



OAI Advisors' Meeting
IMLS National Leadership Grant
The Distributed Library:
OAI for Digital Library Aggregation

May 23, 2006
The Cosmos Club, Washington, D.C.

In attendance:

- **Martha Brogan**, Consultant
<http://www.diglib.org/pubs/brogan0505/>
<http://www.diglib.org/pubs/brogan/>
- **John Carlson**, University of Virginia; DLF graduate student researcher.
- **Timothy Cole**, Mathematics Librarian and Professor of Library Administration, University of Illinois at Urbana-Champaign
<http://www.library.uiuc.edu/faculty/timcole3.htm>
- **Jon Dunn**, Associate Director for Technology, Indiana University.
<http://www.dlib.indiana.edu/~jwd/>
<http://www.dlib.indiana.edu/>
- **Thomas Habing**, Research Programmer for various Digital Libraries Projects, including the Digital Library Initiative (DeLIver) project, UIUC Library's Open Archive Initiative (OAI) projects, University of Illinois at Urbana-Champaign.
<http://dli.grainger.uiuc.edu/idli/idli.htm>
<http://oai.grainger.uiuc.edu/>
<http://web.library.uiuc.edu/grainger/staff/habing.htm>
- **Kat Hagedorn**, Metadata Harvesting Librarian and Manager of the OAIster project, University of Michigan.
<http://www.kathagedorn.com/resume.html>
<http://oaister.umdl.umich.edu/o/oaister/>

- **Martin Halbert**, Director of Digital Programs and Systems, Emory University.
<http://martin.library.emory.edu/>
<http://www.metascholar.org>
- **Christie Hartmann**, Communications and Administration Associate, Digital Library Federation
- **Barrie Howard**, Program Associate, Digital Library Federation.
- **Gail McMillan**, Director, Digital Library and Archives, Virginia Polytechnic Institute and State University.
<http://scholar.lib.vt.edu/staff/gailmac/Gailshp.html>
- **Kenneth Price**, Professor of American Literature and co-editor, *The Walt Whitman Archive*, University of Nebraska, Lincoln.
<http://www.unl.edu/Price/>
<http://www.whitmanarchive.org/>
- **Stephen Railton**, Professor of English and Director of *Mark Twain In His Times: An Electronic Archive*, and *Uncle Tom's Cabin & American Culture: A Multi-Media Archive*, University of Virginia.
<http://etext.lib.virginia.edu/railton/railtonhp.html>
<http://etext.lib.virginia.edu/railton>
<http://jefferson.village.virginia.edu/utc>
- **Bruce Rosenstock**, Associate Professor of Religious Studies, University of Illinois at Urbana-Champaign.
<http://www.relst.uiuc.edu/people/rosenstock.html>
- **Roy Rosenzweig**, Professor of History & New Media and the Director of the Center for History & New Media, George Mason University.
http://cas.gmu.edu/historyarthistory/faculty_staff/biography.php?f=4667
<http://chnm.gmu.edu/>
- **Martha Nell Smith**, Professor of English and Director of MITH, University of Maryland; and General Editor, the *Emily Dickinson Electronic Archives*
<http://www.mith.umd.edu/mnsmith/>
<http://www.emilydickinson.org/>
- **David Seaman**, Executive Director, Digital Library Federation.
<http://www.diglib.org/staff/dmsbio.htm>

- **Crandall Shifflett**, Professor of History and Director of Graduate Studies at Virginia Polytechnic Institute and State University; Project Director, *Virtual Jamestown*.
<http://www.virtualjamestown.org/>
- **Sarah Shreeves**, Visiting Project Coordinator for The Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), University of Illinois at Urbana-Champaign.
<http://ideals.uiuc.edu/>
- **Will Thomas**, John and Catherine Angle Professor in the Humanities, University of Nebraska-Lincoln.
<http://etc.unl.edu/thomas.html>
- **Allen Tullos**, Associate Professor, Institute of Liberal Arts, Emory University.
http://www.ila.emory.edu/faculty_ATullos.htm
- **Simeon Warner**, Information Science, Cornell University.
<http://www.cs.cornell.edu/people/simeon/>
<http://www.openarchives.org/>
<http://arxiv.org/>
- **Perry Willett**, Head, Digital Library Production Service, University of Michigan.
<http://www-personal.umich.edu/~pwillett/>
<http://www.umdl.umich.edu/>
<http://www.dlxs.org/>
- **Jeff Young**, Software Architect, OCLC Online Computer Library Center, Inc.
<http://www.oclc.org/research/staff/young.htm>
<http://www.oclc.org/research/software/oai/cat.htm>
<http://www.oclc.org/research/projects/oai/default.htm>

