

## **Report on the Coalition for Networked Information Fall Task Force Meeting**

**December 14-15, 2009, in Washington, DC**

by Jody L. DeRidder

The Coalition for Networked Information (CNI) is a joint initiative of EDUCAUSE and the Association of Research Libraries (ARL). Its primary task is to promote the use of networked information technology to advance education and research. Within that overarching goal, CNI focuses on three central themes: developing and managing valuable networked information content, developing strategies and collaborations to support the unfolding needs of scholarship and education, and developing technical infrastructure. CNI is completely supported by member institutions, who are invited to send two representatives to each task force meeting: a senior executive from the library, and a senior executive from information technology. Task force meetings are held in spring and fall; the fall meeting is always held in Washington, DC, but the spring meeting travels. Members may be non-profit or for-profit, though they are heavily weighted with higher education institutions, many along the eastern seaboard. In deference to this fact and the current restrictions on travel funding for many of the members, the spring meeting next year will be held in Baltimore, Maryland.

Apart from member meetings, CNI personnel also publish reports on hot topics, occasionally convene invitational or public workshops, co-sponsors other meetings relevant to their agenda, and regularly co-sponsors a conference in partnership with JISC (Joint Information Systems Committee) and UKOLN (UK Office for Library Networking) as part of an ongoing collaboration with these counterparts in Europe and Australia. CNI collaborates with key funding agencies and contributes to National Information Standards Organization (NISO) standards developments and the Library of Congress' National Digital Preservation Task force. In addition, CNI works with Internet2 on advanced networking applications and standards, with CLIR (Council on Library and Information Resources) on scholarly communication, infrastructure and preservation issues, New Media Consortium on the exploration of new technologies in higher education, IMS Global learning consortium on interoperability between learning management systems and digital libraries. These are only a few of the participants with which CNI engages in an ongoing scholarly dialog on behalf of our community.

The format of the Fall Task Force Meeting (2009) included five breakout sessions, each offering at least five presentations from which to choose, an opening plenary session by Cliff Lynch, and a closing plenary by Bernard Frischer (Director of the Virtual World Heritage Laboratory at the University of Virginia). Generous breaks provided time for networking, as did a reception the first evening, and breakfast and lunch on the second day. I also attended the introductory session offered for new attendees, provided by Cliff Lynch and Joan Lippencott. During this session, Cliff indicated that the Blue Ribbon Task Force on Digital Preservation is about to make their final report, which will be of great interest to many of us. He also stated that presentations can be topic-based discussions, and need not be presentations of work underway or finished. It was impressed upon us that all the sessions are meant to be participatory, and that networking and questions are encouraged. One small lively session which precedes the task force meetings and focuses on a hot topic is called the Executive Round Table. Any representative from a member institution can attend by indicating their interest quickly when the topic is announced; generally only the first twenty or so people who respond are invited to attend. The roster fills quickly, generally within three days of the announcement.

## *Opening Plenary: by Cliff Lynch*

In the opening plenary, Cliff provided a broad overview of the issues facing us at the current time. He noted that Joan Lippencott had been studying the uses of library commons, sustainability for libraries and funding issues. He says there's a place here for markets and for circles of gifts. We need to find ways to make them work more effectively. There are some things that it makes sense to think of as common goods.

With the current budget crunch, Cliff expected to lose some members, and indeed, CNI did lose three, which he hopes will return. What he did not expect was to gain six new members, which seems to be a sign of recognition that we must work together to forge solutions for the long haul, especially in difficult times.

The data curation focus has changed into a concern for the full info life cycle, including a new emphasis on reuse of content. Digital preservation work has transitioned from projects to programs. And interestingly, users are beginning to be involved and interested in preservation issues. A new concern is the resilience of digital library preservation. We have discovered, for example that we are unable to build in enough redundancy. We are finding this in more and more of our systems: the bigger it is, the more likely something will be broken.

Special collections are going to play a key role in scholarly work. Due to their unique materials, Cliff predicts that they will be a place of high-profile developments.

A Mellon funded project is underway to support annotation across silos, sharing on both controlled and uncontrolled bases. There must be a hundred annotation projects that developed variations of software, but which were not adopted for one reason or another. There are even more which were local implementations. The lack of this tool is a huge scholarly barrier, so this project is very important.

