OAIster: a “no dead ends” OAI service provider

Kat Hagedorn

The author

Kat Hagedorn is Metadata Harvesting Project Librarian at the University of Michigan and Manager of the OAIster project (http://www.oaister.org/), she maintains the OAIster site and is responsible for Bibliographic Class in the Digital Library Production Service, University of Michigan Libraries, Ann Arbor, MI, USA.

Keywords

Data collection techniques, Digital libraries, Archiving

Abstract

OAIster, at the University of Michigan, University Libraries, Digital Library Production Service (DLPS), is an Andrew W. Mellon Foundation grant-funded project designed to test the feasibility of using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to harvest digital object metadata from multiple and varied digital object repositories and develop a service to allow end-users to access that metadata. This article describes in-depth the development of our system to harvest, store, transform the metadata into Digital Library eXtension Service (DLXS) Bibliographic Class format, build indexes and make the metadata searchable through an interface using the XPAT search engine. Results of the testing of our service and statistics on usage are reported, as well as the issues that we have encountered during our harvesting and transformation operations. The article closes by discussing the future improvements and potential of OAIster and the OAI-PMH protocol.

Electronic access

The Emerald Research Register for this journal is available at **http://www.emeraldinsight.com/researchregister**

The current issue and full text archive of this journal is available at **http://www.emeraldinsight.com/0737-8831.htm**

Library Hi Tech
Volume 21 · Number 2 · 2003 · pp. 170-181
© MCB UP Limited · ISSN 0737-8831
DOI 10.1108/07378830310479811

Background

In July 1999, the process for developing the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, henceforth called the OAI Protocol) was initiated[1]. The protocol was the brainchild of the Open Archives Initiative, started by Paul Ginsparg, Rick Luce and Herbert Van de Sompel, to enable interoperability among metadata providers, i.e. those who wanted a wider audience to be able to access their digital objects, at that point mostly e-print archives interested in open scholarly publication (Van de Sompel and Lagoze, 2000).

The protocol was designed to make this interoperability as easy as possible for data providers and service providers alike. Data providers would conform to a standard set of requests, and register themselves on the OAI site[2]. Service providers would be capable of harvesting (i.e. requesting and receiving) this metadata, aggregating disparate data providers’ metadata, and making the metadata available through a unified portal to end-users (Lagoze and Van de Sompel, 2001).

Digital libraries outside the e-print community were soon involved, since they had difficulties similar to e-print archives in making their collections more widely known. Although generally available through the institution’s library Web site, these collections have typically been difficult to locate and use by end-users. The efforts of envisioning digital collections as cross-institutional digital libraries would be furthered by using Open Archives Initiative requirements (Besser, 2002).

The benefits of interoperability include:

- the opportunity to share methods, results, standards and future endeavors more readily with those in the open scholarly publishing and digital library community;
- the possibility of making those outside their community aware of their efforts and resources; and
- the potential for aggregation of their repositories with other data providers’

The author would like to thank the Andrew W. Mellon Foundation for their generous grant supporting this project, and John Wilkin, Mike Burek, Ingrid Erickson and the DLXS staff for all their hard work and encouragement.

repositories, thereby providing single access points for scholars interested in digital objects.

The Andrew W. Mellon Foundation funded seven projects starting in 2001 to test this latter point – how the new protocol could be used to make repositories' collections more available to end-users. In other words, the projects were designed to demonstrate the capabilities of service providers to aggregate metadata and create a tool for discovery of this metadata by segments of the scholarly community (Waters, 2001).

OAIster, at the University of Michigan (UM), University Libraries, Digital Library Production Service (DLPS), was one of these projects[3]. We tested the feasibility of aggregating metadata that link to actual digital representations, e.g. an Anna Karenina text, the image of a Calder sculpture. Catalogs or other pure reference information were not of interest to us – we wanted to provide a service that allowed end-users to link to all types of digital representations on all topics.

We were not the only ones involved in testing the OAI Protocol. Others include (not limited to the Andrew W. Mellon Foundation funded projects):

- The University of Illinois at Urbana-Champaign (UIUC) initiative, another Andrew W. Mellon Foundation grant project, was designed to test the service provider model by creating a portal specific to cultural heritage, and focused on EAD finding aids[4].
- Two Emory University projects, SOLINET and AmericanSouth.org, both Andrew W. Mellon Foundation grantees, were designed to harvest archives of the Southern experience, utilizing scholars themselves to help design a structure that promotes research, teaching, and communication[5].
- The ARC Cross Archive Searching Service, a large-scale service provider, started early on and allowed advanced searching of many of the OAI-registered repositories. Their efforts have led them to believe that one unified search interface, with controlled values populated across harvested repositories, is possible (Liu *et al.*, 2002).
- The SDARTS project, a protocol and toolkit that was developed to work with collections that can be both SDART-compliant and OAI-compliant (Ipeirotis *et al.*, 2002).

