Preserving State Government Digital Information
2009 All Partners Meeting Summary

Wednesday January 20, 2010
925 L Building; CA Legislative Counsel
Sacramento, California

Attendees

Arkansas History Commission: *Lynn Ewbank*

California Digital Library (CDL): *Stephen Abrams, Tricia Cruise, and John Kunze*

California Legislative Council: *Annie Anderson, Bill Behnk, Diane Boyer-Vine, Linda Heatherly, and Mendora Servin*

California State Archives: *Rebecca Wendt*

California State Library: *Kris Ogilvie*

Illinois Legislative Information Systems: *Tim Rice*

Illinois State Library: *Andrew Bullen*

Kansas State Historical Society: *Matt Veatch*

Library of Congress*: Butch Lazorchak*

Minnesota Historical Society: *Nancy Hoffman, Robert Horton, Carol Kussmann, Charles Rodgers, Shawn Rounds, and Lori Williamson*

Minnesota Legislative Reference Library: *Elizabeth Lincoln*

Minnesota Office of the Revisor of Statutes: *Isaac Holmlund and Tim Orr*

National Conference of State Legislatures (NCSL)*: Jo Anne Bourquard and Pam Greenberg*

Nebraska State Historical Society: *Seth Doty, Andrea Faling, and Gayla Koerting*

North Dakota Legislative Council: *Marilyn Johnson*

North Dakota State Historical Society: *Ann Jenks*

Syntactica: *Dan McCreary*

Tennessee State Library and Archives: *Jami Awalt and Greg Yates*

University of North Carolina*: Cal Lee*

Vermont Joint Fiscal Office: *Richard Reed*

Vermont State Archives*: Tanya Marshall*

Meeting Summary

The 2009 All Partners meeting for the Minnesota lead NDIIPP project was held on January 20, 2010 in Sacramento California. Representatives from each partner state were invited and nine out of ten states were in attendance. The purpose of the meeting was to give an overview of the past twelve months activities with a focus on demonstrating eXist, a native XML database that had been selected as the application to use for building the pilot applications. Bob Horton provided attendees with background information, and Carol Kussmann gave a presentation using eXist (demonstrating the applications that had been developed for the project) and Isaac Holmlund gave a presentation about how eXist could benefit the workflow of the Revisor's Office. During the meeting, participants were also filmed talking about the project or digital preservation in general which will be used to create another podcast.

Meeting Minutes

Presentation by Robert Horton: Project Background, Products and Pilot

- Began with thank you to the California and Minnesota partners who helped arrange the meeting and to the Library of Congress for funding the grant; followed by introductions of where people came from.

- Discussed how important education opportunities were and let people know that a video camera was available for those who would be willing to talk about the project, their part in it, or digital preservation in general. The footage would then be used to create another podcast about this project. The original podcast[1] created by the Minnesota Historical Society about the NDIIPP project was shown.

- Talk about the NDIIPP program in general; a ten year project that started in the year 2000. The Library of Congress is required to produce a 2010 report about the past ten years and suggest ideas for the next phase of the project. Funding for NDIIPP is now a line item in the federal budget, which moves the program from funding grant projects to more of an enabling and guiding program. (per Butch from the Library of Congress) This year, the projects that were awarded grants focused on content areas such as news, government, and Geographic Information System (GIS) data. State studies and summit meetings help guide activities as we move forward. As there is much concern about public policy on the web, the work we are doing on these issues are important.

- The four state NDIIPP grants that were most recently awarded were to Arizona to look at the LOCKSS (Lots of Copies Keeps Stuff Safe) preservation method; Minnesota to study legislative records; North Carolina to study GIS records; and to Washington to explore its digital archives. At the end of these projects, there will be an all states meeting, which will coincide with the Best Practices Exchange in Phoenix this fall.

    - Andrew asked about the scaling limits of LOCKSS, Cal said that there is one, Bob said we do not know the details on the scaling issues.