Text mining in humanities is very hot. There is much debate on how to normalize, and how to computationally extract meaning.

The rate of development of scholarly content is huge, enough that we should pause and reflect on what it means for how useful collections of knowledge can be managed. At a digital curation conference in London two weeks before, a keynote speaker had rattled off these statistics: two new papers a minute are published in biology, five new papers a minute across all of science. Cliff admonished us to think about the implications of this rate of flow for text mining, peer review, even for the future of scholarly communication.

He also sees transitions in education from the traditional course to something else, given the development of new online learning environments. As the latter compete with one another, where will the student spend all his time? Lecture capture will become another genre of scholarly material. How do we understand its context and limitations? What is the future of textbooks? We need to watch the creation of educational materials for high volume classes. We should raise hard questions about roles in the development of learning objects. They are largely being developed outside of educational institutions and being marketed directly to students. This may move toward site licensing.

Mobile devices will begin to have more impact, as something besides mini-laptops. An example of this is the geospatial capabilities being incorporated into more and more mobile devices. Other examples include citizen networks and sensors, and citizen journaling.

Cliff quoted something John Wilkin had said a couple weeks before, about cross-institutional infrastructure. John had distinguished between shared problems and common problems. They're different kinds of problems. Cliff said that the Bamboo project is looking at this in humanities. There are many more examples in the sciences, and more on the way. The Framework program in Europe is

working in the same sorts of issues in e-science and e-research. We need to develop infrastructure components such as gazeteers, and a national and global biographical infrastructure for names.

### ***DataNet Partners Update: DataONE and the Data Conservancy***

The first breakout session I attended was entitled “DataNet Partners Update: DataONE and the Data Conservancy”, presented by Robert J. Sandusky (Assistant University Librarian for Information Technology at the University of Illinois at Chicago) and Sayeed Choudhury (Associate Dean, University Libraries, Johns Hopkins University). Both DataONE (Data Observation Network for Earth) and the Data Conservancy are funded through the National Science Foundation DataNet program, and were the first two projects funded. DataONE focuses on the development of a new virtual organization to support innovative environmental research via a distributed and sustainable cyberinfrastructure. The Data Conservancy seeks to develop the data curation infrastructure necessary to support cross-disciplinary discovery particularly with relation to observational data.

Dr. Sandusky presented the DataONE update; he's on the cyberinfrastructure team. The Principle Investigator on this project is Bill Michener of the University of New Mexico. Sandusky stated that each DataNet effort is supposed to take a 10-year view in all development, and be self-sustaining within 5 years. DataONE has three foci: community engagement, cyberinfrastructure development, and governance and sustainability. There are three coordinating nodes in the US: University of California (Santa Barbara), Oak Ridge National Laboratory, and the University of New Mexico. Only 4½ months into the project, they are already launching three member nodes. The intent is to be diversity tolerant, replicating content across geographically distributed nodes, but with centralized services. Year 1 goals include replication, interoperable metadata search and data retrieval, and basic logging. They've determined they will have 5 broad categories of service API: identification and authorization, object management, discovery and usage, preservation, network services. An Investigator Toolkit will be developed to enable researchers to access and use this content. More information is available from <http://dataone.org> or by emailing Dr. Sandusky at [sandusky@uiuc.edu](mailto:sandusky@uiuc.edu).

Dr. Choudhury ([sayeed@jhu.edu](mailto:sayeed@jhu.edu)) presented the Data Conservancy Infrastructure update. Educause recently posted a video of Choudhury's presentation on this project<sup>1</sup> which he urged attendees to view. There are four working teams, focusing on computer science/information science research needs, the wider impact, infrastructure, and sustainability. The known components of the new infrastructure will include Fedora, arXiv.org, Open Journal Systems and Sakai. To begin, they are working with a subset of data from the Sloan Digital Sky Survey and the Dry Valleys project, 2 data sets that are “finished” or almost so, which reduces changes. The user interface will be an integration of Sakai and the International Virtual Observatory Alliance (IVOA) standards. The prototype scope includes this as well as access API and Fedora infrastructure including Duracloud publishing platform integration (between arXive.org and Open Journal System). They want to support more than one access method, hence the two different systems. We will need to discover what it is to load these huge data stores and find out the ramifications. This will be our first experience as libraries in handling large data sets, and this will *not* be a dark archive. Dr. Choudhury states that data infrastructure will be a success when users don't notice it.

