DLF/IMLS OAI Project Update
& DLF Aquifer Update
DLF Fall Forum 2006

David Seaman, DLF
Katherine Kott, DLF Aquifer
Kat Hagedorn, University of Michigan
Tom Habing, UIUC
Perry Willett, University of Michigan
Liz Milewicz, Emory University
Introduction

- The DLF/IMLS grant is in its final stages and was designed to test 2nd generation OAI tools, services and training opportunities, with guidance from scholars and technical experts: http://www.diglib.org/architectures/oai/imls2004/

- Includes the creation of best practices for shareable metadata, OAI data provider implementations, and tools and strategies for using and enhancing/extending the OAI protocol.

- A report by Martha Brogan on OAI service and data providers: http://www.diglib.org/pubs/dlf106/.
Introduction

- Comprehensive training documents used in several training sessions for potential OAI data providers.
- Two enhanced registries for data providers and service providers.
- Two searchable portals that collect all DLF OAI data providers and all DLF OAI data providers that are using the MODS metadata format.
- An effort at clustering metadata topically for inclusion into the DLF OAI data provider searchable portal.
OAI Scholars Advisory Panel

May 23, 2006
Washington DC
http://www.diglib.org/architectures/oai/imls2004/
Advisory Panel

- John Carlson, University of Virginia
- Gail McMillan, Virginia Polytechnic Institute and State University
- Kenneth Price, University of Nebraska, Lincoln
- Stephen Railton, University of Virginia
- Bruce Rosenstock, University of Illinois at Urbana-Champaign
- Roy Rosenzweig, George Mason University
- Martha Nell Smith, University of Maryland
- Crandall Shifflett, Virginia Polytechnic Institute and State University
- Will Thomas, University of Nebraska-Lincoln
- Allen Tullos, Emory University
Agenda

- Reviewed research progress to date
  - A9
  - Asset action package
  - Best practices in metadata creation
  - Thumbnails
  - Clustering by collection
  - MODS portal
Discussion items

- Dates
- Visualization
- Clustering
- Personal collections
Comments

- Will Thomas suggested that what has been done already outruns what the scholars can do with it.
- Scholars aren't being encouraged to practice digital scholarship and contributing to this is the fact that a great deal of the use of digital objects is focused on pedagogy, as are the tools. But what about scholarly research and production?
- Visualization is important (for different reasons, in different ways) for scholars, students and K-12 pupils
- We need the voice of the scholar matched with the technology.
Full report available at:

http://www.diglib.org/architectures/oai/imls2004/OAIAdvisors060523final.htm
Wiki of OAI Best Practices

http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?TableOfContents

- Shareable metadata and data provider implementations sections developed extensively: soon to be published by DLF
- “Tools and Strategies for Using and Enhancing / Extending the OAI Protocol” section: working on before end of the grant/year
  - descriptions of software packages for data providers and service providers
  - metadata management tools, e.g., UIUC ThumbGrabber, UM OAITransform, CDL Date Normalizer
  - listing of registries for both data and service providers
  - description of extensions to the OAI protocol
OAI Best Practices

Tools and Strategies for Using and Enhancing / Extending the OAI Protocol

OAI Repository Tools

- **[Turnkey Solutions and Packages]** Tools, systems, and other resources for sharing metadata records through the OAI protocol. (This page has been made out of the documentation of the [OAI session] at Chicago NDSL conference 2004.)
- **[Static Repositories & Gateways]** Static repositories are suitable solutions for small, infrequently updated collections of metadata. Registering the metadata collection with a gateway allows it to be harvested.
- **Data Provider and Service Provider Registries and Reference Lists**
  - Listservs ([technical questions], [general information and comments])
  - [OAI Data Provider Registry] The official data provider registry. Includes validation of data repository. Not as extensive as the registry from UIUC.
  - [UIUC OAI Data Provider Registry] The most comprehensive listing of data providers, including the full name of the repository, the OAI baseURL, a description (accessed via [OAIster]), as well as the following.
    - Statistics on use of [metadata schemas (formats)]
    - Statistics on [set names] and [frequency of use of set names]
  - [OAI Repository Explorer] A tool for potential data providers to use in validating their data repository, which also contains a long list of validated repositories.
  - [OAI Service Provider Registry] The official service provider registry. Not as extensive as the registry from UIUC.
  - [UIUC OAI Services Registry] The most comprehensive list of services built around OAI.