Overview

The Open Archives Initiative Protocol for Metadata Harvesting (OAI) has been successful as a protocol to enable the sharing of metadata with which to create digital library services. It is easy to use and can transmit both simple and complex metadata.

This grant-funded project focuses what we have learned from creating and using OAI-based services and provides training, an environmental scan, best practices, and several prototype portals that are informed by scholars' needs:

- a set of best practices for creating OAI records for library use will be completed this summer;
- a major report on and environmental scan of digital library aggregation services by Martha Brogan;
- a training curriculum for OAI data providers;
- a prototype OAI portal for resource discovery containing all OAI records from all DLF institutions;
- a portal based on the subset of records that have the richer Metadata Object Description Schema (MODS) metadata;
- a collections registry for many DLF collections
- and a registry of data providers

Our work is lowering the barrier to creating harvestable metadata and raising our understanding of how to do it well. We expect to see a steady increase in sharable metadata for digital content created in DLF libraries in the near future, and look to the advisory panels to ensure that the services we produce and promote are technically competent and useful to the scholar.

Timothy Cole (UIUC) reminded us that the challenge to libraries is to extend functions beyond the simple retrieval of items; there are tools available or easily imagined that not only find content, but allow one to grab it, bring it into one's native computing environments, and repurpose it. As an example of this, Thomas Habing (UIUC) mentioned experiments with Asset Actions, which will be demonstrated later in the program. As a further example, Roy Rosenzweig (GMU) told us of the researcher tool his group is building -- a plug-in for the Firefox browser that allows one easily to capture and annotate Web-based content into a personal library.

The DLF Portals

Kat Hagedorn (Michigan) presented the DLF OAI Portal, which has all the content from DLF institutions that are OAI data providers, and the smaller one that gathers up the DLF records that use the richer MODS metadata.

New features added to the portals as a result of prior meetings include:

- Boolean operators
- A "language" limiter
- A "dataset" limiter, allowing searches to be limited by five resource types: text, image, audio, video, and dataset
- A bookbag to export metadata for local re-use
- Collections contexts -- rather than institutional ones – now have primacy
- A simple “Google-like” search is the default, rather than the advanced search (which is still available as a choice)
- Thumbnail images of visual items are being gathered up automatically and added to the items records

Machine to machine

A technical discussion ensued about machine-to-machine interfaces to these portals, and Kat Hagedorn confirmed that Search/Retrieval via URL and Search Retrieval Web Service protocols (SRU/SRW) are in place and enables live, federated searching, with an XML file being the output.

Metadata evaluation

Jeff Young (OCLC) pointed to the utility of an evaluation tool for data providers to test values. Martha Brogan (DLF consultant) mentioned that the Open Language Archives Community (OLAC) has a metadata report card that functions in much this manner <<http://eprints.unimelb.edu.au/archive/00001408/>>. OLAC report cards are viewable from: <http://www.language-archives.org/archives.php4>

Date Values

In both technical and access discussions, dates were a common topic – they are of high value as an item with which to limit a search and the metadata too often is variable in the way it expresses them, or silent about what the date represents (When the work was first written? When the edition scanned was published? When the digitization took place?).

Perry Willett (Michigan) asked how one would want to restrict by dates across a large spectrum -- from 5BCE to 2006, for example: By century? By user-specified date range? This found some favor over pre-set century delimiters.