### ***Interoperable Annotation: Perspectives form the Open Annotation Collaboration***

The second breakout session I attended was “Interoperable Annotation: Perspectives form the Open

---

<sup>1</sup> <http://www.educause.edu/E09+Hybrid/EDUCAUSE2009FacetoFaceConferen/InitiativesfromtheNSFsDataNetP/175757>

Annotation Collaboration,” presented by Robert Sanderson (Scientist, Los Alamos National Laboratory: [azaroth42@gmail.com](mailto:azaroth42@gmail.com)) and Herbert Van de Sompel (Digital Library Researcher, Los Alamos National Laboratory: [hdivsomp@gmail.com](mailto:hdivsomp@gmail.com)). The Open Annotation Collaboration (OAC) is seeking to support web-centric interoperable annotation environments. In phase 1 they are integrating AXE Notation<sup>2</sup> and Zotero<sup>3</sup> and exploring existing systems. This is an international effort including at least five institutions<sup>4</sup>, as well as such experts as Tom Habing and Tim Cole at University of Illinois (Champaign-Urbana), Michael Nelson of Old Dominion University, and Ed Summers at the Library of Congress.

One issue already faced is the difficulty defining an annotation; the alpha data model they have now is too generic, and they invite feedback. In this model, both content and target are web resources; they can be any type, format, language, etc. The relationship between the two is called a predicate, where the content annotates the target. One of their basic principles is to focus on interoperability which is based on existing web architecture, thus entities within the model must be identified by HTTP URIs. This provides globally unique identifiers without central system overhead.

Step 2 will include support of a user-entered string as the content of the annotation; in this case there will be no URI, but a URN instead. Text will be captured in the `aoc:body` property. The class used here is a hint to accessing software to not bother trying to dereference the URN.

Step 3 must enable the user to select a portion of the resource as the target of their annotation, not just the whole resource. This will depend upon the W3C Media Fragment URI<sup>5</sup> which should be approved soon. This is based on the current practice of referencing a portion of a web page by appending on a pound hash (“#”) followed by an indicator of the portion of the page desired. In this situation, coordinates will be specified which outline the segment of the page desired, for example: `http://www.here.com#xywh=160,120,320,240` which specifies an X coordinate, Y coordinate, width and height in pixels.

Step 4 will be to capture non-rectangular regions of a resource for annotation. Here they use a context node with a segment description, in any format. A description might be a Scalable Vectors Graphics (SVG)<sup>6</sup> path, an Xpath<sup>7</sup> for XML, or a speaker in an audio track. Work on this topic exists for common cases already, such as in MPEG-7.<sup>8</sup> Segments may include datasets, databases, and non-traditional media.

In Step 5, annotation should be able to have more than one target resource or segment. That is to say, the annotation concerns multiple resources or creates a relationship between them. Multiple content sources can also be modeled, for example when you have the same content in different formats or media, or it was later translated to a different language, location, format, etc.

Step 6 involves making the annotations robust across time with respect to changing resources. For example, a news site front page changes daily. The annotation must apply to the correct version which was originally annotated. Three solutions are under consideration. Solution 1 creates a time stamp used as date/time of the version of content and target resources unless the target version is dated separately. Solution 2 records the time stamp of when the annotation applies to the target. And Solution 3 requires archiving the resource at the time of the annotation, and then relating the annotation to the archived copy.

---

2 <http://thomson.fosterscience.com/Chemistry/Unit3-ChemicalBonding/AXENotation.htm>

3 <http://www.zotero.org/>

4 <http://www.openannotation.org/theOpenAnnotationCollaboration.html>

5 <http://www.w3.org/TR/media-frag/>

6 <http://www.w3.org/Graphics/SVG/>

7 <http://www.w3.org/TR/xpath20/>

8 <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>

Step 7 will focus on making the annotations web available for dissemination in an interoperable manner.

