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

TableOfContents

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 Distributed Library: OAI for Digital Library Aggregation

OAI "Cheat Sheet": A Taxonomy of Rapid OAI Deployment Strategies

This handout provides a taxonomy of strategies that have been frequently deployed in institutions seeking to implement OAI data providers in the context of commonly used metadata formats and digital library systems. Comments are provided to name each of the strategies.

Effort required: While all these strategies certainly require some coordinated work by digital library professionals, the effort entailed for implementation is typically not onerous for most organizations and reaps many rewards through the resulting capability to exchange metadata by means of a standardized protocol. OAI implementation is often misunderstood as requiring major systems development, which is not at all the case. Major overhauls of existing digital library infrastructures are not necessary. OAI implementations is often more akin to a systems patch, mainly requiring the installation of a few programming scripts or a new module onto existing software. Many effective, simple, and quick solutions currently exist for implementing OAI on top of existing infrastructures — solutions requiring only moderate planning and some staff time to deploy. The following taxonomy maps a number of commonly used systems to solutions, and highlights issues to consider when deciding when or whether to employ these strategies.

SEE ALSO: **OAI Implementation: Administrative Planning** (a guide to work and resource planning) **Summary of OAI Metadata Best Practices** (common metadata formats, quality issues in sharing metadata, and best practices for OAI data and service providers) and **OAI Tools** (technology available for generating, converting, managing, and harvesting metadata)

OAI Deployment Taxonomy

Many scripts and software for enabling OAI implementation can be deployed on top of existing digital library infrastructures. Examples of infrastructures commonly found in digital libraries today are listed below (left column) and mapped to OAI solutions that various institutions have found easily and cost-effectively allowed them to implement OAI. Strengths and limitations of each OAI solution are also provided.

Current Infrastructure	→	OAI Solution
XML files in directory structure This is a common situation in many digital library infrastructures, in which XML files maintained in a structured directory hierarchy are indexed by a search engine for public search and display. This is also one of the most flexible infrastructures, and lends itself to many possible OAI implementations. Staff who maintain this variety of digital library infrastructure frequently have developed a strong expertise in one or more preferred programming languages for data wrangling, and often use XML style-sheet transformation tools. The easiest option for proficient programmers is often to write a simple set of scripts that can respond to the six queries defined by the OAI protocol.	→	**Virginia Tech OAI Scripts** (http://www.dlib.vt.edu/projects/OAI) Pros: This frequently deployed toolkit is comprised of free, open source Perl scripts that can be easily configured for many XML formats (for example, works well with TEI headers). The scripts are flexible and simple to adapt to many infrastructures. Cons: The scripts were developed several years ago and are no longer being actively developed or enhanced. **XSL Transforms (or other customized solution using preferred programming language in use at the institution)** Pros: Most flexible option for those who know their own systems well. A set of correctly functioning OAI data provider scripts can usually be written by a competent programmer in a week or less. Cons: Requires a competent programmer.
Z39.50 gateways Many libraries have Z39.50 gateways for systems which hold digital library item records. These systems may include online catalogs or database systems with Z39.50 gateways such as the popular Zebra open source software. Libraries may wish to share MARC records available in these systems via the OAI-PMH.	→	**ZMARC** (http://sunsite.berkeley.edu/~) Pros: This tool was developed by the University of Illinois at Urbana Champaign as a way of providing OAI-PMH access to MARC records already accessible through Z39.50 gateways. ZMARC is free open source software that is straightforward to implement. The software is written in Visual Basic and VBScript and is easy to modify if desired. Cons: Correctly parsing MARC records occasionally requires some troubleshooting.

Related DLF projects:

DLF Aquifer
Bringing collections to light through the Digital Library Federation

document authorship:

METASCHOLAR
An Emory University Digital Library Research Initiative

last revised: 7 November 2005
Mlawicz & Halbert - v. 11

The Distributed Library: OAI for Digital Libra

The Case for OAI

This handout reviews (1) the background and development of the OAI protocol, (2) past a
benefits of increasing the quantity and quality of shared metadata through OAI implement
systems.

Responding to scholarly communities' demands for greater access to pre-publication artic
Archives Initiative (OAI) developed an interoperable, scalable protocol for sharing metada
Protocol for Metadata Harvesting, also simply referred to as the OAI protocol. Scholars' c
awareness of institutional collection holdings, combined with their increased use of online
materials, points to a greater need for institutions to make their holdings more easily sear
the OAI-PMH offers a cost-effective and low-barrier means for institutional metadata shar
demonstrate the benefits to researchers and institutions and the areas for further improve

SEE ALSO: **Project Abstract** (more in-depth background on this IMLS-funded OAI proje

Development of the OAI Protocol

Early demands for electronic repositories of scholarly information came from the sciences
sharing and accessing pre-print publications. The Santa Fe Convention of the Open Archi
agreements for sharing these publications across institutions through its development of t
metadata. As the OAI protocol was being successfully implemented in the scholarly scient
how they might expand their own services into this realm. In 2000, the Andrew W. Mellon
efforts to develop metadata harvesting services in digital libraries; OAI was selected as a
project.

