REPORT ON THE RESEARCH FORUM OF SAA 360°

2011 ANNUAL MEETING OF THE SOCIETY OF AMERICAN ARCHIVISTS

August 23, 2011

Jody L. DeRidder

The day-long research forum preceding the conference consisted of 20 presentations ranging in length from 10-30 minutes and 27 poster sessions. The content covered ranged extensively, from research in particular fields such as archeological record keeping to the impact of social activism on archivist employment. There was a heavy concentration of focus on management of digital content, including scientific data, born-digital contributions to archives, and long-term access. There were also some interesting investigations into leveraging outside input, not only for crowd-sourcing, but also for finding new ways to process archival materials. This report will focus on these areas, and include information from my own presentation on the usability of our mass-digitization methods, in order to address current and upcoming concerns for the University of Alabama Libraries.

INCOMING BORN-DIGITAL CONTENT

RECORDS MANAGEMENT AND ARCHIVISTS' TOOLKIT

Ed Busch (Michigan State University), archives@msu.edu

The Spartan Archive Project [1] is testing Archivists Toolkit (AT) for management of four large electronic record series in university archives. These include the catalog of academic programs, description of courses offered each semester, annual student directory, and schedule of courses.

During accession, permanent records are assigned code A#, scheduled records assigned R#. They generate two reports: scheduled destruction and destroyed. Transmittal forms are scanned and attached to the accession record in AT.

CDS IN A PAPER BAG: DEVELOPING DIGITAL PRESERVATION POLICIES AT AN ARTS COLLEGE

Annemarie Haar, Digital Archivist (Meyer Library, California College of the Arts), ahaar@cca.edu

[1] Spartan Archive Project: http://spartanarchive.wordpress.com/

Their college has a decentralized culture, with extensive duplication of content and inefficiencies. She was hired to create a centralized digital archive to manage content created by all departments. Haar performed a needs assessment of existing practices and expectations, and found many file formats. 75% of the faculty and staff used CDs and DVDs for storage; there were also many USB sticks and external hard drives, and some internal hard drives used for storage, some at home instead of at work.

She realized that she needed item-level description for the content. She involved stakeholders at top level of administration, as it was the only way to get the support they needed to move forward.

Haar also did an assessment of repository software, comparing 15 platforms (open source and commercial) against 30 criteria. She narrowed the choices down to EPrints[2], Resource Space[3], Equella[4] and Canto[5] (2 open source, 2 commercial). They chose Equella (a commercial product). She then did a programmatic assessment to select content for their pilot which would reflect their diverse needs. Her next steps are to load in the content and assess.

GRAMMARS AND PARSERS FOR VALIDATING BINARY FILE FORMATS

William Underwood (Georgia Tech Research Institute), william.underwood@gtri.gatech.edu

There are several validating software tools available to test digital files to ensure the file is compliant with the standards for that type: examples are Droid[6], JHOVE[7] and JHOVE2[8]. However, validation *assumes* that one has identified the file type. Very few types of files are validated by these tools. Unfortunately, the number of file types existing (especially older ones) are numerous.

Underwood used complier-compiler technology to create a scanner for file formats. After succeeding with text-based files, he wondered if it is possible to test binary file formats and identify them from their specifications. If so, can one create a parser? Then, can one use this technology to migrate binary files?

---

[2] Eprints Digital Repository Software: http://www.eprints.org/

[3] Resource Space Free and Open Source Digital Asset Management: http://www.resourcespace.org/

[4] Equella Digital Repository: http://www.equella.com/home.php

[5] Canto Digital Asset Management: http://www.canto.com/

[6] DROID (Digital Record Object Identification: http://sourceforge.net/projects/droid/

[7] JHOVE: JSTORE/Harvard Object Validation Environment: http://hul.harvard.edu/jhove/

[8] JHOVE2: The Next-Generation Architecture for Format-Aware Characterization: https://bitbucket.org/jhove2/main/wiki/Home

There are multiple kinds of binary files. Not only are there directory-based file formats but also chunk-based binary file formats (about 75 of them), IFF[9] for example. Chunks contain sub-chunks, data, and metadata. Started with interleaved bitmap (ILBM[10]).

