REPORT ON SAA 360°
2011 ANNUAL MEETING OF THE SOCIETY OF AMERICAN ARCHIVISTS (SAA)

August 22-27, 2011
Jody L. DeRidder

The first day of this conference only applied to the SAA Council; the second day included various committee meetings and the all-day research forum. The morning of the third day contained orientations, tours, an examination for certification, committees and working groups, and the afternoon was filled with a host of roundtable meetings and forums up until 9 pm. The first plenary was not until the fourth day (August 25th), which meant the primary sessions (70 of them!) for the conference were held on the last 3 days, from 8 am to 9:30 at night (ending Saturday at 4 pm). All but the plenary sessions were joint presentations, with each presenter given about 10-30 minutes each, so a tremendous amount of information was packed into a very short time. Estimates of attendance for this conference ran to 1300; I counted about 1200 people at the first plenary session. This is clearly the single most important gathering for archivists in the United States.

I attended (and presented at) the roundtable meeting for Metadata and Digital Objects (I'm on the steering committee), which included multiple presentations, and also the round table meeting for Archivists Toolkit[1], since we have this in use. The sessions I attended included the plenaries, efforts to leverage controlled vocabulary to improve access, updates on Archivists Toolkit and Archivematica, long term access models, and anything related to managing digital content, whether found on old media in the archives or captured today. There was a tremendous interest at this conference in efforts to effectively capture and manage born-digital content for long-term access. For example, the "Practical Approaches to Born-Digital Records: What Works Today" panel[2] drew about 450 attendees, with standing room only.

This report will focus on these topics, with presentations reordered for coherence, in an effort to address current and upcoming concerns for the University of Alabama Libraries.

Recordings of many of these sessions are available for purchase online.[3]

---

[1] "Archivists Toolkit: for archivists by archivists." (Available from http://www.archiviststoolkit.org/)

[2] Christopher J. Prom, Suzanne Belovari, Melissa Salrin, Laura L. Carroll, Benjamin Goldman, and Seth Shaw. 2011. "Practical Approaches to Electronic Records: What Works Now." Presented at the Society of American Archivists annual meeting, August 26, 2011, Chicago Ill. (Available from http://e-records.chrisprom.com/wp-content/uploads/2011/08/PAERWhatWorksNow.pdf)

[3] Convention Recordings International, Inc. "2011, august 24-27, SAA, Society of American Archivists Conference." (Available from http://www.conventionrecordings.com/catalogs/index.asp?category=339326&count=1)

TABLE OF CONTENTS

PLENARIES

"THEN AND NOW: WOW!"

Dr. Helen Tibbo, President of the Society of American Archivists and Professor in SLIS, UNC Chapel Hill; Scott Simon, National Public Radio (NPR)

Estimated attendance: 1200 people.

The Nazis banned Jazz and Swing music. "Charlie and His Orchestra" took Jazz and Swing music hit tunes of the day and added propaganda to them, then broadcast them in Germany. They had a large audience who listened secretly late at night. The shows were live, with no thought given to archiving. Germans recorded the performances, thinking that the music hid secret information, but most recordings were made by performers themselves after the fact. Only 22 song recordings of this content, out of hundreds, survive today. Why? A post-war German law (still existing) made it illegal to possess Nazi propaganda. The bulk of those recordings were destroyed. The result is a net loss for scholars.

There are a number of forces that work against total preservation. Things even disappear off the CNN (Cable News Network) website and are not retained. If you want to hear something recorded in the 1980's the reels need to be baked. Now we haven't enough space at NPR (National Public Radio) to store CDs. Now, due to space considerations, content has to be purged.

Dr. Tibbo referred to an original interview with a freshman senator from Illinois (who is now president) that was not retained, except for the clips used on the air. That information would be of value today, but who could have predicted it?

There is too much information to store it all. Police now record conversations with citizens. But where does all this info go? Who will manage it? We need a new working definition of what level of information is important enough to collect, if not to keep.

It is not helpful to bury time capsules in the ground – the ground moves. Technologies keep changing. How many generations of formats and software will occur between storage and access? When the time capsule is opened, we may be unable to access the content.

East German Secret Service (Stasi[4]) information was recorded, kept illegally, and is still being revealed. That regime closed 22 years ago. Would destroying the files be morally better? That information gained against consent. Will our archives be overwhelmed by information?

Dr. Tibbo believes in the power of history to sort that out. Clearly she thinks we should store as much as we can possibly manage, but did not address where she thinks those boundaries should be drawn.

[4] Britannica Academic Edition. 2011. "Stasi." (Available from http://www.britannica.com/EBchecked/topic/563751/Stasi)

ON THE OCCASION OF SAA'S DIAMOND JUBILEE: A PROFESSION COMING OF AGE IN THE DIGITAL ERA

David Ferriero (Archivist of the United States) and Dr. Helen Tibbo

Estimated attendance: 900 people.

Ferriero spoke first. He was originally from Duke. He developed New York Public Library's digital strategy, and celebrates the close relationship between SAA and NARA (National Archives and Records Administration) since Roosevelt's administration.

According to Ferriero, the first digital challenge is quantity. They have 20 million emails from the Clinton administration, 240 million from the George Bush era, and already 4 TB of video from the Obama administration.

The second digital challenge is the dizzying array of formats. Laws evolve slowly, while formats change quickly; archivists are required to operate under outmoded rules.

Social media combines size and format challenges and puts them on steroids. For example, in less than 9 months, Facebook added 9 million users. They have a 47% market penetration, with 1 trillion page views in June alone, and 87 million users. Who can keep track of all that content, much less determine what should be archived?

The College Park facility is outmoded. They are developing methods to preserve any kind of electronic documents. The current administration of National Archives believes in using digital media to provide access to content. They are creating guidelines for social media content: if it is unique, reflects policies, is used by agency in work, or if is there a business need for the information, then the digital content is captured.

Ferriero's team handles 34 Facebook pages for niche interests to feature National Archives content, and have a presence on Flikr and YouTube with wide audiences. They now even have a Wikipedia staff member.

In Wikimedia commons, the National Archives have 12 million viewers on a single day; if the content were only available in analog versions, 1000 people at best would have seen it. Citizens have a right to see our documents. The National Archives are under major transformation right now to provide better service and more open access and transparency.

The National Archives depends upon a flow of ideas and input from SAA. This is not an easy time for funding agencies. For example, the NHPRC (National Historical Publications and Records Commission) was cut down to only 1 million dollars for 2012. The Office of Management and Budget has told them that 2013 budgets will be up to 15% lower than that.

Jane Kenemore then introduced Dr. Helen Tibbo, and referred to her 60 page curriculum vitae. Dr. Tibbo received her Ph.D. in Maryland. She has raised over 7 million dollars to research the use of digital records. She recognizes the gap between researchers like herself and archivists, helping to establish the Digital Research Exchange to improve communications. She joined SAA in 1985, and has brought digital curation education to the

SAA programs, including a new certificate for training, which has just been released [described later].

Dr. Helen Tibbo notes that this is the SAA's 75th year. Most work is done by members. This has been a trying year for all of us. Personal finances and retirement funding are tumbling; coworkers and staff members are losing their jobs. Some SAA members have dropped out and are seeking options to reenter the work force. Some archives and special collections have lost more than 50% of their employees in the past 2-3 years. There are funding cuts at the National Archives and at granting agencies.

How do we move forward? Tibbo states that the past and future are in each of us and in all of us. We are the bridge from the profession's past to the profession's future. This is a turning point for decades to come. We are coming of age in the digital era.

We are now faced with managing and preserving a flood of digital records. We will leverage the work done by pioneers in the area from the past few years. It is time to take the foundations of our profession and apply them to digital content. For archivists to transform the world, we must transform ourselves, including our education. The world has changed dramatically. There will always be archival records and manuscripts, but they have and will continue to change in form and extent.

In 2002, total digital capacity in the world overcame that of analog material capacity. In 2007, 90% of our memory was in digital form: over 290 exabytes of information was stored at that point. Far more was sent through communications devices. Content has outstripped storage capacity.

The amount of digital content has grown by a factor of 9 in 5 years, and is expected to grow by a factor of 8 in the next year alone. In contrast, the pool of IT (Information Technology) staff to manage these materials will only grow slightly in the coming years. 99% of information is now born digitally, though most of our archives are currently still analog.

41 years ago, NARA accessioned their first digital records, but little progress was made in the profession at large. In 1989, "Camp Pitt," a colloquium of archivists sponsored first by the Council on Library Resources (CLIR) and then by the NHPRC, determined that digital archiving was both a pressing and a complicated issue for archives. Indeed, it is the greatest challenge in decades. Profound shifts in the creation and management of information renders obsolete some skills in institutions while demanding many new ones.

The Ollie North era and attendant litigation brought to the attention of the country how necessary digital preservation technologies and techniques are. Much progress has been made in the past 15 years. Key projects have addressed how to develop the cyber-infrastructure necessary to manage content. DataNet[5] and NDIIPP[6] have been established. There is much for archivists to do, and much to learn.

---

[5] National Science Foundation, Office of Cyberinfrastructure. 2010. "Sustainable Data Preservation and Access Network Partners (DataNet)." (Available from http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141)

There is a pressing requirement for training. Two studies published last year detail the continued challenges. Education and digital preservation methods must be established. Half of all archival collections have no online presence. Management of born-digital records is in its infancy. Funding has been reduced, and business as usual is impossible.
Lack of space, the influx of born digital material, and digitization are the 3 main problems.

We have a lack of infrastructure for managing born digital materials; and also a lack of expertise and time for planning. Many archivists don't even know what they have. Only half of our institutions have assigned anyone to handle these materials; almost none have collected at scale.

In 2005, a survey showed that almost no one had electronic records programs.
In 2009, 65% responded again: there was almost no change.

A reluctance to take technical courses extends the problem. We suffer from a lack of continuing education. Management and preservation of electronic documents must be taught by graduate programs; this is essential.

How do we as an organization address this?

In the past 18 months, SAA has updated their Guidelines for a Graduate Program in Archival Studies (GPAS), to provide parameters for training new archivists. But GPAs are only guidelines, they cannot be enforced. They do raise the minimum expectations for archival studies programs, and are often used as ammunition obtaining funding and in updating programs.

Almost all records created today are in digital format. There is a clear and urgent need for continuing education for managing digital archives. The SAA strategic plan recognizes this in its goal "to access, manage, and preserve records in all formats." We must develop methods for management from accession through preservation.

The Digital Archives Continuing Education (DACE) task force has just released a certificate and program for continuing education (Digital Archives Specialist (DAS) Curriculum and Certificate Program[7]). There are 4 tiers of studies in the program: "Foundational" courses, "Tactical and Strategic" courses, a "Tools and Services" course, and one "Transformational" course.

Some current SAA courses are right on target; others will be tweaked, and new courses added as needed. Target audiences include administrators as well as workers.

Increasingly manuscript collections contain large digital components. We have a growing number of educational opportunities. Dr. Cal Lee and Dr. Tibbo teach in one of those

---

[6] The Library of Congress. 2011. "Digital Preservation. National Digital Information Infrastructure & Preservation Program: A Collaborative Initiative of the Library of Congress." (Available from http://digitalpreservation.gov/)

[7] Society of American Archivists. 2011. "Digital Archives Specialist (DAS) Curriculum and Certificate Program." (Available from http://www2.archivists.org/prof-education/das)

models at UNC (University of North Carolina, Chapel Hill). SAA is now launching a full-scale effort.

The 4 challenges we as archivists face:

1. Education
2. Planning
3. Securing support
4. Just making it work.

Dr. Tibbo urges all archivists and curators: do something significant before next year's conference to expand your knowledge. Do an internship, take a graduate course, or do research and share your findings. Continue to learn and give back to the archival community. Design your digital repository. Use sound planning. Survey collections with digital content and bring it into your repository. Prepare for ingest and trustworthy work flows. You have to know what your content is and what steps are necessary to be able to get funding support. Keep at it, a little at a time. Take steps. Do something to protect what is important to your users. Share your knowledge with creators so they create more durable content that will last.

What we do now is critical.
Just do it.

ROAD TO THE FUTURE: COLLABORATION AND COOPERATION

Gregor Trinkaus-Randall, incoming SAA President

Estimated attendance: about 150 people.

SAA seeks to represent all members, large and small, rural or city, of any type and level, in all stages of their careers, and at all levels of funding. Many non-members are impacted by SAA as well; everyone needs some help of some sort.

A growing number of archivists are coming out of library science programs instead of history. We are beginning a dialog between the 2 professions that should be built upon. One way is to collaborate at national level, assisting other organizations and presenting on archival issues at other conferences. This must be done diplomatically, thinking outside of the box.

Collaboration in educational efforts could be of value, building on the strengths of each organization. We may be able to help protect funding resources or push for assistance when and where appropriate.

We need more proactive efforts – such as new efforts to protect government records and cultural resources – instead of reactive responses to disasters.

What about at the micro level, such as public libraries, small museums, and historical archives? These organizations are surviving on a shoestring, and maybe operating only

with volunteers. They may not have professional archivists. We need to develop good will on this level as well and lay groundwork for future collaborations.

Cooperation and collaboration is the theme for next year. Katrina Jackson and Rob Spindler will be next year's program chairs, in San Diego. Consider making presentation proposals that address these (or other) types of borders:

- Professions
- Political
- Geographical
- Small/large institution
- Generational
- International

Next year's meeting will be in San Diego, August 6-11 at the Hilton Bayfront Hotel.