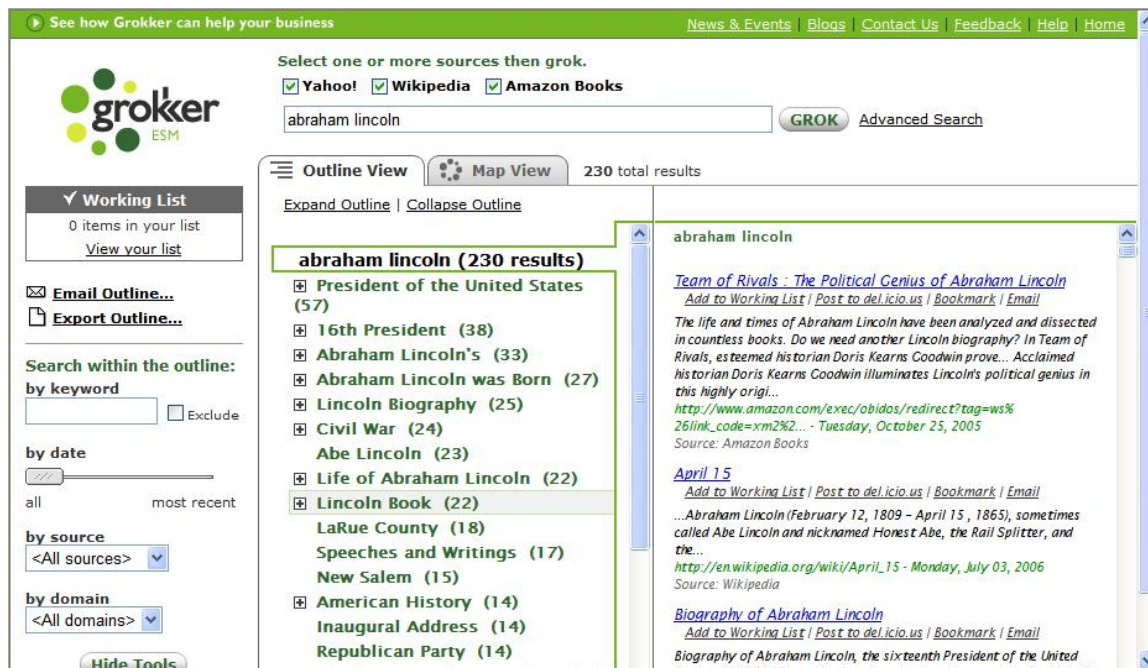
There followed a discussion on how to visualize time and date ranges in search or browse results, and Thomas Habing was later able to offer an example of this from the new DLF Collections Registry. Both the Electronic Cultural Atlas Initiative (ECAI) <<http://www.ecai.org/>> and the United Kingdom’s Cornucopia <<http://www.cornucopia.org.uk/>>, supported by the Museums, Libraries and Archives Council, were pointed to as examples of search by timeline and geography.

Will Thomas raised the issue of how to search by *circa*. Teasing out associations between disparate dates attached to a resource is challenging, e.g., something may be published that discusses a subject from a much earlier date. Will Thomas asked, “What if *no date* is in the metadata record?” The University of Michigan team hasn’t yet discovered an answer for how to deal with this issue.

Services that make use of dates in the metadata are easier to deal with in MODS records, which can capture when a physical item was created or published, when an item was digitized, when a metadata record for that item was created, and so on.