His group thinks that an array of data types is a good method for representing structure of files and of data types. Using this, they manually constructed a binary file grammar, and then were able to generate parsers for IFF and WAVE[11] files (type of RIFF[12]).

Using ANTLR[13] parser generator, they successfully generated a parse tree. The next step is to develop binary file grammars for directory-based formats. For more information, a working paper is available at perpos.gtri.gatech.edu.

BUILDING DATA CATEGORIES AND TAXONOMY TO ORGANIZE TOPIC-SPECIFIC COLLECTION: TEXT MINING FOR NO GUN RI ARCHIVES

Donghee Sinn (University at Albany) dsinn@albany.edu

This presentation was about using text mining to provide organization of access to content. Typical descriptive metadata standards may not be useful for organizing materials based on topic. Sinn worked with the No Gun Ri Massacre collection (from during the Korea War in July 1950). This massacre was first reported in U.S. by the Associated Press in 1999. It was so difficult to organize that they decided to use text analysis.

The corpus tested contains many types of material: 31 archival collections, 23 academic publications, 55 journalistic publications, a government report and a web package. Only English materials were analyzed. The text analyzed included captions, citations, footnotes.

Sinn used TAPoR (Text Analysis Portal for Research[14]). She found top 20 words, top 10 word pairs, and top 10 word triplets. These provided recommended keywords/topics for organizing the content.

"No" is a stop word; this was problematic, given the name of the collection. Sinn created a taxonomy from 20 keywords, word pairs and triplets, resulting in 175 terms. She then divided the keywords by material type, and discovered that more generic terms were in academic content, and fewer in web documents. She then divided up the taxonomy into the areas representing history, research/controversy, political, and general, and created data

---

[9] .IFF File Extension: File Type Interchange File Format: http://www.fileinfo.com/extension/iff

[10] "ILBM" IFF Interleaved Bitmap: http://amigan.1emu.net/reg/ILBM.txt

[11] WAVE Audio File Format: http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml

[12] RIFF (Resource Interchange File Format): http://www.digitalpreservation.gov/formats/fdd/fdd000025.shtml

[13] ANTLR v3 (Another Toll for Language Recognition): http://www.antlr.org/

[14] TAPoR: text Analysis Portal for Research: http://tapor.mcmaster.ca/home.html

categories: people, place, time, activities, topic, genre, object, event, proper names. The amount of text available for analysis matters; keyword extraction is based on frequency. It is probably best to do an analysis by type of content, and then aggregate results.

THE RETURN OF LOST CONTENT: BORN-DIGITAL PROCESSING OF 5.25-INCH FLOPPY DISKS

Karen Ballinger (University of Texas at Austin)

Ballinger reported on University of Texas efforts to capture and preserve electronic records in legacy formats and on obsolete media[15]. They used an old Dell server with 6 hard drives, each of which had a different operating system. At first they didn't even have password, and the system went crashed often. They documented everything on wiki[16], in order to learn from mistakes.

They also used a device side data FC5025 floppy controller[17], and didn't realize at first that it needed a separate power source. Adam Goldberg (an alumni, currently working at Invodo, Inc.) was very helpful. Also the vintage computer community was very helpful in collaboration. They did a disk dump and used the dcfldd[18] command to get a hash value for each file, and also collected the Md5 checksums for each file. They used the disktype[19] command and the Sleuth Kit[20] open source digital forensics analysis tools to identify and analyze volume and file system data. They wrote a giant manual for the lab[21] and developed a functional work flow which included virus scan, bitstream copies, verification of checksums, and accessing the content if possible. After access, they extracted metadata extraction, and then put in the materials online in DSpace[22].

They found that audio and data CDs are encoded differently. 5.25" disks had much variation, and many copy protection problems (all documented on their wiki).