Enhancing & Improving OAI-Shared Records

- **[Metadata Management & Added Value Tools]**
- **[XML Tools]**
- **[OAI Validation Tools]**

Extending the OAI Protocol

- **[Using SRU/W]**
- Collection description schema
- Name authority issues
- Jeff Young's crosswalks
- NSDL safe transforms
“Tools” Best Practices

- Ongoing!
  - in flux at the moment, awaiting assistance from various quarters
  - will continue to add and update as new information becomes available
  - hope that it becomes a comprehensive resource for both data providers and service providers
Requests by Institutions considering OAI Implementation

- **Administrative**: Why implement OAI? What resources are needed (staff, equipment, time)?
- **Technical**: How can we implement data or service providers with existing infrastructure?
- **Metadata**: What are appropriate standards for shareable metadata?
- **General**: Give us clear and brief documents, we don’t have time for tomes.
OAI Training Materials

- Brief documents (1-2 pages each)
- Standalone pieces, focused on one element of the training program
- “See Also” references link the series together
- Current training documents online: http://www.diglib.org/architectures/oai/imls2004/training2/index.htm
The Case for OAI

The implementation and development of the OAI protocol has been driven by several factors. First, OAI provides a framework for sharing digital resources, allowing libraries and other institutions to contribute to a larger, interconnected digital library. This fosters the reuse and discovery of digital materials, which is essential for the preservation and accessibility of knowledge. Second, OAI encourages the use of open standards and practices, promoting interoperability and interoperability. Finally, OAI supports the sharing of metadata, which is crucial for the discovery and understanding of digital materials.

OAI Deployment Strategies

The OAI protocol provides a framework for implementing OAI systems. There are several strategies that can be used to implement OAI systems. One common strategy is to implement OAI at the institutional level, where each institution implements OAI on its own system. This approach allows for greater flexibility and control, but it can also be more complicated and costly. Another strategy is to implement OAI at the repository level, where multiple repositories share a common OAI system. This approach can be more efficient and scalable, but it requires careful planning and coordination.

OAI Deployment Taxonomy

OAI systems can be classified into several categories based on their implementation and functionality. The taxonomy includes the following categories:

1. OAI-PMH (Protocol for Metadata Harvesting) - This is the core protocol of OAI, which provides the basic functionalities of an OAI system.
2. OAI-ORE (Open Archival Information Protocol) - This protocol provides a mechanism for describing the structure and relationships of digital objects.
3. OAI-DC (Dublin Core) - This protocol provides a set of metadata elements that can be used to describe digital objects.
4. OAI-EZ (Easy-to-use OAI) - This protocol provides a simplified interface for OAI systems.

Benefits

Implementing OAI provides several benefits to libraries and other institutions. First, OAI promotes the sharing of digital materials, which is essential for the preservation and accessibility of knowledge. Secondly, OAI encourages the use of open standards and practices, promoting interoperability and interoperability. Finally, OAI supports the sharing of metadata, which is crucial for the discovery and understanding of digital materials.
OAI Training Program

- Combined overview with one-on-one assistance
- Flexible structure supports varying types & levels of expertise
- OAI Training Materials served multiple purposes

* More workshops, more consultation
Successes, Barriers, & Lessons

- Positive response to workshop format
- OAI Training Materials served multiple purposes

-----

- High demand for tools info and training
- Requests for more workshops & consultation
- Resources pose challenge to implementation

