Report on the Digital Library Federation 2013 Forum

4-6 November, 2013, Austin Texas

Jody L. DeRidder

Over the past few years, the Digital Library Federation has opened its ranks to include institutions of all sizes and funding levels. Originally casting itself as "small and agile," seeking to "lead the pack in innovation," this stance has been moderated to be more inclusive. My impression from this conference is that this group is still developing its new identity. Attendance was at an all-time high, with over 360 registrants. Presentations were selected by majority vote prior to the conference. Many of those not selected became poster sessions with brief introductions and "snapshot" presentations lasting only 7 minutes each. As always, the program featured several opportunities for networking between, before, and after sessions.

Internet access was freely available, and Google docs were created for shared note-taking for each session, linked from the schedule.

As this is the leading U.S. conference for academic and national digital libraries, I was dismayed by the lack of focus on linked data and semantic web interoperability, and the constant complaints about lack of resources. Many of the sessions were Hydra-related, as it seems to be the "next big hope for salvation" and the sessions on assessment were extremely popular, as we all need to justify our funding. However, it is my perspective that we need to focus on how to better engage users, in order to attract necessary funding and support. Echoing this perspective, the opening keynote speaker focused on the need for engagement even in digital library systems, the first session I attended focused on impact, and the closing keynote speaker promoted outreach and marketing.

Contents

Day One

Monday, November 5

Opening Keynote

R. David Lankes *is a professor and Dean's Scholar for the New Librarianship at Syracuse University's School of Information Studies and director of the Information Institute of Syracuse. His book,* The Atlas of New Librarianship *won the 2012 ABC-CLIO/Greenwood Award for the Best Book in Library Literature. Lankes is a passionate advocate for librarians and their essential role in today's society.*

A University of Texas Austin librarian introduced the keynote speaker by noting that a rare book which had never left their shelves, had obtained 3421 readers within a week via Google books.

R. David Lankes is a professor and Dean's Scholar for the New Librarianship at Syracuse University's School of Information Studies and director of the Information Institute of Syracuse. His book, *The Atlas of New Librarianship* won the 2012 ABC-CLIO/Greenwood Award for the Best Book in Library Literature. His topic was entitled "The Mortal in the Portal or Why Everything You Learned in Library School is Wrong." When he asks employers what they want from students, they say "common sense."

Lankes quoted Albert Einstein: "Common sense is the collection of prejudices acquired by age eighteen." Common sense tells us:

- Libraries are good and necessary things
- Libraries collect, organize, and provide access to information

Lankes stated that the buildings don't do any of these things. We need to get rid of the idea that libraries do anything; librarians are the ones doing the work.

But if we break that down, it doesn't work very well, even using "librarians". Why do we collect, organize, and provide access to information? And for whom? What is "information"? What is a document, when it changes constantly? An ebook is an app, not a book. What IS a book now?

Our usual definition of data, information, knowledge, wisdom is based on a personal perspective, which varies from person to person. Lankes prefers the concept of knowledge, for it has a more functional definition. The way you see the world is knowledge. It's a fundamentally human process. It comes from engaging in conversations to learn and modify how we see the world. The majority of those conversations are with ourselves. We build the conversation as a series of networks. It is unique to a human. How you understand a concept is unique to you.

Lankes states that librarians collect, organize, and provide access to "conversations and their relations to objects and other conversations," rather than information. What matters is how we've organized information into a series of conversations that make sense.

Per the concept that "librarians are good and necessary things," Lankes brought up his favorite slide, entitled "Mortal behind the Portal" with a notation "(this slide left intentionally blank)." He said it makes him furious every time he looks at it. The slide came from a presentation by Bill Arms, NSDL, October 25, 2002, in the OCLC distinguished lecturer series. Back then we believed everything could be automated. But if so, then what is "professional service?" Lankes contends that:

- Service is important
- The provider is an essential part of the service
- Increasing outcome is knowledge creation
- The professional is guided by a mission

What's our mission? Lankes says that "The mission of librarians is to improve society through facilitating knowledge creation in their communities." We are focused on how to "get people to the stuff" but we need to think more about the ultimate goal.

How do librarians facilitate conversations? First, provide access to the conversations. But that's not enough. We also provide training/knowledge; we increase the knowledge of the person (instructional), to teach people about scholarly communication. We need to provide a safe, secure learning environment. We use social media to find and collect information.

But what is our motivation? The big thing is that ultimately the mission of librarians is to improve society. We seek to change the world. We are not being neutral or unbiased. We have to have a reason why we're doing this. Librarianship is a radical profession.

Radical means: extreme; fundamental (the root, from which new life grows); cool. Librarians are radical change agents, seeking to make things better. It is a noble profession. We are thoughtful, proactive individuals. We need to measure what we do in the framework of why we are doing it. That engagement with people is what matters.

Community is what we're building. Yet most digital libraries are focused on preventing theft, not building community. All these systems and tasks ("engaging in the weeds") are simply tools to achieve a goal. We need to focus on the purpose of the digital library: to improve the world.

There was a question from the audience about "alt-acs" (alternative academics) and how they fit into this view.

Lankes replied that there are three ways to become a librarian: the degree, the job called "librarian," and most importantly, to be a librarian by spirit. These people have the same motivation and goals. We need to open the door wide to this third group and value each other.

As someone goes through our collections, what tools do we offer them to make connections and tell us their story? We need to provide the infrastructure for multiple connections. We need to structure digital library systems to support relationships. Lankes referred us to his new book *The Atlas of New Librarianship* (companion site). We need to get out of our buildings and listen to what others are talking about, and then align things.

It makes no sense to hire in new librarians to change our systems, or to form committees to decide what is innovative. We need to bring in people with other types of education and capabilities, throw away status, and bring in those who have passion, abilities, knowledge beyond our own, and the wherewithal to make real change, the same mission we have -- and we need to change *with* them.

Our digital libraries should support multiple languages and multiple ontologies, and allow our users to filter and connect and relate and build new views. We should engage in an ongoing participatory conversation with those we seek to serve. Let them define the boundaries of what we talk about, but librarians should not be submissive. We have a role to participate in that conversation. We are members of the community and have valuable knowledge and perspectives to share also.

Digital Collections: If you build them, will they visit?

Session Description

How do your cultural heritage organization's digital collections fare in search rankings? Assuming your collections have newspapers from 1915, will a Google search for information about the "Battle of Gallipoli" return results? At the April 2012 Bibliothèque nationale de France International Newspapers Conference, one of the authors examined web traffic rankings and search results for digital newspaper collections at libraries around the world. Both traffic rankings and search results showed that content in cultural heritage organizations' digital collections dwell in Internet obscurity (http://bit.ly/parisinternationalnewspapers).

In this session we re-visit these rankings and results, examining what it means for a digital collection to be successful. Is success only about page views, unique visitors, and bounce rates? If, paraphrasing Trevor Owens (http://crowdstorming.wordpress.com), the mission of a cultural heritage organization is more than random users flipping through the pages of its digital collections, how does one encourage and measure community engagement? Is crowdsourcing "the single greatest advancement in getting people using and interacting with library collections" (Trevor Owens)?

Session Leaders

Frederick Zarndt, IFLA Newspapers Section
Brian Geiger, California Digital Newspaper Collection
Alyssa Pacy, Cambridge Public Library
Joanna DiPasquale, Vassar College
Robert Stauffer, Hoʻolaupaʻi Hawaiian Nūpepa Collection
Meredith Palmer, DL Consulting

View the community reporting Google doc for this session.

Zarndt says that digital historical newspaper collections are typically the most-used collections in libraries with digital text collections. Trove has the largest (National Library of Australia) with over 9,880,000 pages. Yet content in library digital newspaper collections dwells in internet obscurity. Why?

Zarndt selected a search phrase and tried it first in Trove, obtaining 16,231 search results. When he tried the same phrase in Google, the first library-like research result was #18. In first 100 Google search results, not a single result came from the huge digital libraries of that content.