---

[15] Clark, Emily; Criswell, Tiffany; and Locano, Daniela. The George Sanger Game Projects Final Report, 2011: http://repositories.lib.utexas.edu/handle/2152/11318

[16] UT-iSchool Digital Archaeology Laboratory Documentation Collection: https://pacer.ischool.utexas.edu/handle/2081/21808

[17] Device Side Data's FC5025 USB 5.25" floppy controller. http://www.deviceside.com/fc5025.html

[18] Dcfldd: http://www.forensicswiki.org/wiki/Dcfldd

[19] Disktype: http://disktype.sourceforge.net/

[20] The Sleuth Kit: http://www.sleuthkit.org/sleuthkit/

[21] Ballinger, Karen; Cope, Bonnie; Petyak, Jocelyn, Meyerson, Jessica. The Digital Archeology Lab manual (May 2011): http://hdl.handle.net/2081/23283

[22] Published Videogames: https://pacer.ischool.utexas.edu/handle/2081/21815

The Curtis Riggs Zenith Data System Collection[23] was a zenith of conundrums. No matter what they tried, the retrieval would always stop on the 31st disk. Not all of the disks could be captured.

It's very important to plan out testing and be very detailed and organized. The disks were so dirty they were "gunking up" the read heads. Conservation of the actual media is needed.

The future of disk imaging is the ditto command[24], copying exact magnetic flux patterns on the media: the ones and zeros.

They are working with the University of Texas in Austin to package results of extracts and analyses in an XML container.

SPARTAN ARCHIVE: A PROGRAM IN TRANSITION

Cynthia Ghering (University Archives and Historical Collections, Michigan State University)

Ghering reported on a 3-year NHPRC grant collaborative project to develop an electronic records archive for born-digital records and publications[25]. Almost all the money went to information technology (IT) staff. They are studying born-digital records in the present environment, partnering with the Big 10 schools[26]. Their pilot project focused on 4 record series from the registrar's office, in digital form only. They found they have to rethink almost everything. It turned out that this was a hybrid collection (both analog and digital).

They are moving ALL materials from analog to a format-neutral repository. Who should pay for this? They have no virtual storage.

They are using the open source "Records Authority" software[27] created by Denver University, a result of an NHPRC-funded project. This software is for creating, managing, and distributing retention schedules. They are also using Fedora[28], iRODS[29], BagIt[30],

---

[23] Curtis Riggs Zenith Data System Collection: https://pacer.ischool.utexas.edu/handle/2081/21958

[24] Ditto: http://www.manpagez.com/man/1/ditto/

[25] Spartan Archive Project: http://archives.msu.edu/about/spartan_archive.php

[26] Illinois, Indiana, Iowa, Michigan, Michigan State, Minnesota, Nebraska, Northwestern, Ohio State, Penn State, Purdue, Wisconsin.

[27] University of Denver, Penrose Library. Records Authority: http://library.du.edu/site/about/urmp/recordsAuthority.php

[28] Fedora: http://fedoraproject.org/

[29] IRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems: https://www.irods.org

[30] California Digital Library. BagIt File Packaging Format: https://wiki.ucop.edu/display/Curation/BagIt

SIARD[31], ARK[32] and NOID[33]. For digital objects, they are using Islandora,[34] Droid, Jhove and Jhove2.

They created XML schemas for each record series. Clarifying the database structures and data fields for the incoming content was intense and not sustainable. They will be modifying the Fedora Content Model[35] to meet the needs of these record series.

LONG-TERM ACCESS SUPPORT

DOES TRUST MATTER?

Elizabeth Yakel, Ixchel Faniel, Nancy McGovern, Kathleen Fear, Morgan Daniels, Adam Kriesberg (DIPIR, School of Information, University of Michigan)

This group was studying the use of research data from Trusted Digital Repositories (as defined by ISO/TRAC – Audit and Certification of Trusted Digital Repositories[36]); in particular, they studied users of digital content from ICPSR[37], which is part of a bigger project called DIPIR[38] (funded by IMLS). Multiple partners are involved.

