DLF Update on Metadata Services

David Seaman
May 2006

Discovering Distributed Collections: Best Practices for Shareable Metadata
http://www.diglib.org/architectures/oai/imls2004/

The Digital Library Federation (DLF), in partnership with Emory University, the University of Illinois at Urbana-Champaign, and the University of Michigan, is researching, designing, and prototyping a "second generation" Open Archives Initiative finding system, capitalizing on the lessons learned from the first wave of OAI harvesting and using as its raw material collections drawn from across the DLF membership. The aim is to foster better teaching and scholarship through easier, more relevant discovery of digital resources, and to enhance libraries' ability to build more responsive local services on top of a distributed metadata platform.

OVERVIEW

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has proven itself as a straightforward and functional mechanism for sharing metadata across systems for the purpose of building focused services. We have collectively proven the protocol and its utility; however, as Martha Brogan notes, "there are numerous practical, technical and philosophical impediments to the full realization of OAI-based services…." DLF is supported by its first competitive federal grant – a National Leadership Grant from the Institute for Museum and Library Services (IMLS) – as we address these various impediments and move closer to the goal of federated collections across institutions, and the ability to create richer services for our library users. This is a multi-pronged endeavor in which we are solidifying best practices for the creation of metadata about our dispersed collections, which can then be used to build new services such as the following:

1. **A DLF Portal** that allows one to access all items from DLF libraries that have been publicized through the Open Archives Initiative (**1,042,268** records) http://www.hti.umich.edu/i/imls/

2. **A DLF MODS Portal** -- a subset of the full Portal drawing together those records that have the richer MODS metadata that supports much better subject, date, and

geographic navigation (**253,478** MODS records)
http://www.hti.umich.edu/m/mods/

3. **A new DLF Collections Registry** that describes the collections to which many of these items belong (**759** publicly accessible digital collections)
http://gita.grainger.uiuc.edu/dlfcollectionsregistry/browse/

All of this effort is closely informed by a *Scholars' Advisory Panel* and a *Panel of Technical Experts*: http://www.diglib.org/architectures/oai/imls2004/OAISAP05.htm . Their feedback has led to a set of marked improvements as we develop the services above:

- A simpler (Google-like) initial interface;
- The inclusion of thumbnail images of graphical collections into the metadata;
- The close linking of an item to its immediate collection (rather than to its institution);
- More fields for limiting searches;
- A book bag for saving or emailing records;
- Inclusion in A9.com to facilitate simultaneous searching with Amazon, Wikipedia, RedLightGreen, The British Library catalog, and many other OpenSearch services
- And a renewed sense of importance on the permanent nature of the hypertext links and web addresses that we publish.

This grant-funded work is also sponsoring and benefiting from the following:

- An analytical survey commissioned from Martha Brogan of a wide range of services based on aggregated metadata (due Summer 2006)

- A set of OAI Best Practices (co-sponsored with NSDL)

- A set of training documents to help institutions get ready to publish information about their local collections in a form that others can easily gather up and re-use. These materials have been tested in a well-attended day-long workshop in Virginia in Fall 2005, and a second one is planned for Summer 2006 in California.

DETAILED DESCRIPTION OF THE MAIN COMPONENTS OF THE WORK

Metadata setup/harvest/design [Michigan]

Perry Willett, Kat Hagedorn, and Qian Liao at Michigan has completed its setup of the OAI harvester that now regularly goes out to collect all OAI records from all DLF member institutions. This is the bedrock tool for our work. In function and appearance it is modeled on the OAIster service (www.oaister.org), and borrows from Michigan's considerable expertise in running this large OAI aggregation. From this, the portal of the records with richer MODS metadata is created. This subset gives us a better base on

which to prototype more flexible, useful, convenient services. It is this MODS body of records and this interface that we are using as our test-bed in which we are responding to the technical advisors' and scholars' feedback and presenting them interface and service choices.