-----

- 1-page handouts meet many informational needs
- Find new ways to connect OAI consultants with institutions beginning OAI implementation
A Tale of Three Registries

Plus a few other things

By Tom Habing
thabing@uiuc.edu
There was the best of registries for OAI Data Providers…

- **Enhanced metadata**
  - IMLS DCC Collection Description Metadata Application Profile
  - Human cataloging of DLF-member data providers
  - Data feeds for Michigan’s OAIster service
  - Browse by subject terms

- **Improved Export Formats**
  - RSS
  - OAI Data Provider

- **Continued growth** – currently 1164 repositories (~40% growth since start of project)

- [http://gita.grainger.uiuc.edu/registry/](http://gita.grainger.uiuc.edu/registry/)
There was a registry of OAI Service Providers...

- An OAI Service Provider is defined as a system which is reliant on data which is harvested via the OAI-PMH
- The current registry doesn’t strictly comply
- Currently about 100 entries, 40 were collected from the openarchives.org list or harvested via Google, the remainder were manually entered
- Minimal metadata, basic browse and search
- [http://gita.grainger.uiuc.edu/registry/services/services/browse.asp](http://gita.grainger.uiuc.edu/registry/services/services/browse.asp)
There was a registry of DLF-member collections...

- Based on the IMLS DCC Collections Registry
- Data from the previous Michigan registry
  - With some cleanup and enhancements by DLF
- Browse
  - Subject, Place, Time Period, Title, Institution, …
- Basic Search
- OAI data provider for harvesting records
- http://dlf.grainger.uiuc.edu/dlfcollectionsregistry/
There was some other stuff too...

- A9 gateway to the DLF MODS search portal
- Preliminary experiments with Emory’s MetaCombine clustering service
- Automated Thumbnail Generation (Open Source Tool)
- Education and Outreach
  - OAI Training Sessions
  - Distributed Services Registry Workshop at the University of Warwick
  - NSF – DLF – JISC/UKOLN Digital Library Service Registry (DLSR) Workshop
- Reporting in support of DLF Aquifer project
There is some time left for more...

- Allow self-registration in the data and service provider registries
- Turn thumbnail generation into web service
Some Related Links

- http://gita.grainger.uiuc.edu/registry/
- http://gita.grainger.uiuc.edu/registry/services/
- http://dlf.grainger.uiuc.edu/dlfcollectionsregistry/
- http://imlsdcc.grainger.uiuc.edu/
- http://opensearch.a9.com/searches.jsp
DLF Portal

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

- Created to showcase and experiment on the OAI repositories from DLF institutions
- Currently 59 repositories, almost 2.5 million records
- Changes to search interface, metadata tagging, and sorting happened here first
- Based on recommendations from the Scholars’ Advisory Panel
<table>
<thead>
<tr>
<th>Title</th>
<th>Volunteering down Dixie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author/Creator</td>
<td>Ehrgot &amp; Forbriger</td>
</tr>
<tr>
<td>Year</td>
<td>1861</td>
</tr>
<tr>
<td>Resource Type</td>
<td>image</td>
</tr>
<tr>
<td>Resource Type</td>
<td>still image</td>
</tr>
<tr>
<td>Resource Type</td>
<td>Political cartoons.</td>
</tr>
<tr>
<td>Resource Type</td>
<td>Lithographs.</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Note</td>
<td>A satire on Southern recruitment efforts during the Civil War. In a scene before a Confederate encampment, a reluctant civilian (center) is presented with a musket and military coat by two veteran soldiers, one wearing an absurdly tall fur hat. The recruit is restrained from behind by another officer. In the background left a troop of recruits drill; on the right two blacks soldiers, one with rolled pant legs and military coat, look on with amusement.</td>
</tr>
<tr>
<td>Note</td>
<td>The Library's impression of the print was deposited for copyright on July 8, 1861, by the Cincinnati lithographic firm of Ehrgot &amp; Forbriger. It is clearly by the same unidentified artist as two other prints by Ehrgot &amp; Forbriger, &quot;Strayed&quot; and &quot;Ye Conference&quot; (nos. 1861-31 and 1861-32). Two additional prints, not owned by the Library but cited by Weitenkampf, are also listed in the copyright records as Ehrgot &amp; Forbriger registrations. They are &quot;That Feed Won't Do&quot; and &quot;'Wait 'Till the War Is Over.&quot; The copyright imprint, which appears on an impression of &quot;Volunteering Down Dixie&quot; cited by Weitenkampf, appears to have been rubbed or scraped from the Library's impression.</td>
</tr>
<tr>
<td>Note</td>
<td>Weitenkampf, p. 132</td>
</tr>
<tr>
<td>Note</td>
<td>African Americans (portrayed); Army of the Confederacy; Draft (military conscription) and recruitment</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.loc.gov/oc.pnp/cph.3a16238">http://hdl.loc.gov/oc.pnp/cph.3a16238</a></td>
</tr>
<tr>
<td>Institution</td>
<td>Library of Congress Digitized Historical Collections</td>
</tr>
</tbody>
</table>
MODS Portal