One question they raised is: how can contextual information be created and preserved? Their goals were two-fold:

- Bridge the gap between data reuse and digital curation research
- Determine whether reuse and curation practices can be generalized across disciplines

---

[31] Swiss Federal Archives. Archiving of Databases: SIARD Suite: http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en

[32] KDE Utilities. Ark: a program for managing various archiving formats within the KDE environment: http://utils.kde.org/projects/ark/

[33] California Digital Library. NOID: Nice Opaque Identifier (Minter and Name Resolver): https://wiki.ucop.edu/display/Curation/NOID

[34] University of Prince Edward Island. Islandora: building a rich digital repository ecosystem: http://islandora.ca/about

[35] Fedora Commons. The Fedora Digital Object Model, Fedora release 3.0 Beta 1: http://fedora-commons.org/documentation/3.0b1/userdocs/digitalobjects/objectModel.html

[36] Center for Research Libraries. "Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)": http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0

[37] Inter-University Consortium for Political and Social Research (ICPSR): http://www.icpsr.umich.edu/icpsrweb/ICPSR/

[38] DIPIR: Dissemination Information Packages (DIPS) for Information Reuse: http://dipir.org/

What are the significant properties that facilitate reuse? How can we map those as representation information?

There are 3 phases to the project, which lasts from Oct. 2010 – Sep. 2013.
They did an ICPSR survey and have already performed literature reviews. Digital curation literature suggests that the significant properties of scientific data are based on functionality, relationships and appearance. Data reuse literature, however, is focused on the information needed to help users understand the data, and whom and what to trust.

The project group is mapping TDR (Trusted Digital Repository) attributes in TRAC to the concepts. They found that the decisions to reuse are NOT determined by metadata alone. Characteristics that were important to users included trust of the repository and of the data, the delivery context, and the user needs.

The ISO TRAC areas of focus are:

- Organizational infrastructure
- Digital object management
- Infrastructure and Security Risk Management

Significant properties of data include:

- Data context: making data understandable
- Delivery context: making data discoverable and identifiable
- Repository context: found that good reputation of the repository may be based on value of actual dataset as opposed to methods of data management
- User context: Identify designated community and assign knowledge base

Their next steps are to determine what contributes to decision for reuse of scientific data. There are four parameters to weigh:

- Trust in the data (repository context)
- Relevance of the data (user context)
- Quality of the data (data context)
- Ease of use (delivery context)

They will be focusing on trust of 3 sources:

- Data producer
- Repository
- 3rd party endorsing reuse of data (such as Data Seal of Approval[39], faculty member making a recommendation, etc.)

They are going to survey the last 2-3 years of those who published using ICPSR's content. They expect to find that for novice users, the most important measure is trust of the 3rd party endorsing the reuse of data; that for experienced researchers, the trust is more likely to be in the relevance and quality of the data.

BEYOND PRESERVATION TO TRUST: TOWARD AN APPLICATION PROFILE FOR IDENTITY AND INTEGRITY METADATA IN UIRS

[39] Data Seal of Approval: http://www.datasealofapproval.org/

Corinne Rogers (School of Library, Archival and Information Studies, University of British Columbia)

The goal of this project is to develop an application profile grounded in the findings of the InterPARES project[40] that connects theory with functional requirements for authenticity metadata, then to test it in cIRcle[41] (the institutional repository at the University of British Columbia).

The metadata will be both human and machine-readable annotation. This is part of InterPARES, the International Research on Permanent Authentic Records in Electronic Records 3[42].

They are working with the City of Vancouver Archives and Archivematica[43]. Parsimony is one of their functional requirements: only what is necessary and sufficient. Records may be computer generated. They are creating a chain of preservation (COP) model to be captured in PREmis[44] and Dublin Core.

In their view, authenticity is equivalent to the combination of identity and integrity.

LEVELS OF REPRESENTATION IN DIGITAL COLLECTIONS: A FRAMEWORK AND IMPLICATIONS FOR ARCHIVAL RESEARCH

Christopher (Cal) Lee (School of Information and Library Science, University of North Carolina at Chapel Hill): callee@email.unc.edu

We think about computer systems in layers. Manipulation of symbols does not imply understanding. Because meaningful information can reside at all levels, the archivist must decide at what levels to preserve information so as to reflect the content accurately. There is a "lifting problem:" we lift things out of context, so we must preserve the correct amount of contextual information. Fundamental preservation is equivalent to ensuring the conveyance of meaning over time. Digital preservation means we can consistently reproduce it over time within an acceptable range of variability.