Need & Demand

Current popular search engines still cannot locate resources within searchable databases
that contains documents and resources highly valuable to learning communities. As onlin
researchers (and sometimes the only stop), resource-sharing institutions must find new w
accessible. At the same time, the reality of software and hardware costs and the potential
reformatting metadata discourages many institutions from implementing new digital-library
new systems must consider the long-term value as well as immediate demands: what offe
expenditures, and can also be easily integrated with existing systems?

Benefits

The OAI protocol works with multiple systems and users, and its use of Dublin Core (DC)
sharing. Many solutions currently exist for implementing OAI on top of current digital-libra
open-source software available online for free download. OAI's interoperability makes it a
quickly and inexpensively build their collection online, and helps ensure that learning com
institutions can access collection metadata. Exposing richer metadata formats in addition
metadata and enhances awareness of collections.

related DLF projects:

DLF Aquifer
Bringing collections to light through the Digital Library Federation

METASCHOLAR
An Emory University Digital Library Research Initiative

last revised: 7 November 2005
Mlawicz - v. 8

OAI for Digital Library Aggregation

Metadata Best Practices

ich may be used with OAI, (2) quality issues that currently limit the
ncreasing the quality of shareable metadata.

ds on high-quality metadata. Simple or unqualified Dublin Core (DC)
more descriptive metadata formats may be additionally shared via OAI to
ality issues that can limit the reliability and usability of records.
AI data providers and service providers will alleviate these concerns.

g metadata formats into a shareable OAI), **OAI Tools** (technology available
a), and **Summary of the DLF Aquifer MODS Profile** (recommendations
gital collections of cultural heritage materials).

a providers can help identify and eliminate problems with metadata. The
nce Digital Library and the Digital Library Federation) is developing
implementation: http://oai-best.comm.nsdl.org/cgi-
delines for improving the quality of shareable metadata.

ials and to provide the most appropriate views of the metadata.
system, find out their preferred metadata formats and best practices.
ction. (For example, technical metadata is usually not necessary – users

consider whether it makes sense to expose all of the metadata
ook is made available, it usually doesn't make sense to include a
e something unique about each page that merits their inclusion. It makes
e book as a whole.)

ng described. (For example, don't try to use all of the elements of a
only the top level information to the DC record.)

your institution would be able to identify the resources described.

uch easier for service providers to work with 'imperfect' metadata if the
documentation of the decisions and standards used for exposed metadata.

records.
interoperability issues.

related DLF projects:

DLF Aquifer
Bringing collections to light through the Digital Library Federation

document authorship:

METASCHOLAR
An Emory University Digital Library Research Initiative

last revised: 7 November 2005
Mlawicz & Shreeves - v. 13

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/

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/servicesbrowse.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
- http://sourceforge.net/project/showfiles.php?group_id=47963&package_id=159364

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

Record 2 of 3

add to bookbag

Title	**Volunteering** down Dixie
Author/Creator	Ehrgott & Forbriger.
Year	1861
Resource Type	image
Resource Type	still image
Resource Type	Political cartoons.
Resource Type	Lithographs.
Language	English
Note	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.
Note	Title appears as it is written on the item.
Note	[Printed and copyrighted by Ehrgott & Forbriger, Cincinnati]
Note	The Library's impression of the print was deposited for copyright on July 8, 1861, by the Cincinnati lithographic firm of Ehrgott & Forbriger. It is clearly by the same unidentified artist as two other prints by Ehrgott & Forbriger, "Strayed" and "Ye Conference" (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 & Forbriger registrations. They are "That Feed Won't Do" and ""Wait 'Till the War Is Over."" The copyright imprint, which appears on an impression of "**Volunteering** Down Dixie" cited by Weitenkampf, appears to have been rubbed or scraped from the Library's impression.
Note	Weitenkampf, p. 132
Note	Published in: American political prints, 1766-1876 / Bernard F. Reilly. Boston : G.K. Hall, 1991, entry 1861-33.
Subject	African Americans (portrayed); Army of the Confederacy; Draft (military conscription) and recruitment
URL	http://hdl.loc.gov/loc.pnp/cph.3a16238
Institution	Library of Congress Digitized Historical Collections