He then tried the same search in Google news, adding a date range restriction: most of the results come from paid publishers, such as New York Times. Again, in first 100 results, there were no results from their huge collections.

Zarndt said they're not blocking internet traffic, the content is not inaccessible. Checking in Elephind.com (a crawler that indexes historical newspapers) shows results, so we know that the content can be crawled by web search engines.

Zarndt quoted Nat Torkington in his Nov 11 address to the National and State Librarians of Australasia, Auckland, about how digitization project websites are really poor. He then cited research that 89% of colleges students use search engines to being an info search (2% start at library website); 93% are satisfied or very satisfied, and stated that libraries need to make their sites more highly visible in cyberspace. Closing with references to the University College London presentation on Information Behavior of the Researcher of the Future (CLIR report, Jan 2008), Zarndt asked, "How can we market text collections effectively?"

Brian Gieger then reviewed the effectiveness of simple SEO (Search Engine Optimization) strategies (sitemaps.org and robots.txt) and showed that for Cambridge, California Digital Library (CDL) and Vassar, it helps quite a bit. He said he presented at IFLA (International Federation of Library Associations) in Singapore about conventional methods to increase access: use/ collaborate/ publicize in local media, especially in newspapers. He advises we involve the collection users from the start, and notes that genealogists and family historians are the biggest users; most are 50 years old or older.

Geiger closed with the statement that libraries spend far too little on publicity, presentation, and SEO.

Metadata First: Using Structured Data Markup and the Google Custom Search API to Outsource Your Digital Collection Search Index

Session Description

Discussions of building the digital library have centered around using tools like contentDM and Solr/Blacklight to build local search engines for our digital content. This model has served us well and by many measures remains an industry best practice. However, as discussions about making our digital collections findable on the open web have evolved, the work of optimizing data and following SEO best practices has introduced new workflows and pressures to digital library development. In this session, we will explore how this new search-oriented workflow can be incorporated and utilized to provide an effective and efficient local search engine for our digital collections.

This session's focus will be Google Custom Search, a search tool that offers powerful and flexible indexing for a range of content, including the images, documents, and videos that comprise digital collections. Building on Google's search algorithm with Custom Search provides an achievable alternative to other more complex indexing tools such as Solr/Blacklight. We will consider how investing in structured data to create indexable content that can be consumed via a Google Custom Search API client has further advantages over other local search implementations. Specifically, speakers in the session will discuss:

- *Business cases for outsourcing your search index*

- *Gains in SEO and discoverability*
- *Preparing collections for a search engine index (e.g., sitemaps, schema.org HTML markup)*
- *Using analytics available to Google Custom Search to understand the use and indexability of your content*
- *Applying the Google Custom Search JSON API to power your collection search*

Among the benefits of this "metadata first" approach to digital library development, the MSU Library's experience with Google Custom Search has provided an increased understanding of the operations of commercial search engines that crawl and index our content. We have also been able to efficiently align digital library personnel in the work of building interoperable and indexable structured data markup with schema.org and RDFa. And finally, we note the lowering of a significant barrier to creating a searchable collection index and the efficiencies gained in optimizing your digital collections data for commercial search engines.

At the center of this presentation will be a demonstration of how the MSU Library is using the Google Custom Search API to build our local collections search (arc.lib.montana.edu/digital-collections/). The session will include a discussion of how an investment in structured data and indexable content, rather than local search engine development, is a viable path forward for digital libraries.

Session Leaders
Kenning Arlitsch, Montana State University Library
Jason A. Clark, Montana State University Library
Scott W. Young, Montana State University Library
Patrick O'Brien, Montana State University Library

View the community reporting Google doc for this session.

Jason Clark recommends spending time on building your index of what search engines consume, and making sure that discovery is happening elsewhere. He called this "Discovery turned inside-out" in a Duke University Library presentation, at CNI last year.

Foundations of indexable content begin with a software tool that provides:

1. Item pages at stable resolvable URL
2. Standards-based HTML5 markup
3. Structured data markup
4. Navigable architecture with clear design

Scott Young then took over the presentation, starting with "navigable architecture with clear design."

They changed the organization of their content so item pages are not as far from top level. What's good for human users is also good for bots (web crawlers).

They utilized traditional SEO: title tag and metatag description (both unique to each item), sitemaps and robots.txt alignment, and server responses and error pages (to ensure bots can access the pages). He noted that 404 errors and redirects to the home page cause blockage and confusion. He highly recommended use of Google analytics and webmaster tools. Site errors are very important.

Then they moved into use of RDFA Lite, schema.org microdata, and semantic web tools which can make use of this structured data. Right now they are experimenting with Twitter Cards to see how this impacts traffic.

They are using JSON-LD for semantic web support, but are still in early implementations. The focus is on reusing indexed content in order to develop full collection discovery.

Using SOLR with Blacklight (from NCSU) enables faceted search, flexible results, stable URLs, and contemporary design, but the barrier is development time. As an alternative, they turned to using Google custom search in local implementation.

Jason then referred to the web as a database itself. Google is clearly the best indexer. He said that Google Custom Search (GCS) offers three levels of support, and they just bought the highest one, in which Google ships you a server, and you keep it for 3 years to index your content for you.

This enables local discovery by reusing web-scale index; he called it an onramp for their digital collections discovery layer, stating that it is efficient for libraries (small and large). The service leverages an index already optimized for web search, already integrated with a leading commercial search engine. It provides faceted browsing, flexible design, search analytics, recommendations, auto-complete, API, and more are included.

Again, the barrier for GCS is the same as with BlackLight: development time. The API documentation is not great at times. And of course there's a small cost with these services. It's a tradeoff.

GCS efficiencies:

- Build indexable content for bots and humans
- Reuse index locally

They made the display look like Blacklight. The service can generate thumbnails for you. You introduce URL patterns you want and label URLs for faceting. It's easy to use.

There's a pricing model; you get 100 search queries per day for free. For MSU they estimated 20k queries a month, or about $1200 per year as cost. For their discovery layer (Summon) they expect 10k queries a month). It is possible to get a free version for cultural institutions. It comes with rudimentary analytics, but when hooked into web analytics, you get the full breadth of services.

Hunting for Best Practices in Digital Library Assessment

Session Description

Research and cultural heritage institutions are increasingly focused on providing online access to digital special collections and archives. Since funding to these institutions is simultaneously decreasing, we need to strategically focus our efforts, and better understand and measure their value, impact, and associated costs.

However, methods in digital libraries are not yet standardized for identifying user groups; measuring usage, impact, cost and value; obtaining feedback; or analyzing results. As a result, findings cannot effectively be generalized.

What strategic information do we need to collect in order to make intelligent decisions? How can we best collect, analyze, and communicate that information effectively?

Examples of efforts to address these questions will be shared by panelists, along with problems encountered. Audience participants will be asked to help brainstorm how best to standardize evaluation methods. We are testing the waters for the potential of a collaborative effort to build community guidelines for best practices in digital library assessment. By the end of the session, we hope to have consensus on the main areas of need for establishing assessment best practices, as well as at least one actionable idea for moving towards this objective.

Session Leaders

Jody DeRidder, University of Alabama
Sherri Berger, California Digital Library
Joyce Chapman, State Library of North Carolina
Cristela Garcia-Spitz, University of California, San Diego
Lauren Menges, University of North Carolina

View the community reporting Google doc for this session.

Each of the session leaders spoke of their experience and interest in assessment of digital libraries. Sherri Berger focused on measuring impact; Lauren Menges on meeting user needs; Joyce Chapman and Cristela Garcia-Spitz on assessing costs and benefits. My intro was as follows:

“Thirteen years ago, Tevko Saracevic (in Library Trends) said that the six levels of digital library assessment criteria developed thus far were content, technology, interface, service, user and context.

