to partners. Discoverability of their resources has been enhanced, not only through participation in a shared CIC search/discovery service but also through implementation of a number of additional functions helpful to the discovery, manipulation, and comprehension of relevant cultural and educative digital resources. This project was begun with active participation and financial support of ten of the thirteen CIC member institutions. By June 2005, all three of the remaining institutions had joined, bringing participation to 100% and testifying to the usefulness of collaboration in digital library projects. This project can serve as a model for other similar cooperative digital resource projects.

The impact of the collaboration has been different for each participant. The University of Chicago had been interested in implementing the OAI protocol for a long time. The CIC collaboration provided an especially good opportunity to implement a system and to test a number of solutions to adapt to the specific situation. Indiana University had already heavily invested in OAI-PMH and in the creation of high quality metadata and became a leader in the definition of best practices for sharable metadata in the initiative of the DLF and the National Science Digital Library. The University of Michigan has a very large metadata repository resulting from its extensive digitization program led by the library for several years. The CIC metadata portal was an opportunity to split its repository into comprehensive sets by collection and provide relevant descriptions for those sets as a way to improve metadata shareability. Finally, the University of Wisconsin-Madison went from a test repository to a production one including several collections corresponding to OAI sets, including set descriptions. The University of Wisconsin-Madison was particularly interested in the representation of its material in a different context and studied a number of possibilities to allow the CIC service provider to add thumbnails to its data. In the present paper, four CIC data providers present their experience in sharing

their content within the framework of the CIC metadata portal, providing insight on how different institutions handle the priority of content sharing versus local-production requirements, the issues that are raised in this context and the way in which a collaborative project provides a powerful framework to build more efficient content sharing.

IMPLEMENTATION OF AN OAI DATA PROVIDER AT THE UNIVERSITY OF CHICAGO

Charles Blair, University of Chicago

Descriptive metadata for digital objects created by the University of Chicago Library which are intended for harvesting via OAI-PMH exist in two forms: MARC <<http://www.loc.gov/marc/>> and non-MARC. Non-MARC metadata are recorded in a database at the point of digital-object creation, exported in tabular format, recoded automatically as UTF-8 (Universal Transformation Format 8 bits), and converted to XML. For each collection, a cataloger creates mappings from the rich, custom metadata for that collection to both simple and qualified Dublin Core. These mapping are turned into XSLT stylesheets and applied to the XML file for the collection; this results in one *oai_dc* (simple Dublin Core format) and one *dc_qual*-formatted file (Qualified Dublin Core) for each record in the collection. These records are designed to be embedded in an OAI-PMH response, which will add some elements, for example, an *about* element containing rights information. Each record contains a persistent identifier, which allows the Library to keep control of the target of the identifier field even after these records have been harvested.

Items in some collections have MARC records. These are exported in MODS format <<http://www.loc.gov/standards/mods/>> using the MARC to MODS mapping maintained by the Library of Congress. It is planned to create *oai_dc* records from these MODS records using the recently announced mapping from MODS to Dublin Core.

Originally, the thought had been to use GNU E-prints <<http://freshmeat.net/projects/eprints/>> as the Library's OAI provider, because it was being evaluated for another purpose, but E-prints only supports *oai_dc* out of the box, and this project wanted to use as rich metadata as possible. After a review of available software, the DLESE (Digital Library for Earth System Education) OAI Software <<http://www.dlese.org/oai/index.jsp>> was identified. DLESE is funded by the National Science Foundation,

and is closely partnered with the NSDL. This software has some very nice features:

- It is relatively straightforward to set up and use;
- It is well documented;
- It comes with both a provider and harvester;
- It makes it easy to expose collections;
- It allows XSLT to be applied dynamically;
- It works with disk-based files, which is in keeping with the Library's workflow.

In short, there is a lot to say for it. It, however, does have a few drawbacks.

First, it is finicky about file location. That is, if one asks the software to create a set from data in a particular location, after which one moves the data, telling the software to re-index the set from the new location, it will not allow it: One needs to come up with a new setSpec. This means that one must think through very clearly from the start where one wants one's files to live on disk; in an environment where things may move around, that is very inconvenient.

Second, metadataFormat and set are orthogonal concepts in OAI-PMH: One record in one set can exist in more than one format. However, with the DLESE software, if records exist in one set in more than one format, one has to define two sets, one for, say, oai_dc, and one for dc_qual. This is unfortunate, because the protocol is not supposed to work that way.