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

DLF Digital Library Federation MODS

Record 9 of 7360

add to bookbag

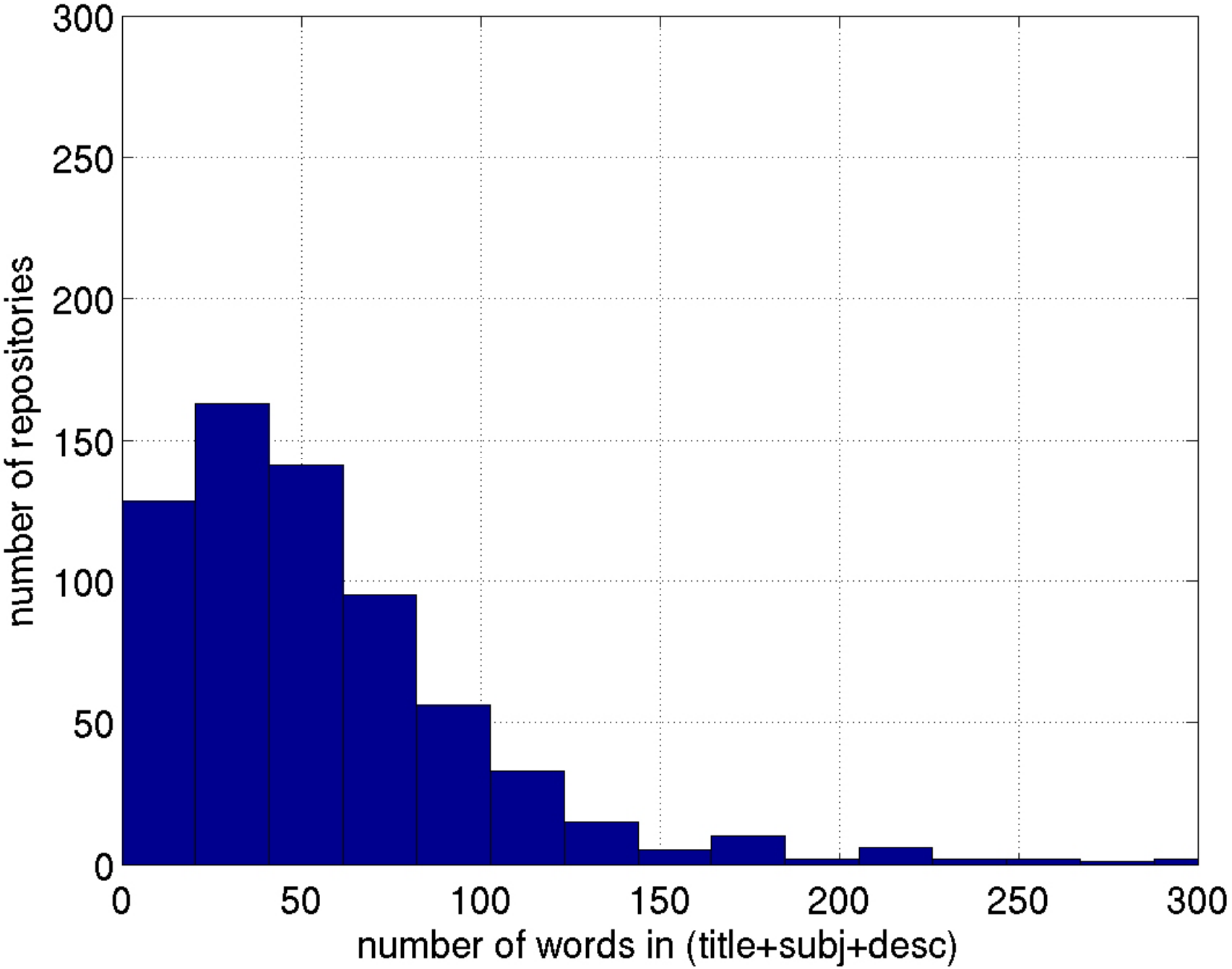
Title	Volunteering down Dixie
Author/Creator	Ehrgott & Forbriger
Type	still image
Issue Date	1861; ||||
Genre	graphic; Political cartoons.
Physical Description	graphic; 1 print : lithograph on wove paper ; 25 x 37 cm (image)
Abstract	A satire on Southern recruitment efforts during the early part of the **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.
Subject: Topical	African Americans (portrayed); Army of the Confederacy; Draft (military conscription) and recruitment
Note	Title appears as it is written on the item.
Note	[Printed and copyrighted by Ehrgott & Forbriger, Cincinnati]
Note	The Library's impression of the print was deposited for copyright on July 8, 1861, by the Cincinnati lithographic firm of Ehrgott & Forbriger. It is clearly by the same unidentified artist as two other prints by Ehrgott & Forbriger, "Strayed" and "Ye Conference" (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 & Forbriger registrations. They are "That Feed Won't Do" and ""Wait 'Till the War Is Over."" The copyright imprint, which appears on an impression of "Volunteering Down Dixie" cited by Weitenkampf, appears to have been rubbed or scraped from the Library's impression.
Note	Use surrogate: either electronic image or Reilly's American Political Prints book illustration.
Note	Published in: American political prints, 1766-1876 / Bernard F. Reilly. Boston : G.K. Hall, 1991, entry 1861-33.
Identifier	LC-USZ62-13955 DLC; hdl:loc.pnp/cph.3a16238
Classification	PC/US - 1861.A000, no. 19
Location	Library of Congress Prints and **Photographs** Division Washington, D.C. 20540 USA
URL	http://hdl.loc.gov/loc.pnp/cph.3a16238
Language	English
Institution	Library of Congress Digitized Historical Collections: Cartoon Prints, American, 1766-1876