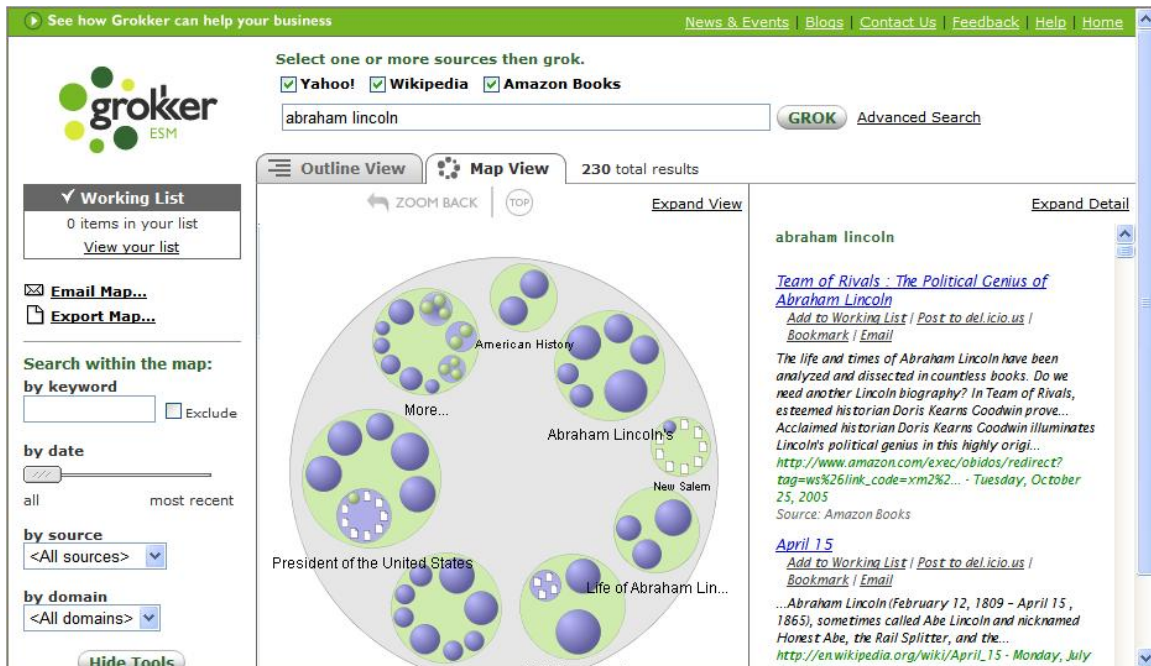
Results visualization and clustering

Can search results be displayed graphically on a data landscape, to show where resources are clustered? This is a visualization function, akin to what can be seen at <http://www.grokker.com/>, where results from a search across multiple sites (Yahoo, Wikipedia, and Amazon Books, for example) can be clustered automatically into topics in a textual list (see below):



The screenshot shows the Grokker search interface. At the top, there's a navigation bar with links for News & Events, Blogs, Contact Us, Feedback, Help, and Home. Below that, the Grokker logo is visible. The search bar contains the text 'abraham lincoln' and a 'GROK' button. The search results are displayed in a list view, showing 230 total results. The results are clustered into topics, with the most prominent being 'abraham lincoln (230 results)'. Other topics include 'President of the United States (57)', '16th President (38)', 'Abraham Lincoln's (33)', 'Abraham Lincoln was Born (27)', 'Lincoln Biography (25)', 'Civil War (24)', 'Abe Lincoln (23)', 'Life of Abraham Lincoln (22)', 'Lincoln Book (22)', 'LaRue County (18)', 'Speeches and Writings (17)', 'New Salem (15)', 'American History (14)', 'Inaugural Address (14)', and 'Republican Party (14)'. The interface also includes a 'Working List' section, an 'Email Outline...' section, and a 'Search within the outline:' section with filters for keyword, date, source, and domain.

Or the same information can be plotted graphically, as clusters on a circular map (see below):



A lively discussion of clustering ensued and centered on how to cluster:

- By institution?
- By collection?
- By free access versus paid access?
- By resource type?

Bruce Rosenstock (UIUC) suggested building a feature to cluster by subject and pointed out that we already cluster results by collection, asking if results can be re-clustered by another type, such as geography, subject, or date. Steve Railton underscored how much more useful the change from clustering results by holding institution to clustering by collection was (a result of our previous feedback session).

A clustering feature would be good to provide on the results page so that the results set could be sorted or clustered by different criteria. Michigan has been looking into clustering functionality.

Will Thomas suggested it would be useful to be able to sort by size of collection, and sorting by geography would be helpful. Gail McMillan (Virginia Tech) asked why "free" vs. "restricted" access were displayed for results and asked for a definition. Kat Hagedorn explained that this is a feature inherited from OAIster, which originally intended to only include free content. There was general agreement that it is useful to see information on objects even if you cannot retrieve them (because they are subscription only, or because the metadata record is of an object in the physical library that is yet to be digitized).