Finally, the software could not be kept running for very long unattended. It would unaccountably "go down." One cannot necessarily pin the blame on the software. DLESE is written in Java and requires Jakarta Tomcat to run: Perhaps there was a problem with the Tomcat version being used, or the interaction between that version of Tomcat and the DLESE software, or the version of Tomcat and the hardware, or perhaps it had something to do with the version of Java and one or all of the above. But the software was installed on two different platforms, one FreeBSD <<http://www.freebsd.org/>> and one Solaris, and the same problems were experienced. However, Tomcat itself was able to run successfully under both of these operating systems, and also under Linux and Mac OS X. All of these uses were in conjunction with Java applications. So the finger of suspicion points to the software, without there being necessarily any proof.

As much time was spent on this problem as was felt could be afforded, given the press of other work; since the software does not come with

source, spending more time with it did not seem like a worthwhile investment. As a result, it was decided to write a new provider, an idea that had arisen several times during this process, though each time it was pushed aside while one kept working with what one had. But in the end it was decided that even if the new software had problems, as initial versions of software invariably will have, one would be in a position to address them readily. In short, there was more confidence and comfort with the long-term maintenance of something completely under one's own control than with feeling at the mercy of code, of which neither the source nor the support were readily forthcoming.

Unlike both E-prints and DLESE, the new software separates indexing data from providing data. As a result, the provider does not care if data are re-indexed to reflect a new location. Also, the new software keeps metadataFormat and set orthogonal.

The initial version of the provider would occasionally time out. Testing locally using a tool such as curl <http://curl.haxx.se/> did not reveal this problem, but testing with the harvester as part of the project did. Profiling the code showed that the provider spent fully 80 percent of its time on one line of code, which opened and parsed a naively constructed disk-based index. The provider was subsequently rewritten to use a RAM-based index; response time with the current version of the provider is now virtually instantaneous.

Support for resumption tokens was not initially included in the provider. They are optional, but testing with the harvester as part of this project showed that it was desirable to include them; not including them from the outset necessitated refactoring the code several times. In the current version of the provider, resumption tokens are an opaque numeric string corresponding to a process, which caches the remainder of the response after each request by the harvester. Processes will time out if a harvester does not complete the transaction before the resumption token expires, and the memory they use will be freed. (In fact, there is a "grace period" after the resumption token is set to expire and before the associated process is actually terminated.)

Working with the project coordinator for the CIC-OAI Metadata Portal, Muriel Foulonneau, revealed several issues with the provider, which were then addressed as part of this project. In addition to the discovery of the timeout problem and the desirability of including resumption tokens, both of which have been discussed above, there were some issues with the validity of the responses in XML terms. Ms. Foulonneau provided helpful advice for addressing these issues, as well as more general advice related to the overall success of the project, such as recommend-

ing the provision of metadata in the richest-available metadata format for any particular collection, and options for indicating thumbnails, discussed elsewhere in this article.

**DEFINING AND DESCRIBING OAI SETS
FOR MORE THAN 250,000
UNIVERSITY OF MICHIGAN METADATA RECORDS**

Kat Hagedorn, University of Michigan

At the University of Michigan, the Digital Library Production Service (DLPS) provides services to the Library by digitizing texts, images and finding aids, and enabling online access to those digitized materials. From the beginning of our digital library development, we have organized our materials around the concept of “collections.” For instance, one of our well-known collections is Making of America (MOA), which contains historical texts written between 1830 and 1870 <<http://moa.umdl.umich.edu/>>. Another is the Bentley Historical Library’s collection of images <<http://images.umdl.umich.edu/b/bhl/>>.

Because we have organized our materials into collections, this was a natural choice when developing our OAI Data Provider service—providing OAI access to the metadata in these collections as OAI “sets.” However, with over 100 collections available, this wasn’t a feasible choice for making our metadata available in a short period of time. Instead, we concatenated the metadata from the text collections and from the image collections and served these large files through our OAI Data Provider tool.

As a member of the CIC, the University of Michigan Libraries was eager to collaborate with the developers of the CIC Portal project. Almost immediately, it became obvious that we would need to create the aforementioned collection-based sets. The CIC Portal team was developing recommendations for set description and collection description, for a variety of reasons:

- *Visibility*: Descriptive information makes it easier for Service Providers to decide which sets to harvest and use in their service.
- *Flexibility*: The set/collection descriptions could be used by Service Providers to enhance searching, by integrating these descriptions into search and browse functions.