DIGITIZING

RAPID CAPTURE IN UNIVERSITY ARCHIVES: A MODEL FOR MASS DIGITIZATION OF INSTITUTIONAL CONTENT

Eric Moore (Lead Archivist for Health Sciences and AUL Archivist, University of Minn.)

Users expect discovery and delivery to coincide. A sustainable model integrated within normal operations is the best means to achieve large-scale digitization.

They developed a programmatic effort to digitize archives. Their focus was mass produced serial publications and records, published by university, not unique, uncataloged, and from the 20th century. Their pilot was called the Academic Health Center History Project. They handled digitization in-house at 7 cents a page, using sheet-feeding scanners and student workers. The same material is now collected digitally, without collection or series selection.

Delivery to users is via an institutional repository (IR) and RSS feeds; but these are not the sole delivery or discovery access points. Downloads are PDFs, bitonal 300 DPI images with OCR content. Metadata is minimal and at the file level.

Some of the material is tied and returned to shelves, or foldered and added back to the archival collections. Mostly, however, the analog content is discarded and recycled. They are only keeping the digital versions. Their commitment is to value, not format. In their perspective, "preservation" does not equal "permanent," and "important" does not equal "unique." Digitization is a recovery of information.

Yesterday morning, Moore said, a man in his 70's called the archives, asking about his father, a pitcher on a baseball team in the 30's. They did a full-text search of the IR and found several items. One was a 1933 press release about the man's father learning how to throw a round-house curve from his own father. That turned out to be his father's signature pitch.

This couldn't have happened outside of this program.

Many collections have a half-life of just a few decades. Moore said we all need to get better about information retrieval and access.

"MEDIUM-SCALE DIGITIZATION"

Dan Santamaria (Seeley G. Mudd Manuscript Library at Princeton University)

Santamaria says they have about 35,000 linear feet of public policy papers and university archives, and 100% of their holdings have been described online in some form since 2008. The main problem is that very little of the content itself is not online. 2 years ago .00011% was online, and now only .00013%.

Patrons want the content online. Open and equitable access in 2011 does not mean requiring patrons to come to reading room 9-5 weekdays. Their goal is to maximize accessibility of collection materials to users. What is the least we can do that is adequate to users needs?

Bill Landis suggested that they find the proletariat of our collections, and figure out what it means to provide access online without enhancing metadata. The effort has not gone well, for cultural, social, and political reasons. Most of their digital content is boutique; they're not even close to Alabama or Minnesota in digital holdings.

Their minimum metadata begins with a barcode as identifier; box, folder, "unittitle" and "unitdate" are added from the finding aid. Digital objects are in PDF format: not a print facsimile, but legible onscreen. This provides navigation of the structure of the file (by delivering the pages in order). They decided that the PDF is the atomic level of description.

They're taking the "low rent approach," which is to say that students work on it in between projects. Their methods are low cost and sustainable, and are based on large-scale digitization. Their results thus far are 12 collections, or about 25 linear feet. This provides some analog preservation support, as analog material for digitized content is *not* available in the reading room any more. Their patrons are very pleased.

Next on their agenda is on demand digitization, different levels of digitization, expanded tracking and metrics, and more advocacy within the institution.

MORE ACCESS TO MORE CONTENT: THE EAD FINDING AID AND OTHER EFFECTIVE TOOLS FOR LARGE-SCALE DIGITIZATION

This panel was chaired by Karen Weiss, Information Resources Manager at the Smithsonian Institute Archives of American Art. All presentations were about leveraging finding aids to provide better access to content and to support large-scale digitization.

THE TRICKLE, THE FIREHOSE, AND THE BUCKET: LARGE-SCALE MANUSCRIPT DIGITIZATION AT UNC CHAPEL HILL

Laura Clark Brown (University of North Carolina, Chapel Hill)

Brown presented on the Digital Southern Historical Collection (SHC). She stated that it's a program, not a project, and with the continued library support in place, it is sustainable, designed to survive tough times. This program is based on digitization on demand, and occasional collection funding when outside funds are available.

Why should we use the EAD finding aid to deliver content? Why is it popular with researchers? They consulted with scholars, who said we shouldn't presume to know what they need. Scholars said we should give them all of the content of the collections and let them decide what they want; no amount of description replaces seeing all the documents. Even non-academic users seem to want to see all of it. Users' expectations for more access to more content outpaces our ability to meet the demand.

Brown is now overseeing a collaborative grant, but it's not making much of a dent in their combined holdings. The main question they're getting from online users of their finding aids is "why won't the folder open?" Users expect the digital content to be there. The EAD format provides flexibility and the nimbleness for us to digitize many collections. The UNC (University of North Carolina) projects are often hybrid – mass digitization as well as boutique. Since they have multiple streams of content, they needed a programmatic approach to delivery based on finding aid instead of on project.

Content appears online in CONTENTdm[8] as soon as it is digitized. They have JavaScript embedded to locate content from the EAD already online.

They receive multiple requests per week as to what to digitize. Usually within days they have the requested content online. They make no item-level division; all content in a folder is scanned together, with no documented relationship between one page and another. This is nearly identical to the reading room experience.

The main critique is the sparseness of the metadata, which is at the folder level, pulled from the finding aid. Criticism comes from colleagues, not users. They worry about context for the materials in CONTENTdm, and users' unfamiliarity with the finding aid.

Research is work: it is work in the reading room, and it is work online. However, making it possible for that research to take place online makes a huge difference to patrons.

CHEAP, QUICK, AND PRETTY: MASS DIGITIZATION OF LARGE MANUSCRIPT COLLECTIONS[9]

[8] OCLC. 2011. "CONTENTdm: Digital Collection Management Software." (Available from http://www.contentdm.org/)

[9] DeRidder, Jody L. "Cheap, Quick, and Pretty: Mass Digitization of Large Manuscript Collections" [Powerpoint], part of a panel presentation on "More Access to More Content: The EAD Finding Aid and Other Effective Tools for Large-Scale Digitization" at Archives 360°: 75th Annual Meeting of the

Jody DeRidder (The University of Alabama Libraries)

DeRidder presented on the NHPRC-funded project for digitizing the Septimus D. Cabaniss papers[10], and started by noting that we've all heard about leveraging EAD finding aids for access to digital content. The approach raises six critical questions:

- How difficult is it?
- What does it look like?
- How long does it take?
- What does it cost?
- How effective is it?
- What's missing?

DeRidder sought to answer these questions from the context of their work at Alabama. She stated that the difficulty is minimal; in this approach, file names need to include the box and folder number, as well as the sequence for delivery. Scripts provide the linking into the finding aid based on this information. Screenshots were displayed of the finding aid in Acumen[11], an item-level display in Acumen, and another item displayed in the software developed by the grant project, available open source.[12]

A comparison of time in work flow between the usual item-described content and this mass digitization method showed this new approach required 47% less time (4.34 minutes per scan as opposed to 8.25 minutes per scan). Cost is 68% cheaper (less than $0.80 per scan as opposed to $2.47 per scan) and experienced researchers prefer this access method to digital content via the finding aid. Their usability study found that participants new to digital collections found access via the finding aid easier (42% less time, 27% fewer clicks, and 12% more success than when searching through item-described content). However, foreign students found the terminology of the finding aid difficult.[13]

DeRidder called for improved usability of the EAD interface, addressing terminology, availability of a "search in page" feature, and hierarchical navigation links. She also stated that we need to establish the learnability of this interface by testing the same users over multiple sessions. (More information will be available in an upcoming article in *American Archivist*.)

ACCESS TO A LEGACY: DIGITIZING THE ARCHIVES OF THE JOHN F. KENNEDY PRESIDENTIAL LIBRARY

Society of American Archivists, held in Chicago, IL, 22-27 August, 2011. (Available from http://jodyderidder.com/writings/presentations/SAA2011/CheapQuickPretty.ppt)

[10] University of Alabama Libraries. 2010. "Septimus D. Cabaniss Papers Digitization Project." (Available from http://www.lib.ua.edu/libraries/hoole/cabaniss)

[11] University of Alabama Libraries. 2010. "Acumen Digital Library Software." (Available from http://sourceforge.net/projects/acumendls/)

[12] University of Alabama Libraries. 2010. "Software for You!" (Available from http://www.lib.ua.edu/wiki/digcoll/index.php/Software_For_You%21)

[13] DeRidder, Jody L. "Providing Access to Digitized Content Via the Finding Aid: A Usability Study" [Powerpoint], at the SAA Research Forum of Archives 360°: 75th Annual Meeting of the Society of American Archivists, in Chicago, IL, 23 August, 2011. (Available from http://jodyderidder.com/writings/presentations/SAA2011/DeRidderResearchForum.pptx)

James Roth and Erica Boudreau (John F. Kennedy Library)

Roth stated that they began digitizing holdings in 2006, and launched their website 5 years later. They have a tremendous amount of content which continues to grow. Their path is not easy, partnering with multiple institutions and working with an uncertain budget. They are using DC (Dublin Core[14]) metadata and Documentum[15] for content management, which has a capability of 30 TB (they have 19 TB now). Their finding aids are in multiple formats, as they failed in 2003 to move to the EAD. Instead they use information in Documentum to generate a hyperlinked finding aid.

Erica Boudreau stated that they exported metadata used for indexing on their website, and then began to use it to export EAD-tagged XML (adding the DAO linking element). In their display they digitize to the folder level.

Karen Weiss presented for the Smithsonian Archives of American Art: "Making the most of your access: leveraging the EAD Finding aid to increase digitization and access to collections." They took an archival approach to digitization and access. Thus far, 112 collections have been scanned completely, which amounts to 14 TB of data, composed of 114 million images.

Microfilm defined the archives for 50 years; now it is the EAD.

They developed ColdFusion software and a relational database to capture and manage the finding aid content. The software parses the EAD, and creates object stores in the database. This allows flexibility for display. Each series has a representative thumbnail. They provide folder-level titles, access and digitization (all items in a folder are digitized together without boundaries). The "more about" tab in their interface contains the fully indexed finding aid in locally-installed Google software for search and retrieval. Breadcrumbs are used for navigation.

The scanning directory structure is based on boxes. There may be multiple files with the same file name in different directories; the directories provide the context of box and folder. After scanning, the external drive is handed off for processing. Then folders are generated, an archivist reviews the results, and finally they launch the site. After launch, they ingest the TIFF files into their preservation system. They use the OpenText "Artesia" digital asset management system. [16] They have reverse-engineered the use of EAD to move content into storage, where materials are held by collection, keeping series, boxes, and folders together. Storage includes descriptive, technical and administrative metadata, and they adhere to the perspective of "balance in rights management"[17] which is based on fair use principles and a take-down policy. Researchers are grateful.

---

[14] Dublin Core Metadata Initiative. "Dublin Core Metadata Element Set, Version 1.1." 2010. (Available from http://dublincore.org/documents/dces/)

[15] EMC². "EMC Documentum." (Available from http://www.emc.com/domains/documentum/index.htm)

[16] OpenText. "Digital Media Group." (Available from http://digitalmedia.opentext.com/)

[17] OCLC. 2010. "Introduce Balance in Rights Management." (Available from http://www.oclc.org/research/activities/rights/)

They are planning to incorporate digitization on demand.

LEVERAGING CONTROLLED VOCABULARY TO IMPROVE ACCESS

LINKED OPEN DATA – LIBRARIES ARCHIVES MUSEUMS (LOD-LAM)

Jenel Farrell, Digital Archivist, Minnesota Public Radio / American Public Media
(Presented at the Metadata and Digital Object Round Table Session)

Institutions are beginning to copyright their metadata, especially EAD (Encoded Archival Description[18]) finding aids. This is a barrier. We are urging OCLC to support open access to metadata.

To support linked data, it is necessary to have unique URIs (Uniform Resource Identifiers) for each resource, and RDF[19] to represent your metadata on the web. Once this is done it is possible to use assertions to embed relationships between resources in tags. The foremost method for this is to leverage SKOS (Simple Knowledge Organization System[20]), and Sparql[21] for querying data.

DBPedia[22] mines the data out of Wikipedia, which can then be used to make your own content more valuable by creating relationships with information on the web that already exists.

To support linked data, remember these guidelines:

- Make your content available on the web under an open license, as structured data.
- Use non-proprietary formats.
- Use URIs to identify content.
- And link data to provide context.

THE SOCIAL NETWORKS AND ARCHIVAL CONTEXT PROJECT (SNAC): EAD-CPF[23] AT WORK

Daniel V. Pitti, University of VA; Ray Larson, University of California at Berkeley; and Brian Tingle, California Digital Library (CDL)

---

[18] Library of Congress. "EAD: Encoded Archival Description." (Available from http://www.loc.gov/ead/)

[19] W3C Semantic Web. 2004. "Resource Description Framework (RDF)." (Available from http://www.w3.org/RDF/)

[20] W3C Semantic Web Activity. 2009. "SKOS Simple Knowledge Organization System." (Available from http://www.w3.org/2004/02/skos/)

[21] W3C. 2008. "SPARQL Query Language for RDF." (Available from http://www.w3.org/TR/rdf-sparql-query/)

[22] DBPedia. 2011. (Available from http://dbpedia.org/About)

[23] EAD-CPF. "Encoded Archival Context: Corporate Bodies, Persons, and Families." (Available from http://eac.staatsbibliothek-berlin.de/)

The EAC/CPF (Encoded Archival Context Corporate Bodies, Persons, and Families) has now been formally adopted as a standard.