Bookbag

The bookbag is new and the result of earlier feedback. It allows one to download or e-mail the contents of an item as a plain text file. Bruce Rosenstock pointed to EndNote for the purpose of easily importing bibliographic citations into scholarly prose. This would be a welcomed feature for the bookbag function.

Will Thomas suggested that the biggest issue holding back digital scholarship is the inability to build one's own data collection – a personal digital library of material. Now scholars want to use tools that reach across open collections. Greenstone allows one to collect resources from disparate sources. AskJeeves gives users the ability to download URLs, save bookmarks for working off line, and allows annotation. Roy Rosenswieg pointed us to GMU's Scribe tool < <http://chnm.gmu.edu/tools/scribe/>> for note-taking.

Thumbnails

Thumbnails for image or video items are now being included in the portal that is based on MODS, thanks to the "Thumbgrabber" tool developed at UIUC. It uses a listed URL for an image or video item, crawls the resource on its Web page, and grabs a thumbnail of the largest image on the page (assuming this to be the image of the object being cataloged). In this manner it can work without any necessity to liaise in person with the site that publishes the object. UIUC is working with the University of Michigan on the technical issues remaining with this process, but even now it is showing real success, as you can see from the MODS-based DLF Portal. The MODS standard has added a field to list a URL for a thumbnail.

Go to the DLF MODS portal at <http://www.hti.umich.edu/m/mods/> and search for "civil war" – the first item shows the thumbnail generator in action:

Title	Medal of honor men [graphic]; Have received medals of honor in United States Army and Navy
Author/Creator	Du Bois, W. E. B. (William Edward Burghardt), 1868-1963, collector.
Type	still image
Issue Date	[ca. 1900]; 1900
Genre	graphic; Portrait photographs-1890-1910.; Photographic prints-1890-1910.
Physical Description	graphic; 15 photographic prints.
Abstract	Portraits of 15 African American soldiers and sailors including Sgt. John S. or W.(?) Lawson, Milton M. Holland, Robert A. Pinn, Sgt. Brent Woods, Powhatan Beaty, Corporal Isaiah Mays, Sgt. John Denny, James H. Harris, Dennis Bell, Thomas R. Hawkins, Sgt. William Carney, Christian Fleetwood, Pvt. James Daniel Gardner, Sgt. Alexander Kelly, and Sgt. Thomas Shaw.
Subject: Topical	African Americans; Military service; Military personnel; Union; History; Military personnel; Union



Subject: Geographic	United States
Subject: Time Period	1860-1870; 1860-1870; Civil War , 1861-1865
Note	Individual portraits have copy negatives in range LC-USZ62-118551-118565.
Note	Plate 69.
Note	Daniel Murray Collection (Library of Congress).
Note	No known restrictions on publication.
Note	Surrogate available as microfilm in P&P Reading Room.
Note	Exhibited: American Treasures, Library of Congress, 2005-06.
Identifier	2001695575; LC-DIG-ppmsca-08997 DLC; LC-USZ62-91683 DLC; hdl:loc.pnp/ppmsca.08997; hdl:loc.pnp/cph.3b38021
Classification	LOT 11931, no. 69
Location	Library of Congress Prints and Photographs Division Washington, D.C. 20540 USA
URL	http://hdl.loc.gov/loc.pnp/ppmsca.08997
URL	http://hdl.loc.gov/loc.pnp/cph.3b38021
Language	English
Rights	No known restrictions on publication.
Institution	Library of Congress Digitized Historical Collections: African American Photographs Assembled for 1900 Paris Exhibition

Session 2

DLF Collections Registry and A9.com

The University of Illinois at Urbana Champaign (UIUC) is hosting the DLF Collections Registry <<http://susanowo.granger.uiuc.edu/DLFCollectionsRegistry/browse/index.asp>>. John Carlson (DLF graduate student researcher) added GEM subject headings to the records for these collections. You can redirect your browser to a collection's host site from the records. The Registry can be browsed by subject, geographical area, and date ranges. The date ranges can be visualized using a tool Thomas Habing developed. Soon this metadata will be harvestable using OAI.