- *Granularity*: The opportunity, with a new schema recommendation, to provide more detailed information within the set-description container (e.g., `dct:format` for the metadata format(s) used for the set).

As we had always intended to provide our sets based on our collections, these proposed recommendations gave us the incentive to start this process.

In order to understand our choices for providing set description in the manner we did, some background on DLPS practices is necessary. DLPS develops a suite of software, the Digital Library eXtension Service (DLXS), designed for the building and mounting of digital libraries. Each of our collections is developed and run under DLXS, and each collection's metadata (i.e., the collection title, the host URL, the available sorting functions, and the range of searchable dates) is contained within a database that the DLXS engine refers to for every search within our environment.

Since each of our collections has a handcrafted Web interface, much descriptive information about collections had never been codified as metadata elements within the collection-metadata database. In other words, proposed set/collection description metadata elements such as `dct:abstract`, `dct:accrualPolicy` and even `dc:identifier` (as unique from the host URL) were not available in the database for each collection [Set-description proposal for DLF best practices, based on the Dublin Core-Collection working group (Foulonneau and Shreeves, 2005)].

In addition, for each of our text and image collections, item-level metadata needed to be transferred to our metadata-only processing system. This system, called our Bibliographic Class, contains and processes just the metadata associated with our collections (as well as purchased bibliographies). Roughly, only one-third of our item-level metadata had been transferred by the time we were ready to conform to the recommendations of the CIC Portal team. Staff time commitments, as well as DLXS code changes that needed to be effected, slowed this process down further.

Because of this, we decided to only add one new metadata element to our collection-metadata database—`dc:description`—so that we could add summary information about the collection, access restrictions, and the collection URL within the one metadata element. While this was not CIC recommended practice, it gave us the opportunity to make minimal descriptive information available through our OAI Data Provider tool.

Differences from the DLF-NSDL Discussion Paper on OAI Sets

A discussion paper on set descriptions has been developed for the DLF-NSDL best practices on OAI and sharable metadata. This includes a proposal based on the use of the DC-Collection format (still under construction) for metadata collections (in OAI sets). The proposed framework would entail further discussion in-house about the interaction between our Data Provider tool and the collection description database. Since our Data Provider tool was developed by us, and is an integral part of the DLXS system, we may be able to make changes to the tool without affecting the database itself.

However, we may not be using all the proposed metadata elements exactly in their recommended form. For instance:

- We recognize the usability of having a textual summary of the collection in the `dct:abstract` field, however, we would find it far easier to instead link directly to the home page of the collection in a `dct:references` element (e.g., `dc:identifier`), where that type of description resides.
- The size of our sets (i.e., number of records) is already coded in the `completeListSize` attribute on the `resumptionTokens` for each set. Since our Data Provider tool uses our search engine to get this number, this is more accurate than the human-modified metadata element in the collection-description database. We are likely to continue to use the `resumptionToken` attribute instead of populating a `dct:extent` element.
- Accrual information is difficult to adhere to. Often a collection that we consider closed to new items will be reopened by new grant funding that we could not foresee. Development of a standard vocabulary, even a vocabulary that contains vague information (e.g., sporadically updated), might provide the kind of information OAI Service Providers find beneficial.
- Audience is not an issue that we have spent resources to discover. The majority of our materials are focused on the academic scholar, and while we could populate each set with a `dct:audience` value of “academic,” it might not be the correct approach.

The value of providing more descriptive information for Service Providers to work with has been incentive enough for us to better organize our OAI metadata records. The added benefit of working within the CIC

Portal, with the portal team for assistance and encouragement and as a system test bed for our changes, has been invaluable.

USE OF OAI SETS AND THUMBNAILS AT THE UNIVERSITY OF WISCONSIN-MADISON

Peter C. Gorman, University of Wisconsin-Madison

Using OAI Sets to Expose Collection Descriptions

The University of Wisconsin Digital Collections (UWDC) is organized into large “umbrella” collections into which new content is continually added. The descriptions of these collections must therefore evolve as they expand, so the UWDC Center sought a method for providing collection descriptions that would allow them to be updated with as little manual or duplicate effort as possible.