OVERVIEW

Daniel Pitti provided us with an overview. SNAC[24] is a cross-institutional research project funded by the National Endowment for the Humanities, which began in May 2010 and will end April 2012.

The EAD metadata available contains a mix of the description of records with description of creators of records and people evident in the records. This project facilitates the separation of the description of people, to enhance access and understanding of users of libraries, archives, and museums.

They are using authority control of forms of names; EAD doesn't support alternative forms of names. This will make description more flexible; one need only describe a person once for multiple collections and repositories. For example, Walt Whitman need only be described once for 71 different repositories.

This will enable cooperative authority control and integrated access to cultural heritage materials. Also we can now create a social/historical context, documenting the interrelations between people, which provides greater context for understanding content.

The Library of Congress (LC), Online Archive of California (OAC), Northwest Digital Archives (NWDA), and Virginia Heritage are all contributing EADs to the project.
Authority records used are:

- Name Authority Cooperative (NACO) / Library of Congress Name Authority File (LCNAF)[25]
- Getty vocabulary program: Union List of Artist Names (ULAN)[26]
- Virtual International Authority Files (VIAF)[27]

The process involves extracting EAD-CPF records from existing EAD-encoded archival descriptions (this portion of the work performed by partners at the University of VA), then matching them against one another and against existing authority records. Once matched, records for same entity are merged (by Berkeley partners).

---

[24] Social Networks and Archival Context Project (SNAC). 2011. (Available from http://socialarchive.iath.virginia.edu/prototype.html; demonstration: http://socialarchive.iath.virginia.edu/xtf/search)

[25] Library of Congress. "NACO: Name Authority Cooperative Programs of the FCC." (Available from http://www.loc.gov/catdir/pcc/naco/)

[26] Getty Institute. Union List of Artist Names Online. (Available from http://www.getty.edu/research/tools/vocabularies/ulan/index.html)

[27] OCLC. "VIAF: The Virtual International Authority File." (Available from http://viaf.org/)

Then team members can enhance the EAD-CPF by normalizing, adding alternative entries and titles to VIAF and historical data to ULAN. One key challenge is when there are 2 or more people with same name, or two or more names for same person.

California Digital Library partners are creating a prototype historical resource and access system which will combine historical data and social-professional networks, with links to archive, library, and museum resources. Challenges include problems with the content of EADs. They are of widely varying quality, including in the number of names, in formation of the names, and in the categorization of names. Many names are given but not identified as such; the most important of these are in the biographies/history ("bioghist" tag) and in correspondence description.

So far, extraction has focused on names tagged as names; project staff is moving toward locating names not identified as such. Unfortunately, natural language processing does a poor job of picking out names inverted and in alternate forms. Archival descriptions document interrelations among people. It's easy to get out what's in the "controlAccess" section tagged as names. But "unitTitle" sections, for example, also often contain names. Sometimes the names are in direct order, indirect order, mixed with text, or contain other anomalies. Similar problems are in the name-rich biographical sketch ("bioghist" section) and elsewhere.

EAD-CPF is based on ISAAR (CPF)[28]. It includes a biographical-historical description to identify names, and provides context. This requires more description than just control of names, for it includes occupations and birth and death places, for example.

The use of subject headings has a high biographical historical value when attached to the fonds[29] of a person. The relationship of the person to the resources can be captured.

They have extracted a bunch 197 records so far, but have not yet analyzed the results. The depth of analysis and quality of description varies widely. These finding aids were created without SNAC in mind. For example the Library of Congress finding aids include a large number of authority-controlled names. However, the OAC and NWDA finding aids contain far less.

The next step is to refine the extraction processing, incorporating some natural language processing methods. They will be verifying the type of name, massaging poorly formed names, identifying names in strings, and providing context.

Beyond the project, the intent is to build a National Archive Authorities Infrastructure, funded by IMLS for 2 years, Oct. 2011-Sept. 2013. 140 scholarships will be provided to EAC-CPF SAA workshops, in order to create a cooperative.

---

[28] International Council on Archives. "(ISAAR (CPF). International Standard Archival Authority Record for Corporate Bodies, Persons, and Families, Second Edition." 2004. (Available from http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf)

[29] SAA Glossary of Archival and Records Terminology. "Fonds: The entire body of records of an organization, family, or individual that have been created and accumulated as the result of an organic process reflecting the functions of the creator." (Available from http://www.archivists.org/glossary/term_details.asp?DefinitionKey=756)

Another proposal being written is SNAC II—a proposal to expand SNAC to incorporate a lot more data. This will involve NARA, Smithsonian Institution, MARC records from WorldCat, and more finding aids.

Daniel Pitti welcomes others to share their finding aids with this project.

MERGING, MATCHING AND ENHANCING CONTROLLED VOCABULARY

Ray Larson then spoke for the portion of the work on SNAC being performed by Berkeley partners. They are merging together all the disparate records from various sources. All the matching and enhancing is up to them.

They are using Cheshire search[30] and primarily using VIAF for reference as it subsumes Library of Congress personal names, but not all corporate names. Then they address authority control, identifying creator entities and referenced entities such as correspondents, recording the name or names used by and for them, and applying a rule-based heading or entry formation and control. This is an attempt to provide a consistent set of descriptions for use.

A problem they've encountered is the proliferation in forms of names. Even Books in Print is only semi-controlled, for example. For example when looking up Goethe in that resource, one will encounter "see also" that leads to "see also" that leads to "see also."

Also, many authors have the same name: an example is John Muir. The library answer is to put together all the variations of a name together. They start by assuming identical names are the same person and merge the resources. Then they search authority files, primarily VIAF, for both authoritative and non-authoritative forms. A non-authoritative match is considered as a candidate match for the authoritative form.

Those names that are flagged as likely to be the same are then merged if possible. They extract the authoritative form of name (defaulting to the LC version if there is not one). The data is then combined, retaining all source record identifications and information, and they output a merged EAD-CPF record.

Total extraction thus far consists of 123,920 "unique" names. There was a significant collapse in size by merging information from multiple repositories.

Up to this point, they have been using exact matches as a base, with the assumption that LC cataloging is being followed; that's a problem. The abbreviations, spacing, punctuation, completeness and alternate rules create difficulty in string matching. So they are modifying their matching algorithms to try to catch more matches. For example, there are variant Romanizations, initials vs. names, omitted first name, inversion versus un-inverted, various combinations in spelling, translation, use of II, I, etc. In some curious entries they're not

[30] "Cheshire3 Information Framework." From the website: "Cheshire3 is a fast open source XML search engine. Given a set of records, Cheshire3 can extract data into one or more indexes. Indexes support common operations such as ranked search, faceted search, browsing, and result reordering." (Available from http://www.cheshire3.org/)

sure if the entry is a person or corporation or place or what, as there is no first name and no associated dates. They are also trying not to overmatch, either.

First they need to know what's failing and why. In this next stage, they will do a random sampling and detailed evaluation. They may be able to solve problems by using contextual clues from EAD records, more sophisticated matching for phonetic variants, or additional normalization before attempting the match and merge. They are also testing new merging methods, such as using Systems Network Architecture and merging with Freebase[31] and the Internet Movie Database.[32]

The approach under consideration includes looking at birth and death dates and the distance of this name from that name. They use the "shingle sequence," or "shingle language model" for approximate matching. The method is to take three letters from each name and match them; even if misspelled, the probability is that the two may be related. To do this, each name is considered as a circle; where it ends, it starts again. A related approach is to use a decision tree using string distance: counting the number of edits needed to modify one name into another. These are widely available metrics for matching.

Their corpus average is 71.7% TPR (true positive rate for positive matches) with 17% FPR (false positive rate). There is no single way to do it; they need a sequence of matching operations. Once records are merged, they are passed on to Brian Tingle (CDL) for search and display.

SEARCH AND DISPLAY

Brian Tingle, of California Digital Library, presented on "Discovering Historic Social Networks: Prototype Historical Resource Demonstration."[33]

They developed personas for their target users:

- Randy the Ph.D. researcher, studying biographies
- Connie the contributor; her archive donated records, and she wants to know how this site is useful for their patrons
- Quincy is a library student working with the project
- Adele is doing authority work during processing
- Lenny ("the linkhead") likes linked data, and wants to be able to mine the links that are established via the project

The interface is viewable at http://socialarchive.iath.virginia.edu/xtf/search.
It includes facet tabs across the top to limit results to a type of records (Person, Corporate Body, Family, or All). Mousing over the search button allows the user to locate the advance search option to limit results to a section of the standard. Every page has the option to see

[31] Google. 2010. "Freebase: An entity graph of people, places, and things, built by a community that loves open data." (Available from http://www.freebase.com/)
[32] IMDB.com, Inc. 2011 "IMDB: The Internet Movie Database." (Available from http://www.imdb.com/)
[33] Tingle, Brian. 2011. "Discovering Historic Social Networks: Prototype Historical Demonstration." Presented at Stanford University and again at SAA. (Available from http://www.slideshare.net/tinglebrian/snac-dh2011june2011)

the XML[34] source file. There's an auto-complete function that runs across all names to enable the user to see what matches. Also the search function includes spellcheck; it will look for similar names. Within search results, one can narrow the result set by facets. The display allows users to track back to the finding aid where the names were found.

They're using HTML4[35] microdata[36] in the chronological listings so can later they can try a SIMILE[37] timeline or something similar. The related entries section includes finding aids mentioning the person, and the other people and corporate bodies associated with the entry. Also they have linked WorldCat[38] entries to each name, and plan to add Dbpedia links and more.

They are utilizing RDFa[39] encodings of OWL[40] "sameAs"[41] in the HTML of each web page to link different entries about the same person or thing. Graphs are generated using JavaScript.

They spent a lot of time trying to do this with RDF, but instead wound up using the TinkerPop Graph Stack[42] which is based on a property graph model, and the GraphML file format[43] and RDF Sail[44] support. The TinkerPop Gremlin software[45] proved very useful for analysis and manipulation of multi-relational graphs. It's an easy way to get started, and a gateway into using linked data. It comes with "doghouse" software for browsing graphs, which contains a "gremlin" query which will show you which names are most connected. They're using a "Rexster" multi-faceted graph server[46] for display. It only shows the top 50 most popular people for a node, and shows them in a radial graph. When the user clicks on another name, the graph reforms. It's all in JavaScript. The display stops at 5 levels out to keep from exploding; it's a really nice visualization. Changes to the interface will be additions of an information box to read about the person whose name one is centered on, and improved search.

---

[34] W3C. 2011. "Extensible Markup Language (XML)." (Available from http://www.w3.org/XML/)
[35] W3C. 1999. "HTML 4.01 Specification." (Available from http://www.w3.org/TR/html401/)
[36] W3C. 2011. "HTML Microdata." (Available from http://www.w3.org/TR/microdata/)
[37] Massachusetts Institute of Technology. "Semantic Interoperability of Metadata and Information in unlike Environments (SIMILE)." (Available from http://simile.mit.edu/)
[38] OCLC. 2011. "WorldCat." (Available from http://www.worldcat.org/)
[39] W3C. 2008. "RDFa Primer: Bridging the Human and Data Webs." (Available from http://www.w3.org/TR/xhtml-rdfa-primer/)
[40] W3C. 2004. "OWL Web Ontology Language Reference. (Available from http://www.w3.org/TR/owl-ref/)
[41] W3C. 2004. Ibid., subsection 5.2.1: "owl:sameAs." (Available from http://www.w3.org/TR/owl-ref/#sameAs-def)
[42] TinkerPop. "TinkerPop Graph Stack." (Available from http://www.tinkerpop.com/)
[43] GraphML Working Group. "The GraphML File Format." (Available from http://graphml.graPh.D.rawing.org/)
[44] OpenRDF.org. "Interface Sail." (Available from http://www.openrdf.org/doc/sesame2/api/org/openrdf/sail/Sail.html)
[45] TinkerPop. "Gremlin: Defining a Property Graph." (Available from https://github.com/tinkerpop/gremlin/wiki/Defining-a-Property-Graph)
[46] TinkerPop on Github. "Rexster." (Available from https://github.com/tinkerpop/rexster/wiki/)

Tingle recommends the use of the open source "eac-graph-load" software[47] for processing the EAC records into either a database or a graphML file.

Special thanks go out to Ed Summers (Library of Congress), who used Python to create RDF using FOAF[48] and Aaron Rubinstein's vocabulary.[49] Tingle plans to publish a version of the graph as linked data also; and a branch of XTF[50] software used for searching EAC records is now available.[51]

Silvia Mazzini, of the Italian "Regesta.exe SRL"[52] developed an ontology that represents the EAD-CPF in RDF.[53] Tingle will be studying this, and will release a version in this ontology.

There will be an EAD to EAD XSLT[54] transform forthcoming from VA; and record merging code will be forthcoming from Berkeley.

Pitti states that they will be normalizing dates and geographical names, and developing timelines and mapping features. For incoming EADs, they will request inversion of names in name tags, normalized.

For more information, the slides are available online (in Macintosh "key" form) at http://www.slideshare.net/tinglebrian/snac-saaaug2011try-3-keynote.

INCOMING BORN-DIGITAL CONTENT

BORN DIGITAL "BABY STEPS"

Jackie Dooley, OCLC Research

Dooley reported on a new project describing ways to archive born digital content. Administrators often don't know why or how archivist's skills and expertise are broadly relevant to management and preservation of born-digital content, and the archivists need training and tools and need to know where to start.