To date, however, digital library evaluations have been largely focused on interface and user levels. Content and context levels receive little attention. Moreover, most of the criteria used are merely borrowed from the domains of traditional library and information retrieval system.

In 2009, Ying Zhang analyzed these 6 levels into multiple aspects, and then asked groups of developers, administrators, librarians, users and researchers to rank each aspect in terms of importance. He then compared the results across all these groups to determine the top criteria across the board.

However, I contend that the differences between the groups should guide the assessment of digital libraries. If what is most important is the user's perspective or the researcher's perspective, then those criteria are the ones that should be tested. However, a full assessment of any digital library system may require testing from several different points of view. Obviously, if it's too difficult to manage and support the system, that's critical; also if it's too costly. And those would fall into the developer's and administrator's points of view.

Zhang *also* noted which of the aspects had not yet been tested in published studies, and that concerns me. In following in the footsteps of traditional library and information retrieval systems, I believe we are overlooking areas that are critical to evolving digital libraries, and the evolution of our users, and of how they access and use our systems.

Especially now as we begin to move into support of linked data, it seems clear to me that we need to establish guidelines as to what one should consider in assessing a digital library, from any of the perspectives that could be considered critical. And that is why we are here today. I hope you will work with us to take the first steps towards mapping out digital library assessment. There is no reason that this should continue to be an area of confusion for any of us. Digital libraries are no longer in their infancy, and all of us need to be able to make wise choices with our resources these days. Assessment is critical to making those choices. So thank you for coming! I urge you to share your ideas and your thoughts, and help us step forward on this path."

At this point, we asked the attendees to select their preferred focus for the "small group" discussions: we had set up flip charts and areas for user needs, costs and benefits, and demonstrating impact, with additional flip charts and areas for other focuses related to digital library assessment. The questions for breakout session number 1 ("What are the challenges?") were:

- What assessment have you done in this area?
- What methods/approaches have you used?
- What issues have you encountered?
- Where have you experienced the most difficulty?

As we had close to 200 attendees, I created a division of the two largest groups into separate sections, to enable better discussions. I facilitated one of the extra groups, and obtained a volunteer to facilitate the other.

After about 25 minutes, we shared the highlights of the discussions of each group, and went on to breakout session number 2 ("Where do we go from here?"), with the following questions:

- What do you think would be the best approach to solving this challenge?

- What can we do to support that (guidelines, new tool, best practice, etc.)?
- What would you be willing to be involved in?
- How can we engage others?
- Who will be the leaders of this effort?

Because the discussions were so intense, we were unable to finalize results from this session, so instead we collected names and emails (over 50) of those wanting to further this discussion, and thanked everyone for their contributions. We are determining the best method to carry forward this effort at this writing.

Embedded Semantic Markup, schema.org, the Common Crawl, and Web Data Commons: What Big Web Data Means for Libraries and Archives

Session Description
Search engines are reaching the limits of natural language processing while wanting to provide more exact answers, not just results, especially for the mobile context. This shift is part of what has spurred progress in how data can be published and consumed on the Web. Broad and simple vocabularies and simplified embedded semantic markup is leading to wider adoption of publishing data in HTML. Libraries and archives can take advantage of new opportunities to make their services and collections more discoverable on the open Web. This presentation will show some examples of what libraries and archives are currently doing and point to future possibilities.

At the same time as this new data is being made available, only a few organizations have the resources to crawl the Web and extract the data. The Common Crawl is helping to make a large repository of Web crawl data available for public use, and Web Data Commons is extracting the data embedded in the Common Crawl and making the resulting linked data available for download. This presentation will share data from original research on how libraries currently fare in this new environment of big Web data. Are libraries and archives represented in the corpus? With this democratization of Web crawl data and lowered expense for consumption of it, what are the opportunities for new library services and collections?

Session Leader
Jason Ronallo, North Carolina State University Libraries

View the community reporting Google doc for this session.

There are three major types of embedded semantic markup:

- Microformats
- RDFa (Lite)

- Microdata

Ronallo introduced us to Common Crawl (http://commoncrawl.org/), saying that "Common Crawl builds and maintains an open crawl of the web that can be accessed and analyzed by everyone." 12% of their content has embedded structured data.

Instead of triples, many now use N-Quads, which adds context. The fourth parameter is a reference to the provenance, or the content about which the triple is created.

He compared the amount of semantic markup available from NCSU peer institutions (see presentation). Penn State had the most; 18% of their pages have some sort of semantic markup. Some institutions do not have their library sites crawled at all.

Penn State U is using:

- Hcard (people, places, orgs) (most common; a microformat)
- Hcalendar (events; also a microformat)
- Open graph protocol (Facebook-supported RDFa) (next most common)

He asked for sitemaps, so he can include more content in his research, and noted that NCSU only has 4 pages in the common crawl right now.

Enhancing an OAI-PMH Service Using Linked Data

Session Description

For over ten years the Sheet Music Consortium has been harvesting metadata using the OAI protocol and providing user services at http://digital2.library.ucla.edu/sheetmusic/. With the support of IMLS planning and leadership grants the latest iteration of the portal maps all metadata to MODS (rather than DC), invites users to add structured metadata to records, offers metadata downloads, and provides metadata mapping and Static Repository services that facilitate the participation of smaller and less technically able institutions. Despite these enhancements the problems inherent in metadata harvesting projects persist, including variant metadata standards, inconsistent application of standards, and varying levels of authority control. Utilizing the user-supplied metadata infrastructure and Linked Open Data principles and standards, SMC has initiated a project to both improve the normalization of the Consortium's metadata and expose enhanced metadata as Linked Open Data (LOD), thereby expanding its impact and our ability to share data more widely and effectively, both directly with our users and through automated systems.

The Consortium's strategy for publishing trustworthy linked data leverages the user-supplied metadata layer of the data repository, which is maintained separately from the harvested data. Text analysis tools such as Open Refine and Voyeur/Voyant are used to group data and assist in the identification of

appropriate normalized forms, which are then written to the user-supplied metadata layer that forms the basis for publication of LOD records.

Our presentation will address and compare the challenges and possibilities for publishing LOD for creators, titles (works), publishers and subjects. Then we will discuss a pilot project that focuses on normalizing publisher information and exposing that as linked open data.

While this case study is focused on sheet music, the methods discussed are generally applicable in the context of harvested metadata.

Session Leaders
Stephen Davison, University of California, Los Angeles
Elizabeth McAulay, University of California, Los Angeles
Claudia Horning, University of California, Los Angeles

View the community reporting Google doc for this session.

Horning states that resources are not always available for authority work at point of description or aggregation. Some important elements (publishers, for example) not traditionally subject to authority control.

In addressing the challenges of aggregated metadata, one of the most difficult is the ability to aggregate "works" (composer and title), due to variations in practices by contributing institutions. Many agents do not have entries in national authority files.