The Open Annotation Alpha Data Model was released 1 December 2009. During its development they studied Annotea, LEMO, DiLAS, Fab4, Pliny, Google Sidewiki, Flickr annotations, Richard Newman's tag model (plus 3 extensions), Common Tag Model, Henry Story's tag model, and many annotation systems. Annotea is REST-supported, Google Sidewiki uses ATOM. Most, however, are completely proprietary. The OAC recommends NO protocol, though existing systems are tightly coupled. Should restrictions be needed on who can retrieve or use annotation, OAC has decided that should be supported via server authentication, rather than included in the framework development.

For more information, visit <http://groups.google.com/group/oac-tech>, <http://openannotation.org>. (Notably, during the question and answer period, Dr. Ed Fox stated he has a Ph.D. student who has been working on this issue for several years in collaboration with someone in Oregon.)

### ***Everything Old is New Again: Newspapers and Auction Catalogs in the Age of Web 2.0***

This was a joint presentation of two separate efforts to involve users in tagging online digital content and correcting OCR (Optical Character Recognition) text for the material (also online).

The first presentation was “Many Hands Make Light Work: Public Collaborative OCR Text Correction & Annotation” by Keith Jeffers (Director, IT Services, National Library of Australia: [kjeffers@nla.gov.au](mailto:kjeffers@nla.gov.au)), about the Australian Newspaper Digitisation Program<sup>9</sup>. They developed an open source software which enables users to correct the OCR, tag, and comment on content as well. In a single year (still in beta), they've had half a million testers and thousands of suggestions. There have been only two instances of vandalism (both Russian brides for sale); they've added software in the background to watch for this.

Relatedly, just last week, they released Trove, a new discovery service which offers a single search box over a wide variety of content sources, even cross-institutional. On the entry site<sup>10</sup>, at the time of this writing, the claim is made that 2107 searches have occurred in the last hour, 28671 newspaper text corrections today along, 7702 items were tagged this week, and 418 comments were received this month. Users can either log in to edit, tag, or comment, or they need to enter “captcha”<sup>11</sup> values to prove they are human. (An example is provided here: <http://trove.nla.gov.au/ndp/del/article/12000553?searchTerm=higher%20education>). The article sought is highlighted within the page, the words sought are highlighted within the article, and corrections can be made for any line in the OCR, which is presented side by side with the image. Users can view last set of changes; administrators can view the history of changes. Some people are interested in monitoring changes, so as it has turned out, administrators have not needed to change anything.

Participants like it; they view it as a challenge, and they appreciate the trust extended to them. The response has been enthusiastic, and Jeffers noted some very impressive numbers. Already over 354000 articles have been corrected. They're making a point of encouraging participants and creating a team spirit. Already they use Flickr to capture images of participants, and they will be creating user profiles that will work with Facebook, to show the statistics of what each participant has contributed, to build a competitive spirit and provide visible appreciation. When asked how the site was publicized, Jeffers stated that they posted it to a couple of genealogical sites, but otherwise, just put an invitation on their front page.

---

9 <http://www.nla.gov.au/ndp/>

10 <http://trove.nla.gov.au/>

11 <http://www.captcha.net/>

Currently there are nearly a million pages in service, 10 million articles. Newspapers are delivered in JPEG or PDF. Keith Jeffers built the architecture and storage; Rose Holley managed the project. Only the OCR is outsourced. Their OCR is 85- 90% good. Asked for 95%, but they think it's only 87% correct. They check 10% for quality control. Supposedly the outsource agency is using multiple OCR engines and looking for correlations between them.

The Aussie software depends on Lucene for indexing, MySQL for metadata retrieval and storage, use an EMC San, and Linux with SSD disks. The software is available open source.

The second presentation was “Enhancing Digitized Auction Catalogs Using Community Editing Tools” by Lesley Goodwin (Project Coordinator, JSTOR). JSTOR is a product of Ithaka. They also have others, notably PORTICO). This project, which is Mellon funded, is another community editing platform, but designed for researchers and curators instead of the general public. The JSTOR version supports “softlinking” identifying keywords and phrases that can be used to link to other resources, and this is not specific to auction catalog data. Their process involves OCR, then zoning the OCR, and analysis to recognize and zone the hand written annotations. They use machine learning to correctly segment the components of a catalog. Feedback has been very light. Users tended to complete the shorter catalogs.