The OAIster proposal

The OAIster project as proposed was designed to establish a broad, generic information retrieval resource pointing to publicly available digital resource representations, mostly provided by the research library community.

We expected the proposed service to:

- Begin finally to reveal "hidden Web" digital resources in a way that they are not now revealed. Digital resources are often hidden from the public because a search engine, like Google or Altavista, cannot get past an institution's search forms or CGI scripts to the databases that store the resource information. While the resources are accessible through the institution's interface, end-users need to know where to look for this interface (Bergman, 2001).
- Have no dead ends. End-users would not retrieve just the metadata about resources – they would have access to the online representation of those resources. For instance, instead of just viewing the catalog records of a slide collection of Van Gogh's works, end-users would be able to view images of the actual works.
- Provide one-stop "shopping" for end-users interested in digital resources. OAIster would be accessible to anyone who needed to use digital objects and would encompass as broad a collection of resources as possible (i.e. with no subject parameters).
- Be easily findable and viewable. The middleware we use to index these resources makes this possible. DLPS offers the Digital Library eXtension Service (DLXS) framework to produce large and varied digital collections[6]. The components of this framework include tools for mounting collections (i.e. class middleware) and XPAT, a powerful SGML/XML aware

search engine. This framework would allow us to quickly and efficiently host OAIster.

The aggregation of digital resource metadata would be achieved through harvesting OAI-compliant data repositories' metadata. We would collaborate with UIUC, using their harvester, and be an early release site for their tool. On our end, we would need to develop mechanisms to store and manipulate the harvested data in a system enhanced for information retrieval.

The OAIster project would utilize DLXS middleware to transform the harvested metadata into a standard format (DLXS Bibliographic Class)[7]. The Bibliographic Class DTD elements are designed to handle relatively flat bibliographic information, i.e. collections such as encyclopedias, reference texts, catalogs, and other uses of stand-alone metadata[8]. In this respect, this class would work effectively to encode the harvested metadata that would be indexed and made searchable through XPAT.

The manipulation (i.e. transformation) tool we developed, and the framework for OAIster itself, would be made available to DLXS framework customers in the release following completion of the project.

Our methodology

Starting in December of 2001 and continuing through February of 2002, we familiarized ourselves with the OAI landscape – those working on the OAI protocol, those developing OAI-enabled repositories, those becoming service providers who would harvest those repositories, those building tools to provide and harvest, and those interested in the open archives, digital libraries and free scholarship movements in general. From February through June 2002, we developed our harvesting and transformation system, performed user testing and designed our search interface. Since July 2002, we have been harvesting on a regular, periodic basis, making this metadata searchable through our Web interface. As this article is being written, we are planning a second set of user testing and interface improvements.

Our harvesting and transformation system

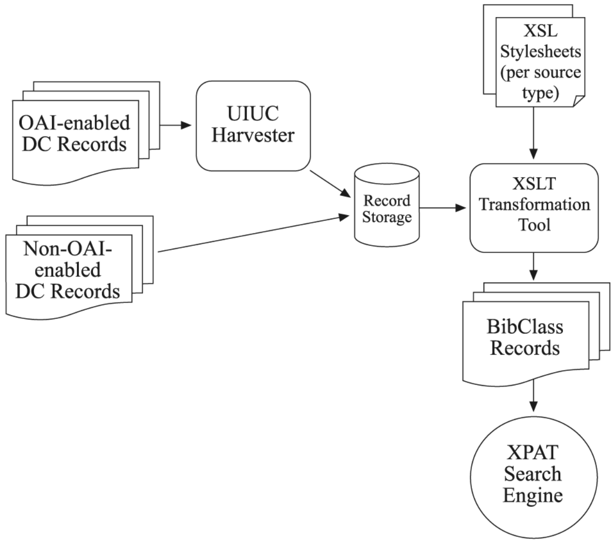
Gaining familiarity with OAI and its environment gave us the impetus to develop the most appropriate system for our purposes. Figure 1 illustrates the design of our system.

Simply put, we expected to harvest Dublin Core (DC) encoded metadata in XML format from OAI-enabled metadata repositories, use XSLT to transform that DC metadata into DLXS Bibliographic Class encoded metadata, our native format, index this metadata and make it available to end-users through an interface that used the XPAT search engine. (Non-OAI-harvested DC records are discussed in the following section.)

To harvest the metadata, we used a harvester developed at UIUC for our joint purposes. A Java version of the harvester was designed by UIUC to be used by us in our Unix-based environment. Prior to the Java version being built, we created our own testing harvester to harvest a sample of repositories, in order to begin developing the transformation tool we would need to handle the harvested metadata. We are currently running version 2.0B4 of the Java harvester, as created and tested by Joanne Kaczmarek, Yuping Tseng and Thomas Habing of UIUC and available via SourceForge[9].