The IMLS Digital Collections and Content Registry allows clustering by type of collection. There is an effort to integrate an item-level and collection-level record, so you can switch between the two.

What metadata has to be exposed to be harvested into the DLF Collections Registry? UIUC is acting as a broker to expose collection-level records with OAI. Item-level records are not always associated with collections. At UIUC they are using sets records to expose information about collections. In the OAI world, there isn't always a use of sets as

a collection-level record. Sets are a logical breakdown of overall groups of items, but aren't necessarily intentional collections. There is some time being spent to help identify what items belong to a collection, and what collections have extant items. The technical registry shows how terms like “published” are used to identify sets.

Thomas Habing also presented on the experimentation with open-search protocol developed for A9 <<http://www.a9.com/>>. Using the A9 protocol enables federated searching of the DLF MODS Portal along with Amazon, Wikipedia, and hundreds of other information sites. Service provider: In order to do this, all he had to do was register a URL template. User: When you register with A9, you can customize the list of databases that will be searched when you execute a federated search with A9.

An A9 search of “Civil War” across Amazon Books, Wikipedia, and the DLF MODS Portal.

The screenshot shows the A9 search interface with the search term 'civil war'. The search results are organized into three columns:

- Book Results:** Shows 1 - 10 of about 225,267 results. The top result is 'The Official Military Atlas of the Civil War' by U.S. War Dept., George B. Davis, Leslie J. Perry, Joseph W. Kirkley, and Calvin D. Cowles (13 October, 2003) - Barnes & Noble. A second result is 'Civil War Sites: The Official Guide to Battlefields, Monuments, and More' by Civil War Preservation Trust (25 February, 2003) - Globe Pequot.
- Wikipedia Results:** Shows 1 - 10 of about 8,220 results. The top result is 'civil war' with a summary: 'For other uses, see [civil war \(disambiguation\)](#). See [list of civil wars](#) for individual examples. A **civil war** is a [war](#) in which parties within the same [culture](#), [society](#) or [nationality](#) fight for political power or control of an area. Some civil wars are also categorized as [revolutions](#) when major societal restructuring is a possible outcome of the conflict. An [insurgency](#), whether successful or not, is likely to be classified as a civil war by some [historians](#) if, and only if, organized armies fight conventional [battles](#). Other historians state the criteria for a civil war is that there must be prolonged violence between organized...
- Digital Library Federation MODS Portal:** Shows 1 - 10 of about 8,220 results. The top result is 'Medal of honor men [graphic]: Have received medals of honor in United [...]' with details: '[...] States **Subject: Time Period:** 1860-1870; 1860-1870; **Civil War**, 1861-1865 **Note:** Individual portraits [...]' and a small image of a medal. A second result is 'Arms of ye Confederacie' with details: '[...] Flag of the Confederacy; Plantations and planters; South Carolina, in the **Civil War**; Slaves and slavery **Note:** Title [...]' and a small image of a flag.

Session 3

Best Practices Work

Tim Cole discussed identifiers, invoking some of the work done by DLF Aquifer, which is creating a central repository for tools to search, capture, and manipulate content. The Asset Actions work that the DLF Aquifer group is doing now provides an applet tool -- The Virginia Collector Tool -- that allows one to export and annotate images. See <http://rama.granger.uiuc.edu/assetactions/index.asp> for more details. A version that works with textual materials as well as images is in the planning stage.

Sarah Shreeves (UIUC) reported on the *Best Practices for OAI Data Provider Implementations and Shareable Metadata* <<http://oai-best.comm.nsdlib.org/cgi-bin/wiki.pl?TableOfContents>>, work supported by this IMLS grant and the National Science Digital Library that dovetails with the DLF Aquifer work and their *DLF MODS Implementation Guidelines for Cultural Heritage Materials* <http://www.diglib.org/aquifer/DLF_MODS_ImpGuidelines_ver4.pdf>. Together these guides are prescriptive, based on what we have learned working with OAI. The audience for this information is libraries and vendors, and the best practices work has already informed some of the training curriculum developed by Martin Halbert and others at Emory.