The software is supported by Djatoka (open source JPEG 2000 server), mySQL, and the Thrift server developed by Facebook, which is part of Apache now. The front end is Plone based, written in Python. The site is open access through April, but then will be restricted to JSTOR subscribers: <http://Auctioncatalogs.jstor.org> On this site, as you mouse over a word it is highlighted in the OCR and vice versa, and you can zoom to sections of the image. In addition, you can merge “lots” which are segments of text, rather than simply correcting line by line.

Both implementations support simple html in comments, but not the upload of images. Neither is yet tracking “corrected” versus “uncorrected,” but both track “changed” versus “unchanged.” One commenter pointed out that the corrections could be datamined for patterns which could enable the software to become partially self-correcting. The JSTOR group is looking for partners.

### ***ARTstor Shared Shelf Initiative: a Networked Image Management Platform***

ARTstor, eight partner colleges and universities, and the Society of Architectural Historians have engaged in a joint initiative for the management and sharing of digital images, called “Shared Shelf.”<sup>12</sup> Institutional partners include Colby College, Cornell, Harvard, Middlebury College, New York University, Yale, The University of Miami, and the University of Illinois at Urbana-Champaign. The intent is to allow images created by individuals, by institutions, and those in ARTstor, to be combined, managed, and used, without local on-site infrastructure. ARTstor hosts the implementation.

James Shulman, President of ARTstor, presented first, followed by short presentations by some of the involved institutions. ARTstore hosts content for 141 institutions, which catalog their materials using multiple types of systems and tools. Because of the variety, updating hosted collections is problematic. Participants wanted tools for batch or individual asset upload. Users need to be able to share materials across their own campus, and across multiple systems.

The components of the Shared Shelf system include a cataloging system, controlled vocabularies, digital asset management, and publishing. Cataloging relies on the VRA Core<sup>13</sup> schema with additional

---

12 <http://www.artstor.org/what-is-artstor/w-html/services-hosting.shtml>

13 <http://www.vraweb.org/projects/vracore4/>

optional fields. After ingest, users can publish selections of content to other places, for example to Flickr. The software is still in Beta, but will be in version 3 of Beta in spring 2011. To stay informed, send an email to [sharedshelf@artstor.org](mailto:sharedshelf@artstor.org).

Pauline Saliga, Executive Director of the Society of Architectural Historians represented their Architectural Resources Archive (SAHARA)<sup>14</sup>. Their organization is located in Chicago, and they publish the Journal of the Society of Architectural Historians (JSAH) which will be online next year for the first time, in multimedia format. The reason they joined the Shared Shelf initiative was to have user contributed shared online resources, with global coverage. Starting in January they will have editorial tools. Members' personal collections are unvetted, but the shared one is peer reviewed. This will offer a new working paradigm for visual resources librarians and scholar editors; they will be able to work in teams to edit images and metadata. They are hoping for a huge influx of new images from users. Guest accounts are available starting in January 2010 at <http://www.sah.org>.

Clem Guthro, Director of Libraries at Colby College (Waterford, ME) spoke next. Theirs is a small, private 4-year liberal arts college. They received a ¼ million grant from the Davis Educational Foundation for development of a collection of images to support faculty teaching and student learning. They have been using Luna Insight, which Guthro says has serious limitations, and Filemaker Pro, with four scanning stations in two different locations, ½ time visual resources librarian, ½ time visual resources curator, and 8-10 students per semester. The Luna Insight work flow is cumbersome; Filemaker Pro helps, but now they have two separate systems to support. They need a better work flow and cannot afford to run multiple platforms. They have 45000 images in Luna. Their faculty and staff want an integrated image resource.