---

[47] Google. "eac-graph-load: eac to graphML; rexter extensions; RDF scripts." (Available from http://code.google.com/p/eac-graph-load/)
[48] W3C. 2010. "FOAF Vocabulary Specification 0.98." (Available from http://xmlns.com/foaf/spec/)
[49] Aaron Rubinstein. 2011. "Arch: an RDF vocabulary for describing archival collections and the names associated within them." (Available from http://gslis.simmons.edu/archival/arch/index.html)
[50] California Digital Library. 2011. "XTF: eXtensible Text Framework." (Available from http://xtf.cdlib.org/)
[51] Google. 2011. "XTF-CPF: XTF branch for the Social Networks and Archival Context Project." (Available from http://code.google.com/p/xtf-cpf/)
[52] Regesta.com. 2011. "New Media & Historical Heritage: Regesta.exe." (Available from http://www.regesta.com/info/)
[53] Francesca Ricci and Silvia Mazzini. 2010. "EAC-CPF Ontology." (Available from http://templates.xdams.net/IBC/ontology/eac-cpf.rdf)
[54] W3C. 1999. "XSL Transformations (XSLT) Version 1.0." (Available from http://www.w3.org/TR/xslt)

Dooley reported on part of a 2009 OCLC survey on born-digital issues called "Taking our Pulse: the OCLC Research Survey of Special Collections and Archives"[55] in which they found that the 3 most challenging issues are lack of space, born-digital materials, and digitization. Half of all the gigabytes reported were held by only 2 institutions. The top 13 libraries held 93% (this does not include NARA). Assignment of responsibility for born-digital content has been made in only 55% of the institutions who responded. 13% of them placed that responsibility with their special collections and archives departments. Of the born-digital content already held, there were more photographs than university records. 69% said the major impediment to managing this content is lack of funding followed by lack of planning, then lack of expertise.

Recommended actions based on this survey included:

- Define the characteristics of born-digital materials that warrant management as special collections
- Define a reasonable set of basic steps for initiating a program
- Develop use cases and cost models for selection, management and preservation of born-digital archival materials

Dooley recommended that we all read the final report from the Blue-Ribbon Task force.[56] Penn state has done some use case studies[57] that are of interest as well.

Target audiences are research library directors and higher administration, archivists and special collections librarians, and other research library specialists in IT, collection development, digital libraries, institutional repositories, metadata, and web development.

Dooley argues that archivists have relevant skills and expertise in the following areas: appraisal, donor relations, deeds of gift, intellectual property rights, legal issues, privacy and confidentiality, authenticity, provenance and context, hierarchical organization of content, collective metadata, and the understanding that preservation impacts permanence.

She recommends that archivists:

- Make friends with information technology
- Promote their skills
- Consider new aspects of donor negotiations and agreements
- Identify the basic nature of content currently held
- Get as much education in this realm as is possible

Dooley states that the first steps to take with born-digital content are:

---

[55] OCLC Research. 2010. "Taking our Pulse: The OCLC Research Survey of Special Collections and Archives." (Available from http://www.oclc.org/research/publications/library/2010/2010-11.pdf)

[56] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information." 2010. (Available from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf)

[57] Patricia Hswe, Michael J. Furlough, Michael J. Giarlo, and Mairéad Martin. 2011. "Responding to the Call to Curate: Digital Curation in Practice at Penn State University Libraries," The International Journal of Digital Curation, 2:6, 2011. (Available from www.ijdc.net/index.php/ijdc/article/download/191/256)

- Inventory what you have
- Learn basic "do no harm" file management steps
- Transfer the content from physical media to secure storage

PRACTICAL APPROACHES TO BORN-DIGITAL RECORDS

Chris Prom (Assistant University Archivist and Associate Professor of Library Administration, University of Illinois, Urbana-Champaign) chaired the panel called "Practical Approaches to Born-Digital Records: What Works Today." He edits a blog on "Practical E-Records."[58]

The Garnter Hype Cycle[59] was developed in a project funded by the Fulbright commission. Visibility is represented by the Y axis and maturity by the X axis. The message is that hype massively increases visibility, creating a peak of inflated expectations. The trough of disillusionment represents the maturity of the project. Then comes the slope of enlightenment, followed by the plateau of productivity.

The dark cloud before us now the OAIS (Open Archival Information System) reference model[60] with tools mapped onto the concepts. William Killbright said that NASA called in their friends to develop OAIS.

Are we doing anything perfectly? No. We need collaborative leadership. Prom stated that the topic this morning is recycling ideas from other people, starting with convincing people to turn their stuff over to us.

THE DONOR INTERVIEW

Suzanne Belovari (Archivist for Reference and Collections,Tufts University) presented on "The Donor Conversation RE: Hybrid Collections (Physical and Electronic)." It is likely that all collections that come into our repositories in the future will be hybrid. We must adjust workflows, properties, and policies. Examples of recent efforts include the "Digital Lives Research Project"[61] for personal digital collections, the Mellon-funded AIMS project[62] to develop an inter-institutional model for stewardship of born-digital content, and JISC's

[58] Chris Prom. 2011. "Practical E-Records: Software and Tools for Archivists." (Available from http://e-records.chrisprom.com)
[59] Gartner, Inc. 2011. "Gartner Hype Cycle: Interpreting Technology Hype." (Available from http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp)
[60] Consultative Committee for Space Data Systems. 2002. "Reference Model for an Open Archival Information System (OAIS)." (Available from http://public.ccsds.org/publications/archive/650x0b1.PDF)
[61] British Library. 2009. "Digital Lives Research Project." (Available from http://www.bl.uk/digital-lives/index.html)
[62] University of Virginia Library. 2011. "AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship." (Available from http://www2.lib.virginia.edu/aims/)

Paradigm project[63] in the UK. After assessing these, Tufts adapted Chris's paradigm to their needs.

Some of the general questions and information included asking the donor about their primary or core identities (such as scholar, teacher, etc.) which would help determine which materials and formats were most important from their point of view. When content was donated, staff would go through it, list which items have archival value from archivist point of view, and share this list with the donor, creating a conversation. Between them, they determined what should be in the permanent record.

Originally the donor had (almost always) only seen himself as a single identity, and had failed to see all his other roles and core identities. Outside of his donation, he wasn't so constrained. Like many donors, he considered his digital content ephemeral and somehow less important. However, each component impacts the others. In selection, they considered sensitive materials, weeding, virtual death and will, and transfer of materials.

Organizing digital material is crucial for future researchers. Collaboration in managing backups may be an indicator of the importance of content.

IDENTIFYING CONTENT

Gabriela Redwine (Harry Ransom Center, University of Texas at Austin) presented on "Preserving Born-Digital Materials"

Redwine reported that in the Ransom Center[64], they have mostly literary papers, including 8 personal computers and 2200 disks. 46 holdings contain digital media. They used to pull content off disks file by file, but now they are collecting the entire image of the disk. The quantity and variety of files on hard drives is far beyond that on a floppy disk or CD. They are using a free demo version of Forensic Toolkit[65] and are storing text files of collected information with their digital content.

Redwine reports that forensic investigators are doing similar work, so a number of archivists are now using forensic tools. Examples she gave include:

- Digital Lives Project[66] of the British Library: Jeremy Leighton John
- futureArch[67] project of the Bodleian Library: Susan Thomas
- Digital Forensics at Stanford[68] : Michael Olson

[63] JISC. 2007. "PARADIGM: The Personal Archives Accessible in DIGital Media project." (Available from http://www.paradigm.ac.uk/)
[64] Gabriela Redwine, Harry Ransom Center, The University of Texas at Austin. 2010. "Ransom Edition, Spring 2010: Preserving Born Digital Materials." (Available from http://www.hrc.utexas.edu/ransomedition/2010/spring/borndigital.html)
[65] AccessData Group, LLC. 2011. "Forensic Toolkit (FTK) Computer Forensics Software." (Available from http://accessdata.com/products/computer-forensics/ftk)
[66] British Library. "Digital Lives Research Project." (Available from http://www.bl.uk/digital-lives/)
[67] University of Oxford. 2011. "Bodleian Electronic Archives and Manuscripts: futureArch." (Available from http://www.bodleian.ox.ac.uk/beam/projects/futurearch)
[68] Stanford University Libraries & Academic Information Resources. "Digital Forensics @ Stanford University Libraries." (Available from http://lib.stanford.edu/digital-forensics)

- Digital Records Forensics project[69]: Luciana Duranti

Cal Lee is teaching how to use forensic tools at the University of North Carolina, Chapel Hill (UNC).[70]

Redwine was recently involved in a collaborative Mellon-funded project[71] which resulted in a CLIR report on these efforts.[72]

Redwine notes that the presence of hidden and deleted files on disk image makes things more complex. The donor or creator may not even know the files still exist. She raises some important questions:

- How do we work with creators, dealers and other custodians to ensure all parties are well informed? They may have donated more than they thought they did.
- How do we develop policies robust enough to deal with future changes in technology?
- How do we deal with digital media that are already in our collections, which were accepted before we developed policies for rights and access?
- How do we protect our repositories and depositors while still meeting needs of researchers? The latter may want access to disk images, which may contain files that should be redacted.

Ideally, Redwine states that the archivist would use forensic tools to view content with the donor or creator before donation, to sift and make decisions prior to accession.

What constitutes a violation of privacy? This is context-dependent. Many donors have little understanding about what's in the disk image. There's a difference between capturing everything and keeping everything. Appraisal with creator's input is ideal, as it enables the archivist to carve out what should be retained and effectively document how it should be managed.

SELECTING AND APPRAISING

Peter Chan (Digital Archivist, AIMS project[73], Stanford University) presented on "Using Forensic Software to Assign Metadata to Born Digital Archives."

---

[69] "Digital Records Forensics Project." (Available from http://www.digitalrecordsforensics.org/)
[70] University of North Carolina, School of Information and Library Science. "Christopher (Cal) Lee." (Available from http://www.ils.unc.edu/callee/)
[71] Maryland Institute for Technology in the Humanities. "Computer Forensics & Born Digital Content in Cultural Heritage Collections." (Available from http://mith.umd.edu/forensics/)
[72] Matthew Kirschenbaum, Richard Ovenden, and Gabriela Redwine, with Rachel Donahue. 2010. "Digital Forensics and Born-Digital Content in Cultural Heritage Collections," Council on Library and Information Resources. (Available from http://www.clir.org/pubs/reports/pub149/pub149.pdf)
[73] University of Virginia Library, "AIMS Born-Digital Collections."

In the inter-institutional Mellon-funded AIMS project for management of born-digital content, Chan is using forensic software to assign metadata to born digital archives. Currently he is working with a variety of formats. They have 160 floppy disk drives, both 3" and 5.25" and even punch cards. There are over 2500 items, but 800 are duplicates due to backups. File formats identification is a challenge! Jhove[74] only identifies limited file formats. Droid[75] cannot recognize Microsoft WordPerfect files.

Viewing the files is another challenge, which must be met before we can assign necessary descriptive metadata. Many files were created with obsolete software. Quick View Plus[76] can allow viewing of many of the file formats, but it has no functions for assigning metadata. How do we annotate?

How about search? Microsoft (MS) Search 4.0[77] recognizes some MS Office, PDF, and ASCII[78] but not WordPerfect or Lotus 1-2-3. It has no pattern search; how do we locate sensitive information such as social security numbers, credit card numbers, etc.?

Also, how do we generate reports? We need an integrated tool to perform all this.

AccessData FTK[79] is a forensics software Stanford is using to do all this. It identifies over 300 file formats and has an XSL-FO (Extensible Stylesheet Language Formatting Objects[80]) output. XSL-FO includes formatting and is not good for ingest. They wrote XSLT to transform this to XML content file by series and file. This provides a graphical user interface (GUI) to allow assigning series to ranges of content and creating labels.

ACCESSIONING

Ben Goldman (Digital Programs Archivist at the American Heritage Center, University of Wyoming) manages digital collections. He talked about a Rushdie collection, in which they extracted almost 1300 manuscript files out of over 9000 user-generated files.

Goldman states that OAIS is a reformulation of the traditional archival management functions. There are adjectives we never want used to describe archives: undercounted, undermanaged, and inaccessible. But unfortunately, these generally apply in the digital arena at present. They are trying to articulate initial steps to get started.

---

[74] JSTOR. 2009. "JHOVE: JSTOR/Harvard Object Validation Environment." (Available from http://hul.harvard.edu/jhove/

[75] Sourceforge. 2011. "DROID (Digital Record Object Identification)." (Available from http://sourceforge.net/projects/droid/)

[76] Avantstar. 2011. "Quick View Plus 11 Standard Edition." (Available from https://avantstar.com/metro/home/Products/QuickViewPlusStandardEdition)

[77] Microsoft. "Windows Search 4.0 for Windows XP." (Available from http://www.microsoft.com/download/en/details.aspx?id=23)

[78] AsciiTable.com. "ASCII Table and Description." (Available from http://www.asciitable.com/)

[79] AccessData Group, LLC. 2011. "Forensic Toolkit (FTK)." (Available from http://accessdata.com/products/computer-forensics/ftk)

[80] W3Schools.com. "XSL-FO Tutorial." (Available from http://www.w3schools.com/xslfo/default.asp)

Accessioning involves pulling data off old media formats and uploading them, but where? Network storage, secure and redundant, is the first solution.

As far as the administration is concerned, the media is already accessioned, but that only makes sense if you call a media an item instead of a container of items. The files on the disks had not been accessioned at all.