Best Practices **[UIUC]**

Members of DLF institutions have worked closely with others in the National Science Digital Library community at NSF to take what we have learned from building services with harvested OAI and to feed it back to the community as a set of best practices for data providers. The group consists of Caroline Arms (Library of Congress), Tim Cole (UIUC), Naomi Dushay (NSDL/Cornell), Muriel Foulonneau (UIUC), Kat Hagedorn (Univ. of Michigan), Martin Halbert (Emory University), Diane Hillmann (NSDL/Cornell), Arwen Hutt (Univ. of Tenn.), Martin Kurth (Cornell Univ.), Ann Lally (Univ. of Washington), Bill Landis (CDL), Bill Moen (Univ. of North Texas), Clay Redding (Princeton Univ.), Jenn Riley (Indiana Univ.), David Seaman (DLF), Sarah Shreeves (UIUC), Robert Tansley (HP-DSpace), Jewel Ward (USC), Simeon Warner (Cornell), and Jeff Young (OCLC).

A working draft of the *DLF/NSDL OAI Best Practices* is available on the project wiki -- http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl -- the product of many conference calls, meetings, circulated documents, and community feedback. A key recommendation is to move to MODS as the core preferred metadata scheme – an additional level of detail needed to build the sort of services our users tell us they want. The documents are undergoing final revisions and we expect them to be submitted to DLF for copy-editing, layout, and print publication (in addition to electronic publication) in late May 2006.

Training and Curricular Development [Emory]

We recognize the need to offer training in OAI provision for those institutions which need it, and Emory University is focused on developing curricular materials and training sessions that introduce OAI Best Practices to institutions committed to deeper engagement with shareable metadata and the "next-generation" services they allow. For this phase of the project, Emory University organized and conducted the following:

- a Focus Group session (DLF Spring Forum, April 2005), in which training materials and a workshop format were introduced to DLF affiliate institutions who then commented on the content and form of the materials and offered suggestions for additions and change;

- an OAI Implementers Workshop at the 2005 DLF Fall Forum in Charlottesville, VA (November 10, 2005), and is beginning work on a website for collecting, updating, and publishing OAI training materials.

- Another OAI Workshop is also being developed, to be held at Stanford University in August 2006.

DLF institutions committed to deeper engagement with shareable metadata and the "next-generation" services they allow were each invited to send two individuals to this workshop: one who was responsible for generating and maintaining metadata (e.g. a metadata librarian) and one who was responsible for implementing the OAI technical infrastructure of the institution. The fourteen workshop attendees represented eight DLF institutions: University of Minnesota, North Carolina State University, Yale University, Stanford University, Johns Hopkins University, University of Virginia, New York Public Library, and the Library of Congress.

Emory University has created a series of curriculum materials for use in OAI best practices training. This curriculum series includes eight separate documents that together provide a concise set of materials for training institutional teams in best practices for OAI implementation http://www.diglib.org/architectures/oai/imls2004/training/:

a. **DLF OAI Implementers Workshop [Agenda]**

b. **OAI Administrative Planning**
Aspects to consider when allocating resources for OAI implementation, including personnel, technology, monetary costs, and time.

c. **Summary of the DLF Aquifer MODS Profile**
Recommendations for implementing MODS with cultural heritage materials, as an example of a more descriptive format for metadata sharing.

d. **OAI Tools**
Technology for implementing OAI, from scripts to systems. Addresses three main tool categories in metadata sharing process: data providers, digital library systems w/ OAI capabilities, and harvesters.

e. **OAI Cheatsheet: A Taxonomy of Rapid OAI Deployment Strategies**
Types of digital library infrastructures and strategies for deploying OAI on top of them. Offers comparison with own institutional structure and how it might be.

f. **Project Abstract**
Summary of the DLF IMLS OAI Best Practices project, including project goals.

g. **Case for OAI**
Background of the OAI protocol's development, why it's needed, and who benefits from its implementation.

h. **OAI Metadata "Best Practices"**
Common metadata formats and their use with the OAI protocol, focusing on quality issues and guidelines for improving shareable metadata.

Scholar/Expert Groups

One of the resources we were determined to have from the beginning was a pair of review committees, one technical and one scholarly. The technical team has been a resource especially in the early planning stages and will be convened again soon as we go into our second year. They include Jeff Young (OCLC), Simeon Warner (Cornell), Gail McMillan (Virginia Tech), Sarah Shreeves (UIUC), Qian Liao (U Michigan) Tim Cole (UIUC), and Jon Dunn (Indiana).

The OAI Scholars' Advisory Panel is equally important to us. It is critical and useful to have a group of scholars informing our service and design decisions from early in the project: the advisors include Ken Price (U Nebraska, Lincoln, English), Roy Rosensweig (GMU History; Center for History and New Media), Bruce Rosenstock (UIUC, Religious Studies), Steve Railton (UVa, English), Martha Nell Smith (U Maryland, English), Randy Shifflett (Virginia Tech., History), Will Thomas (UVa, History), Allen Tullos (Emory, English), and John Willinsky (U British Columbia Department of Language and Literacy Education).