Digital materials reside in various levels of information. The levels of representation include:

- aggregations
- objects
- in-application rendering

---

[40] InterPARES Project: International Research on Permanent Authentic Records in Electronic Systems: http://www.interpares.org/

[41] The University of British Columbia. cIRcle: UBC's Information Repository: https://circle.ubc.ca/

[42] InterPARES 3 Project: http://www.interpares.org/ip3/ip3_index.cfm

[43] Archivematica Open Archival Information System: http://archivematica.org

[44] PREMIS: Preservation Metadata Maintenance Activity: http://www.loc.gov/standards/premis/

- file as viewed through the file system
- file viewed as a raw bitstream
- sub-file data structure
- the bitstream through input/output equipment
- the bitstream as encoded on the physical medium (ones and zeros)

The vectors of interest that form the relationship between stakeholders and these information layers include the abilities to control, access, and destroy. These vectors are not mutually exclusive.

SCIENTIFIC DATA

MANAGING SHARED DIGITAL RESEARCH DATA IN FEDERATED STORAGE CLOUDS FOR HIGHER EDUCATION

Richard Marciano (University of North Carolina)

They deployed a federated data infrastructure across 4 partners: Duke, UNC, NC State and RENCI[45] called the TUCASI data infrastructure Project (TIP).[46] This was a 2-year project that ended in June 2011, funded by the Triangle Universities Center for Advanced Studies, Inc. It leverages other funding too: NSF, NARA, IMLS, etc. They used iRODS, TRLN[47], and Shibboleth[48]. In their project, only one copy was made of each other's content, in a round-robin style.

They determined that establishment of data policies are crucial:
1) Cross site and inter-institutional
2) Data access and modification policies
3) Preservation and curation (data lifecycle evolution)

Applicants for funding from NSF, NIH and NEH granting agencies must now address data management:
1) Required data management policy
2) Data management support
3) Institutional support

In their project, they had researcher/technologists and librarian/archivists working together, and determined that adequate personnel support is essential. They spent $2.5 million on hardware, but did not fund personnel: this was a mistake.

---

[45] RENCI: RENaissance Computing Institute: http://www.renci.org/

[46] Triangle Universities Center for Advanced Studies, Inc. (TUCASI) data-Infrastructure Project (TIP): http://www.renci.org/focus-areas/project-archive/tucasi

[47] Triangle Research Libraries Network (TRLN): http://www.trln.org/

[48] Shibboleth: http://shibboleth.internet2.edu/

DEMISTIFYING THE DATA INTERVIEW

Jake Carlson and Eugenia Kim (Purdue University Libraries) [poster]

A clear definition of data is essential for setting the context of the interview. Their selected definition is from OMB Circular A-110[49]: "Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings."

Curators need a reference guide for the data lifecycle. Each researcher may have their own definition of data lifecycle that may not apply to other research, and some data may not require curation. Their selected approach was a broad categorization of the following stages: Raw, Processed, Analyzed, and Published.

THE CREATION AND MANAGEMENT OF RESEARCH DATA SETS AND LABORATORY NOTEBOOKS IN UNIVERSITY LABORATORIES OF SOUTH KOREA

Jihyun Kim (Department of Library and Information Science, Ewha Womans University) [poster]

They found much resistance among researchers to adopting Electronic Lab Notebooks (ELN). Researchers preferred to write by hand. Learning to use the ELN reduced efficiency of work. It was difficult to bring laptops into the research environments and type there.

To the researchers interviewed, preservation of primary records meant keeping everything, organized by date of creation and creator's name.

PARTNERSHIP BUILDING IN THE SOCIAL SCIENCE DATA COMMUNITY

Peter Granda and Jared Lyle (ICPSR) [poster]

They are following the framework of Myron Gutmann and Ann Green: "Building Partnerships Among Social Science Researchers, Institutional-Based Repositories and Domain-Specific Data Archives."[50] Most data accesses are about 50 MB or less, but they are looking into streaming video for access. They have not needed to provide DVDs or hard drive transfers.