For digital content, we must take adequate steps. These are all part of accessioning:

- Virus check
- Descriptive metadata
- Data about formats
- Checksums
- Begin documentation that records management and preservation actions over time

Goldman compares each of these steps to parts of the archival definition of accessioning for analog materials. They serve to lay the groundwork to put the content out there for users.

- The virus check is like the archivist's first pass over materials.
- Descriptive metadata is equivalent to trying to gain intellectual control and lay groundwork for processing.
- Collecting data about formats is very similar to documenting analog formats, as it helps establish preservation and access plans.
- Checksums and documentation build on the provenance aspects of accessioning.

Goldman demonstrated using virus check software, then the Duke Data Accessioner[81] software. This captures descriptive metadata, identifies formats, grabs checksums, and puts it all in an XML document which they store with the materials in same folder. They also use an MS Word document accession form, with unique identifiers for each accession. It's a very basic, practical system, which enables them to determine the extent of digital holdings, and have them stored safely, not on fragile media.

ARRANGEMENT, DESCRIPTION, AND ACCESS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Melissa Salrin (Fulbright Scholar and Visiting Archival Operations and Reference Specialist, UIUC) is currently dealing with 70,000 files in many different folders. Salrin presented on arrangement, description and access at Illinois, and referred to the issues raised by the Office for Intellectual Freedom[82] case study.

Their workflow is file system based, shuffling records between unprocessed, in process and processed. One of the first things they do is to expunge personal information. They use

[81] Duke University Libraries. "Data Accessioner." (Available from http://library.duke.edu/uarchives/about/tools/data-accessioner.html)
[82] American Library Association. "Office for Intellectual Freedom." (Available from http://www.ala.org/ala/aboutala/offices/oif/index.cfm)

"Firefly SSN Finder"[83] to scan and identify files containing credit card numbers, social security numbers and such. They use TreeSize Pro[84] to find file size and formats. They delete duplicates and files not worth archiving, and then use the Duke Data Accessioner. The resulting files are moved to a preservation folder which has tiered level of access (read only for most). They generate access copies, and store both nearline and online, dependent upon restrictions.

An example collection is the Stanley Smith papers[85], which are made web-available using Archon.[86] The nearline holdings will not contain the files with access restrictions. Sometimes online access files will be created for selected content. Their E-records viewer is a PHP-based directory browser[87].

Of their lessons learned, the most important is that documentation is key. We need to make the process transparent. Documents must be saved to the unprocessed folder until processing is complete, then moved to the preservation folder. End-users need to be kept informed as to the status of collections. For example, they provide a warning on the PHP page that many file formats are out of date and are accessed at the users' own risk.

Here's the advice Salrin offers:

- Always process analog records first, since that provides crucial context
- Be prepared for time consuming work with the digital content
- For access, additional description is needed to show how documents are related.
- Privacy and copyright issues abound.
- Users must trust that online content is only a selection, not the complete record of content.

They are still exploring how to manage access to nearline content.

EMORY UNIVERSITY

Laura L Carroll, former Manuscript Archivist at Emory University presented next on the Salmon D. Rushdie papers as an example of personal papers. New York Times reported on this effort as a model for digital file management.[88]

---

[83] University of Illinois at Urbana-Champaign, Campus Information Technologies and Educational Services, "Social Security Number Remediation Program." (Available from http://www.cites.illinois.edu/ssnprogram/)

[84] JAM Software. "TreeSize Professional, v5.5. The Powerful, Graphical Manager for the Hard Disk Space." (Available from http://www.jam-software.com/treesize/)

[85] University Archives, University of Illinois. "Stanley Smith Papers, 1957-2006." (Available from http://www.library.illinois.edu/archives/archon/?p=collections/controlcard&id=10857)

[86] University of Illinois at Urbana-Champaign. "Archon: The Simple Archival Information System." (Available from http://www.archon.org/)

[87] University Archives, University of Illinois at Urbana-Champaign. "Stanley Smith Papers, 1957-2006 (Series 15/5/50)" as an example implementation of E-Records viewing method. (Available from http://www.library.illinois.edu/archives/Electronic%20Records/index.php?dir=University%20Archives/1505050_Stanley_Smith_Papers/)

[88] Patricia Cohen, The New York Times, 15 March 2010. "Fending Off Digital Decay, Bit by Bit." (Available from http://www.nytimes.com/2010/03/16/books/16archive.html?ref=salmanrushdie)

Carrol said that the first concern is to retain provenance and original order. Rushdie is a novelist with privacy concerns, and an international figure. His donation contained a substantial amount of born digital material, including several Macintosh computers, and also about 100 linear feet of paper material.

At the time of donation, they established restrictions on portions of his papers, but did not even look at the content on the computers before the donation. The established restrictions shaped much of the workflow for the rest of the project. For example, Rushdie did not want born digital content made available on the web. It is only accessible in the reading room.

First they isolated the files to be restricted. To do this, they had to look at almost every file, because there were so many, and the donor agreement called for it. They grouped materials into broad categories such as personal papers, writings, journals, etc. Luckily Rushdie was meticulous in file naming of versions and drafts, which was very helpful.

The archivists assigned series that mirrored what was used in the paper collection. All decisions were documented separately. One learns much about a creator by how they organize and name things.

For access, content was categorized according to the following:

- "As is" means the files can be released as is for both the emulation environment and the database
- "Redaction" means the files will be available for database access but not for the emulation environment
- "Restricted" means there will be no access in either environment
- "Emulation only" means the content will only appear in the virtual environment and will not go into the database.

They compared the contents of each file with the initial assessment and added subseries levels, changing the initial assessment as necessary. Example subseries include: fiction, non-fiction, scripts, other writings, and general correspondence. Then they sorted by series and subseries definitions.

The next step was to import the information into the offline Fedora[89] structure, which provides secure encrypted access online, with a data entry interface currently being created for archivists. This will enable them to provide information about content rather than just listing it.

We all need flexible and collection-specific approaches to deal with hybrid collections. Her advice: don't panic. Just do it.

DUKE UNIVERSITY

[89] Fedora Commons, Inc. "Fedora Commons Repository Software." (Available from http://fedora-commons.org/)

Seth Shaw (Electronic records archivist at Duke University, and a teacher for SAA) presented next, on "Practical Approaches to Born-digital Archives: Access."

Access is one of the bigger challenges we face. What is the most practical solution? His answer is, "It depends." We each have a different institutional context and different materials. What are you providing access to? The original disk? The original bit streams? Normalized copies? (This is the most common approach.) An emulated environment? Each approach requires a different mode of access.

What do your users expect? Most expect all of it online now, full text searchable, a la Google. This is not practical; we're not there yet.

What can you actually do? Only when you realize the limitations of your resources can you make practical decisions.

You have to choose. Out of the options of fast, cheap, and good: pick two. For access, Shaw has yet to find anything that matches all three.

What's most important for your institution? How much time do you have? How much do you have in the way of resources? Do you want to provide context and rendering and make it user-friendly? You can do it but it won't be fast or cheap.

Shaw provided some web archiving example options, and commented on them:

- Adobe Acrobat as an access mechanism for PDFs is (in Shaw's opinion) usually lousy. Your results may vary.
- HTTrack[90] allows download of local copies of the site and a browser (not good but not bad). It also allows for a local copy of the site in a virtual machine. It's not fast and requires continuous migration of the emulation hardware and software.
- Heritrix[91] and Nutchwax[92] (open source software used by the Internet Archive's Wayback Machine[93]) are not fast or easy. It is the gold standard for web archiving right now, but is not for the faint of heart.
- Archive-It or WAS (Web Archiving Services)[94] is a subscription service portal for archiving websites, but it's not very cheap. Duke University uses it.

For "My Documents" type of incoming material, Shaw outlines the following access method possibilities:

- The most practical approach is emailing the requested content to the users who ask for it, once it's been described. Most researchers don't ask for large volumes of materials.

---

[90] Xavier Roche and others. 2011. "HTTrack Website Copier: Free Software Offline Browser." (Available from http://www.httrack.com/)

[91] Internet Archive. 2011. "Heretrix. Heretrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project." (Available from http://crawler.archive.org/)

[92] Internet Archive. 2009. "NutchWAX (Nutch + Web Archive eXtensions)." (Available from http://archive-access.sourceforge.net/projects/nutch/)

[93] Internet Archive. 2011. "The Wayback Machine." (Available from http://www.archive.org/web/web.php)

[94] Internet Archive. "Archive-It." (Available from http://www.archive-it.org/)

- For larger amounts of content, utilize local media: load it onto DVD and allow access in the reading room.
- Provide a local computer station. They are setting this up now. It will need to include whatever hardware/software support the researcher needs to access content in obsolete formats. Most researchers don't want to travel to your reading room, but we have to make choices.
- For web server access: put the files online and point to them. This is fairly easy to do, even including installation of software to search over content.
- Use a document management system: Sharepoint[95] or other existing systems can be leveraged
- Purchase support via third party systems.

What they are doing may not be eye-candy but it will work. He expects it to improve over time. They are restricting access, using Shibolleth[96] for security, providing specific access to certain researchers. Digital rights management is an arms race. At some point we have to trust researchers that they won't distribute this out. Shaw suggests that we make researchers sign statements, but not let the rights management problem paralyze us. Select an approach and do it; we can always improve over time.

QUESTIONS AND ANSWERS

Q: Do you use aggregate or single file descriptions?
A: It depends on the collection. This is MPLP ("More Product, Less Process") [97] for electronic records; do just enough for what you're dealing with. Go as far as is necessary to provide adequate access to your materials.

Q: What about university business records which require permanent storage?
A: A lot of those are in enterprise storage systems, which ensure no loss until migration. They are safer where they are now than by coming to the archives. If we get them, we scan for sensitive information and flag those.

Q: How do you ensure data integrity when moving content from one system to another? Or are you considering it?
A: The InterPARES project[98] has taught us that authenticity is key. We can checksum files and folders while the content is still on the original disk, and set up write blockers to prevent modifications to files. With enterprise systems, the content is stored in database fields, which makes it a bigger challenge to ensure integrity. Use a file system checker that verifies that the checksum doesn't change during the move.

---

[95] Microsoft. "SharePoint 2010. (Available from http://sharepoint.microsoft.com/en-us/Pages/default.aspx)
[96] Internet2 Middleware Initiative. "Shibboleth." (Available from http://shibboleth.internet2.edu/)
[97] Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *American Archivist*, 68, no. 2 (Fall/Winter 2005): 208-63.
[98] "InterPARES Project: International Research on Permanent Authentic Records in Electronic Systems." (Available from http://www.interpares.org/)

Q: If the content has already been selected out by the donor and put on a flash drive and sent to you, it's already out of context. How do you handle this type of situation?
A: Handle this the same way as with analog content provided out of context: impose some sort of order or arrangement. Maintain authenticity. It's worse when they email content to you; all you have then is the discrete object. Assign it to an existing folder or series if you can. Make a record of how it was accessioned.

Q: How to you deal with resistance from server administrators to transferring content from dubious sources?
A: We find that server administrators are, in general, completely confused about the needs of archiving materials. For example, Duke University has an internal auditing department that had never before encountered what we were dealing with; we scared them. But we were able to work with them to allay their fears, and obtained new contacts. We accession onto a local machine and clear out all security issues before transferring the content to university hosted storage.

Shaw provided a final warning: don't expect electronic records to be correct. Compare the content against other electronic records.

SKELETONS IN THE CLOSET: ADDRESSING PRIVACY AND CONFIDENTIALITY ISSUES FOR BORN-DIGITAL MATERIALS

Erin O'Meara, Gabriela Redwine, Bonita L. Weddle

This was a well-attended session, with about 200 in the audience.

PRIVACY CONCERNS IN THE LAND OF PUBLIC RECORDS: MANAGING APPROPRIATE ACCESS TO ELECTRONIC RECORDS

Erin O'Meara (Electronic Records Archivist, University of North Carolina (UNC), Chapel Hill

UNC is a public university subject to state's public records law (GS 132[99]). Some records may be excluded from public inspection. The ultimate goal is to balance access and privacy.

There was a recent data breach in their School of Medicine; breast cancer patient information was obtained by a hacker. New complex far-reaching data security policies were then instituted. As a result, there were numerous retirements and departures of campus administrators, a new chief of security hired, and a change in the work culture.

Their approach is to work with the record creator to identify groups of records that may contain sensitive or confidential information, but this is not reliable.

They scan records with a bulk extractor or with general command-line regular expressions to identify sensitive markers, such as social security numbers, credit card numbers, or keywords surrounding sensitive records like Personnel, grievance, etc. They are hoping for better tools.

[99] North Carolina General Assembly. "Statutes, Chapter 132: Public Records." (Available from http://www.ncga.state.nc.us/enactedlegislation/statutes/html/bychapter/chapter_132.html)

They need to determine the level of appropriate accessibility to records; perhaps by series.

They provide 3 levels of access:

- Public, on the web
- Public only in reading room
- Restricted to administrative reference (reading room or digital access copies sent to office)

Regardless of the status, material is still listed in the finding aid; access restrictions are indicated if they exist. Access control can be down to the file level. They use the "CRUD" access layers (Create, Read, Update, or Delete). Administrators have all these rights. Curators may have either CRU or CRUD rights for specific collections. Patrons can read specific collections. What is available to the public is only open content (for read-access only). Access control is via metadata in SOLR[100] search implementation over FOXML (Fedora Object XML[101]).