The transformation tool we developed is written in Java which in turn uses XSLT to transform DC records to Bibliographic Class

Figure 1 Design of harvesting and transformation system



records. The function of the Java code is as follows:

- collect individual harvested records stored in directory trees in our file system into large files ready for transformation;
- automate the above over all repositories;
- filter out records that do not have digital objects associated with them;
- normalize the DC element Resource Type and populate the Bibliographic Class NORM element with the normalized information;
- add in institution information to the Bibliographic Class INST element;
- count records and provide quality of data feedback;
- convert UTF-8 to ISO-8859-1, which is required for the XPAT search engine; and
- use XSLT to transform DC records which have been preprocessed by Java into Bibliographic Class records.

Our first step in developing the transformation tool was creating a mapping between DC and Bibliographic Class, i.e. correlating a DC element with a Bibliographic Class element. For example, the DC Creator element was mapped to the Bibliographic Class L element, which we then labeled "Author" for the searching interface. Table I indicates the mapping we performed.

All metadata values are displayed "as is," without modification. At present, the DC Contributor and DC Relation elements are not being mapped or displayed. We expect to re-visit the element mappings and their display labels soon.

The tool parses the retrieved metadata to find only those records that have DC Identifier elements with values that contain valid URLs. For example, we are not interested in making a record with only the DC identifier "vrc1006840" available, but are interested in making a record with the DC identifier "http://hdl.loc.gov/loc.gmd/g3195.ct000379" available. We separate out these records because we only want to provide records that point to actual digital objects.

After filtering, the tool uses a normalization table to transform DC resource type values such as "book" and "paper" to the normalized value "text," and values such as "illustration" and "picture“ to the normalized value "image." The table was manually created from a retrieval of unique DC resource type values among all harvested records. The resulting normalized value is placed in the Bibliographic Class NORM element, specifically created for this purpose. Consequently, in the interface, end-users are able to limit their search to those records in which the NORM element value equals, for example, "text."

Table II shows an original DC example record, and the Bibliographic Class counterpart after transformation.

After transformation, the tool counts the number of records per repository that were selected. This provides us with numbers we can place on our Web site, for end-users to see how many records are available per repository and how many records are available in total[10].

At present, we manually process the indexing of the transformed records in our system. Once the records are indexed using appropriate data dictionary and data region files, we test and then make this data available through a CGI interface, where it is searchable using the XPAT search engine.

Design and launch of the search interface

In order to design the appropriate interface, and to be sure that desired features and functionality informed actual development of the interface, we spent considerable time discovering the needs of end-users.

Our first step was to develop an online survey that could assist us in determining what end-users might want from a system like OAIster. We were interested in what sorts of digital resources end-users were interested in when working online, what they looked for but were not able to find, and some of the problems they ran into when looking for information online. A summary, plus the original questions and raw results, is available on the OAIster site[11]. We received 591 responses over the month that the survey was open.

Our second step was to design a potential user interface, based on the many years of work behind the interfaces of the DLPS collections, that would effectively showcase the aggregated metadata we harvested. Our designs started on paper and were subsequently discussed, changed and developed in DreamWeaver. We

Table I Mapping between OAI and DC elements and Bibliographic Class elements

Original elements	Example value	BibClass element	Displayed as . . .
OAI elements			
Identifier	oai:VTETD:etd-92398-135228	ID attribute for A (i.e. complete record) element	For internal use
Datestamp	1998-10-23	DT attribute for A (i.e. complete record) element	For internal use
DC elements			
Title	Estimating exposure and uncertainty for volatile contaminants in drinking water	K	Title
Creator	Sankaran, Karpagam	L	Author
Subject	Civil and environmental engineering	SU	Subject
Description	The EPA recently completed a major study to evaluate exposure and risk associated with a primary contaminant, radon and its progeny in drinking water (EPA, 1995). This work . . .	AA	Note
Publisher	Virginia Polytechnic Institute and State University	T	Publisher
Date	1998-10-23	YR	Year
Type	Text	TYPE	Resource type
Format	Application/pdf	FMT	Resource format
Identifier	http://scholar.lib.vt.edu/theses/available/etd-92398-135228/	URL	URL
Language	en	LANG	Language
Rights	I hereby grant to Virginia Tech or its agents the right to archive and to make available my thesis or dissertation . . .	X	Rights

Table II Comparison of original DC record and transformed Bibliographic Class record