http://www.hti.umich.edu/m/mods/

- Created to encourage OAI data providers to create MODS records
- Currently 5 institutions make some or all of their sets available in MODS
- (DSpace just made MODS available, so potentially many more to come.)
- Highly detailed records-- MODS:DLXS
- Research into addition of thumbnails
<table>
<thead>
<tr>
<th>Title</th>
<th>Volunteering down Dixie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author/Creator</td>
<td>Ehrgott &amp; Forbriger</td>
</tr>
<tr>
<td>Type</td>
<td>still image</td>
</tr>
<tr>
<td>Issue Date</td>
<td>1861;</td>
</tr>
<tr>
<td>Genre</td>
<td>graphic; Political cartoons.</td>
</tr>
<tr>
<td>Physical Description</td>
<td>1 print : lithograph on wove paper ; 25 x 37 cm (image)</td>
</tr>
<tr>
<td>Abstract</td>
<td>A satire on Southern recruitment efforts during the early part of the Civil War. In a scene before a Confederate encampment, a reluctant civilian (center) is presented with a musket and military coat by two veteran soldiers, one wearing an absurdly tall fur hat. The recruit is restrained from behind by another officer. In the background left a troop of recruits drill; on the right two blacks soldiers, one with rolled pantlegs and military coat, look on with amusement.</td>
</tr>
<tr>
<td>Subject: Topical</td>
<td>African Americans (portrayed); Army of the Confederacy; Draft (military conscription) and recruitment</td>
</tr>
<tr>
<td>Note</td>
<td>Title appears as it is written on the item.</td>
</tr>
<tr>
<td>Note</td>
<td>[Printed and copyrighted by Ehrgott &amp; Forbriger, Cincinnati]</td>
</tr>
<tr>
<td>Note</td>
<td>The Library's impression of the print was deposited for copyright on July 8, 1861, by the Cincinnati lithographic firm of Ehrgott &amp; Forbriger. It is clearly by the same unidentified artist as two other prints by Ehrgott &amp; Forbriger, &quot;Strayed&quot; and &quot;Ye Conference&quot; (nos. 1861-31 and 1861-32). Two additional prints, not owned by the Library but cited by Weitenkampf, are also listed in the copyright records as Ehrgott &amp; Forbriger registrations. They are &quot;That Feed Won't Do&quot; and &quot;Wait 'Till the War Is Over.&quot; The copyright imprint, which appears on an impression of &quot;Volunteering Down Dixie&quot; cited by Weitenkampf, appears to have been rubbed or scraped from the Library's impression.</td>
</tr>
<tr>
<td>Note</td>
<td>Use surrogate: either electronic image or Reilly's American Political Prints book illustration.</td>
</tr>
<tr>
<td>Identifier</td>
<td>LC-USZ62-13955 DLC; hdl:loc.pnp/cph.3a16238</td>
</tr>
<tr>
<td>Classification</td>
<td>PC/US - 1861.A000, no. 19</td>
</tr>
<tr>
<td>Location</td>
<td>Library of Congress Prints and Photographs Division Washington, D.C. 20540 USA</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.loc.gov/loc.pnp/cph.3a16238">http://hdl.loc.gov/loc.pnp/cph.3a16238</a></td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Institution</td>
<td>Library of Congress Digitized Historical Collections: Cartoon Prints, American, 1766-1876</td>
</tr>
</tbody>
</table>
Topical clustering