Initial discussions with researchers covered these topics:

- intellectual property rights issues

---

[49] The White House. Office of Management and Budget: Circular A-110 Revised 11/19/93 As Further Amended 9/30/99: http://www.whitehouse.gov/omb/circulars_a110

[50] Green, Ann and Gutmann, Myron P. 2006. "Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain-Specific Data Archives." http://hdl.handle.net/2027.42/41214

- long term preservation planning
- access controls
- confidentiality
- file format options
- metadata standards

Access to and stewardship of data and metadata over the long term are the responsibility of the domain repository. This project seeks to provide a shared interface to access the stored data of multiple repositories.

They printed a "Guide to Social Science Data Preparation and Archiving"[51] and are developing data citation standards.

COPYRIGHT ISSUES

HOW ARCHIVISTS VIEW COPYRIGHT

Jean Dryden (College of Information Studies, University of Maryland)

Dryden did both a survey and interviews. She found that archivists are confident in their understanding of copyright, but their actual knowledge is questionable. 50% said it's up to the users to figure out copyright, though most think it's their job to educate (the survey and interview results disagree).

Archivists are ambivalent in the assessment of risk. 40% are willing to take a risk rather than strictly comply with copyright.

INVOLVING OTHERS IN CREATING CHANGE

EVERYONE A CURATOR: EVALUATING THE IMPACT OF "SOCIAL METADATA" ON LIBRARIES, ARCHIVES AND MUSEUMS

Helice Koffler (Manuscripts & Special Collections, University of Washington Libraries)

Koffler spoke on behalf of the RLG (Research Library Group) Social Metadata Working Group;[52] members consisted of 21 partner staff from 5 countries (began in 2009). They worked with OCLC and the University of Australia to evaluate the impact of crowdsourcing technologies on libraries, archives and museums. Their charge:

1) Leverage user contributions to metadata
2) Address issues at network level

They have generated 3 separate reports.

---

[51] Inter-university Consortium for Political and Social Research. 2005. "Guide to Social Science Data Preparation and Archiving." http://hdl.handle.net/2027.42/61289

[52] OCLC. The RLG Partners Social Metadata Working Group: http://www.oclc.org/research/activities/aggregating/group.htm

1) Environmental scan
2) Analysis of site manager survey
3) Recommendations

The first one will be released Sept. 2011.
An example of moderation of incoming metadata can be found in the Plateau People's Web Portal[53]. Registered users can add tags and annotations; tribal council members review and manage the input content.

In their survey, almost half of the sites incorporating social media contribute content to the site, not just descriptive metadata. One example of this is the Australian Newspaper project[54]; another is the University of Michigan Islamic Texts[55].

Only 2 sites said spam was much of a problem. Abusive behavior was sporadic. It helps to make user activity visible and to reward their effort with recognition.

[More information was found online in a document authored by Smith-Yoshimura of OCLC Research[56]]

The site manager's survey found that the highest priorities are building communications and increasing traffic; also 60% wanted to use the contributed metadata to improve description.

Recommendations, in short are these:

1. Go ahead and do it!
2. Consider how to integrate the metadata.

COLLABORATIVE CREATIVITY: THE RADCLIFFE WORKSHOP ON TECHNOLOGY AND ARCHIVAL PROCESSING

Mary O. Murphy, Manuscripts Archivist at Harvard (Radcliffe Institute) and Anne Sauer (Digital Collections and Archives, Tufts University)

[53] Plateau Peoples' Web Portal: http://plateauportal.wsulibs.wsu.edu/html/ppp/index.php

[54] National Library of Australia. Australian Newspapers Digitisation Program: http://www.nla.gov.au/ndp/

[55] University of Michigan, MLibrary. "Collaboration in Cataloging: Islamic Manuscripts at Michigan": http://www.lib.umich.edu/special-collections-library/clir-islamic-manuscripts-project

[56] Smith-Yoshimura, Karen. 2010. "Social Metadata for Libraries, Archives, and Museums." Presented at the 2010 DLF Fall Forum: http://www.diglib.org/wp-content/uploads/2011/01/SocialMetadataforLAMs.pdf

Marilyn Dunn[57] wanted to bring together professionals from multiple fields to find ways to provide maximum access to archival content sustainably. They needed new approaches to defeat backlogs and support digital processing. The planning committee for the workshop was drawn from Harvard, MIT, Tufts, JFK Library and the Massachusetts Historical Society.