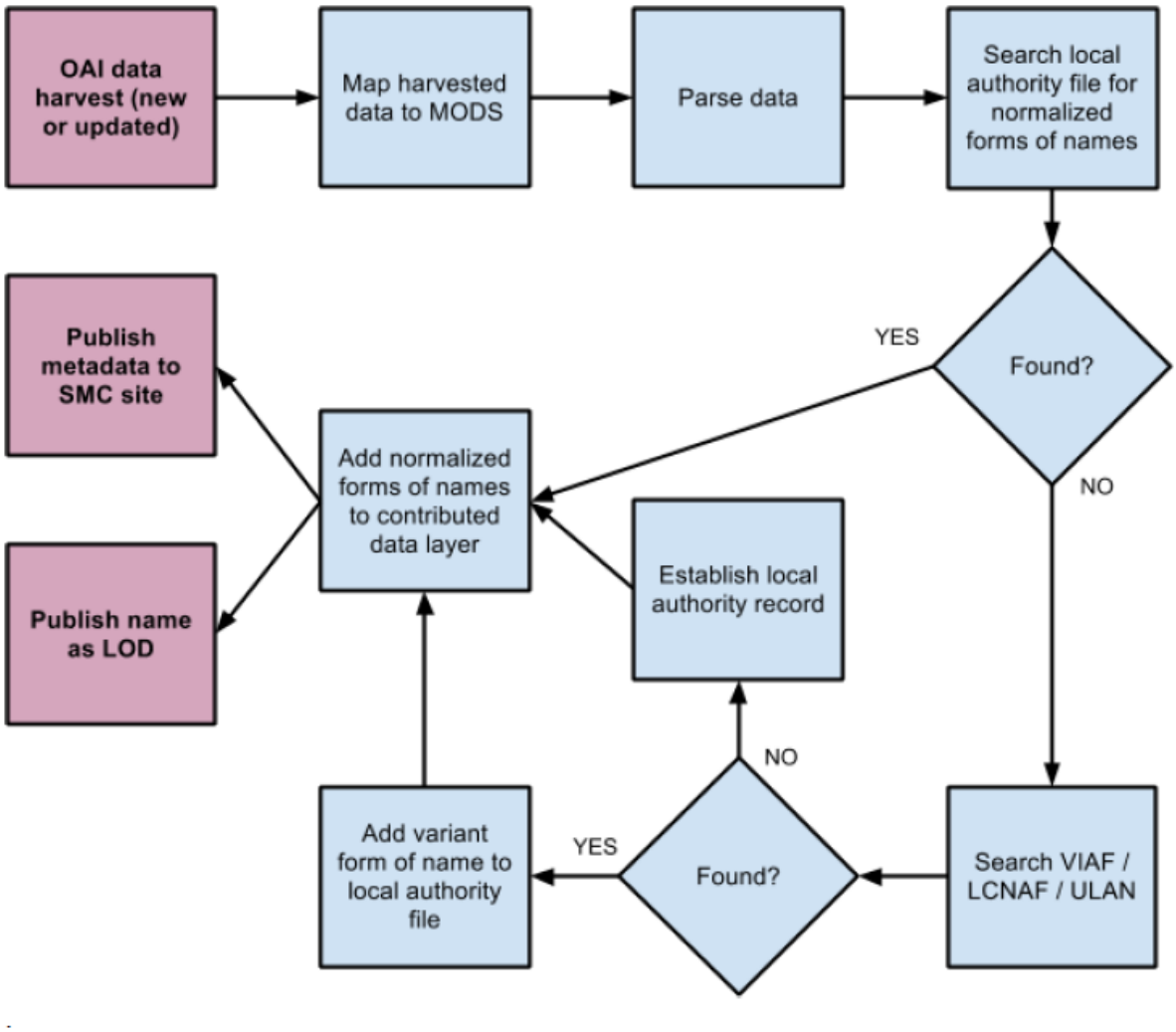
Titles are difficult to define. Some use first line of text, some first line of chorus; the same music may be published under different titles, and the same title may be used for multiple songs.

Access points they want to add include: creators (LCNAF), titles/works (FRBR), subjects (LCSH, TGM), and publishers. What authority is available for publishers? Some are in LCNAF, simply because they "authored" works such as catalogs. They consider adding publishers to be enriching the collection after the fact, as a pilot project. On published items, publisher names and locations change frequently. Roles of composers, lyricists, pubs and performers are more interrelated than in many other forms of publication. However, Linked Open Data enables us to enrich bibliographic information and create actionable metadata.

In terms of infrastructure, new data is not easily written back to contributing institution. User supplied metadata is kept separate. Associations between harvested and contributed metadata could be lost upon reharvesting.

Here's their proposed name authority workflow:

Figure 1 Proposed name authority workflow



(Davison, Stephen; Sugiyama, Yukari; McAulay, Elizabeth; & Horning, Claudia. (2013). Enhancing an OAI-PMH Service Using Linked Data: A Report from the Sheet Music Consortium. UCLA: Retrieved from: http://escholarship.org/uc/item/50c9g411, p. 17.)

Pathways to Stimulating Experiential Learning and Technological Innovation in Academic Libraries

Session Description

Technology units within academic libraries are continually searching for new ways to leverage technology to improve digital collections and library services to students and faculty. Oregon State University Libraries & Press (OSULP) and the University of California, Los Angeles (UCLA) Library have discovered that in many cases the students themselves are best positioned to develop innovative solutions to the challenges they face.

These two libraries have pursued different approaches to promoting and providing opportunities for student leadership in technological innovation. UCLA created Simul8, a team of developers made up exclusively of students who work in an agile programming environment to develop ideas for services, and design interfaces by and for students. Much of their activity to date has focused on mobile and web app development for collection access and data visualization of digital library collections. OSULP also works in an agile manner, but utilizes a more directed approach that includes student programmers guided by Library staff, and collaborations with academic departments that enable students to develop technological innovations via capstone projects and internships. Examples include web sites and apps that build on the content of monographs published by OSU Press, as well as online tools for students to secure study rooms and tables in the Library.

At both OSULP and UCLA, a culture of collaboration with students underpins these programs. This session will offer examples of projects we've created with our student programmers, the principles behind these strategies for student engagement, and the tools and infrastructure utilized. We will discuss how this approach supports digital content development and experiential learning for students, as well as the benefits and pitfalls of engaging students in this kind of mutually beneficial enterprise.

Session Leaders

Shan Sutton, Oregon State University
Evviva Weinraub, Oregon State University
Todd Grappone, University of California, Los Angeles
Kevin Rundblad, University of California, Los Angeles

View the community reporting Google doc for this session.

Oregon State U is using students in graphic design, web services, programming and public services support. They put them in charge of projects, and set aside at least 20% of the student time for learning from existing faculty/staff. They offer capstone projects for CS and EE, which are invariably selected, and ask these students to build useful prototypes. OSU donated $1200 for the creation of responsive design sites, such as room reservations.

The students are required to use agile programming methods, open source software, OSULP (Oregon State University Library and Press) development guidelines, code review, programming and standards. The students work with clients, and are supervised and reviewed by the library programmer. The hope is that some of these students may someday become library donors.

Pros and cons: the students bring new perspectives, excellent ROI (return on investment), and this effort enables the library to facilitate student learning beyond classroom support. Cons: recurring training, potential for overreliance as replacement for permanent staffing, and the fact that student employee budgets are easy targets for cuts.

Todd Grappone spoke next to introduce the effort at UCLA, called Student Driven Innovation (Siml8 Group @ucla library). They have a mobile/web focus. Selection of projects are experience driven. They use a startup model, and consider innovation as simply part of the process. They've been doing this for 5 years.

Kevin spoke next with more detail. They were looking for how to engage students, and stay abreast of what apps they are using. They wanted to create an ongoing hacker culture. In forming the structure, "being effective" is the only rule. Strength is in creating. New experiences are built from new thinking. The students needed a more organic structure than the library mode, and they've determined that working in the same fashion the students do is even more important.

Students are P2P (peer-to-peer) mode, impromptu and self-organizing. Information flows in all directions. Their orientation embodies the "Netflix" culture: freedom and responsibility. One needs to embody respect for expertise in order to support innovation. They started with 5 students and two founding concepts:

#1: leverage student skillsets. So they started with, and continue to focus on mobile web development.
#2: innovation smeared over time. You can't plan everything up front; you have to let it play out. They build student value by providing a platform for their skills and giving them ownership over projects.

As a result the students go off to some great companies, because they did something cool and useful for us.

How the group works is important: independently on laptops, whenever they want. The group meets weekly, but otherwise has flexible hours. The supervisors listen to them and guide them. It's important to pick students who have motivation, and motivation to learn. The focus is on collaborative coding and experimentation, using github to share and version code.