DC record before transformation	BibClass record after ttransformation
<record>	<A
<header>	ID="oai:lcoa1.loc.gov:loc.gmd/g3195.ct000379"
<identifier>oai:lcoa1.loc.gov:loc.gmd/g3195.ct000379</identifier>	DT="2002-06-06T18:07:03Z">
<datestamp>2002-06-06T18:07:03Z</datestamp>	<B><K>Meet-konstige vertoning van de grote en merk-waardige
<setSpec>gmd</setSpec>	zons-verduistering.</K>
</header>	<L>R. & J. Ottens.</L>
<metadata>	<L>Panser, Simon.</L></B>
<dc>	<E><T>S.l.</T><YR>1748</YR><X/></E>
<title>Meet-konstige vertoning van de grote en merk-waardige	<G><AA>A Sporting Chance exhibit, 1979. DLC</AA></G>
zons-verduistering.</title>	<I2><SG><SU>Solar eclipses–Maps.</SU></SG></I2>
<creator>R. & J. Ottens.</creator>	<J><URL>http://hdl.loc.gov/loc.gmd/g3195.ct000379</URL></J>
<creator>Panser, Simon</creator>	<FMT/>
<subject>Solar eclipses–Maps.</subject>	<LANG>dut</LANG>
<description>A Sporting Chance exhibit, 1979. DLC</description>	<TYPE>image</TYPE>
<publisher>S.l.</publisher>	<TYPE>cartographic</TYPE>
<date>1748</date>	<TYPE>map</TYPE>
<type>image</type>	<NORM>image</NORM>
<type>cartographic</type>	<NORM>image</NORM>
<type>map</type>	<NORM>image</NORM>
<identifier>http://hdl.loc.gov/loc.gmd/g3195.ct000379</identifier>	<INST>Library of Congress American Memory Project</INST>
<language>dut</language>	</A>
</dc>	
</metadata>	
</record>	

then wanted to test this provisional design with end-users, in face-to-face interviews with them, to discover which elements of the design worked, and which elements had their limitations.

We tested with nine participants, who ranged from experts in searching digital collections to novices. (Raw results are available on request.) We asked these participants to look at mockups of the provisional interface in a Web browser. The mockups consisted of a search page (Figure 2), a page where they would select a particular repository to search in, and a search results page.

Our questions were as open-ended as possible without being directionally vague. For example, we asked "What do you think this page is for? What do you think you can do here?" and "How would you go about finding images of Monet's 'Water lilies' series using this page?" We also asked participants to tell us if there was anything they found lacking on the mockups by writing their answers on an actual paper copy of the mockup.

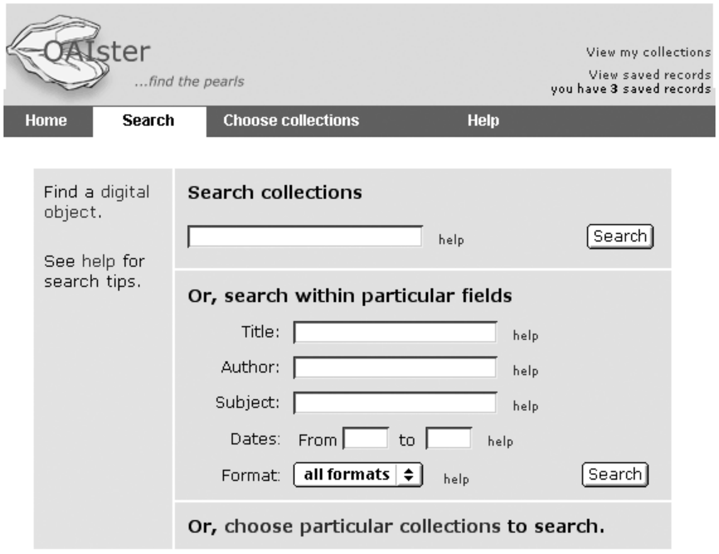
At the end of the session, we had the participants perform some paper prototyping – in this case, indicating which metadata elements they would want to see on a search results page by placing cards containing those elements on a blank printout of the interface. For example, a user may have decided that they wanted to see Author, Title, Publisher and Subject in each record on a search results page. They selected the cards with these words on them and placed them on a piece of paper containing the outline of a blank search results interface. Figure 3 illustrates the paper prototyping process.

The results of the testing informed improvements to the search interface, which we made subsequent to the testing. We launched the search interface on June 28, 2002 with 274,062 records from 56 repositories. Bug fixes and small improvements were made during the months of July and August, 2002, as well as many more repositories harvested, and at present we make nearly a million records from over 100 repositories searchable. We expect both the number of records and the number of repositories to increase considerably as more institutions become OAI-compliant.

Results of our work

Clearly, one of the results of our work has been a proven method for aggregating metadata from a number of quite varied repositories, and making this searchable by end-users. This section reports our analysis of these efforts, and some of the issues we have run across during this project.

Figure 2 Search page mockup tested in-person with participants



OAIster
...find the pearls

View my collections
View saved records
you have 3 saved records

Find a digital object.