Much of the work on the DLF MODS Portal, including simple search, image thumbnails, the bookbag feature, A9 integration, and list of collections to which results belong, are the result of a series of conversations with our Technical and Scholarly Advisory Panels.

Web service: Thumbnails for Browsing (UIUC)

A feature of strong interest that was identified at the June 2005 Scholars' Advisory Panel was the ability to preview thumbnails of image resources along with their associated metadata. Based on thumbnail generating technology that was developed for UIUC's CIC-OAI Harvesting Service Project, a thumbnail service is being developed to generate thumbnail images for use with the DLF MODS portal at the University of Michigan. For an example see **Appendix A**.

We have currently harvested about 165,000 thumbnail images (~1.5 GB) from Indiana University, University of Chicago, and the Library of Congress. We plan to turn this into a formal Web service such that we can generate the thumbnails for particular collections, and then service providers can harvest them from us using the OAI-PMH or possibly other means.

Web Service: OpenSearch (UIUC)

Because many attendees of the June 2005 Scholars' Advisory Panel indicated an interest in the Amazon A9 search portal based on the OpenSearch protocol, we implemented a prototype A9 gateway to the DLF MODS portal that is being developed at the University of Michigan. This gateway allows the DLF MODS portal to be searched via the A9 aggregator service or other similar OpenSearch portals (examples include searching our

Portal, Amazon books, Wikipedia, the British Library's public catalog, and RLG's RedLightGreen services simultaneously).

The A9 search portal can be accessed at http://www.a9.com – a range of additional channels can be found by choosing More Columns and selecting the "library" key word. The DLF MODS Portal appears as a choice. See **Appendix B**. The OpenSearch description file for this gateway is at http://gita.grainger.uiuc.edu/opensearch/dlf_mods/dlf_mods.xml.

Registry Building **(UIUC)**

We have an Experimental OAI Registry at UIUC of use principally to builders of OAI services: http://gita.grainger.uiuc.edu/registry . The most significant recent additions to the registry are rich, human-generated collection descriptions for many of the DLF member OAI data providers, including description of select subsets. These data are browsable via the registry Web interface or as XML which conforms to the DC Collection Description profile. Over 180 new OAI data providers have been identified and added to the registry. Some have also been removed because they were no longer operational or they were duplicates. This brings the total number of working repositories in the registry to 1,027.

In addition, "OAI Repository Cataloging Procedures and Guidelines" are now available from the registry web site. This document, authored by the project's original graduate assistant, describes the procedures he developed for generating collection descriptions for the DLF member OAI repositories.

Describing and Enhancing our Collections Descriptions **[UIUC]**

The project's library graduate assistant has completed the cataloging of nearly all of the DLF-member OAI repositories. This includes 33 repositories and 155 individual collections exposed as OAI sets within those repositories. A detailed report of the cataloging procedures and guidelines that were followed is forthcoming. Nearly all of the repository description is based on the collection description schemas that were developed for the IMLS Digital Collections and Content project which is also ongoing at UIUC. A number of the import/export XML schema and processing scripts that have been developed to support this project will be back-ported to support the IMLSDCC project, as well as other registry projects at UIUC.

This same graduate assistant has performed a significant amount of registry maintenance. This includes manually adding new OAI repositories as they are found, and periodically running the maintenance scripts which automatically maintain the metadata about each repository as well as discover new repositories from other existing registries and via Google searches. The GA has also performed a significant amount of manual database cleanup, such as folding duplicate records into a single record, and deleting "dead" repositories from the registry.

Summary of Outputs

DLF Portal: http://www.hti.umich.edu/i/imls/

DLF MODS Portal: <http://www.hti.umich.edu/m/mods>.