45 participants came from 20 institutions:

- 51% from technical/design sectors
- 42% from libraries and archives
- 8% other sectors

The rules were:

- Brainstorm out loud and in group.
- NO whining.
- No "yes, but" – don't tell us it can't be done.

Communications were assisted by distributing use cases ahead of time, based on actual collections and problems. Two narrative scenarios were developed – one for each day of discussion. The first day focused on maximizing discovery, and the second on streamlining processing. The workshop was held May 16-17, 2011, and began with a multi-institutional display of collections by their archivists, to familiarize non-archivists with the types of content and purpose of the work. During the workshop, participants worked in groups of 10; each group had a table leader and a reporter. Cliff Lynch did the wrap up session.

The ideas that emerged focused on outcomes, on the product, not the process. Most of the technologists proposed digitization of everything before arrangement and description.

Questions raised about this include:

- How?
- By whom?
- And at what level of quality? (Just for staff access?)

One suggestion was that early processing notes could perhaps be verbalized and recorded, then transcribed using voice-recognition software.

One strong suggestion was to have an assembly-line approach to the work, matching skills to the job. For example, archivists should not be refoldering; students can do that. Make sure archivists make best use of their time. A Google Books representative was there, and his comments were a big eye-opener.

Workshop suggestions included that they use automation for:

- Confidentiality filters (use a spam filter and configure it to recognize such patterns as social security numbers and drivers licenses)
- Name recognition (via clustering and ranking of OCR'd digital files)
- Speech recognition (of processing notes)

---

[57] Radcliffe Institute for Advanced Study, Harvard University. Marilyn Dunn, Executive Director of the Arthur and Elizabeth Schlesinger Library on the History of Women in American and Librarian of the Radcliffe Institute. [Available from http://www.radcliffe.edu/about/leaders_dunn.aspx]

- File numbering (to tie item to original location in the box)
- Analog to digital association
- Crowdsourcing (to capture metadata)

The risks of completely changing their approach to incorporate these suggestions include that the archivists:

- Could fall behind even more
- Could appear inefficient or ineffective
- Could lose resources
- Could frustrate users, donors and themselves.

The question was raised: is incremental change enough? It is daunting to completely change the workflow all at once.

The archivists' next steps are these:

- Continue the conversation
- Test the workshop ideas
- Disseminate information to staff
- Get coaching from experts
- Integrate sustainable methods.

PROVIDING ACCESS TO DIGITIZED CONTENT VIA THE FINDING AID: A USABILITY STUDY

Jody DeRidder (University of Alabama Libraries)

In a grant project partially funded by the NHPRC, the University of Alabama Libraries developed a low-cost model and supporting open-source software for implementation. This approach enables low-cost digitization of even large manuscript collections, providing online access to material that otherwise may never be digitized. The grant project included a usability test which compared the resulting interface to a similar collection delivered with item-level descriptions accessed outside the finding aid. At an estimated cost of 79.5 cents per page, our mass-digitization method costs less than a third of our usual item-level description access.

They sought to measure efficiency, effectiveness, satisfaction and the critical element of learnability.

Efficiency was defined by the two measures of time on task and total number of steps (clicks) required for a participant to successfully complete a task. Effectiveness was measured by whether or not a task was completed successfully.

They measured satisfaction by both the users' overall perceived difficulty of each interface (on a 1-5 scale), as well as the total number of positive versus negative comments about an interface recorded during testing. For learnability, they examined improvements from task 1 to 4 for both interfaces, in the following variables: time to first click (which may indicate indecision), total time to locate content, number of steps, and success in completing a task.

Experienced researchers have already been shown to prefer the finding aid as interface[58] [59]. Both Scheir[60] and Chapman[61] found that novice users experienced a learning curve during exposure to finding aids, gaining confidence and ease with time. For our study, participants were primarily novice users.