See help for search tips.

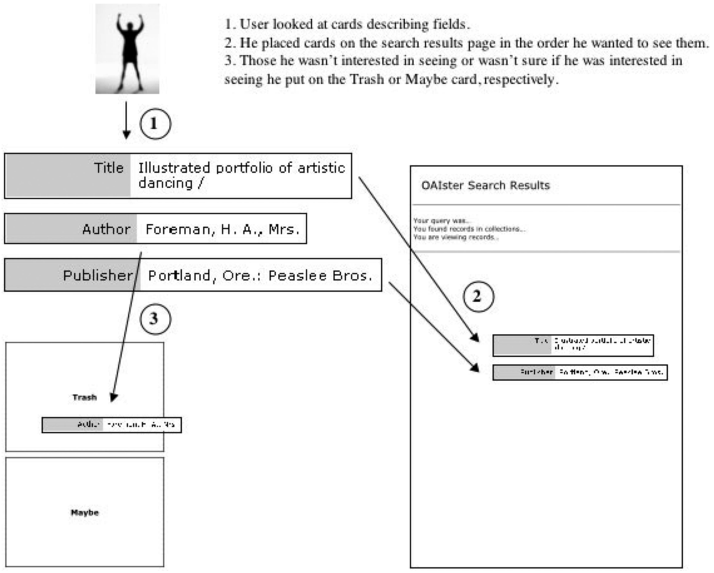
Search collections

Or, search within particular fields

Title: help
Author: help
Subject: help
Dates: From to help
Format: all formats help

Or, choose particular collections to search.

Figure 3 Paper prototyping process



1. User looked at cards describing fields.
2. He placed cards on the search results page in the order he wanted to see them.
3. Those he wasn't interested in seeing or wasn't sure if he was interested in seeing he put on the Trash or Maybe card, respectively.

1
Title Illustrated portfolio of artistic dancing /
Author Foreman, H. A., Mrs.
Publisher Portland, Ore.: Peaslee Bros.
2
3
OAIster Search Results
Your query was...
You found records in collections...
You are viewing records...
Trash
Maybe

Analysis of harvesting efforts

In conjunction with UIUC, we have developed a method that can be used to harvest OAI-enabled metadata. The harvester tool is currently available to others, as noted before. The transformation tool will be made available for use in a subsequent release of DLXS middleware, along with the framework for OAIster itself. All of this is open source, and can be potentially modified by other service providers.

We avoided most of the sticky issues associated with the use of the protocol, as we did not develop the harvester tool. However, by using the tool, we became aware of some inherent problems with harvesting, in general:

- We were not initially able to re-harvest a repository from scratch. This appeared to be more a limitation of the UIUC harvester than the protocol itself. Mistakes on our end, and discrepancies in the number of files reported as harvested and the number actually harvested for some repositories, made it clear that we needed to re-harvest certain repositories. In order to re-harvest a repository in its entirety, its administrative metadata must be removed from the harvester database. We developed a Perl script designed to clean out the administrative metadata from the harvester database, which UIUC has included with version 2.0B4 of the Java harvester.
- In over 10 per cent of the repositories currently being harvested, we have encountered XML validation errors. Some data providers have not been strict in conforming to the UTF-8 encoding standard, and our harvester would fail as a result when gathering records from these repositories. If we requested that the harvester gather records under loose validation, we would receive non-encoded records as well as normally encoded records, although in general we would receive more records in total than if we had not harvested using loose validation.
- Scheduling harvesting can also be challenging, as long harvesting efforts can often end up overlapping, and thus cause problems with memory-intensive, concurrent processes. At times we would receive out-of-memory errors, and a harvest would fail as a result. Since we do not currently have an automated process to inform us when new records have been added to the file system, we need to manually add, remove, and add schedules again when needed.

We expect that a majority of these problems will be fixed as the OAI protocol becomes more popular and OAI tools become more robust. More tools will be developed, tested and modified, the OAI protocol will be changed as needed, and data providers will find it necessary to clean up their data.

Repositories vary in terms of the types of records they offer. They differ in digital object formats (e.g. text, video), academic levels (e.g. graduate student theses, peer-reviewed articles), and topics (e.g. physics, religious studies), among others. And, the repositories vary significantly in the quality of their metadata, including their use of DC. Although they all must be at the least DC encoded before being OAI compliant, institutions use certain elements more frequently (e.g. hundreds of DC Subject elements for one record) and/or ignore other elements completely (e.g. often there is no DC Rights element used).

The normalization of the DC Resource Type element was our attempt to standardize some of the metadata, so that end-users could search more effectively using this element. Admittedly, the method we used for normalization is not perfect, since each time a new repository, with potentially different varieties of values for DC Resource Type, is added to our service, the normalization table must be expanded. The normalization of metadata will eventually require the use of a thesaurus or controlled vocabulary, and automated methods for gathering the normalization table.