[1] http://discussions.mnhs.org/collections/2009/10/good-government-through-digital-infrastructure-and-preservation/

- Project Outcomes: Capture, preserve and provide access to "at-risk" digital content from state legislatures; Test the model in Minnesota; Determine capacity of other states to adapt the model; Promote the results through education and outreach; Connect to national cyber-infrastructure

- Preservation Issues: Continuity of Operations (COOP), disaster recovery; legal framework (record laws, litigation, discovery); increased public attention and expectations; complexity of systems (email, Records Management Applications, web, web 2.0) (North Carolina recently completed a policy on best practices for social media usage)[2]; costs (need to have justifications of why we need money); lack of a good model.

- Although the medium is different, these are familiar challenges. There may be a cultural shift in skill sets as we move from analog to digital but we still need to look at what we want to collect and how we can do it better.

    - Discussion on social media, working with people, and expectations.
        - Tanya said that Vermont is finding that with the addition of social media more and more people have roles or interest in controlling the information and what to do with it. It is becoming more and more difficult to define roles and responsibilities, and people do not always want to cooperate.
        - Butch responds that archivists and librarians can take this opportunity to let more and more people know that archivists and librarians do have some of the answers to these questions.
        - Tanya would be glad to do so; however they are often dependent on if other people/departments want to work with them. They can't be forced to work together.
        - Andrew says that Illinois has steered away from Web 2.0 because they do not want to expose the state to litigation (if for some reason data that was provided was incorrect). Having only one avenue for people to get at the information may reduce opportunities for the public but we want to make sure the data is correct.
        - Mendora says that in California they concentrate on making sure the data is accurate and then make raw data available. It is then up to the users to do whatever they want to do with the data. The newspapers for example create polls, and do so often. It is our job to concentrate on making the data accurate, how users twist it is up to them.
        - Andrew asks how California keeps the data accurate.
        - Mendora responds that for example in the Legislature's voting database, it is possible for a vote to be recorded incorrectly, so it is possible that the wrong data can get out there. However there are lots of people out there who can tell you that the info is wrong before it becomes public. If the wrong data has become public, we update the records as soon as possible.
        - Bob comments that the paper, "Government Data and the Invisible Hand"[3] discusses the idea of providing data to the public who have more skills and time

[2] Best Practices for Social Media Usage; December 2009. http://www.records.ncdcr.gov/guides/best_practices_socialmedia_usage_20091217.pdf

[3] Government Data and the Invisible Hand, 2009; http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1138083

to do the mashups they are interested in. The people know what they want and will create better mashups than government agencies.

- Cal points out (as well as the Mashups White Paper) that it does not matter if you provide raw data or not, the second information is available on the web, mashups can be done. Users will get the data by screen scraping if the raw data is not provided, and there is no way to prevent the screen scraping. So it is better to provide the raw data in the first place.
- Bill mentions that California was sued over not providing useful information on the public website. In response the Legislative Information System now provides a MySQL database full of data to analyze. (This is what the newspapers use for their polls.) People in California are also very interested in analyzing campaign contributions but his office does not control that data.
    - Andy asks how the MySQL is provided. Bill responds that it is an Oracle database that is published in a MySQL format. Mendora explains that the Legislative Information Council has many databases in multiple formats which makes it hard to get all the data in one place, so the information is combined and dumped into the MySQL format and made available.
- Bob reminds people that there are now federal rules to make information available on sites such as data.gov.

- Assumptions: We need to collaborate with others and keep up with what else is going on a national level. We will need to adapt to some standards. We will have to have a rigorous appraisal and return on investment. The use value of electronic records will be important. There must be a focus, and develop a manageable workload. As archivists and librarians we will have to learn some new things or change some institutional goals or strategies. Any solution must be sustainable.

- Ongoing Process: Standards, Migration, Conversion, Innovation, and Collaboration. These are not onetime issues, they are ongoing. Standards change. Software and hardware will need to be migrated or converted over time. It is important to continue to follow innovations and research and understand that collaborations may also change over time.

- Legal Framework: The current laws for public records and electronic records are not very clear. There are multiple laws that do not define public records in the same way. Redefining public records will be important, but messy.

- Conceptual Framework: Access is a key concern. The point of preservation is to make records available for use; the most useful records are therefore the most valuable and the most important to preserve. Data consultancies are important because providing guidance to the records creators will help them produce better systems – systems that result in better preservation outcomes. We can do a gap analysis to map where a records system is now and what it will need to change achieve this goal. We will need to utilize both national and international guidelines and standards. Research papers, podcasts and other education and outreach efforts will promote use of standards. Policies, laws, and technology information all need to be disseminated. We want to put electronic records issues into a larger context. Archivists can promote and persuade, but cannot compel records creators and managers to

implement best practices. Instead, we have to help them see how these practices are in their best interest.