There is a tension between the shift to minimal processing with paper records yet doing more review with e-records. New training and tools are needed, as well as new staffing models.

O'Meara raises several questions:

- Who should be doing this work?
- What are the standards or best practices?
- How do we know the material is ready to be delivered to the public without opening every single file?
- How do we consistently address redaction or restrictions?
- We need a corpus of sensitive patterns to search for in archives or manuscript collections.
- Can we learn from mass digitization efforts regarding privacy and copyright review?

We need to think creatively about rapid and minimal redaction; the level of liability is quite different.

PERSONAL PRIVACY AND FREEDOM OF INFO IN THE DIGITAL AGE: CHALLENGES AND STRATEGIES FOR GOVERNMENT ARCHIVES

Bonnie Weddle (Coordinator, Electronic Records Unit, New York State Archives)

[100] Apache Software Foundation. 2007. "Lucene: Apache Solr." (Available from http://lucene.apache.org/solr/)

[101] Fedora Project. 2005. "Introduction to Fedora Object XML (FOXML), Fedora Release 2.0." (Available from http://fedora-commons.org/download/2.0/userdocs/digitalobjects/introFOXML.html)

Weddle says they don't always get information with the records that show up on their doorstep. In addition, they are being handed policy decisions by politicians that are difficult or impossible to implement. There's a records management/IT divide.

On one hand, there's an expectation that everything should go online, without consideration for privacy issues. On the other hand, there's increased litigation, and deferral of key IT investments. They are balancing competing mandates for access: Freedom of Information Law (FOIL[102]) against the Personal Privacy Protection Law[103].

Their approach: if not restricted, they expose. Some do the opposite: if not mandated, they protect it.

Other laws that impact their work include civil practice law and rules, criminal procedure law, executive law, and general business law, such as the expanding enforcement of the "Martin Act" for financial industry investigations[104] which has existed since 1921. There are state archives and department law memorandums of understanding to consider also, regarding attorney work products and client-attorney communications.

Identifying attorney work products is not easy for non-attorneys, and attorneys themselves do not always agree what the definition encompasses.

Redaction is the process of removing information that is legally restricted, sensitive or classified from records before making them accessible.

An example is the papers from the Department of Law (Office of Attorney General) Elliot Spitzer, composed of 1000 cu, feet of paper and 5 GB of e-records. A cursory review revealed resumes, employee health information, security infrastructure information, and attorney-client communications, in two email archives of high-ranking deputies. These were in PST (Personal Storage Table) files.

They struggled with what approach to use for redaction. Conversion to paper is not feasible, plus content is not then available in digital form. If they convert content to PDF and redact electronically, some metadata is altered. Adobe Acrobat Professional versions 8 and 9 have a built-in tool to support automated and manual redaction, which is widely used and as yet uncracked. It allows for saving "marked for redaction" versions.

Their initial steps were to import a copy of the PST archive into Microsoft Outlook, and to conduct keyword searches on a folder by folder basis. They then converted the result sets to PDF. This reduced 23,000 messages to only 3,000. Training session and supporting materials were developed. Reference services now identify restricted information and the electronic records unit redacts the content.

---

[102] University of the State of New York, New York State Education Department. "Freedom of Information Law." (Available from http://www.oms.nysed.gov/foil/)
[103] New York State Archives. "Public Officers Law Article 6-A: Personal Privacy Protection Law." (Available from http://www.archives.nysed.gov/a/records/mr_laws_po6A.shtml)
[104] Robert A. McTamaney, Washington Legal Foundation. "New York's Martin Act: Expanding Enforcement in an Era of Federal Securities Regulation," Legal Backgrounder, 18:5, Feb. 28, 2003. (Available from http://www.wlf.org/upload/022803LBMctamaney.pdf)

They use a multiple team approach. One team includes an electronic records unit archivist, records manager or reference-rotation archivist. The second review team (which includes the head of reference and a designated reference-rotation archivist), review the work of the primary review team, and consult with the department of law team and the state educational department office of council.

During the review process, they burn files that are marked for redaction to read-only CD disks and send them to the department of law for final review.

Redaction remains labor-intensive. It means more work for professional staff, and less for clerical staff. The review team approach works. Communication with counsel is essential.

E-redaction creates copies which must be managed properly. But we need better tools: proximity search capability and social network analysis. Legal proceedings take precedence and may override everything else.

Weddle's recommendations:

- Know your records creators. E-records are records, and must be accessioned and managed. Several contained viruses.
- Know the laws, particularly the FOIA, HIPAA[105] and FERPA[106]/Buckley Amendment[107]. Also know your applicable state legislation with regards to personal privacy, data breach notification, freedom of information, and other statutes.
- Consult with others as needed, including government bodies and legal counsel
- Be assertive.
- Know the technology. Some software removes or obscures restricted information. Some are better than others. Using Adobe Acrobat to draw black boxes over restricted information NEVER works. One application they tried rendered social security numbers visible as images instead of deleting them. Tools are only as good as their users. Research and test.

NEW AND DEVELOPING TOOLS AND SOFTWARE

THE FUTURE IS NOW: NEW TOOLS TO ADDRESS ARCHIVAL CHALLENGES

Dr. Mark Conrad (NARA), Kenton McHenry, (National Center for Supercomputer Applications), Maria Esteva (Texas Advanced Computing Center), Wm. E. Underwood, Jr. (Georgia Tech Research Institute)

This session was extremely well attended, with about 230 people, standing room only. Interestingly, all the presenters have Ph.D.'s in Computer Science.

---

[105] U.S. Department of Health & Human Services. 2011."Health Information Privacy: HIPAA." (Available from http://www.hhs.gov/ocr/privacy/)
[106] U.S. Department of Education. 2011. "Family Educational Rights and Privacy Act (FERPA)." (Available from http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html)
[107] Electronic Privacy Information Center. "Family Educational Right to Privacy Act (Buckley Amendment): 20 USC S. 1232g." (Available from http://epic.org/privacy/education/ferpa.html)

THE ISDA[108] TOOLS: COMPUTATIONALLY SCALABLE FILE MIGRATION SERVICES TO KEEP YOUR FILES CURRENT[109]

Kenton McHenry (National Center for Supercomputer Applications)

The problem is that there is an abundance of file formats, many proprietary. It seems as if every software creates its own formats. All have unique ways to represent their data. McHenry has counted 144 3-dimensional formats, but there's many more. We need to convert all these files to an open/standardized format that is likely to be supported for some time. We cannot support them all; however, conversion is lossy (some content is lost in conversion).

In 2009, NCSA (National Center for Supercomputing Applications[110]) developed Polyglot[111] as an extensible, scalable, and quantifiable method of converting between formats. Using the AutoHotKey scripting language, they created GUIs around other applications. Polyglot has a simple work flow referred to as an Input/Output (I/O) graph. PolyGlot also allows them to open a file before and after conversion and make comparisons to determine the amount of loss.

McHenry described various ISDA (Image Spatial Data Analysis Group) file migration tools:

1) They created a "Conversion software registry" utilizing any and all available third-party software. It is a database listing input/output formats. About 2050 software conversion tools are covered so far. The workflow is a simple I/O Graph: users can select input and output formats and find the conversion software needed. Right now access is by browse instead of search.
2) Software servers using "Imposed code reuse:" they wrap third-party software to provide API (Application Protocol Interface) –like access so the software service can be called. This allows them to share the functionality of many kinds of software over the web, replacing the original interface with one that is uniform across all the included software. Tasks are carried out on a remote machine. This allows any desktop application to become a cloud based service. Basically, this is software functionality sharing. Users can select an application they currently support, select the task, input format and output format, and the conversion is done remotely, via point and click functionality. This will be available by URL as well, so computer applications can make use of the service, using curl[112] and wget[113] calls. This service

---

[108] National Center for Supercomputing Applications. "ISDA: Image, Spatial, and Data Analysis Group." (Available from http://isda.ncsa.illinois.edu/drupal/)

[109] Kenton McHenry, Rob Kooper, Luigi Marini, and Michael Ondrejcek. 2011. "The ISDA Tools: Computationally Scalable File Migration Services to Keep Your Files Current." (Presentation available from http://www.slideshare.net/NARACAST/the-isda-tools-computationally-scalable-file-migration-services-to-keep-your-files-current)

[110] University of Illinois. "NCSA: National Center for Supercomputing Applications." (Available from http://www.ncsa.illinois.edu/)

[111] Image Spatial Data Analysis Group. "Polyglot." (Available from http://isda.ncsa.illinois.edu/drupal/software/polyglot)

[112] LinuxManpages.com. "curl." (Available from http://linuxmanpages.com/man1/curl.1.php)

[113] LinuxManpages.com. "WGET." (Available from http://linuxmanpages.com/man1/wget.1.php)

is very robust; the server doesn't die if things go awry. If the software can transform the file to something that can be viewed via the browser, it will.

3) "Polyglot" listens for software server broadcasts, catalogues what they do in terms of input/output, identifies conversion paths, and chains together software. It is a monitoring service that manages multiple server sources, and can jump across software packages and machines to find what's available.
4) "Versus" compares file content before and after conversion. It will be a Java library/framework. They have yet to add measures to determine loss during conversion, but are using weighted information. They're working to determine which conversion preserves the most for each type of file. Multiple types of measurements are used, such as the light fields measure, which emphasizes shape through silhouettes, and the spin image measure, which emphasizes shape through relative vertex positions. No one measure is perfect; different measures are more important for different applications.

There are other tools as well: image utilities, 3-D utilities, and a CyberIntegrator.[114] Software development is being funded by NARA.

MAPPING ARCHIVAL PRACTICES TO VISUALIZATION

Maria Esteva (Texas Advanced Computing Center)

(Much of this presentation is covered in a recent online article.[115])

They are developing methods to allow archivists to examine and process large electronic records collections using visual analytics: data analysis methods, visualization and interactivity. Visualization provides discoveries and inferences. These are tools to assist users in making decisions, as they highlight significant features, shown as abstract visual features. The focus is on narrowing information down without losing detail.

Esteva provided a scenario: a visual finding aid. In this visualization, the color of blocks (each block represents a directory of content) indicates the amount of content relevant to the search term. Tools on one side provide statistics. When homing in on a subset, the size of a different colored segment shows the second search term results. The tool allows her to select the subcollection containing the most content related to her search terms.

In the content under this analysis, there was information captured as tags in file names. Based on these tags and the directory structure, the software categorized the content and assigned different colors to different categories. One can visually see which collections contain more content, and how structured they are. The display also includes tree structures to assist in browsing.

---

[114] ISDA. "Cyberintegrator." (Available from http://isda.ncsa.uiuc.edu/cyberintegrator/)

[115] Maria Esteva, Weijia Xu, Suyog Dutt Jain, Jennifer L. Lee, and Wendy K. Martin. 2011. "Assessing the Preservation Condition of Large and Heterogenous Electronic Records Collections with Visualization," *The International Journal of Digital Curation*, 1:6, 2011. (Available from www.ijdc.net/index.php/ijdc/article/download/162/230)

Formats were classified into 20 classes (including unknown files), assigned colors and sizes that indicate the amount of content in each file format. Archivists can view one directory's information at a time.

To show the levels of preservation risk (utilizing Stanford's criteria[116]) they did a proof of concept. While mousing over each directory, information about the content can be viewed on the right side, with each level of risk color coded and sized by amount of content.

Arrangement was made using terms extracted from labels and filenames, and was based on four criteria: spatial, naming, temporal and sequential. For description, they classified content in categories again, and assigned colors. The display was very complex at this level, and required creation of a list of terms which was not user friendly.

For complex records, the archivist needs to determine what *is* a record. This software assists in visualizing the structure and characterization of files and identifies patterns. Once the boundary of a record is determined, one can assess preservation risk for the record as opposed to the formats.

Behind the scenes, the software extracts metadata, puts it in a relational database management system (RDMBS), and assigns classes and categories. Then it is possible to perform queries, aggregations, data mining, statistical calculations, and regular expression matching. The software is a combination of data transfer, computing, and display systems.

The visual representation is pixel based rendering, working with percentages, not one-to-one representation. To assess accuracy, they built in multiple ways to cross-reference information.

Esteva showed an image of someone interacting with information in 6 displays on a wall-sized screen, dragging and moving things around using gestures (not touch). She called this a vision of the archivist work of the future.

TOOLS FOR FILE TYPE AND RECORD TYPE IDENTIFICATION[117] [118]

William Underwood, Jr. (Georgia Tech Research Institute)

Archivists need the capability to identify file formats for:

- ensuring compliance with record transmittal agreements

[116] Richard Anderson, Hannah Frost, Nancy Hoebelheinrich, and Keith Johnson. 2005. "The AIHT at Stanford University: Automated Preservation Assessment of Heterogenous Digital Collections," D-Lib Mangazine 11:5, December 2005. (Available from http://www.dlib.org/dlib/december05/johnson/12johnson.html)

[117] William Underwood. 2011. "Tools for File Type and Record Type Identification." (Presentation available from http://perpos.gtri.gatech.edu/presentations/Underwood-Archival-Tools-SAA-Chicago-2011.pdf)

[118] William Underwood. 2010. "Grammar-Based Recognition of Documentary Forms and Extraction of Metadata," International Journal of Digital Curation, 5:1 (2010). (Available from http://www.ijdc.net/index.php/ijdc/article/view/152)

- accessing files
- conversion to current or standard file formats
- archive extraction
- password recovery and decryption, and
- repair of damaged files.