They have developed several apps, such as:

- iphone and android apps with features like 101 top checkouts
- slide-out navigation,
- autosuggest for searches,

- course reserve
- stackMap.
- Turn tablet into e-reader using twitter bootstrap.
- Off campus and cell connection detection, automatically gives VPN (virtual private network) links.
- Responsive design for both mobile and tablet
- Stashd: a bookmarklet to save anything you're looking at online. It uses JavaScript. The resulting collection is searchable and organizable. They want to build this out to include interoperability with Zotero and similar citation software.

"If you get the culture right then most of the other stuff … will happen naturally on its own." (Tony Hsieh, CEO of Zappos)

Day Two

Tuesday, November 5

Snapshots Group A

Snapshots are 7-minute presentations meant to engage and energize the audience.

Presenters are asked to give a dynamic overview of their topic in a quick timeframe, with up to 24 slides. Snapshot presentations are grouped together based on an over-arching theme or idea, with time for audience questions at the end of each session block.

View the community reporting Google doc for all of the Snapshots.

Using Hadoop to Process Big Data at Scale

Roy Tennant, OCLC Research

Apache Hadoop provides a parallel processing infrastructure using MapReduce techniques that enable organizations to process massive amounts of data quickly on commodity hardware. How Hadoop and HBase are being used to effectively process this data will be covered, as well as lessons learned and things to consider when using these tools to process "big data". Examples of services created from this processing will be highlighted, as well as specific techniques OCLC Research has found useful within a Hadoop environment.

OCLC has been implementing MapReduce processing since 2005 using Python code. In 2012 they moved to larger cluster work with Hadoop and Hbase. They use Java, Python and Perl to work with

Hadoop. For all this work, it's necessary to think of it as a two-step method: map, and reduce processes. However, you don't have to have a reduce component; you may only need map processes.

Once a process flow has been created, they can add to the content at will and Hadoop will manage it. Hadoop expects Java, but you can string the processes together with different programs using the "streaming" option. Jobs are best run in shell script, as they require multiple parameters.

The data sits on Hadoop Distributed File system (HDFS). A shell script calls Hadoop, giving it instructions. Output can be written back to the HDFS. String searching can be done in a few minutes across 7 million records using this process.

When using "reduce" component (which basically summarizes), this adds another step before writing out the content. They just chain it into the process. One can also output to standard error and watch/interact with the process as it happens.

Hadoop comes with a certain number of things you can do to monitor the processes. Reducing can start before mapping is complete (this is controllable). You can get a report of the status of all jobs in progress and can also see output and status of a single job. Hadoop gives a running report of progress.

Worldcat identities is all Hadoop work. Marc usage in Worldcat came from this effort.

The Growing Impact of R in E-Scholarship

Harrison Dekker, University of California, Berkeley Library Data Lab

This Snapshot will explore how the principles of reproducible research are helping drive development in the R community. It will also identify and discuss the functionality of specific R packages pertinent to the digital library community and where opportunities for collaboration exist.

R is a statistical math language, with some overlap with Python, and a couple others. R was created for and by statisticians; it came out of a language called S. R is heavily used in academia, and growing in industry. Compared to other stats packages, R is used across disciplines. You don't have to know how it works; it's a black box. You can write some code and distribute it through the software network (CRAN: Comprehensive R Archive Network). There's about 5,000 packages on that network. Increasingly they are also available on github. R is Cross-platform; you can even make it available through a browser from a server. R-studio is an IDE (integrated development environment) for it, commercial quality but free, and this is making it more popular.

There are bundles of packages (called "task views") many of which support use of "markdown" in which you include your R code and then render to HTML or PDF, to show tables/outputs in your document. If any data changes, just recompile your code, and everything is updated. Most of the graphics functions are very strong, providing publication quality in a number of formats. All your work can be done in code,

so you can integrate your analysis with creation of visualizations. R Studio was also designed with reproducibility in mind. It gives you a way to manage your workspace. You can post the web documents directly to your website. ROpenSci project is a group of ecologists looking to create a unified network for uniting scientists with data from data repositories. Usgsd is similar – for working with major government survey content.

Opportunities: build APIs to our repositories. Build R metadata tools (such as r2ddi). Promote it to researchers.

This is awesome as a platform for scientist/coders to build and share tools. It's great for accessing, manipulating, analyzing, visualizing and publishing data.

Preserving Software at Scale: The Stephen Cabrinety Collection

Michael Olson, Stanford University Libraries
Douglas White, National Institute of Standards and Technology

This Snapshot will describe the current status of the Cabrinety Software preservation project including work done to date, media degradation statistics by format, and data modeling of software for ingestion into our Digital Repository. Descriptions of the technologies used by NIST to capture software and photographs, and the lab workflow will be included. Deviations when encountering problematic software formats will be discussed. Details of metadata storage in the NSRL database will be provided. We will outline opportunities for partnerships and software preservation activities with other academic research libraries, and also summarize the importance of this data set to disparate communities.

They generate metadata to support this content for legal reasons. Now they are managing over 14,500 types of computer software. They capture images for forensic work. Cabrinety contains games from 27 operating systems, and multiple formats. Thus far they have 900 media images, and have a 17% error rate on those (pretty high). Sometimes generating an image is very difficult. For example, they could not generate a complete, consistent media image from a cartridge until they found a ROM (Read Order Mark) checksum in the header.

They just received their first batch of data from NIST (National Institute of Standards and Technology). Thy have an 83% success rate without modification or intervention. They can increase that 9% with human intervention, but 8% of media have many (> 10%) sector read errors.

For photography, they use RAW data and convert it. Cameras are much faster than scanners, and provide better quality images. They need to automate more. Hardware for legacy media is a real issue, especially for Apple. They even have audio cassettes containing software. And there can be failures even within the media.

They are doing data modeling for a repository now, and will create PURLs and integrate content into the Stanford catalog when rights allow.

Developing a Hydra Head without Fedora

Declan Fleming, University of California, San Diego

We have a locally developed, RDF-based DAMS [digital asset management system] at UCSD and we wanted to add a lot of new features to the front end UI. Instead of building from the ground up, we looked at open source community efforts and chose Hydra as a platform. We will present then discuss our strategy and architecture.

Ruby supports a test driven development focus. They needed a lot of flexibility too. They were a Drupal shop but needed more. They didn't know Ruby/Rails, but learning was quick. They are now keeping metadata in an RDF triple store rather than data streams. This data model enables access to other metadata at the point of need, reducing duplication of storage and supporting linked data and semantic web capabilities.

They provide open access to their DAMS ontology and data model, API, DAMS manual.

They are hoping that Fedora 4, which is to be RDF-based, will make their work easier. UCSD put Hydra on their own DAMS instead of on Fedora due to past investment. They then emulated the 25 Fedora calls that Hydra uses. Hydra only uses 20-25 of Fedora's 200+ REST API calls. UCSD spoofed them. This is significant and is now informing Fedora futures.

Automated Interactions With WorldCat: A Look at OCLC's WorldCat Metadata API

Terry Reese, The Ohio State University

Earlier this year, OCLC released through their developer's network specifications to a metadata API that provided direct access to both read and write data into WorldCat. As libraries look for more and more ways to automate mechanisms to proliferate metadata about the materials that they are capturing and preserving to wider audiences – OCLC's Metadata API encourages sharing with the cooperative by removing previous barriers to automated access.

This snapshot will look at a sample implementation of the Metadata API using MarcEdit, as well as specific feedback and techniques for working with the API.

There is a developer's network for this. They have fairly limited available operations so far, but can create/read/update bib data. There are 2 platforms, old and new, and they don't work well together. There's very little documentation. Terry was the first to try to make it work. There's no sandbox; you're working in real time. So right now there's test records in WorldCat that he put there to ensure that insertions work. They are live immediately online (not a good thing).

These API open up several possibilities including building pipelines between our repository systems and WorldCat, developing localized interfaces for metadata entry outside the library, and further automating tech services processes like batch record ingestion. There are, however, authentication challenges for those not in the community.