Analysis of search logs and user testing

The results of our user testing revealed interesting facts about how end-users search for information online, and specifically digital objects they seek in services like OAIster. The most interesting survey result was end-users indicating they were most interested in online journals and reference materials when they went online to look for information, but that these were the digital resources they were unable to find. End-users also noted they were not able to retrieve the resource itself (i.e. they were only able to get to metadata), that they encountered a variety of problems with searching, and that they did not know where to begin to look for information.

In addition, the in-person user testing helped us develop more appropriate labeling and descriptive text in the interface, helped us understand which processes were difficult for end-users to understand (e.g. the repository selection page was quite cumbersome for most end-users we tested with), and how we could better arrange useful information on the page (e.g. where best to place the results summary).

In particular, the paper prototyping indicated that end-users did not often differentiate between a "short record format" (fewer fields displayed in an individual record) and a "long record format" (more fields displayed). This led us to develop a search results page that had all the elements we were mapping immediately displayed to the user. In retrospect, we realized that some of these fields were very lengthy (e.g. the DC Description element), and we will work on making this easier for end-users to handle in a future round of improvements.

Preliminary analysis of search logs (Table III) for the first 33 days of the service's operation gave an indication of how many end-users were using the search interface (number of accesses, not number of sessions), the heaviest use of the interface by certain academic institutions, the percentage of accesses in which end-users tried Boolean searching and tried using search limiters, and the types of search terms being used.

The Boolean AND operator could be used in the simple search interface by adding a term in the first search box at the top of the page (e.g. "aquaculture") and then a term in the "Keyword" box (e.g. "fish"). Both these boxes search all the indexed fields. In this roundabout fashion, which was admittedly poorly designed on the interface, end-users could search more than one word, and not as a phrase. We expect to work on making this easier for end-users to perform.

End-users had the option to limit their search to certain fields – Title, Author, Subject and Resource Type. ("Resource Type" is the normalized field.) An example of the use of a limiter is entering "duisburg" in the first search box and "grimm" in the "Author" search box.

Search terms that end-users entered varied widely, and uncovered a number of interesting issues:

- Misspellings, e.g. "blue swede shoes." In many search engines, end-users have come to expect their misspellings to be taken care of by the system itself. For instance, entering "beuaty beast" at Amazon.com gets results for "Beauty and the beast."
- Multiple words strung together, e.g. "east detroit halfway." This seems to indicate that end-users expect to search the system as they would a Web search engine, such as Google, with multiple words searched separately. A more obvious method for performing Boolean AND searching may alleviate this problem.
- Limiters used after trying one word or phrase, e.g. the user's first search = "bibliographic instruction"; the

Table III Preliminary analysis of first 33 days of OAIster search logs

	Months in 2002	
Types of statistics	**June (only last three days)**	**July**
Total number of accesses	689	8,321
Top five institutions using OAIster	University of Michigan = 164 University of Virginia = 25 North Carolina State University = 25 Cornell University = 20 University of Oxford = 15	University of Michigan = 317 Boston College = 94 State University of New York, Buffalo = 55 Glasgow University = 32 Northern Arizona University = 27
Percentage use of Boolean AND (on three sample days – 1, 18, 30 July)	n/a	2.2 per cent (20 out of 905 total searches)
Percentage use of search limiters (on three sample days – 1, 18, 30 July)	n/a	8.1 per cent (73 out of 905 total searches)

second search = "bibliographic instruction" plus the "Author" field limiter "sutherland." End-users may be retrieving either too many records to handle or retrieving too many to sort and display.

After launch, we gathered anecdotal evidence on the successes and limitations of our service. Several end-users were disturbed by our choice to limit the number of records retrieved to 1,000 (sorting a large number of records is demanding for the system, and, in the interest of time, we launched without making unsorted record sets available). Others found the interface difficult to use because of the small fixed font or because it was difficult to determine how to formulate multi-word searches (rather than phrase searches).

In general, end-users were pleased that there was a service like this available. We received positive comments such as "Splendid service, and I will promote it widely!", "An excellent resource – I have already made good use of it twice this morning!", and "I think it's a great service – and a wonderful site to use to illustrate the power of the OAI effort." At this point, many of these comments are coming from researchers in the digital library environment, as the service has not yet been widely promoted to end-users.

Larger questions and issues

Some issues we have encountered, if not necessarily handled, in our project are of a wider scope and need to be discussed by a larger community. These include duplication of records harvested, types of restrictions on metadata or the digital objects themselves, granularity of records, and authenticity and authority of the digital objects.