Dean Kraft (Chief Technology Strategies for Cornell) shared that they have a large number of independent image collections in Luna and various ad-hoc repositories. The library just took over managing visual resources. Existing library digitization faculty uses Pictor for work flow management. Shared shelf allows a common system for all teaching images. This avoids maintenance costs for Luna, and there is no need to create new metadata/work flow system. They are fine with an externally hosted provider they trust. Cornell is currently migrating 15 Luna collections, and determining the cost to migrate the ad hoc collections. They are also considering non-art images (such as biology slides). Their concerns include copyright/fair use issues, ensuring Shared Shelf work flow will support their digitization project needs ( it has to work for faculty) and leveraging their existing Shibboleth authentication system.

The University of Illinois at Urbana-Champaign (UIUC, presented by Beth Sandore, Associate University Librarian) College of Fine & Applied Arts Slide Library, has 350,000 analog slides, 20,000 digital ones with limited access, and then they have 20,000 more in CONTENTdm collections. UIUC has a large decentralized campus, and scattered visual resources. They want a single platform on which their users can access all their visual content, along with that of ARTstor. Currently they have a federated search tool that cross-searches all the silos, but they must maintain the database connectors and scripts, and thus must change it constantly.

Tracy Robinson (Head, Office for Information Systems, Harvard University Library) spoke about the visual resources at Harvard. Their cataloging is done in Olivia, and their visual resources catalog is VIA (Visual Information Access, locally grown). They are in the process of migrating legacy data from 20-25 different cataloging instances across campus. All of their systems are homegrown, so this move to a single hosted system will include downsizing and cost savings.

---

14 <http://www.sah.org/index.php?src=gendocs&ref=HOME&category=Sahara%20HOME>

## *Geographic Tools & Digital Collections*

This was a dual presentation of the University of North Carolina (UNC), Chapel Hill, and Brigham Young University (BYU). Both institutions have been developing user tools and mash ups to make digital content more compelling and more useful for researchers and educators alike.

Natasha Smith (Head, Digital Publishing Group, Carolina Digital Library and Archives) and Richard Szary (Director, Louis Round Wilson Library, and Associate University Librarian for Special Collections) spoke first, for UNC. They have been working this past year with both the spatial and temporal history of North Carolina, using GIS (Geospatial Information Systems) technology in digital libraries. One such effort includes collections in “Documenting the American South”<sup>15</sup>, another is the maps within the Carolina Digital Library and Archives (CDLA)<sup>16</sup>. They found their users wanted more maps, but of course maps are difficult to display online as they are very large. Nonetheless, UNC decided to try to digitize and publish online every printed map published prior to 1923. Already they have 2354 online.

To manage the large size for digitization, they use a vacuum table, which sucks the map flat, and then can be rotated perpendicular to face a camera some distance away. Using this methodology, it requires only 5 minutes to capture a map. Their oldest map dates to 1484 and is called “La Florica.” Their longest map is 31 feet long.

Within “Documenting the American South,” there is a collection is called “Going To The Show,<sup>17</sup>” which is the brainchild of Professor Robert C. Allen (UNC-Chapel Hill). This collection contains a wide range of multimedia content related to attending moving picture shows in North Carolina. Each of these documents is being mapped for browse and retrieval by location. For this project, UNC digitally stitched together the Sanborn fire insurance maps from 1896-1922 to represent each entire downtown area, a manual process using Photoshop. Then they georeferenced them, so they can be used with Google API. They had to identify point coordinates for venues and street corners as well. They use Zoomify to display, and Google Maps as an image viewer. (An example of Asheville, North Carolina can be viewed here: <http://docsouth.unc.edu/gtts/map/asheville> ). They use a “ticket” tab to indicate a location where they have digital content. Hovering a mouse over the tab brings up more data for the related content for that address, and the user can click through to view it.

UNC had to add lots of content, and relied heavily on the use of students of professors who were interested in the project. They had to assign bounding coordinates, and then layered historic maps over Google maps. By zooming in to the Sanford maps, the user can alter the opacity to see the current structures, and by downloading and using Google Earth options, the user can obtain a current street view which makes the existing buildings appear to rise up through the Sanborn maps.

The display shows the coordinates of the bounding box the user selects. They are moving toward spatial search, and want to connect all the content for a particular region. However, to do so, everything must have a geocode reference.