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


300
250
200
150
100
50
0
number of repositories
0
50
100
150
200
250
300
number of words in (title+subj+desc)

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/

Advanced Search - b

Simple Search

Advanced Search

Browse by Subject

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas tortor lacus, iaculis eu, vestibulum a, vulputate non, urna. Maecenas lobortis ullamcorper odio. In hac habitasse platea dictumst. Nam vitae arcu in odio accumsan vulputate.

Search within fields (help)
Use a word or phrase, e.g., diploma*, fancy dress

Keyword | And
Keyword | And
Keyword

Limit by: (help)

Resource Type

☐ all types ☒ text ☐ image ☐ audio ☐ video ☐ dataset

Subject Area Arts & Humanities

Sub Categories

ALL
African Studies
African-American Studies
American and Canadian Studies
Anthropology & Archaeology
Architecture
Art & Design
Art History

Sort results by: (help) Title

Search

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 OAIster 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

Bringing collections to light through the Digital Library Federation

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

DLF Aquifer Development for Interoperability Across Scholarly Repositories

Framework for Implementation

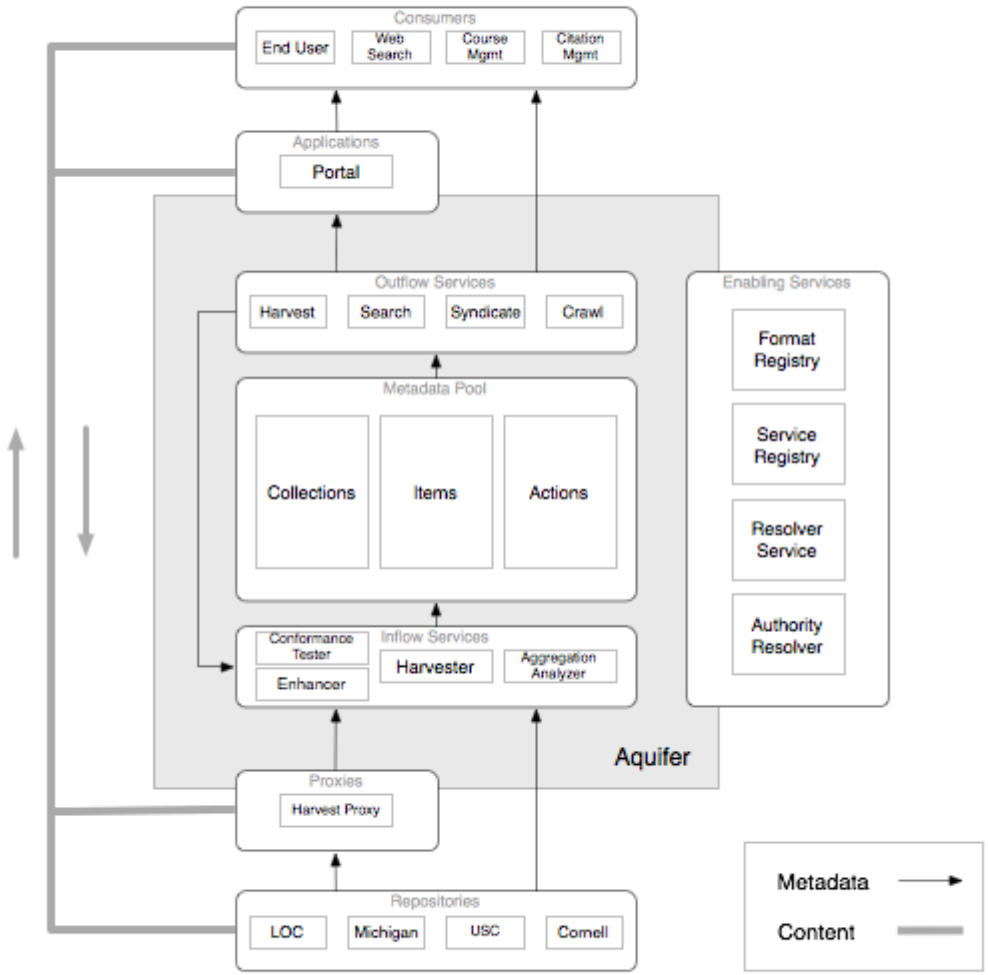
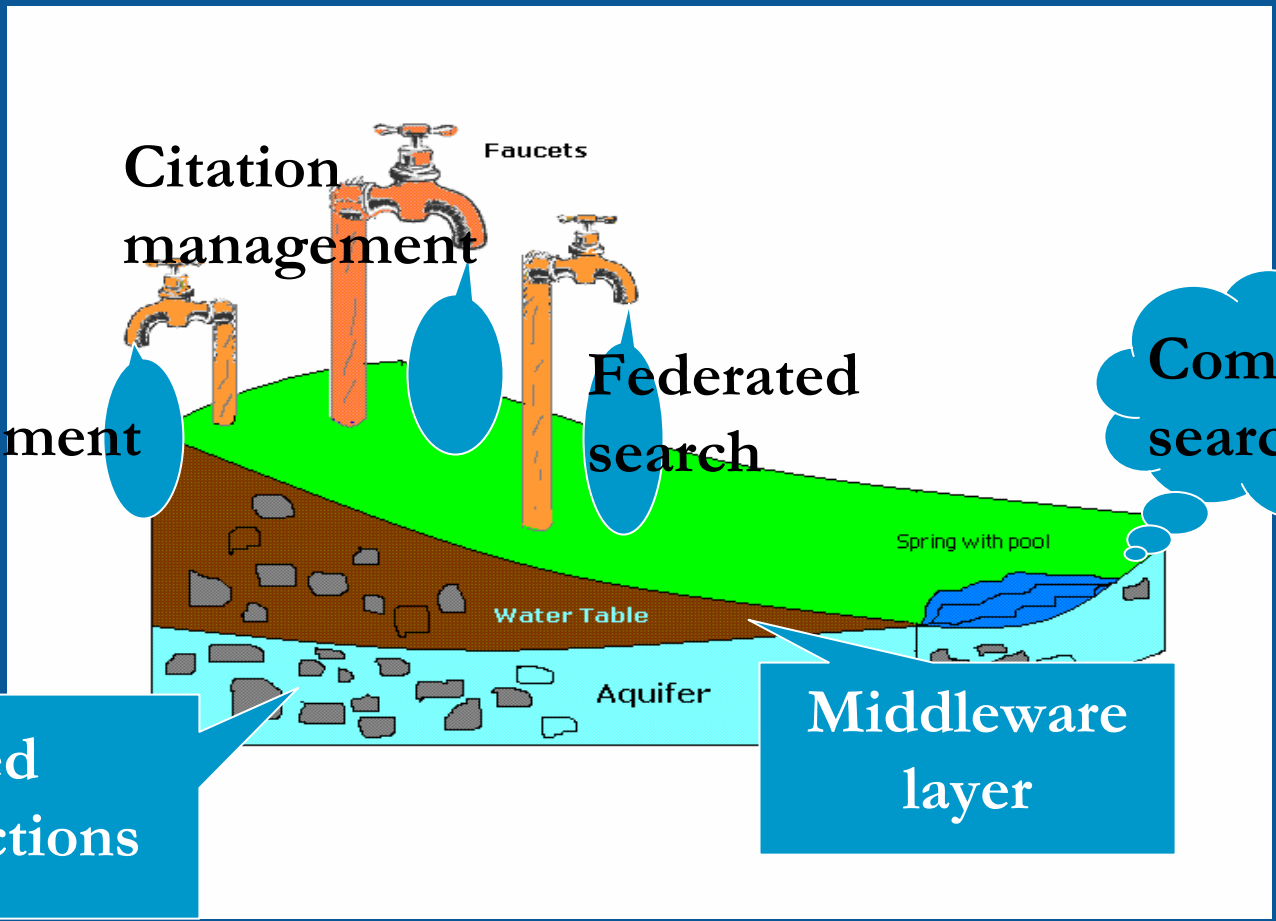


Figure 2. Detailed View - showing example components

Aquifer metaphor



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