The test consisted of 4 known-item searches, repeated for each of two similar collections accessible in different ways. The item-level described digitized items in the Jemison collection could be searched using a search box, while the Cabaniss items were accessible via links embedded in the EAD finding aid.

Overall, participants required an average of 35% less time and 48% fewer interactions with the item-level described collection than with the finding aid as web interface. This indication of reduced efficiency is to be expected without a search option in the finding aid page, as results may only be obtained via browsing. Success rates via the item-level search interface were 7.5% higher.

Participants as a whole clearly prefer the item-level interface, by a factor of 3:1, though 40% of those with a background in history, and a third of those with special collections experience or without digital library experience preferred the finding aid interface. Marked differences in efficiency were evident in both interfaces for participants for whom English is a second language, but the difficulty was more pronounced in the finding aid interface (51% more time and 10% less success in the item-level interface; 41% more time and 13% less success in the finding aid interface). 80% of the participants for whom English is a second language preferred the item-level interface.

Interestingly, those without previous digital collection experience found the finding aid interface significantly easier than those who claimed familiarity with the more traditional digital library interface. The EAD interface took them 42% less time, 27% fewer clicks, and provided 12% more success.

---

[58] Cory Nimer and J. Gordon Daines III, "What Do You Mean It Doesn't Make Sense? Redesigning Finding Aids from the User's Perspective," *Journal of Archival Organization* 6, no. 4 (2008), http://dx.doi.org/10.1080/15332740802533214

[59] Tim West, Kirill Fesenko, and Laura Clark Brown, "Extending the Reach of Southern Sources: Proceeding to Large-Scale Digitization of Manuscript Collections," Final Grant Report for the Andrew W. Mellon Foundation, *Southern Historical Collection, University Library, University of North Carolina at Chapel Hill*, June 2009, http://www.lib.unc.edu/mss/archivalmassdigitization/download/extending_the_reach.pdf

[60] Wendy Scheir, "First Entry: Report on a Qualitative Exploratory Study of Novice User Experience with Online Finding Aids," *Journal of Archival Organization* 3, no. 4 (2006), http://dx.doi.org/10.1300/J201v03n04_04

[61] Joyce Celeste Chapman, "Observing Users: An Empirical Analysis of User Interaction with Online Finding Aids," *Journal of Archival Organization* 8, no. 1 (2010) http://dx.doi.org/10.1080/15332748.2010.484361

This bodes well for future acceptance of this method of web delivery, and corroborates Chapman's findings that "groups that showed the most significant improvement over time were novice participants and Internet users with a beginning proficiency level."[62]
If indeed an interface is more learnable, we would expect to see statistically significant improvements in effectiveness and efficiency from task 1 to 4 for all users in each interface separately. Although improvement from task 1 to 4 did occur 62.5% of the time in favor of Jemison, these differences were not statistically significant.

More tests are needed on the finding aid interface to determine what actually helps users. Suggestions from the research[63] include:

- replacing archival terminology
- Providing search in page feature
- Providing navigation links for sections of the finding aid on the left

This should be followed by learnability tests for novice users that span multiple sessions[64].

Finding aids present digital materials in the context of the collection, and hence provide far more information to be sifted than content described solely on the item level. Efficiency and effectiveness measures should not be applied in comparing the EAD interface with item-described content. The result is a comparison of apples and oranges.

What is truly at issue here is learnability, particularly for novice users and those for whom English is a second language. Modifications to the display and terminology should be tested to verify that these changes increase access and learnability.

By increasing the ease of use and verifying the learnability of the finding aid interface, we will be better positioned to leverage this low-cost digitization method to provide online access to large manuscript collections.

---

[62] Joyce Celeste Chapman, "Observing Users," *Ibid*.

[63] Elizabeth Yakel, "Encoded Archival Description: Are Finding Aids Boundary Spanners or Barriers for Users?" *Journal of Archival Organization* 2, no. 1 & 2 (2004), http://dx.doi.org/10.1300/J201v02n01_06.

[64] Tom Tullis and Bill Albert, *Measuring the User Expereience: Collecting, Analyzing, and Presenting Usability Metrics* (Burlington, MA: Morgan Kaufmann, 2008, 92-94.