- Effort to create smaller sets of metadata for ease of end-user retrieval
- Scholars’ Panel thought topically was best to work on initially
- Working with University of California Irvine researcher David Newman, who has a “Topic Modeling” tool
Vocabulary

- Tool builds a vocabulary from the metadata values in Title, Subject and Description (oai_dc)
  - scientific repositories contained more consistent and enriched metadata (from previous experiment)
  - task for this prototype will be building topics using non-scientific and scientific repositories together
Learning topics

- Tool uses vocabulary to “topic model” or cluster groups of like metadata records together
  - a. with a small number of records to learn what the clusters/topics are
  - b. with all the records to classify them according to the learned topics
Incorporate in records

- Result will be metadata records containing topic labels
- Search and browse will be re-built to incorporate searching by topic labels
  - clusters/topics will be mapped to UM Library’s High Level Browse classification
  - e.g., as used at http://www.lib.umich.edu/ejournals/
DLF Portal: Re-engineered

- All work will be prototyped in the DLF Portal by the end of the year
- Prototype will kickstart re-design of OAIsster in spring 2007
- Suggestions on interface design welcome!
Conclusion

- Importance of best practices to transmit what we have learned (the hard way)
- Training helpful to transform the willing into active participants
- Tools and services critical – the registries and portals are useful even while in prototype stages
- MODS rules
- Close contact with scholars save us from ourselves.
DLF Aquifer Highlights

Katherine Kott
Aquifer Director
DLF Fall Forum
November 10, 2006
Where we are

• Clear purpose within DLF mission
• Key outcome
  – Making content available where scholars work
    • Creating schemas and communities of practice for digital library interoperation
    • Developing a better understanding of what services content consumers want and need
Figure 2. Detailed View - showing example components
Upcoming work

• Two year project to begin April 2007 adds dedicated project staff
• Implements architectural framework
• Implements *American Social History Online* collection through four channels
  • Course management system
  • Federated search
  • Commercial search service
  • Citation management software
• Includes assessment
Accomplishments and activities

• Redesigned the DLF Aquifer website
  – Products and experiments listed on Products web page
    • Final release of MODS implementation guidelines
    • Framework diagram and explanation
  – Publications in D-Lib Magazine
    • General DLF Aquifer article June 2006
    • Asset Actions article October 2006
• Cornell, USC and Yale joined
DLF Aquifer desiderata

• Funding for metadata remediation and enhancement
• Funding for cross platform tools integration or development
Within the working groups

- Adding collections
  - Submission agreement
  - Gap analysis report
- Melding two product experiments (portal and asset actions)
- Creating capacity for usability and assessment best practices work, built on existing DLF efforts
- Developing and documenting best practices for collaborative development
Connection with other DLF initiatives

• Services framework
  – Glossary extraction
  – Collaboration between services framework and technology/architecture working group prompted work on
    • Architectural development
    • Development method

• Synthesizing gap analysis work with OAI/IMLS grant findings
In conclusion

• Working on things DLF does well
  – Building communities of practice
  – Recommending best practices
  – Incubating innovation

• Continuing to map DLF Aquifer within the broader organizational context

• Accomplishing core work of the initiative through the upcoming two year project
Thank you…

For more information

http://www.diglib.org/aquifer/

kkott@clir.org