Since 1998, the UW Libraries have created and maintained Dublin Core-based records for digital collections encoded as HTML <meta> elements within the collections’ Home pages, the values derived originally from the Libraries’ catalog record for a collection, and maintained in situ as the collection grows. Since Wisconsin’s OAI infrastructure equates “Sets” with “Collections,” these DC records provided an obvious source for OAI Set Descriptions. A script was written to harvest the metadata from the collection Home page, transform it into OAI-compliant XML, and write it to a configuration file used by the OAI Data Provider in response to a “ListSets” verb. When the collection’s metadata is updated in the HTML page, those changes can be automatically propagated into the OAI Set Description.

Providing a Link in Metadata Records to Thumbnail Images

As discovery services built on aggregated metadata mature, it is natural to expect that they will embody features common to local repositories. With image content, for example, users are accustomed to seeing a surrogate of the digital object along with the metadata record in search results displays. While this is a standard feature of local systems providing management of digital objects, it is somewhat more difficult for metadata-aggregating services to harvest (or construct) unambiguous references to thumbnail images or other surrogates for digital objects located in remote repositories.

The CIC Metadata Portal Project considered several strategies for providing references to thumbnail images:

1. Encode the reference in unqualified Dublin Core, as an instance of `dc:relation`, or use Qualified Dublin Core to extend the relation to “`isVersionOf`.” Since in either case this element could be used to express a number of different relationships, the link’s specific purpose as a thumbnail reference would have to be made explicit in the element’s content.
2. Create a new element to contain the thumbnail reference, or include a thumbnail-reference element from an existing metadata schema, such as that created by the National Library of Australia (NLA) for the PictureAustralia collection (National Library of Australia. PictureAustralia Metadata Schema <URI:<http://www.pictureaustralia.org/schemas/pa/>>).
3. Derive the thumbnail reference, if possible, from a resource URI supplied in the `dc:identifier` or other element.

The first option, though attractive in not requiring modification of the standard DC metadata schemas, was ruled out, as the resulting value strings would mix element refinement semantics with value data.

Though labeled strings representing structured values are permitted by the DCMI Abstract Model (Dublin Core Metadata Initiative. DCMI Abstract Model, Appendix A. <URI:<http://dublincore.org/documents/2005/03/07/abstract-model/>>), particularly in combination with a syntax-encoding scheme, it was considered more in keeping with the philosophy of DCMI to express the element refinement directly rather than encapsulating it in a string value that would have to be parsed by the metadata aggregator.

It was decided, therefore, to import the National Library of Australia’s digital-object-element namespace into the CIC Qualified DC schema. The resulting `<thumbnail>` and `<location>` elements are defined as refinements of `dc:relation` and `dc:identifier`, respectively, making their semantics explicit without requiring string parsing by the harvesting service. This approach has the added advantage of leveraging existing practice, enhancing its potential sustainability and scalability (Figure 1). The new elements can be used by data providers either by referencing the NLA’s namespace explicitly, or by using the corresponding elements imported into the CIC Qualified DC schema.

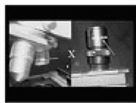
FIGURE 1. Thumbnails from the University of Wisconsin Science Collection are Located on the University of Wisconsin-Madison Library Server and Represented on the List of Results of the CIC Metadata Portal

Jump to Records: 1 | 26 | 51 ... 201 ... 401 ... 626 ... 826 ... 1051 ... 1251 ... 1451 ... 1676 ... 1876 ... 2101


Record 1 of 2102

Title	Side view of the microscope used in General Botany taught at the University of Wisconsin-Madison.	
Author/Creator	Clayton, Michael W.	
Type	Image	
URL	http://digital.library.wisc.edu/1711.d/SSRectIDSearch?repl1=Science&repl2=Science.1.1.bib	
Collection	The Science Collection	

Record 2 of 2102

Title	View of objective and ocular lenses showing their individual magnification.	
Author/Creator	Clayton, Michael W.	
Type	Image	
URL	http://digital.library.wisc.edu/1711.d/SSRectIDSearch?repl1=Science&repl2=Science.1.2.bib	
Collection	The Science Collection	

Record 3 of 2102

Title	Cross section of an alfalfa (Medicago) stem - prepared slide	
Author/Creator	Clayton, Michael W.	
Type	Image	
URL	http://digital.library.wisc.edu/1711.d/SSRectIDSearch?repl1=Science&repl2=Science.10-1-1.1.bib	

For data providers with the capability to define arbitrary metadata schemas, providing links to object surrogates is fairly straightforward. Existing structural metadata or repository-specific algorithms can be used to create the thumbnail references as <thumbnail> elements in the context of a CIC Qualified DC record (<http://cicharvest.grainger.uiuc.edu/schemas/QDC/2004/07/14/CICQualifiedDC.xsd>).