Martha Brogan reported on her 2003 work *A Survey of Digital Library Aggregation Services* <<http://www.diglib.org/pubs/brogan/>> -- and the follow-up work to this -- *Contexts and Contributions: Building the Distributed Library* -- which she is currently completing as part of the IMLS grant. The current report ties in to the grant work by inform our continuing efforts “to foster better teaching and scholarship through easier, more relevant discovery of digital resources, and a much greater ability for libraries to build more responsive local services on top of a distributed metadata platform.” The new report

highlights major developments affecting the ecosystem of scholarly communications and digital libraries since the last survey and provides an analysis of “OAI implementation demographics,” based on a comparative review of repository registries and cross-archive search services. Secondly, it reviews the state-of-practice for a cohort of digital library aggregation services, grouping them in the context of the “problem space” to which they most closely adhere.... [T]he report investigates the purpose, function and challenges of next-generation aggregation services ... [and situates] these services in a larger context and to understand how they fit into a multi-dimensional and interdependent ecosystem supporting the worldwide community of scholars. Finally, the report summarizes the contributions of these services thus far and identifies obstacles requiring further attention to realize the goal of an open, distributed digital library system.

This time, there are about 40 providers examined, and new concepts such as cyberinfrastructure, open access publishing of scholarly articles, and the effect of Amazon, Yahoo, and Google on information seeking behaviors are being addressed.

Speaking of the DLF portals and related publications in general, Bruce Rosenstock suggested that this work needs to be marketed more widely, as too little of what we are discussing today is known to the scholars who we aim to serve.

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? We need the voice of the scholar matched with the technology. Libraries have outrun the scholars in the

humanities, and collectively we need to find ways to discuss the needs of (and the path to) digital scholarship in the humanities.

Randy Shifflett (VPI) suggested that changes in scholarship are currently technology driven. The technology is a tool for generating content. This may be true in the humanities, but some disciplines, e.g., astro-physics, have been transformed. Much scholarship is based on a journal economy.

In response, Roy Rosenzweig demonstrated the SmartFox tool, a “personal library” plugin for the Firefox web browser that can capture data and metadata from websites, being developed at GMU.

Session 4

Looking Forwards

In closing, we took a while to reflect on what we have seen in this day-long session and what we need to be doing, moving forwards.

Steve Railton (UVA) commended the project team on their impressive technical work, and suggested the next step has to be to train scholars, to help them re-imagine scholarship.

Roy Rosenzweig pointed out that graduate students are embracing technology and open content, but too often what is lacking is content that –once accessed – can be processed in more complex ways. We can search and find, but still cannot do very much with what we find except to use it in the fashion and with the tools that the publisher or library thinks fit.

Jeff Young said there needs to be more ways to decouple bundled services to extend their use by scholars – and to make them simple to use. People should pursue protocols because the benefits far outweigh the initial confusion about how they work or are implemented.

David Seaman asked, would it be sufficient to use Google as the OPAC for your library, if all the holdings from your library were available there? Bruce Rosenstock suggested the online resources page available from his library is the worst search interface. If he could have something available through his library gateway like what A9 is doing, that would be great. We should be able to provide HTML pages of things we'd like to share.

Steve Railton requests different ways of visualizing returns, e.g., dates and clusters. Long textual lists are rarely the best format for displaying results. Visualization speaks volumes to K-12 learners as well. Randy Shifflett observed that visualization is a breakthrough technology and technique for historians, and suggested that increasingly our students have a strong visual sense of the world.

Persistent identification is still an issue. Even legacy identifiers that could be scraped from a Web page could be useful. These could be mapped into global persistent identifiers. The stability of the identifier is necessary to develop the systems for reliable discovery-to-delivery services (D2D). What is the library's role to enable scholars to capture and re-purpose content? How can scholarship be built around increasingly granular bits of content?

In closing, Gail McMillan requested another term for “metadata remediation”. When you have limited resources, then some metadata is better than none. Metadata enhancement may be a better way to refer to it, as we look at what can be fixed after a metadata harvest and what cannot be done at that stage.

With no further business, and with a real sense of accomplishment, the Advisory session closed.