Underwood provides definitions:

- A file format is a set of rules for encoding and decoding data or computer instructions in a file.
- A file type is a class of files with the same file format.
- A file format signature is invariant data in a file format that can be used to identify the file type or format of a file.

External file format identifiers are file name extensions, metadata stored in the operating system, MIME types[119] and the PRONOM[120] Persistent Universal Identifier (PUID). The Linux "file" command and Magic File[121] are likely the most widely used tools for file format identification. The Magic Number is the term used for the concept of an internal file format signature. The "file" command applies tests for Magic Numbers contained in the Magic File to files to determine file type and relevant metadata.

Limitations of this approach include:

- It's difficult to update the tests
- Tests may give conflicting results and must be properly sequenced.
- Test for Magic Numbers are not a one-to-one match with file types.
- It tests output metadata as well as file type
- Its tests for character set and language of text files needs refinement
- Only a few tests for MS Windows file types are included
- Tests for magic numbers have not been rigorously tested.

Underwood has developed extensions for the "file" command and Magic File to overcome these limitations. He showed an example of a Magic Test for Broadcast Wave Format Version 1 (for which he built an interface). With the results of their testing and software development, they are helping to improve the PRONOM registry. Thus far, they have added about 200 signatures, and about to send another 50. Collaboration will continue.

The motivation of their research is twofold:

- Metadata extraction is a critical aspect of the ingestion of textual e-records into digital archives and libraries

---

[119] Internet Assigned Numbers Authority. "MIME Media Types." (Available from http://www.iana.org/assignments/media-types/index.html)
[120] The National Archives. "PRONOM: the Technical Registry." (Available from http://www.nationalarchives.gov.uk/PRONOM/Default.aspx)
[121] LinuxManpages.com. "MAGIC." (Available from http://linuxmanpages.com/man4/magic.4.php)

- Metadata is needed to support description of individual e-records and aggregations of these records and to support search and retrieval of records.

There are multiple document types. NARA calls them record types. A "documentary form" consists of both the intellectual form and physical form:

- The intellectual elements are those terms or semantic categories that are common to a document type.
- The intellectual form is the rules that characterize the possible combinations of intellectual elements.
- Physical elements are the physical attributes of the intellectual elements
- physical form is the rules that characterize the layout of the physical elements.

He includes several other definitions as well.

The question is, how do we recognize these automatically, when using text extraction?

He manually constructed grammars for document types, augmented with semantic rules which identify the metadata, for extraction. He then creates a parse tree based on the semantics of the document. Metadata is extracted for item description and indexing; this is used to generate item descriptions, folder descriptions, and series descriptions.

After developing the rules and grammars for 14 documentary forms, he merged the grammars and converted them to SUPPLE Parser[122] [123] Notation and added in semantics. Then this functionality was added as an option to PERPOS[124] for automatically describing contents of containers.

Underwood then tested and evaluated the software against 14 document types. Of 159 documents, 106 were correctly recognized; most of the ones not recognized were ones for which they did not have a grammar (results of tests are online[125]). Document recognition is dependent on the information extraction part which performs at 92% accuracy, but needs to be improved.

They are continuing this work. It's not yet practical. They need a better model and theory for how to automatically recognize the intellectual content of material.

Questions yet to answer:

- Can the intellectual elements of documentary forms be learned without a teacher?

---

[122] University of Sheffield. "General Architecture for Text Engineering (GATE): SUPPLE Parser." (Available from http://gate.ac.uk/sale/tao/splitch17.html#sec:parsers:supple)
[123] Robert Gaizauskas et al. "SUPPLE: A Practical Parser for Natural Language Engineering Applications." (Available from www.dcs.shef.ac.uk/intranet/research/resmes/CS0508.pdf)
[124] Georgia Tech Research Institute. "Presidential Electronic Records Pilot System (PERPOS); Phase 2." (Available from http://perpos.gtri.gatech.edu/)
[125] Georgia Tech Research Institute. 2007. "PERPOS: Information Extraction." (Available from http://perpos.gtri.gatech.edu/ie/)

- Can grammatical induction be used with examples of a document type to induce a grammar automatically?
- Can this be extended to include physical elements of documentary form and layout?

Droid lists only 813 file types. We need to be able to recognize many more.

ARCHIVEMATICA[126]

This session was entitled "Practical Approaches to Born-Digital Records: What's Coming Next," but its focus was on a single software package.

Although this session was held on Saturday, the last day of the conference, approximately 300 people attended. The Archivematica software under development was described first, and then beta-testers from three different institutions reported on their experiences with it.

OVERVIEW

Peter Van Garderen (Artefactual Systems Inc., Vancouver)

Artefactual Systems is in British Columbia, and consists of 8 employees, most with library science degrees. They are creating open source solutions and providing consultative services. They make their money off of consulting, and adding things and tweaking what they've already developed open source, for clients. It's a sustainable approach. They based their system (Archivematica) on the OAIS model, and will have a beta release at end of this calendar year. This started out as a proof of concept. Three years ago, there was no Hydra[127] project and no Archives Space[128] in development. Since there were no solutions, they sought to fill a gap.

Archivematica software is released under a Creative Commons[129] license. There're plenty of open source utilities floating around, so they mapped available tools to the OAIS concept slide and started stitching it together and filling in gaps.

Archivematica is an integrated stack of open source applications. They created Debian[130] packages for the tools, and keep them up to date with each release to support the software. It's an alternative to the Fedora based approach, which has high overhead. Van Garderen states that "We were doing microservices before they were called microservices."

Archivematica uses a "watched directory approach" for ingest. As soon as files are dropped into the directory, scripts collect checksums, extract metadata, and run various microservices, which are selected by the user. All services are optional, and they include virus checks and quarantines and a rights dialog kit. They extract zip and tar files, and

---

[126] Artefactual Systems Inc. "Archivematica Open Archival Information System." (Available from http://archivematica.org/wiki/index.php?title=Main_Page)
[127] "Hydra." (Available from http://hydraproject.org/)
[128] "Archives Space." 2011. (Available from http://www.archivesspace.org/)
[129] "Creative Commons." (Available from http://creativecommons.org)
[130] Software in the Public Interest, Inc. 2011. "Debian: the Universal Operating System." (Available from http://www.debian.org/)

capture a snapshot of the original directory structure, because it has meaning in itself. There are no standards for accession records yet, so they're using the Archivists Toolkit version.

There are three basic preservation strategies:

- Emulation
- Migration
- Normalization, which is to apply migration of content at ingest to an archival format

Archivematica does normalization (though it can be turned off), retains a copy of the original, and generates web derivatives on ingest.

Either the ingest works or doesn't; errors are output into another watched directory. They use XML configuration files to chain workflows together. Content goes from ingest to access in a manner compliant with OAIS.

The intent is to manage complexity as simply as possible. Archivematica provides a simple interface; there is no need to be familiar with command-line options or complex metadata. A 12-minute screencast with a focus on ingest is available online.[131]

They work with limited budget, limited scope, and tight deadlines. With each iteration, they are adding functionality and usability.

To use, install Archivematica on a client machine, bring the files into that system via system transfer or off hard drives: the software requires a local file manager.

They are running it on Ubuntu Linux, but you can run it in a virtual machine on other operating systems (some issues have been encountered with this approach). The entire system can be run from a USB key. Archivematica includes about 25 open source tools at this point. The Master Controller Program (MCP) server hands off services to other processing clients, to speed the work.

Artefactual Systems does not want to offer the storage system. Mostly they are working with network storage, but are looking ahead to clouds and LOCKSS methods of storage as well. They also don't want to be the metadata management and web discovery tool. Storage and access are not included.

Archivematica creates a submission information package (SIP), hands off the archival information package (AIP) to storage, and also creates the dissemination information package (DIP) for the consumer. They are working on monitoring and syncing with stored content, hoping to leverage OAI[132] or AtomPub[133] for this. They are also working to synchronize with format registries.

---

[131] Artefactual Systems. "Archivematica." (Available from http://www.youtube.com/watch?v=dFPtDA4nAPY)
[132] "Open Archives Initiative." (Available from http://www.openarchives.org/)
[133] "Atom Publishing Protocol." (Available from http://atompub.org/)

There is much legacy rescue work needed on born-digital content in archives. Much work must happen pre-ingest (they call it transfer processing).

What's the best method to approach that? Right now there is no nice neat import. Most of what is deposited is in chaos. They're considering a pre-ingest processing workstation, and are looking at Curator's workbench[134] and digital forensics tools. The constraint is that they would need to bundle this into the Archivematica stack and include it in the dashboard, which is a challenge. The end goal is to provide high-quality preservation packages.

They do apply normalization strategies, and some conversion to open formats. They keep the original, and create PREMIS[135] records for everything. MODS, EAD, and Dublin Core (DC) metadata are accepted. They will create a data entry form for DC, but no more. They are using METS metadata[136] to tie all the content together.

No existing applications are doing rights metadata well yet. The goal is for the rights metadata in the accession to come along for the ride.

They also need to figure out what to do with unstructured transfers of digital documents. The user needs to be able to right-click and create SIPs, but then the archivist is required to make sense of all the content. There's no intellectual control. At some point there will be physical file system constraints.

Van Garderen thinks a good approach would be to add a microservice providing an index of the folder selected, pulling keywords and pattern matching for privacy/security sensitive information, and collecting a list of PDFs included that have not been OCR'd (processed with an Optical Character Reader to extract text for search support). It should be possible to feed in institution-specific codes for which to search. Then a file visualization analysis could, for example, provide relative sizes of subdirectories to help the archivist determine what should and should not be included in the storage.

There will be a way for institutions to sync to new preservation rules and formats in their own registry. The Artefactual registry will interoperate with and piggyback on other format registries.

The City of Vancouver Archives, University of British Columbia Library, the Rockefeller Archive Center and others are using the system already; there are 20-30 pilot testers.

There have been over 10 workshops in last year, though the system is still in alpha, with the next (4th) release in December 2011 (the first was in Feb. 2009).

2011 development priorities include:

- Building dissemination package uploaders for CONTENTdm and possibly XTF

---

[134] UNC Libraries. "Curator's Workbench." (Available from https://github.com/UNC-Libraries/Curators-Workbench)

[135] Library of Congress. "PREMIS: Preservation Metadata Maintenance Activity." (Available from http://www.loc.gov/standards/premis/)

[136] Library of Congress. "METS: Metadata Encoding & Transmission Standard." (Available from http://www.loc.gov/standards/mets/)

- An ingest uploader from DSpace[137] exports and EAD exports from Archivist's Toolkit
- Format registry integration with the Open Planets Foundation[138]
- Email with attachments preservation plain
- Indexing of the archival package
- Development of Rest[139] APIs & URIs.[140]

ARCHIVEMATICA AT THE CITY OF VANCOUVER ARCHIVES

Glen Dingwall

Dingwall says they have been using Archivematica since its beginning in 2008, for the city's electronic records and document management system. This included content from the Olympic games, which was almost all digital. They have networked cache storage; about 50 TB. Their installation of Archivematica is still kept in a test environment. They started with a virtual appliance on a PC desktop, then a local area network on 4 machines. There was a noticeable improvement when taken off virtual installation: much quicker, and no crashing. Archivematica functions as a pipeline – nothing sits in it long, so if it goes down, they haven't lost anything. They're testing workflow, file normalization, and scalability.
For example:

- Does the workflow make sense?
- Does the dissemination package have everything it needs?
- Can we take a submission package with arbitrary components and run it through successfully?

Lately they are testing file normalization, which involves making test sets of all kinds of documents. Next they will test scalability. The next collection to test has over 25 TB of content.

Thus far, there has been a failure to remove some files during ingestion, and hidden files were ingested by accident.

Future work includes:

- Setting up external normalization paths
- Submission information package creation
- Exception handling

MS Word is currently transformed to PDF using Open Office[141], which takes two steps (less than ideal). The transformation is lossy, especially if there is a lot of formatting in the original document. Utilizing external tools for transform may be better.

---

[137] "DSpace." (Available from http://www.dspace.org/)
[138] "Open Planets Foundation." 2011. (Available from http://www.openplanetsfoundation.org/)
[139] Microformats Community. 2009. "Rest/URL: URL Conventions." (Available from http://microformats.org/wiki/rest/urls)
[140] Network Working Group. 1998. "Uniform Resource Identifiers (URI): Generic Syntax." (Available from http://www.ietf.org/rfc/rfc2396.txt)
[141] Open Office. 2011. "Open Office." (Available from http://ooodocs.org/)

When dealing with private donors, the incoming digital content is unstructured. We need tools to create submission information packages (SIPS) from unstructured collections, add access rights and restrictions, and any other metadata necessary to provide context and usability. Archivematica can't do that for you.

Other necessary improvements would include the ability to pull problem files out of the submission set, instead of the current situation of having the entire process shut down if there's a problem with a single file. Also Archivematica must be able to normalize email attachments while retaining connections between the email and the attachments.

Dingwall says that Archivematica is not an out-of-the box solution. Users don't need an IT (Information Technology) background, but they do need technical literacy and willingness to test and explore and learn. It is not a total preservation solution, but just seeks to fill a gap, to help put together something we are capable of preserving longer.

ARCHIVEMATICA AT THE INTERNATIONAL MONEY FUND (IMF)

Paul Jordan

The IMF installed Archivematica in December 2009, using it for testing and prototype development. They didn't have a lot of formats at first, but this past year was much more interesting, as they are using subsets of real collections from actual donors. They woefully underestimated the number of documents they would get; the files date back to 1980, and they couldn't identify many with JHOVE or DROID. Some of the files are older than Jordan is. They have no idea what they are, as they have no file extensions.