However, not all data providers have the ability to provide metadata in any format other than unqualified DC, and may have a limited ability to preprocess data values in response to harvesting requests. For that reason, aggregation services may need to consider the third approach defined above, processing object URIs using platform-specific algorithms documented by the data provider.

**THE BENEFITS OF COMMUNICATION
BETWEEN SERVICE PROVIDERS
AND DATA PROVIDERS**

Jenn Riley, Indiana University

As mentioned earlier, the CIC-OAI project has benefited greatly from an open dialogue between project participants. The OAI initiative, from the beginning, has focused on the idea of “community,” knowing that sharing of metadata is more effective when some shared semantics are applied. OAI promotes community implementations by allowing metadata formats supplementing simple Dublin Core, actively maintaining resources for implementers on the initiative Web site, providing friendly mailing lists where implementers and potential implementers can ask questions, and fostering an environment in which development of the protocol is done openly with a great deal of community involvement.

As seen by the discussion of the implementation of a new data provider at the University of Chicago, bringing new participants to OAI has been a major accomplishment of the CIC-OAI project. The addition of three new OAI data providers during this project demonstrates the power of community within the OAI environment. The critical mass of ten existing data providers contributing to this project at its inception allowed the staff at UIUC to develop a service provider that demonstrated the benefit to the CIC community of sharing metadata via OAI, and contributed heavily to the decision of the three remaining institutions to commit to the project.

The CIC-OAI project has provided great benefit to institutions with existing data providers as well. Data providers with limited experience planning for use of their metadata in a shared environment often do not realize the need to tailor metadata exposed via OAI for this shared environment. Native local metadata can be problematic in the shared environment by missing critical context, inclusion of metadata intended for local administrative needs, and inclusion of system-specific metadata (see, for example, Hutt and Riley, 2005; Cole and Shreeves, 2004). Communication with a service provider can help to teach data providers about the value of creating true shareable metadata and start to mitigate this problem. Similarly, service providers can communicate to data providers which metadata elements they can and will normalize for pooling with metadata from other institutions, so that the data provider can focus instead on elements the service provider does not normalize.

To this end, UIUC staff managing the CIC Metadata Portal have developed a set of guidelines intended to help data providers better

prepare their metadata for the shared environment (<http://cicarvest.grainger.uiuc.edu/dcguidelines.asp>). As the project has progressed, data providers have increasingly made use of these recommendations, and the recommendations have been iteratively added to and edited, as metadata has been harvested and strategies for indexing have been developed. As participating data providers see the developing portal and the improved retrieval possible with more robust and structured metadata, many have returned to existing records or made changes to procedures for generating OAI records for newly added collections. The CIC geographic browse interface tested by the portal is perhaps the most striking example of this phenomenon. Geographic browsing is not common among library databases; metadata creators in libraries rarely are able to see the results of careful application of geographic terms to metadata records. The demonstration of the benefit of geographic headings gained by the geographic browse interface of the CIC portal serves as a strong incentive to data providers to expend the effort to apply these headings completely and consistently (Figure 2).

Improving the quality of shared metadata is an ongoing activity. CIC institutions tend to have growing digital library collections, requiring periodic additions of new records to their data providers. Participants can see the collections exposed by others and use this information as a factor in prioritizing adding new collections to their OAI data providers. Each institution can learn from previous experience and improve their exposed metadata with each new batch of records added to a data provider.

Perhaps the greatest benefit gained by collaboration in OAI is guidance on the implementation of optional parts of the OAI protocol. For data providers, implementing the required aspects of the protocol is easy. Data provider software presumably already handles these functions, and as they are required for participation, they obviously are important for basic interoperability. Optional features, however, provide a much greater challenge. Many data providers make use of turnkey content-management software, such as ContentDM (<http://www.contentdm.com>), to provide OAI capability. These software packages may not implement various optional parts of the OAI protocol, including set support, metadata formats supplementing simple Dublin Core, and <about> containers. Data providers using these packages do not have the option of implementing any optional features their software does not allow. For data providers implementing open-source software packages, however, the situation is somewhat different. Often, CIC institutions have

FIGURE 2. Browseable Maps Providing Access to CIC Resources. (The Maps were Created by <<http://www.yourchildlearns.com/>>)



Reprinted with permission.

sufficient programming staff to add any optional features of the OAI protocol not already present in their open-source data provider software.