A newly funded (Library Services and Technology Act) project entitled “Driving Through Time: The Digital Blue Ridge Parkway in North Carolina,”<sup>18</sup> will create a Blue Ridge Parkway geoportal. Another project funded by an National Endowments for the Humanities Start Up Grant will be called “Main Street, Carolina”<sup>19</sup> will create a web-based digital toolkit to allow libraries, schools, museums, etcetera, across North Carolina to document their downtowns. Digital content will be compiled and

---

15 <http://docsouth.unc.edu/>

16 <http://cdla.unc.edu/>

17 <http://docsouth.unc.edu/gtts/>

18 <http://www.lib.unc.edu/blogs/morton/index.php/2009/06/driving-through-time-project-funded/>

19 <http://iah.unc.edu/chat/festival/projects/mainstreet>

underlain by Sanborn maps, highway maps, and county soil maps. UNC expects to have an alpha version available by mid spring of next year. It will require PHP5 and SQLite 3, and will have a minimal imprint, using Javascript API to allow users to include digitized content from NC's website. Four pilot sites have already been selected.

Scott Eldredge (Digital Initiatives Program Manager: [se@byu.edu](mailto:se@byu.edu)) presented for Brigham Young University (BYU), which is the largest privately owned church funded school of higher education (owned by the Church of Latter-Day Saints), located in Provo, Utah. A little known fact is that 75% or more of their students speak two or more languages, which has an impact on their digitization work. Their project is called "MappifY"<sup>20</sup> and it is hosted by the Harold B. Lee Library, and funded by Angel Partners. This project builds a geographical palette on which to build collections.

Users can select collections from a drop down list and will be zoomed in to where content can be found. Content is geolocated with push pin icons and blue bubbles with counts in them (where there are clusters of content); clicking on either zooms in and lists the images; hovering gives a thumbnail. Data is stored in mySQL harvested from CONTENTdm. Clicking on an image takes the user into CONTENTdm to view the content. The coordinates in the dc:coverage field in the metadata was harvested into the mySQL.

One item of interest here is 3500 maps of 1930s Germany, which overlays old maps on the current Google view. By using the opacity slider, patrons can see the differences between then and now. The next phase will include working with how to geolocate diaries and journals, where there are multiple pages for a single location, or multiple locations on one page. Right now they're developing this on a test server, which shows the text and the image for every location on the author's traveled path. PHP and Javascript were used for the interface, Google Earth and Google Maps, ARCGIS software for creating KML and TFW files and georeferencing, PKP Open Archive Harvester for gathering the metadata, mySQL for storing it, and a GNIS database as well.

One thing they noticed is that Flickr and Google use different scales for zooming; consistency is necessary. Other problems they faced were:

- Where do you place the photo?
- How much accuracy do you attempt, at what cost?
- Where do you place portraits?
- How do you deal with dates and date ranges?

Dr Brandon Plewe, assistant professor of geography is heavily involved in this project, and multiple others. BYU lost programmers since the hiring freeze began, so this project is not currently in development.

Both institutions agreed to multiple other problems faced, such as archaic and misspelled names, and changing boundaries of counties. Space is a huge issue at UNC. BYU uses an engineering scanner instead of a camera, due to skew. They had most of the content previously digitized, but it had been archived on CDs, and they had a tremendous failure rate when trying to retrieve the archived images. They finally gave up trying to identify which ones were still good, and finally rescanned them all.

When asked how long it takes to georeference, the estimate given was 12 minutes per photo or 5 minutes per map. Tito Sierra (Associate Head for Digital Library Development Digital Library Initiatives, North Carolina State University) was in the audience and he recommended the use of a software created by New York Public Library called "Map Warper Beta"<sup>21</sup> for georectification. Anyone

---

20 <http://www.lib.byu.edu/DigitalMaps/>

21 <http://warper.geothings.net/>

can create an account. Sierra said Josh Greenberg is the contact for this; they are crowdsourcing it.

### *Closing Plenary by Bernard Frischer*

This talk was entitled “Beyond Illustration: New Dimensions of 3D Modeling of Cultural Heritage Sites and Monuments,” and it focused on work performed by the Virtual World Heritage Laboratory at the University of Virginia, where Dr. Frischer is the Director.