We have encountered duplication of records in two ways. We have harvested nearly every OAI-enabled data provider repository, including those that aggregate a range of original data provider repositories (henceforth called "aggregator providers"). Because of this, we have records in our system from original repositories and from aggregator providers. As an example, a search performed in OAIster for "double-well Duffing oscillator" retrieves two records, exactly the same, but one was harvested from the arXiv.org Eprint Archive repository (an original repository) and one was harvested from the CiteBase repository (an aggregator provider). The decision of whether or not to harvest from aggregator providers is made more complex because these providers also contain records that are not currently available through OAI channels, and they do not always contain all the records of a particular original repository.

Aggregator providers can be very useful for service providers who are unable to complete harvests of certain repositories, as was the case for us in our initial attempts to harvest the arXiv.org repository. However, if an original repository is both a data provider and a service provider, there is the danger that they could (potentially unknowingly) harvest their own data from the aggregator provider. The use of the OAI Provenance element could alleviate these problems, however it is not currently widely used by data providers and most harvesters do not have the capability to utilize it.

Our strategy has also been to include records not available from OAI-enabled data providers, what we call "snapshots" of metadata since we cannot use a harvesting schedule to gather new and modified metadata automatically. We have run into duplication in this manner as well. One example is records we harvest from the British Women Romantic Poets (BWRP) project at the University of California Davis. We at DLPS currently host SGML-encoded versions of these records[12], which were included in OAIster as part of our own harvested collections. (We are also a data provider and harvest our own collections.) In order not to serve duplicates, we decided to remove the DLPS hosted records from OAIster and provide only the original UC Davis BWRP records.

Rights and restrictions associated with metadata, and associated with the digital objects themselves, are also an issue. For instance, in our role as data provider, we provide metadata of our own collections that contains digital objects restricted to certain communities (e.g. CIC institutions), and metadata restricted due to contractual obligations with the originator. We do not use OAI to provide metadata describing the latter, but we do use OAI to provide metadata for the former, when such metadata exists.

There are issues surrounding inclusion in OAIster of metadata pointing to restricted digital objects, not the least of which is that it may confuse end-users if they attempt to get to a digital object and are prevented from doing so due to access limits. However, there are reasons for making this metadata available. Limiting access to records because the digital objects are restricted to a certain community ends up limiting that community's total access (if it is the only access the community has to these digital objects). At DLPS, we have digital objects that fit this pattern, and we are in the process of including clarified rights information in the records for these objects.

Rights information in the harvested metadata (i.e. that supplied by the data provider in the DC Rights element) varies widely, and consequently makes the choice of inclusion or exclusion of restricted digital object metadata difficult. Currently, there is no standardized method of indicating that digital objects are restricted. We would like to see an OAI protocol "restricted/public" toggle element implemented and then tested by data providers and service providers.

Granularity or specificity of digital objects will become more of an issue once more metadata is available. We foresee that it will confuse the user to be able to access separate records of the scanned artwork of the Peggy Guggenheim collection, but only be able to access a single record of a book of Emily Dickinson poetry and not records for each of the poems. With so many harvested records, it is not possible to manually determine the specificity of each record and mirror that in the interface appropriately (e.g. a hierarchical display). The future of the protocol may allow more automated approaches to solving this problem, especially if the OAI Set element is more fully realized and used. (The currently defined Set element allows a data provider to indicate which groupings a record belongs to, e.g. a "low temperature physics" set.)

During our work, we saw very few instances of repositories that were not authentic, in that the majority were collections of digital objects developed by the institution providing them, or mirrors of another institution's digital objects. However, determining whether these digital objects were authoritative was more difficult. Who can say whether they are the definitive or official digital objects? We believe that a reasonable method is for the data provider to indicate this in its records, and that if the data provider has taken the steps to enable and register the repository, the service provider should consequently trust that the digital objects are authoritative. As stated by the Research Libraries Group (RLG):

> A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future (RLG, 2002).

It is our belief that data providers fall into this category.

Future of OAIster and OAI

The OAI protocol has increased in popularity in the year and a half since it was developed. In one month's time (June-July 2002), over ten new OAI-compliant repositories were registered on the official Open Archives Initiative site. We agree with Peter Suber in his quote regarding e-print archiving and the OAI initiative:

> If you've been following the progress of the FOS [Free Online Scholarship] movement for any number of years, you'll agree that no other single idea or technology in the movement has enjoyed this density of endorsement and adoption in a six month period (Suber, 2002).

The statistics noted earlier indicate that the OAIster project has been quite popular in the short time it has had a public search interface. This indicates that a service of this type is potentially quite useful for scholars. We expect it to become even more popular as more repositories are added and as the service is more widely publicized.