Ongoing communication with the UIUC staff harvesting and processing metadata for the CIC-OAI project has allowed data providers to prioritize their work, making informed decisions about which optional features to implement. The project's developing interest in making use of collection-level descriptions, for example, has prompted several data providers to include information about a collection in the OAI <setDescription> element. Similarly, the close relationship

between the service provider and data providers in the CIC-OAI project has resulted in the exposure of metadata formats other than simple Dublin Core by several data provider participants. Although the OAI protocol requires all records be exposed in simple Dublin Core, the protocol allows, and in fact, encourages, metadata in other formats to be exposed to meet community needs. Very few data providers, however, make use of this optional feature. The problem is of the “chicken and egg” variety—data providers have little incentive to expose supplementary metadata formats because few service providers make use of it, and service providers have little incentive to develop new procedures for making use of supplementary formats, because few data providers expose them for harvesting. The CIC-OAI project was able to break this cycle by dialogue—both sides realized the benefit to all of taking the step to use metadata more robust than simple Dublin Core, and mutually agreed to do the work required to realize this benefit. Several data providers chose to expose metadata in MODS (<http://www.loc.gov/standards/mods/>), as they were becoming familiar with this emerging format through local initiatives. Others exposed qualified Dublin Core according to a schema the CIC portal staff developed, using three relevant XML Schemas published by the Dublin Core Metadata Initiative.

Indiana University (IU) was one of the CIC institutions that loaded MODS records into their OAI data provider over the course of the CIC-OAI project. Staff in the IU Digital Library Program, who maintain an OAI data provider for digital library collections created and disseminated by the University, were interested in implementing MODS to learn about this up-and-coming format and its potential for use in other digital library projects, especially as a native-metadata format to describe photograph collections. Mapping a custom local format for a richly described collection of 14,500 slides to MODS for OAI exposure seemed to be a low-barrier method for quickly learning the MODS syntax along with its strengths and weaknesses. Through talks with CIC-OAI project staff, it became apparent that providing MODS through OAI could be beneficial not only to IU, but also to the CIC-OAI project and to users of the CIC metadata portal. The experiment has been a success. Indiana University is currently in the final stages of planning for a MODS implementation for sheet music collections and is in the preliminary stages of planning for the use of MODS in at least one other project. They will soon have MODS records for other collections available in their OAI data provider. CIC-OAI project staff has harvested MODS records from two CIC institutions and uses the richer MODS records to provide better access to portal materials.

The close relationship with a service provider does present challenges to a data provider, however. Tailoring metadata to assist one service provider may be detrimental to other service providers with different expectations. Any customizations a data provider makes to its metadata must still be understandable to other service providers. Emerging standards and best practices for metadata in a shared environment should serve to educate both data providers and service providers in these issues and more clearly define the boundaries in which specific agreements can operate.

The benefits from collaboration in the CIC-OAI project have not all resulted from formal documentation and demonstration. Many of the project partners have developed informal relationships as a result of collaboration on this project. These individuals serve as resources for the others, giving opinions on queries at any level of complexity regarding OAI or related issues over email or the phone. The personal relationships developed have greatly enhanced the value of the CIC-OAI collaboration by allowing each institution to think of the others as partners rather than simple content or service providers. Each has become invested in the successful outcomes achieved together.

INSTITUTIONAL COLLABORATION AS A STARTING POINT FOR EFFICIENT CONTENT SHARING

Muriel Foulonneau, University of Illinois at Urbana-Champaign

The CIC-OAI project has provided an opportunity for partner institutions to test state-of-the-art techniques designed to handle and represent their metadata in a shared, interoperable context. Each partner has brought something of their unique experience to this joint community effort, and through this project, each partner has had an opportunity to help shape and guide the evolution of digital libraries.







Innovative Interfaces and New Research Threads

The CIC-OAI metadata portal collection-item interfaces aim to improve discoverability of resources through the use of collection-level descriptions in concert with item-level descriptions (Foulonneau et al., 2005). In both the DLXS-based and the SQL server-based interfaces a number of functionalities have been implemented that take advantage of collection-level descriptions. Collection-level descriptions are used (among other things) to enhance the filtering of search results and to

augment metadata record displays in search result pages. These interfaces help put aggregated item-level metadata records back in original context by restoring links between item-level metadata records and collection descriptions. This facilitates end-user discovery and (perhaps more importantly) end-user selection from search results and browse listings of metadata records. Collections are also recognized as useful resources on their own, albeit at a different level of granularity (see Figure 3). Further experiments are underway to exploit (for metadata enrichment) the full text of resource items and associated content provider Web pages.