Their 3D model is called “Rome Reborn,” in which they have sought to create scientifically accurate three-dimensional renderings of ancient Rome. Three problems face them now:

- 1) Real-time use with secure remote rendering.

The average personal computer can't handle this, and most implementations will not allow downloads anyway. The issue there is property rights. How do we give hi-resolution access but not allow download? The solution they are exploring is called “Scanview,” developed by Dr. David Koller, of Stanford and UVa. The way this software works is to enable the user to download a cheap version which is then draped in more detail via streaming. They are using the Ennea system to improve on this, and the resultant combination should be out next year.

- 2) Collecting, preserving and disseminating the models on the internet.

There is no online digital repository for 4-D models. The closest is Google's rather simplistic KML model. Currently under development is a repository called “SAVE” which will be a peer-reviewed alternative.

- 3) Using them as tools for further research and discovery.

We need to make tools that are useful via the web. Most interesting right now is “OpenSimulator” which is the open source version of Second Life. Content can be ported into it and out of it. Currently Frischer is working with SRI International. With their assistance, researchers can meet inside model simulations and discuss issues about the models from across the world in real time. One of those issues is what to do with areas of the models where there is insufficient data.

### *Conclusions and Recommendations*

The DataNet efforts will sooner or later impact the University of Alabama, as the researchers and scientists on campus begin to seek out the support of cross-institutional cyberinfrastructure to support their work. While I do not expect this within the next handful of years, it is important for us to be aware of the strides these groups are making, and at some point to insert ourselves into one of the partnerships in order to pave the way for future change. These efforts bear monitoring, as they will no doubt lay the groundwork for massive change in higher education, particularly with regard to scientific endeavor.

The shared infrastructure of Shared Shelf (ARTstor) is interesting but troubling to me, as it isolates the images from all other related content, and places an uncomfortable amount of power over content access and use in the hands of ARTstor. While I certainly appreciate the need to downsize and share infrastructure, as well as to increase access across multiple silos of content, handing over so much control to an outside party seems to me to be setting the stage for trouble.

The Open Annotation Project will, I believe, solve the problem of interoperable annotation, so I don't think we need to extend effort in that regard when considering new tools for scholars. I also don't think we are in any position to be developing 4D models, though I'm certainly glad someone is, particularly on platforms that are open source, unlike Second Life. Open Simulator might be worth exploring for

use here on campus, but funding for such development would need to be covered with grants or other endowments.

The GIS developments were quite spectacular, and heavily dependent upon in-house programmers, far beyond our current staffing. While I found their development quite stunning and compelling, I am aware of the costs that went into such major efforts. The research options were indeed deeply extended, and I was pleased to see the “Main Street, Carolina” project being developed, as it would engage users in contributing content to build a shared digital library open to all.

Projects which involve user participation are the most fascinating to me, and I think most applicable to where we need to focus our energies. The more interactive we can make our offerings, the more engaged our patrons will be with our content, and the more they will feel they have a stake in making sure UA Libraries has what it needs to succeed. I was deeply impressed by the massive outpouring of enthusiastic participation in the Australian Newspaper Project. Part of their secret was to allow anyone to participate; another part of their secret is to recognize participants for what they bring to the table, and encourage them. I think it would bring us tremendous visibility to engage in a similar project here, perhaps as a multi-institution consortium as a grant-funded project. Certainly there are pre-1923 newspapers available in Alabama which have not yet been digitized, and others that may be available on microfiche. Since the Australian software is open-source, the Trove interface could be used cross-institutionally to provide a shared search interface, while the OCR-correcting software could be installed at each institution able to support it.

Outside of the sessions, I had the opportunity to converse with Tito Sierra of UNC, who is championing the development of an infrastructure for mobile computing. I also chatted with Thorny Staples, who is now working with DuraSpace, the joint effort of DSpace and Fedora, Tim DiLauro of Johns Hopkins who is deeply involved with the work to support scientific data sets, John Butler of the University of Minnesota Libraries, and many others. The networking time is valuable for making the University of Alabama more visible, more involved, and more responsive to changes in the field and the needs of others. I am grateful for the opportunity to have attended the CNI Task Force Meeting. It was a terrific experience, and the information gathered here could well inform our future.