- Policy Framework:
    - Document decisions and transactions; record laws
    - There is a lot of concern on accountability, transparency, and Freedom of Information; many organizations are pushing the material out but we also need to understand what it is the people want.
    - Privacy: On the flip side of pushing out the information is a concern about privacy. In the past, public information was harder to get and not many people took the time and effort to get the information (practical obscurity), now there is concern about invading people's privacy.
        - Tanya asked if another concern was really revenue and losing a source of money that if the data is easily accessible they can't charge for it anymore
        - Bob responded that often the fees that have been charged for retrieving records often did not cover the cost
        - Tanya expressed a concern that even if the records fees did not cover the copying cost, agencies would still resist loosing the cash flow.
        - Bob acknowledged that inefficient record systems persist because agencies cannot agree on systems even if they would ultimately save money. This happened in Minnesota when the state CIO unsuccessfully tried to implement a standard document management system across state government. Bob also recalled serving on an electronic records task force that recommended switching from paper to electronic records for real estate transactions. Bob believes this proposal failed because it required fewer workers and would eliminate jobs that decision makers did not want to cut.

- Sustainability: In order for this to be sustainable, we must come up with appropriate solutions, there is not a one size fits all solution. It must be easy to integrate into work routines. We will need to set some priorities, control costs, show use value (access, transparency, open govt) and continue collaborations.

- Archivist Role: As archivists we can define our niche by adding value to records. What is it that we can do for others, what do we specialize in? In addition to adding value to records we are good at facilitation, bringing people together, collaborations and getting people to talk about the issues. We do not need to be the ones with all the answers. With paper records, archivists provide materials to lots of researchers that facilitate the discovery of the significance of the objects; we can also do this with Web 2.0. In addition we understand the issues surrounding long term preservation.

- Project Partners: Project partners include the Minnesota Historical Society, Minnesota Office of the Revisor of Statutes, the Minnesota Legislative Reference Library, the California Digital Library, the National Conference of State Legislatures, Thomson Reuters and others from the private sector, and the states of Arkansas, California, Illinois, Kansas, Mississippi, Nebraska, North Dakota, Tennessee, and Vermont.

- Project partners are unique and solutions must be appropriate to each of their environments.

They all have different audiences, missions, priorities, initiatives, capacities, and resources.

- Common Ground: However even with different backgrounds there is common ground. The ideas of providing trustworthy (accurate, useful and valid) information; providing records with enduring value, making records accessible, and working together on standards are of a concern for all. NCCUSL is working on a model law for authentication and these ideas came from one of their drafts.

- Process: How are we working on all of this… we have lots of meetings with each other as well as with national organizations; we document our experiences on Basecamp and on the public website; research our partners' topics of interest; use re-granting to help further research topics; developed a pilot project; continue to evaluate each step.

- Lessons we're learning: We have learned to take things one step at a time, the project is in perpetual beta, we will never be quite done as new options come along new avenues will be taken. Things will never be perfect or complete. Working with partners and working with disparate data sets will continue to be in the works. Resources of partners are constantly changing; funding is not stable within the recent budget crisis. We have learned that partners are often paying 'Constant partial attention' (having so many things going on that you can't give full attention to any one of them) to all of their responsibilities. User's expectations are that they will have access to materials over time which equals preservation. Success then equals providing content with increasing functionality which enhances the value of the content.

- Practical Outcome: Storage ------> Preservation There is a line drawn between storage and preservation. Storage is not enough to ensure preservation, and there is no best solution for long-term preservation. What we need to do is to find the sweet spot in middle using the most appropriate solutions and options for policies, standards, partners, technology, and model. As options evolve and change over time we can move closer to the goal of full preservation, but as we do this we need to make sure the solutions are not burdens.