In order to facilitate the identification and selection of resources in the CIC aggregation, thumbnails and thumbshots (thumbnails created out of Web page snapshots) have also been added, with thumbnails either provided by data providers (e.g., the University of Wisconsin-Madison, see Figure 1) or generated by the service provider at Illinois

FIGURE 3. List of CIC Collections Available on the CIC Metadata Portal

University of Chicago		
<p><u>American Environmental Photographs, 1891-1936: Images from the University of Chicago Library</u></p> <p>This collection consists of approximately 4,500 photographs documenting natural environments, ecologies, and plant communities in the United States at the end of the nineteenth and the beginning of the twentieth century. Produced between 1891 and</p>	4519 items	
<p><u>Archival Photofiles</u></p> <p>The Archival Photographic Files (the Photofiles) constitute a separate record group in the University Archives containing most of its photographic holdings. Some 60,000 in number, the photographs are arranged in five series which visually document</p>	8968 items	
<p><u>The First American West: The Ohio River Valley, 1750-1820</u></p> <p>This collection consists of 15,000 pages of original historical material documenting the land, peoples, exploration, and transformation of the trans-Appalachian West from the mid-eighteenth to the early nineteenth century. The collection is drawn</p>	731 items	
University of Illinois at Chicago		
<p><u>First Monday Journal</u></p>	586 items	
University of Illinois at Urbana-Champaign		
<p><u>American Library Association on-line archives</u></p> <p>Includes photographs of past ALA conferences and images of library buildings around the country</p>	246 items	
<p><u>Illinois Historical Maps Online</u></p> <p>Includes maps charting the last 400 years of historical development in Illinois and the Northwest Territory and topographic maps of Illinois.</p>	417 items	

through a specific application that captures images on data providers' Websites and automatically creates thumbnails (*Thumbgrabber* application developed by Tom Habing and Muriel Foulonneau at the University of Illinois at Urbana-Champaign <http://sourceforge.net/project/showfiles.php?group_id=47963&package_id=159364> described in [Foulonneau et al., 2006]).

An interface also displays results through browseable maps. Geographic coverage is identified in either subject or coverage fields of metadata records or at collection level. The CIC resources cover 175 countries, with resource items in above 80 languages. The browseable maps of four continents show the variety of resources that are available on the metadata portal.

Further research is currently underway to identify the characteristics of metadata records that perform best in the portal and to help define guidelines for shareable metadata for the CIC consortium. Results from the CIC-OAI project are also informing the ongoing DLF-NSDL initiative to create best practices for OAI and shareable metadata (<http://oai-best.comm.nsdll.org/cgi-bin/wiki.pl>).

The project has helped to strengthen the relationships between partner institutions and further joint research activities and can serve as a model for similar projects within other consortium. Results have so far been presented in the United States at the Joint Conference on Digital Libraries 2005, in Taiwan at the International Conference on Digital Archive Technologies 2005, (Foulonneau, 2005) and in Europe at the European Conference on Digital Libraries 2005. The project has made a significant contribution and sparked interest at national and international levels. This project represents one of the ways the CIC consortium is contributing to the evolution of the field of digital libraries.

Data Provider Improvements Have No Impact Without Service Provider Improvements

In 2003, Martin Halbert et al. (Halbert, 2003) analyzed the “barriers to adapting the [OAI-PMH] protocol” in academic institutions and suggested that collaborative projects create incentive for individual institutions to share their metadata. The CIC metadata portal is among the projects that have emerged to encourage institutions to more efficiently and effectively share their digital content on the Web. The collaboration mechanism relies on allowing data providers to immediately see and assess the results of any improvements in their content sharing framework.

The simultaneous collaboration on both data provider practices and service provider implementation helps to break the “chicken and egg” problem between data and service providers and stimulates the implementation of new functions and services. Data providers have greater incentive to create thumbnails if they know that a service provider will take advantage of and exploit the thumbnails created. Similarly, until data providers can be assured that a service provider would harvest and exploit MODS metadata records, they are disinclined to make MODS records available, in spite of a clear community consensus that richer metadata formats, like MODS, are desirable.