Conspicuously absent operations in the current OCLC API: record validation, anything to do with record validation, record locking, service status, and user validation.

Testing Omeka Core for Digital Library Services

Session Type: Presentation

Session Description

The Omeka web publishing software has been successfully used for a wide variety of digital exhibits and digital humanities initiatives. These valuable sites have often been implemented as standalone initiatives, however, heavily tailored towards an individual project's needs. As academic libraries seek digital library infrastructures that are sustainable, scalable, and easy to maintain, a number of them are testing Omeka as a platform for hosting a variety of disparate projects within a single technical infrastructure.

This panel will bring together speakers from the libraries at the University of North Carolina at Chapel Hill, Indiana University, the Ohio State University, and Duke University to discuss their pilot testing of Omeka through various formal and less formal means. Through a moderated, interactive question and answer format, speakers will address the goals of their respective Omeka pilots, the structure of the experimentation they have done, lessons learned, and any plans that they have developed for turning pilot tests into fully supported production services. A representative from the Center for History and New Media at George Mason University will draw threads from the various presenters together and discuss future development directions for Omeka that will advance its utility for multi-project production environments.

Session Leaders

Jenn Riley, University of North Carolina at Chapel Hill
Patrick Murray-John, Center for History & New Media

Leslie Barnes, University of North Carolina at Chapel Hill
Will Cowan, Indiana University
Tschera Connell, Ohio State University
Melanie Schlosser, Ohio State University
Will Shaw, Duke University

View the community reporting Google doc for this session.

The Omeka Development team manager is Murray-John. Omeka was originally developed for small institutions without tech departments. Now, it's being used by Europeana, Grateful dead archives and DPLA.

Will Cowan at Indiana U, Bloomington presented on "student life at IU." IU archives manages the content, their tech team manages the system. They already had special collection content in Fedora; but moved it, and added newly digitized content (using spreadsheets) into Omeka, using the exhibit function for the War of 1812.

Tschera Connell is the director of content services for OSU. She applied for the grant, tested creation of customized exhibits, and determined support for major exhibits.

Liz Mileqicz represented Duke U instead of Will Shaw, who builds the many Omeka Instances. They hosts Omeka only for specific exhibits. They're talking about a new direction, experimental, to determine where to build services. They want to use it to innovate undergrad curriculum, working with faculty and students, in a pedagogy focus, involving students in curating collections, describing and arranging objects. They would be pulling students into librarian type work. Their A/V lab is creating a Sonic Booms of the South exhibit, many of which use NeatLab. Their Wired Lab is also using Omeka. They want to use it more in cross-collaborations. Challenge: having to create instances is challenging as is setup for faculty/students. The learning curve is steep.

Jenn Riley (now at McGill in Montreal) spoke about UNCC efforts to scale up support for digital humanities work, using Omeka. They wanted a place where the faculty have ownership without technology being a big issue. They set up accounts for the faculty. They also set up a pilot for library exhibits of content already digitized. Leslie Barnes then spoke for UNCC. She worked on a literary atlas using Neatline, mapping different editions and such, and 2 pedagogical projects. One is end of year projects; she's using Omeka to handle field notes from the semester and as publishing platform.

Humanities Data Curation in the Library: The preservation of Digital Humanities Research Now and to Come

Session Type: Research Update

Session Description

Academic libraries have been critical partners on a number of early digital humanities research projects, and several pioneering digital humanities archives still endure today. This paper examines four case studies of digital humanities projects that are being maintained in libraries, and how the infrastructures developed by the academic libraries for these projects shed light on the academic library's role in curation of digital humanities projects.

The author conducted in-depth interviews with scholars, librarians, and project managers who oversee four pioneering and long-maintained digital humanities research initiatives: the Victorian Women Writers Project, William Blake Archive, Walt Whitman Archive, and the MONK Project. For each of these cases, the paper considers the curation workflows developed by the project staff, the digital preservation infrastructure that currently exists, the personnel and financial support that sustains the project work, and the steps being taken by project managers to improve the data curation infrastructure and workflow for data curation in the project.

The paper analyzes the current data curation workflows of the case studies, as mapped against the Digital Curation Centre Lifecycle, in order to propose a needs assessment of how humanities data curation can be realistically sustained by libraries.

The paper ultimately argues that if the preservation and curation of digital humanities projects is to critically involve libraries, it necessitates a transformation in the ways in libraries conceptualize collection development and they must develop a strategic integration of humanities data into library content management processes.

Session Leader

Harriett Green, University of Illinois at Urbana-Champaign

View the community reporting Google doc for this session.

Green compared three different systems: Walt Whitman Archive (Nebraska), Victorian Women Writers Project (Indiana) and Valley of the Shadow (Virginia). She reported on the types of data, the software versions, and more. Each one in order was increasingly advanced in terms of the level of data curation. There was an immense amount of work required not evident on the front end. It is necessary to map out the cost of the process from the beginning. Every decision faculty make impacts the work of high level data creation. It's impossible to have a standardized mass process for this as each project is different.

She advises a needs analysis of resources, personnel, and training needs prior to beginning such a project. Each one required time, resources, tools and infrastructure that were not available. This type of project needs special funding and support, as well as guidance for theoretical infrastructure. Building interoperability is another step, requiring more qualified personnel and more funding. Some of the requirements include a data scrubbing team, metadata work, programming, and project management

time and resources. A host of librarians needed training, or an influx of those with the skills needed was required.

For principles of data curation, she recommends the Sustaining Digital Scholarship report (levels 1-5) which covers everything from metadata creation through the preservation of the entire project. She recommends checking out Data Curation Profiles and the DCC curation lifecycle.

These efforts must be a collaborative and customized process.

Digital curation must be included in the evaluation rubric for digital humanities projects; recommended reading is the Mellon-funded "Supporting Digital Scholarship" final report. Digital humanities projects require long term planning, and consideration of digital curation as a core function of the library, with sufficient financial and other resources, as described by Walters and Skinner in the ALR report "New Roles for New Times".

Fedora 4: Updates from the Community

Session Type: Presentation

Session Description

Fedora 4's principle objectives are to provide a robust platform to support emerging data management needs, with native support for management and publishing via linked-data, in a package that is dramatically streamlined to deploy and support, and that preserves Fedora's traditional strengths (flexibility, extensibility, durability, community). This session will present Fedora 4's emerging architecture and functionality, highlighting the new capabilities of the system. As importantly, it will present Fedora's revitalized development, community and governance framework, with an invitation and opportunity for the digital library community to join and help shape this critical effort in these still early days.

Andrew Woods is the Fedora Tech lead. Version 4 will store content in RDF. The key capabilities will be: audit, authorization, content modeling, durable storage, large files, linked data, search, transactions, versioning – plus non-functional requirements. It will support batch operations, disseminators, metrics, multi-tenancy, and OAI as well.

The development process is use-case driven, and completely dependent upon the community. They have frequent releases and rely on acceptance testing. Developers volunteered from various institutions focus only on Fedora for 2-week sprints; what is done is impacted by skills, and institutional interests. It's a fairly agile process; every 2nd or 3rd sprint, there's a new release, and then they seek feedback.

As of August 2013 they have had 6 sprints, using 13 developers (from 10 institutions), 6 of whom were new to Fedora. On average they have 4 developers per sprint. They've also had 1 hacking event and are planning 2-3 more. If anyone wants to get involved, they're still in the early days of Fedora 4, so one can still have an impact in the direction development takes. Examples of institution-driven focus of development:

- UNCC focuses on authorization.
- UCSD on RDF and the persistence model.
- U of New South Wales on OAI.

There are multiple stakeholder opportunities: developers, advisors, acceptance testing, and development of use cases. If you contribute developer resources, you can be part of the advisory committee. Those interested in Fedora need to test the functionality as it's being developed. They need more community engagement. They hope to have a release in early next year, dependent upon an increase in community support.