DLF Collections Registry:
http://gita.grainger.uiuc.edu/dlfcollectionsregistry/browse/

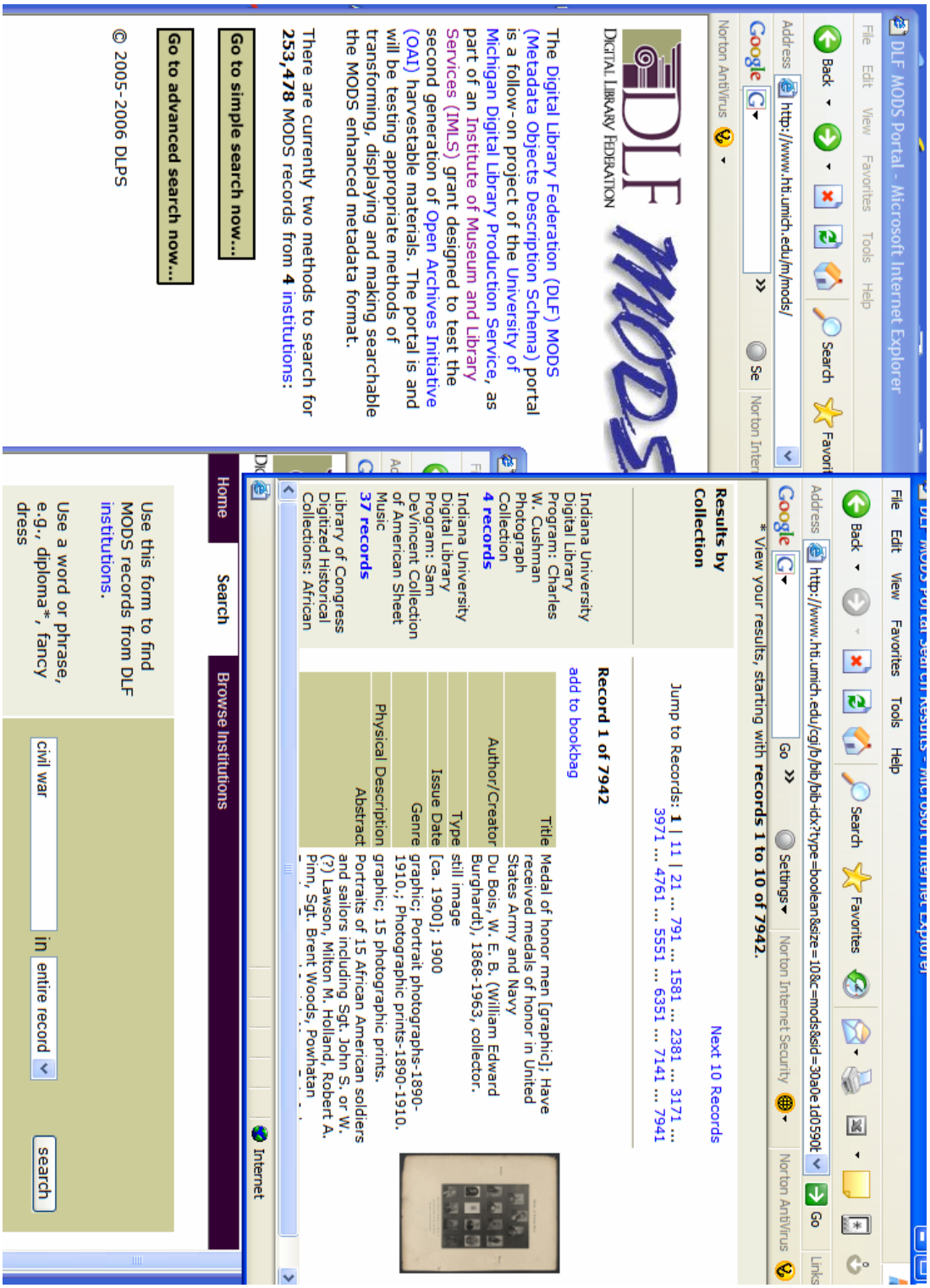
OAI Repository Cataloging Procedures and Guidelines:
http://gita.grainger.uiuc.edu/registry/CatalogingOAIRepositories.pdf

Current versions of OAI Training documents:
http://www.diglib.org/architectures/oai/imls2004/training/

OAI Best Practices, co-developed with NSDL: http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?OAI_Best_Practices

APPENDIX A

DLF MODS Portal showing a simple "Google-like" search for the phrase Civil War -- note thumbnail view, BookBag personal collections feature, and results summarized by collection. Page two of this Appendix shows the result of clicking on the thumbnail.



APPENDIX A
continued

APPENDIX B

A9 Service showing simultaneous searching of Amazon, Wikipedia, and the DLF MODS Portal (with thumbnail images visible)