Incentive for data providers to improve the metadata they share and to make available components, like thumbnails and collection descriptions, will grow only as service providers demonstrate that such features can improve the visibility and discoverability of resources in an interoperable, aggregated context. Service providers can do so only when data providers make such improvements available. The CIC collaboration has proven an efficient ground to test innovation together, discover best practices, and advertise contributions for the future of digital libraries.

Received: December 12, 2005

Revised: January 31, 2006

Accepted: February 6, 2006

AUTHORS' NOTE

The authors wish to acknowledge the libraries of the following participating CIC member institutions for providing metadata and participating in the discussions on metadata sharing: University of Chicago, University of Illinois at Chicago, University of Illinois at Urbana-Champaign, Indiana University, University of Iowa, University of Michigan, Michigan State University, University of Minnesota, Northwestern University, Ohio State University, Pennsylvania State University, Purdue University and the University of Wisconsin-Madison. This work was supported by a grant from the Committee of Institutional Cooperation's Center for Library Initiatives.

REFERENCES

- Cole, Timothy W., and Shreeves, Sarah L. “Lessons Learned from the Illinois OAI Metadata Harvesting Project.” In Diane I. Hillmann and Elaine L. Westbrook, eds. *Metadata in Practice*. Chicago: American Library Association, 2004.
- Foulonneau, Muriel and Cole, Timothy W. Strategies for reprocessing aggregated metadata. In *9th European Conference on Digital Libraries, ECDL 2005, Septem-*

- ber 18-23, 2005, Vienna, Austria. (Proceedings Series: Lecture Notes in Computer Science.) Heidelberg: Springer-Verlag. <http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/11551362_26>.
- Foulonneau, Muriel, Cole, Timothy W., Habing, Thomas G., and Shreeves, Sarah L. (2005). Using Collection Descriptions to Enhance an Aggregation of Harvested Item-Level Metadata. In *JCDL 2005: Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries [Denver, June 7–11]*. New York, Association for Computing Machinery: pp. 32-41. < <http://doi.acm.org/10.1145/1065385.1065393>>.
- Foulonneau, Muriel and Shreeves, Sarah L., Describing OAI Sets: A Discussion Paper, July 2005, http://comm.nsd.org/download.php/623/set-description_ver6.doc [August 2005].
- Foulonneau, Muriel. Metadata aggregation for digital libraries, June 2005, <http://www.iis.sinica.edu.tw/ICDAT05/FILES/foulonneau.pdf> [August 2005].
- Halbert, Martin, Kaczmarek, Joanne, and Hagedorn, Kat. Findings from the Mellon Metadata Harvesting Initiative, in Traugott Koch, Ingeborg Torvik Sølvberg (eds.) *Lecture Notes in Computer Science, Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003 Proceedings*, Springer-Verlag GmbH, Feb 2004, pp. 58-69, <http://www.springerlink.com/app/home/contribution.asp?wasp=6590eb7adfc74a07acf9a25e6c518fcd&referrer=parent&backto=issue,7,47;journal,780,2099;linkingpublicationresults,1:105633,1>.
- Hutt, Arwen and Riley, Jenn. Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data Providers of Cultural Heritage Materials Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries, 2005 <http://doi.acm.org/10.1145/1065385.1065447>.
- Muriel Foulonneau, Habing, Thomas G., and Cole, Timothy W., Automated Capture of Thumbnails and Thumbshots for Use by Metadata Aggregation Services, in *D-Lib Magazine*, vol 12 #1, January 2006 <<http://www.dlib.org/dlib/january06/foulonneau/01foulonneau.html>>.
- National Library of Australia. *PictureAustralia Metadata Schema* <URI:[http:// www.pictureaustralia.org/schemas/pa/](http://www.pictureaustralia.org/schemas/pa/)>.
- Shreeves, Sarah L., Knutson, Ellen M., Stvilia, Besiki, Palmer, Carole L., Twidale, Mike B., and Cole, Timothy W. (2005). Is 'quality' metadata 'shareable' metadata? The implications of local metadata practice on federated collections. In H.A. Thompson (Ed.) *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, [April 7-10, 2005, Minneapolis, MN]*. Chicago, IL: Association of College and Research Libraries: pp. 223-237.
- Stvilia, Besiki, Gasser, Less, Twidale, Mike, Shreeves, Sarah L., and Cole, Timothy W. (2004). Metadata quality for federated collections. In *Proceedings of ICIQ04–9th International Conference on Information Quality. Cambridge, MA*: pp. 111-125.