Snapshots Group B

Forward to Libraries: Experiences Connecting Digital Libraries, Local Libraries, and Wikipedia

John Mark Ockerbloom, University of Pennsylvania

The Forward to Libraries service invites users researching a topic, author, or work online to discover related resources in local libraries, or other digital libraries and websites. It can take users to their favorite library's relevant offerings in a single click. In this session, I will describe ontological, technological, and cultural challenges I dealt with in implementing the service on The Online Books Page and Wikipedia, and how adoption has spread since its initial introduction. Following that will be an open discussion on how we can better connect our libraries, and attract inquisitive users to relevant resources in our collections.

John Mark Ockerbloom has developed open source code which, if used, will forward users from Wikipedia to your library; in Wikipedia it's described in "Forward to Libraries"… the service will link subjects of search to a search for local library content (presentation).

The Built Works Registry: Large-Scale Collaborative Data Curation

Margaret Smithglass, Columbia University

The Built Works Registry Project (BWR) seeks to establish unique identifiers for architectural works and the built environment. Funded by an IMLS National Leadership Grant, BWR is a collaborative effort between the Columbia University Avery Architectural and Fine Arts Library, ARTstor, and the Getty Research Institute. The process of developing technical infrastructure, curating/securing/ analyzing seed content, and developing geo-coding strategies highlight the myriad challenges and benefits of cross-

institutional collaboration. This session will present technical and data frameworks and the collaborative data curation model being developed in anticipation of the BWR launch in 2014.

Built Works Registry (BWR) focuses on collaborative large scale data curation. For content, they focused on "the built environment" – architectural structures. For these materials, there is no equivalent ISBN or ISSN, and no coordinated object identifier registry. This project seeks to construct that registry, with unique IDs and trusted data records. They are adding geolocations and records about the structures as well as variant names support.

Artstor uses the Shared Shelf image management system; BWR is a subset of Shared Shelf. It is not fee based at all, however. For content to be included, it must have an ID number, normalized name, and geolocation to be included. They will have a freestanding website with APIs for download and open participation. Already they have content from multiple locations, including Harvard and Cornell. 51 collections are being disambiguated and deduped now. Their goal is a registry of 100K items by launch in the fall of 2014. Their authority is coming from Getty Research Institute, and this effort will be part of the Getty Research Institute's planned Cultural Objects Name Authority (CONA).

Fedora Plus Curators Makes Quality Metadata

Ann Caldwell, Brown University Library
Joseph Rhoads, Brown University Library

Creating MODS records for uncataloged collections can lead to insufficient and sometime inadequate information. Collection curators are the logical ones to supply information. This Snapshot describes Fedora-based editor designed for non-technical users that allows editing of XML records, revision of those records and reingestion. It also describes the training process developed for curators.

Brown set up a shibboleth login for curators, assigning each collection to a curator to enable them to access the records. The curator can then see all the elements for the MODS and can add/modify/delete elements, subelements, attributes. Submitted changes go to metadata librarians for review. They can accept or reject the changes (they still need to build a "hold" so the metadata librarian can go talk with the curator). (From the demonstration, it appears this software only allows a one-by-one record access/ change.) This software is using JQuery XML editor from UNC, Solr, and Django.

Day Three

Wednesday, November 6

Digital in Context: Designing a digital program for the humanities

Session Type: Working Session

Session Description
Collectively, we have all been working on the digital aspects of libraries for well over 20 years. We have digitized and made accessible hundreds of millions of objects, collected and published faculty outputs through institutional repositories, and are increasingly working closely with faculty members on digital research projects which often involve visualizing data, and creative digital outputs from otherwise 'traditional' research.

Most of this work, however, has been reactive – designed in response to funding calls, private donors, or urgent needs such as responding to government guidelines. As a result, we have collected a very large and rapidly expanding set of project outputs—collections and websites—many of which are unsustainable. Now is a good time to step back from our activities and think programmatically. Looking just at the Humanities, this working session will be a focused, purpose-led discussion of how we go about building a coherent digital library program to support the Humanities. Using Oxford's Bodleian Libraries as one example of many – but with input from all participants welcome – the session will focus on the key characteristics needed to build a sustainable and coherent digital program to support the humanities.

Specific topics will include:
- *Characteristics of successful programs*
- *Practicalities of building programs*
- *The Evolution of a program*

Much of the session will be informal and focused on discussion, but the group will be tasked with coming up with real outputs in each of these areas. Working session attendees will, for example, be asked to consider how their work fits into the mission and strategic aims of their institutions and how these considerations can make for a stronger program.

Session Leader
Christine Madsen, Oxford University

View the community reporting Google doc for this session.

At Oxford, they have well over 20 years of experience creating digital stuff, resulting in hundreds of millions of things. Most of their work is "reactive" -- driven by funder and government mandates, donations, and grants. Results are scattered, unsustainable outputs, all over the board: websites, collections, data, and publications. But what are our services? How do we organize this to make it much more sustainable?

There are 104 libraries at Oxford: 40 college libraries, 1 Bodleian, and 30 Bodleiean as well, and various others. Over 11 million things, 34k linear meters of special collections (21.12 miles, over 111k feet). They have 526 FTE (full time employees), 38M pounds budget. They provide core electronic services to all these libraries, and currently manage about 10-15 active projects (not all digital library; many are digital humanities). They're currently building core digital infrastructure.

They work with faculty and librarians to build digital tools that help answer research questions, and seek to provide online, thorough, innovative access to library collections. Only 3% of the operational budget

goes to support digital library systems, and 90% of that goes to catalog support. They depend heavily on outside funding.

They focus facets on people, places, and locations using linked data. Thus if you look up Harlib, Samuel, 1600-1662, you can see the number of letters sent/received and in which he is mentioned in over time, etc. This enables them to do some interesting visualizations, mapping movement of people across Europe… and more. But they want to add more data, to complete the picture of what people were doing then.

Right now they have 9 services and 88 collections, running on 51 different platforms; this is not sustainable. Adding project after project does not create a program.

Why create a program? We need priorities due to limited resources, need to link activities directly to institutional mission and strategy, and need to justify our continued existence. We need to think forward and think big.

Why focus on the humanities? "For the humanities, the library is their laboratory." A program looks a lot like a digital library. We have built collections and resources, but not a digital library.

Recently, libraries have been about information (last 120 years). We have been working in support of information, and have forgotten about the people coming in and what they're doing with it. Libraries support knowledge creation. Libraries are where people take existing information and create new knowledge: this is scholarship (in business it's called innovation).

Libraries are a set of services that support knowledge creation. A digital program is a temporary arrangement to give content some meaning and order for others. A digital program will have characteristics, components, practicalities and problems.

She assumes the program must be digital, physical, and hybrid; that it must be an integrated approach to how people experience libraries.

Characteristics should include: flexible, inclusive, user focused, evolving.

Components include: systems, infrastructure, policy, services.

Two audiences she's thinking of: those looking from the inside out (librarians building and supporting this) and the outside looking in (scholars, administrators).

View from the inside

- Components: Infrastructure: reformatting, preservation, storage.
- Systems (tools): mining, curation, annotation, management, visualization.
- Policy: ownership, openness, retention, curation
- Services: digitization, discovery, preservation, access, management, curation

But in looking this over, it's clear there's something fundamentally missing…

View from the outside in

Can we map service, systems, and policies? This seems to be never ending. What's missing is: teaching and outreach. Who is our audience?

They are missing staff with particular skills, and ongoing research partners (not temporary soft money support).

Closing keynote

Early riser, devoted oceanite, and advocate of radical neutrality, ***Char Booth*** *explores the integration of pedagogy, research, technology, and design in libraries. Char is Head of Instruction Services and E-Learning Librarian at the Claremont Colleges Library, and is on the faculty of the ACRL Information Literacy Immersion Institute. Char blogs at info-mational and tweets @charbooth. Her publications include the Ilene F. Rockman Instruction Publication of the Year-winning Reflective Teaching, Effective Learning: Instructional Literacy for Library Educators (ALA Editions, 2011) and Informing Innovation: Tracking Student Interest in Emerging Library Technologies (ACRL, 2009).*

This presentation was on Information Privilege: Critical Approaches to Access and Advocacy. She spoke about the dark side of privilege, and noted that information privilege comes in many forms. The primary focus of this presentation was to say that we need more advocacy, outreach, and marketing.