Another problem they have is scale. They have 2.5 million discrete lines of text simply listing the files in the network system. This list by itself was so huge it was unmanageable. Transfer times were long for moving content off external media, and only about a quarter of some the media is still readable, due to data loss.

This is a good reason to go out and talk with donors to archives: we need better transfer routines for digital records. At IMF they have very limited IT resources, but a power user was able to get Archivematica working. He says the software is not yet ready for production. It needs support for PDF/A (archival version of PDF) and email attachments. It is very flexible, and supports any work flow.

Classification is a huge issue. At IMF, they are very security conscious, and have a full time declassification archivist. No paper content can be made available until classified content is removed and verified the removal; it will be the same with digital content. Currently Archivematica automatically creates a dissemination package (DIP) upload with derivatives. They need another step added, to remove classified content. That part is reasonably easy, as it could be set up as a microservice. Beyond that, things become more problematic.

Some documents they know will have classified status; that's easy. Older records are more difficult. They have 40-50 TB of data which is up to 27 years old. Documents on their share drive may have no notation. Email may have one classification and the attachment another. This is a problem, as they don't want to release classified information unintentionally.

They plan to add classification steps, and are hoping provenance will help determine the level of classification. They are also hoping to isolate subdirectories that should be classified, with the expectations of going back to do item-level classification later. Jordan is also hopeful that the upcoming full-text indexing in Archivematica will speed this process for identification.

In paper collections, they add a withdrawal notice where something is redacted. Jordan is not sure how to manage this in the digital environment.

The more classified the content, the more restrictive the data storage needs to be. They will need a large amount of secure storage space for data processing.

They're considering a digital withdrawal notice for classified information; a researcher may subscribe to it, so when it is declassified, the researcher will be notified with a link to the document. This functionality will provide real-time documentation of which documents researchers are interested in.

ARCHIVEMATICA AT THE UNIVERSITY OF ILLINOIS ARCHIVES

Angela Jordan

Jordan said they are a small institution with little IT support. They had two failed attempts in installation using virtual box on an older computer. They finally installed it on a separate computer, and developed guidelines for installation.

They also developed a template for work flow. There were some problems on input. They evaluated Archivematica against 9 record series. When it worked, it worked beautifully. If a single file failed, the entire folder failed. Error messages were not very useful. This impacted their ability to ingest content in a timely manner, as they had to ingest file by file.

Sometimes it would stall on ingest; the delay could last several hours or a day, with no recognizable reason. They had to stop it sometimes and try again.

Jordan said they chose not to include it in their production, because currently when it fails, it fails badly. Once the issues are worked out, Archivematica could be useful.

Jordan noted that many small institutions lack the hardware or technical capabilities for this software. It requires lots of RAM memory and updated computers. Installation requires command-line work and is not user-friendly. It is best run from a dedicated server, which most small institutions won't be able to provide.

Jordan suggests that we need a collaborative effort between archivists and others to develop sustainable solutions for small institutions, such as a hosted process that institutions could contract with on an annual basis.

CHANGES AND ADD-ONS FOR ARCHIVISTS TOOLKIT

ARCHIVESSPACE UPDATE

Mark Matienzo (Digital Archivist in Manuscripts and Archives, Yale University, and Technical Architect for the ArchivesSpace project) and Katherine Kott (Manager of Strategic Digital Projects and Organizational Development, Stanford University, and Development Manager for ArchivesSpace) reported on the ArchivesSpace initiative.[142]

They had a planning grant for the merger between Archivist's Toolkit and Archon[143] for the last 7 months of 2009. Now they have a Mellon funded planning grant that began in January 2010. They have developed draft specifications technical planning and design, and performed a survey regarding implementation and deployment.

They've found that the majority of respondents have 3 or fewer archives, and that institutions want the software on their own servers, with little technical support.
A sustainability and business model desiderata was developed by Beth Sandore (UIUC), Luc DeClerck (UCSD) and David Millman (NYU).

Project dimensions involve developing a new unified system, creating tools and a support service to migrate existing implementations of Archon and Archivist's Toolkit to the new system, and selection of new home. They may or may not use the same code. Delivery and access and web management will come from the Archon implementation.

They will consider the option of setting up multiple instantiations from same hosted repository. There will be a version for consortia as well as stand-alone versions. ArchivesSpace will have a fee-based membership structure and governance structure.

In 2012 they plan to do the core programming and create a governance body. In 2012-2013, they'll be doing quality control, beta testing and rollout, and create documentation and migration methods.

Archivists involved are Chris Prom (University of Dundee), Scott Schwartz (Sousa Archives and Center for American Music), and Brad Westbrook (UC San Diego Libraries), though there is a long list of people involved.
They have set up an ArchivesSpace Google group[144] to involve participants in the technical review process, and in the planning and integration process.

END TO END: AUTOMATING DIGITAL OBJECT WORKFLOW

Jennifer Waxman and Nathan Stevens (New York University)

---

[142] ArchivesSpace. 2011. "Building a Next-Generation Archives Management Tool." (Available from http://www.archivesspace.org/)
[143] University of Illinois, UIUC Library. "Archon: the Simple Archival Information System." (Available from http://www.archon.org/)
[144] Google. 2011. "Google Groups: ArchivesSpace." (Available from http://groups.google.com/group/archivesspace/)

NYU (New York University) is using Archivists Toolkit for digital objects, importing work via a plugin[145], which auto-generates handle-based URIs. Their publication unit associates URIs with derivative files according to a work order, and publishes the finding aids. The work order is a tab-delimited text file which includes the resource identifier, persistent identifier, and is human-readable. Each URI will be a landing page (this is still under construction).

An example collection already online is the Harry Randall: Fifteenth International Brigade Photograph Collection.[146] They had two sets of photos digitized by vendors at different times. File names of derivatives are based on the container indicator of the analog instance. The input file contains a list of derivatives and directives of how to create digital images. The total process takes less than 5 minutes.

ATREFERENCE[147]

Marissa Hudspeth, Rockefeller Archive Center

Hudspeth reported on an effort to develop reference service functionality add-on to Archivist's Toolkit, customizing open source software for the archival community. The intent is to replace the discovery functionality for managing patron registration and duplication services. This software consolidates any duplicate patron information stored, and automates data capture and manipulations for researcher services.

There are 5 phases of development:

1) Patron registration, which tracks research visits, publications, funding, etc. This is available now.
2) Duplication services are under development. It manages requests for duplication, fees and shipping rates, calculates cost estimates, tracks duplication requests, creates invoices, and tracks the number of copies requested by patrons. It should be out by the end of December 2011.
3) Retrievals, bar-coding and use tracking. This will track use of materials by patrons, automate chargeouts, and add barcode functionality to accession records and folder and item levels of resources. It will provide the ability to electronically submit retrieval requests at each level (based on the finding aids). This is due to be out in the summer of 2012.
4) Reference requests will manage off-site requests and schedule appointments to access reading room. In addition it will store a history of requests, etc. (The date available is as yet unknown.)
5) Web interface and personalized user accounts for patrons. This will enable a patron to do their part of registration and requests via the web so archival staff don't have to do it for them. (Again, the date available is as yet unknown.)

---

[145] Archivist's Toolkit. 2011. "BatchDO2 Plugin." (Available from archiviststoolkit.org/node/ 246)
[146] New York University, Tamiment Library & Robert F. Wagner Labor Archives. "Harry Randall: Fifteenth International Brigade Photograph Collection ALBA PHOTO 011." (Available from http://dlib.nyu.edu/findingaids/html/tamwag/randall_photo.html)
[147] Github Social Coding: Rockefeller Archive Center. "ATReference." (Available from https://github.com/RockefellerArchiveCenter/ATReference/wiki)

Hudspeth said that they particularly want feedback on the statistics and reports archivists need to generate, and asked for samples. Members of the audience suggested that they add tracking of the length of visits, and the amount of time an archivist spent with the patron on each visit. Researchers' publications are also not currently linked to which collections were researched.

Four of the institutions represented in the audience have installed it and are using it. Once installed, users must get updates from the ATReference github site instead of from the Archivists Toolkit site. When asked how the upcoming merger of Archon and Archivists' Toolkit will impact this software, Hudspeth seemed at a loss. Since the merged software may not even use any of the original code, there may be a huge impact.

ISO STANDARDS FOR CERTIFYING TRUSTWORTHY DIGITAL REPOSITORIES ISO/DIS 16363 AND ISO/DIS 16919

Marc Conrad (NARA Archives Specialist, Applied Research Division)

Conrad reported on the current status of the development of ISO standards (16363 and 16919). The proposals by the working groups are being reviewed by the CCSDS (Consultative Committee for Space Data Systems), and Conrad notes that in this presentation, he is only representing himself.

The effort began with the development of the OAIS reference model in 2002, followed by the RLG report on "Trusted Digital Repositories."[148] Then in 2007, OCLC, NARA and the Center for Research Libraries published the "Trusted Repositories Audit and Certification Criteria and Checklist (TRAC),"[149] and then a draft recommendation for requirements for bodies providing audit and certification for those trusted digital repositories.[150]

Why?
Basically, because everybody loved OAIS and claimed they were compliant, but there were no metrics for it and no established understanding of it.

CCSDS is doing the final editing and then it will be published as ISO standards. Test audits in June and July 2011 included 3 institutions in Europe and 3 in the US. Two of these were large space science repositories, two were large digital libraries, and one of the remaining

[148] Research Library Group - OCLC. 2002. "Trusted Digital Repositories: Attributes and Responsibilities." (Available from http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf)
[149] "Trustworthy Repositories Audit & Certification: Criteria and Checklist." 2007. (Available from http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf)
[150] CCSDS. 2010. "Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories: Draft Recommended Practice, May 2010." (Available from http://wiki.digitalrepositoryauditandcertification.org/pub/Main/WebHome/RequirementsForBodiesProvidingAuditAndCertification-SecRev1.doc)

two was the Kentucky Digital Libraries and Archives. More information is available online.[151]

ISO 16363 is based largely on TRAC, and has 3 parts: organization infrastructure, digital object management, and infrastructure and security risk management. The hierarchy of ISO standards is concerned with good auditing. This one (ISO 16919) exists to ensure good practices when carrying out the audits using ISO 16363. The recommendation versions sent to the ISO will be available free from ccsds.org (the first one is already available[152]).

The next steps are for the working group that develops standards constituted as the Primary TDR (Trusted Digital Repository) Accreditation Board (PTAB) to set up a multinational organization and work out legal constraints, conduct audits and train auditors around the world, and to keep standards current. After all, ISO standards expire.

Soon if you claim OAIS compliance, look out.

The audience had many questions.

In Europe, there are 3 possible levels of audit approval:

- Bronze: Data seal of Approval – a monitored self-audit
- Silver: Bronze self audit with 16363 or German standard metrics and published evidence
- Gold: Have an outside team perform audit in accordance with ISO standards

When asked how it's possible to assess onsite the software integrity and business rules, Conrad stated that they don't audit software. They audit people, infrastructure, policies, and how the infrastructure is used. The audit requires that the repository is meeting the needs of designated community; whatever the agreement is with that designated community is determines what the results should be, such as what percentage of content still available after X number of years.

It took the 6 sites audited an **average** of 400 hours to prepare for the test. All needed improvement. There will be a cost for auditing, including travel. It will be a cost recovery model, with some things amortized. An audit requires at least 2 auditors on site for 2 days. Conrad is not sure how many will want to go through the process. None of the repositories he's ever worked in get a passing grade.

It will be helpful to use the standard to form the basic framework when forming a new program, but an institution will need to go beyond that for preservation.

---

[151] David Giaretta. 2011. "Welcome to the Digital Repository Audit and Certification Wiki." (Available from http://wiki.digitalrepositoryauditandcertification.org/bin/view)

[152] CCSDS. September 2011. "Audit and Certification of Trustworthy Digital Repositories: Recommended Practice. CCSDS 652.0-M-1." (Available from http://public.ccsds.org/publications/archive/652x0m1.pdf)

There is no training yet specified for auditors. Higher level ISO standards require an annual review, and a recertification every 3 or 5 years. The ISO standard will be more specific than TRAC.

CONCLUSIONS

Mass digitization of large collections is a cost-effective and valuable tool for providing access to content. By leveraging this approach and implementing item-level description for select content based on research value, we can provide better service to our patrons, who are increasingly expecting online access to our materials. I highly recommend investigation into methods of leveraging existing metadata for better access and usability, as evidenced in the SNAC project; we should begin by contributing finding aids to this effort. Collaborative projects such as these will improve access points and our ability to meet patron needs. Increasingly it appears that metadata efforts should be strategically focused on that which can be leveraged for linked data and other semantic web implementations.

Archivists both here and abroad are facing a major shift into managing digital records. Given that history did not end with the advent of digital records, we need to have a realistic and well-planned team approach to management of born-digital content coming into our special collections (and if possible, our archives). The approach should be informed by efforts already expended by others, and appropriate personnel should be encouraged (and funded) to obtain necessary training. If the research value of our distinctive special collections is to be a major focus of libraries in the days to come, it behooves us to selectively obtain and curate important digital and mixed collections regarding recent historic events, groups, institutions, and individuals. The complexity of the task recommends careful planning and implementation, but should not sway us from meeting the challenge.