In the immediate future, before the official end of the project, we expect to make some improvements to the current OAIster service:

- Make it easier to search using at least Boolean operators.
- Provide more effective sorting of results, and introduce ranking mechanisms.
- Offer full results instead of limited results (retrieval of up to 1,000 records), instead of

forcing end-users to perform a new, more limited, search.
- Provide more compact displays of records, in particular the often-lengthy DC Description element, by offering the full display of an element or record on a separate page.
- Allow end-users to revise their current search.
- Let end-users see which records retrieved are from which institutions.
- Make it possible to search within a particular institution's records.
- Research methods for pointing end-users to "best answers" in their search results.

Our hope is that we can further improve upon OAIster at a future date. The following are more substantial issues that we would like to develop:
- Determine a method for handling duplicate records.
- Normalize more elements, such as DC Language.
- Provide high-level topical (or similar) browsing capabilities, perhaps drawing on the OAI Sets functionality.
- Work with UIUC on data mining research to offset issues related to metadata inconsistency.
- Target particular audiences within the research community.
- Collaborate with other projects that could benefit from using OAIster, e.g. giving professors the ability to find digital objects while developing their courses online in a learning object environment.

While we are hopeful that there will be more and more varied service providers in the near future, we are somewhat concerned that offering a multitude of venues for finding digital objects will become confusing to end-users. In our survey, end-users indicated that they were frustrated with not knowing where to start to look for something.

Our wish is to alleviate this potential problem by developing a community of service providers and data providers who discuss the issues noted earlier, coordinate efforts, and ensure that our services remain usable and understandable for any end-user looking for digital objects. We expect to be a significant part of this community in the months to come.

Notes

[1] www.openarchives.org/OAI/openarchivesprotocol.html, The Open Archives Initiative Protocol for Metadata Harvesting.
[2] www.openarchives.org/Register/BrowseSites.pl, Registered Data Providers.
[3] www.oaister.org/, OAIster.
[4] http://oai.grainger.uiuc.edu/index.htm, The University of Illinois at Urbana-Champaign, Open Archives Initiative Metadata Harvesting Project.
[5] www.americansouth.org/, AmericanSouth.org, a joint project of Emory University and ASERL.
[6] www.dlxs.org/, University of Michigan Digital Library eXtension Service.
[7] http://docs.dlxs.org/class/bib/bib-index.xml, Bibliographic Class Documentation.
[8] http://docs.dlxs.org/class/bib/bib-dtd.xml, The Bibliographic Class DTD.
[9] http://sourceforge.net/project/showfiles.php?group_id=47963, UIUC OAI Metadata Harvesting Project File List.
[10] www.oaister.org/o/oaister/viewcolls.html, OAIster Harvested Institutions.
[11] www.oaister.org/o/oaister/surveyreport.html, OAIster Survey Summary Results.
[12] www.hti.umich.edu/cgi/t/text/text-idx?c=bwrp, British Women Romantic Poets (BWRP) hosted at the University of Michigan Digital Library Production Service.

References

Bergman, M.K. (2001), "The deep Web: surfacing hidden value", *The Journal of Electronic Publishing*, Vol. 7 No. 1, August, available at: www.press.umich.edu/jep/07-01/bergman.html

Besser, H. (2002), "The next stage: moving from isolated digital collections to interoperable digital libraries", *First Monday*, Vol. 7 No. 6, June, available at: www.firstmonday.org/issues/issue7_6/besser/

Ipeirotis, P.G., Barry, T. and Gravano, L. (2002), "Extending SDARTS: extracting metadata from Web databases and interfacing with the Open Archives Initiative", *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, OR, USA, 13-17 July, pp. 162-70, available at http://doi.acm.org/10.1145/544220.544254

Lagoze, C. and Van de Sompel, H. (2001), "The Open Archives Initiatives: building a low-barrier interoperability framework", *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, VA, USA, 24-8 June, pp. 54-62, available at: http://doi.acm.org/10.1145/379437.379449

Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M.L., Knudson, F. and Holtkamp, I. (2002), "Federated searching interface techniques for heterogeneous OAI repositories", *Journal of Digital Information*, Vol. 2 No. 4, 21 May, available at: http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/

RLG (2002), "Trusted digital repositories: attributes and responsibilities: an RLG-OCLC report", May, Research Libraries Group, available at: www.rlg.org/longterm/repositories.pdf

Suber, P. (2002), *Free Online Scholarship (FOS) Newsletter*, 8 August, available at: www.topica.com/lists/suber-fos/read/message.html?mid=1607391538&sort=d&start=38

Van de Sompel, H. and Lagoze, C. (2000), "The Santa Fe convention of the Open Archives Initiative", *D-Lib Magazine*, Vol. 6 No. 2, February, available at: www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Waters, D.J. (2001), "The Metadata Harvesting Initiative of the Mellon Foundation", *ARL Bimontly Report*, No. 217, August, available at: www.arl.org/newsltr/217/waters.html