Collection Assessment in a Collaborative Environment: BHL and DPLA

Session Type: Working Session

Session Description

The Biodiversity Heritage Library (BHL) and the Digital Public Library of America (DPLA) work collaboratively with their partners to make scientific and culturally significant resources openly available to the world. The BHL (with a core consortium of fifteen libraries and hundreds of contributors) and the DPLA (with seventeen partners, including the BHL, who themselves represent over 600 libraries, archives, museums, and historical societies) continue to grow within the US and internationally.

Challenges:
The BHL collection consists of more than 60,000 scanned full-text biodiversity-related book and journal titles with attached OCR and the DPLA aggregates nearly 3 million records from diverse catalogs, local databases, and multiple metadata formats. Because
of the variety of metadata, deep analysis of these data sets has proved to be a significant challenge.

Simple questions like "how many entomology or illustration records are there?" are difficult to answer as not all records include the exact term in either the subject heading or genre descriptions, or in fact, anywhere in the record.

Especially challenging for the BHL is that Library of Congress Subject Headings seldom connect directly with the biological taxonomies behind the literature rendering analysis of the collection's strengths and gaps incompatible with the needs of its core user group of biodiversity scholars. Likewise, the DPLA brings together subject headings, genre terms, and format descriptions from a variety of specialized thesauri and controlled vocabularies, both widely accepted standards and local constructs, challenging true subject analysis and usage beyond keyword search.

Needs:
Both BHL and the DPLA are interested in developing a detailed view of each collection's subject coverage to locate gaps to identify additional materials for digitization, and to connect with specific audiences and new partners that can fill these content voids. In addition, the DPLA is interested in automatically identifying the controlled vocabularies or thesauri for subject terms (or genres, formats, etc.) that come from our partners without an URI, pointer, or other indicator.

The two organizations are interested in discussing possible applications that might address some of these challenges:

- *Visualization tools to drill down from broad terms (e.g., Trees–North America) to more specific terms (e.g., Pinus banksiana) using thesauri specific to biodiversity and connecting them to the LCSH hierarchies.*
- *Vocabulary identification tool that might "lob" subject headings at open controlled vocabularies to associate terms and grab URIs*
- *Other ideas proposed by session attendees*
  The working session will start with brief presentations from BHL and DPLA representatives, followed by a discussion of common challenges. The last two hours will provide time for a mini "hackathon" to experiment with subject heading datasets and conceptualize prototypes for potential tools to satisfy collection assessment challenges highlighted by the discussion.

Session Leaders
Constance Rinaldo, Harvard University
Mark Phillips, University of North Texas

View the community reporting Google doc for this session.

Connie spoke about the Biodiversity Heritage Library (BHL) and Mark about the DPLA (Digital Public Library of America). BHL is a consortium of 15 natural history, botanical libraries and research institutions. They have an open data repository of taxonomic names and bibliographic information, open access, with full text for legacy biodiversity literature. It's an expanding global effort (primarily US

and the UK, but includes a few others as well). All the global nodes are just nodes; there's no legal entity.

Their core principles include delivering content where users are already working – via other biodiversity websites and taxonomic resources, and social media platforms. They involve users in selection and technical development activities: scanning locally, coordinating globally. They have an open data policy, and they provide data exports, APIs, OAI, and stable URLs.

Visualization software they use includes:

- JournalMap
- Better life index bit.ly/1c4…
- Altmetric
- Tableau
- Worth it (? unable to find more info)
- Visualizing article performance

For taxonomic manipulation of names, they depend on:

- Euler project
- Drupal taxonomy API

They need metadata reconciliation, gap analysis, and automation. The content is served via both Europeana and DPLA.

Mark then spoke about the University of North Texas and DPLA.

The Portal to Texas History is big state collaborative project, with about 325k items. It also incorporates lesson plans and such, with lots of outreach around K-12. UNT digital library is the traditional library, with ETDs, concert recordings, and other institutional repository content (about 100k items).

The Gateway to OK History is primarily newspapers; they're about to add lots of images in next 2 years. UNT provides their infrastructure and tools. UNT also manages the Texas Heritage Online search interface, aggregating collections into that. So UNT is working with a variety of partners, tools, and schemas, utilizing the funnel approach: many funnels into one.

They are working on identification and classification. But it's hairy; for example: names. How can they tell if a name is a personal name, event name, location name, or organization name? Different types of names need different management. Personal names sometimes are inverted, for example. They need more information about the name to know if they can apply generic rules like this. It's really hard to map this stuff at scale. The metadata is poor, as some of the vendors are creating metadata and scans for only 7 cents an image.

They worked with the 2008 End of Term Web Archive.

Questions with which they are struggling:

- How do we remove items we know we're not interested in? Or isolate the ones we are?
- How do we make the data work harder for us?
- How do we identify the areas of the collection that need the most attention?
- Can we programmatically identify candidates from large collections of documents that are most likely to be in scope with our collections?
- How can we leverage user interaction with our systems to provide hints on where we should focus our attention?

They're using name entity extraction software, GeoNames, Google and Bing API.

For name authority, they are relying on VIAF, LCNAF, Wikipedia, freebase, and local authority data.

Other tools they're using:

- Mallet for topic modeling
- Wea – machine learning
- Gephi – graph exploration
- NLTK (natural language tool kit)
- Hadoop
- R for stats and NLP (natural language processing).
- Leafletjs.com for mobile-friendly interactive maps
- OS js (cloud platform).
- D3 (data driven documents)
- Prototype.js
- Google chart API
- Open layers (maps for the web)
- Google fusion tables for visualization

He also recommends Stanford's Java library for natural language processing (it's 'pretty good') and talked about the UNT Name App they developed for reference, disambiguation, and storage of name records.

He recommends an exploration of how Europeana uses linked data to build a business case for implementation.

Conclusion

Multiple tensions were evident in this conference. UCSD's use of RDF triples for metadata storage, and then emulating Fedora calls in order to use a Hydra head, has stimulated a major shift in direction for Fedora development. Moving away from a "record" focus to "relationships" focus between items, concepts, dates, entities, and locations makes it possible to provide far better findability and browsing of content online. The Fedora 4 development, however, seems to be pressed for developer support. I

found it troubling that when I raised the question of what challenges Fedora3 users will experience in migrating to Fedora4, the response from the presenter seemed to indicate it might be difficult. This implies to me that Fedora support may wane.

Other approaches to providing RDF-type functionality include embedding semantic markup in web presentations, and extracting entities using natural language programming tools and large scale data software (like Hadoop), and then building services and visualizations using those (OCLC research, UNT, NCSU, and the Sheet Music Consortium).

There's increased focus on visualization of content, to assist users in finding, browsing, and using content, and a very real pressure to better market our content, and make it available to users wherever they may be on the web. The concern for return on investment was evident in the tremendous attendance for the assessment-focused working sessions (one of which was my own).

Another approach to better ROI was the use of students: creating ongoing developer teams manned by students, overseen by library programmers, as evidenced by UCLA and Oregon State U.

While some institutions are suggesting the use of Omeka for digital humanities and student engagement, voices from other institutions who have either managed many projects on demand, or studied the points of pain in managing digital humanities projects online recommend far more careful approaches.

Based on this experience, I recommend that we analyze our existing metadata for entities, relationships, locations, and dates which, once cleaned up, would enable us to provide visualizations, browse capabilities, and improved search, faceting, and retrieval, as well as the ability to easily support semantic web capabilities and web search engine support.