    - Andy asks if we are talking to Richard Pearce-Moses in Arizona about standards. Bob comments that the four states have had minimal contact but the metadata that we are using can be cross-walked to multiple other metadata sets.
    - Andy asks if the goal is to be able to have a site where all 50 states can be searched at one time. Bob replied that the NCSL and Thomson Reuters have both researched the concept and have found issues with it. What we need to do is have a situation where states are sharing enough information about their data so that developing a data interchange model becomes easier. More people have moved to XML bill drafting systems, but it does take time, money, and attention so change will not happen overnight, but states seem to be moving slowly in the direction of using XML.

- Graphic slide: A rough workflow diagram with an archive in the middle who receives content from the Legislature and by means of web harvesting. The archives can then provide item and batch level access to records and also work with a repository on the preservation of records.
    - Andy asked about what it is we want to harvest off the web. Bob responded that we are looking at other legislative content such as state agency reports, mandated reports, and committee meeting minutes. He also mentioned that there are legal policy concerns about the material that is made available.

- Andy also asked about privacy issues. Bob responded that these types of records should not cause privacy concerns because the records are already public records
- Marilyn asks about if a search is performed, what results do I get, everything? Bob responded that it could be possible, but in reality it will be hard to facilitate this. There will need to be enough common data across state lines and across data types before this can happen and data normalization is a huge cost. Thomson Reuters spends a lot of time and money on this.
- Tanya comments that this model is very similar to a regular 'paper' archives model. Bob agrees and goes back to the point that there are familiar challenges. The partnerships and tools may be different, but we still need to build a technology framework that is sustainable, can automate exchanges, and reduce burdens on any one entity or function. It would also be possible for the legislature to work directly with the repository (we would need to add an arrow on the graph to indicate this).
- Marilyn asks about evaluation. Who will be evaluating this? Outside users or only people on this project? Bob responds that yes we want a sustainable model, it seems like legislative staff would be a group who could assist in evaluation, for example the Legislative Reference Library. Bill comments that one way to sell a program is that it is not about preservation but disaster recovery; this has helped in California.

- Progress: We have done a lot of research and completed many white papers on access, records management, digital audio/video, legislative history, xml usage, mashups, authentication, etc. We have also tried to provide education through the use of handouts, podcasts, and publications (The NCSL is currently working on a digital preservation publication with us.) We developed a core schema to be used with a wrapper around legislative bills as we understand most states have an investment in their schemas and will not want to revise it; the wrapper and schema allow their data to be gathered as is. We have developed various applications using eXist, and XML native database and will see a demonstration on it after lunch.

- MHS Next Steps: We hope to integrate non-xml content into our pilot project and learn what is practical and what is not. We will work with the California Digital Library (CDL) on import and exporting for preservation. California and Kansas may work with us to test the system further. We would like to continue to find ways to automate the process/workflow. We will continue with education and help provide a gap analysis to partner states. We will create a toolkit with information on what we did and resources that will be helpful to other states. We will also evaluate our choices and compare alternate options.

- Partner Next Steps: We hope that you continue to follow us and learn with us. You can share content with us, learn how to adopt the models to your own environment, take part in gap analysis, and evaluate our decisions.

- Questions
  - Andy asked about the web-harvesting tool. Bob indicated that we will be using the Web Archiving Service (WAS) with CDL. John and Trisha commented that WAS uses Heritrix just like Archive-It, but that they feel it is easier to use and there are controls on it that allow the data to not be public; it was designed for a more academic setting. CDL would be happy to provide the group with test passwords to try it out for themselves. Cal pointed out that both web archiving applications use the same tools, just different management systems.

- Shawn stated that one part of diagram that we did not talk about was authentication. It will be important to establish the chain of custody or you can no longer say that the information is authentic. Diane said that the Government Printing Office is using PKI technology. Andy commented that IL is using MD5 hashes and keys.
- Bill asked about the preservation aspect of this project. Bob said that we will be working with the CDL on this topic. We knew that we could not build or sustain a preservation capacity on our own. It does not make sense to have each state build their own digital repository.

Lunch Break

- Viewed the Library of Congress' newest podcast about digital preservation. The teenage perspective on digital preservation; what they know and what they are concerned about.

Demonstrations

- eXist Pilot Project Demonstration; Carol Kussmann, Minnesota Historical Society
  - Carol demonstrated the applications that had been built for the pilot project architecture. These included the Requirements Manager, Glossary Manager, User Manager, User Story Manager, FAQ Manager, the state collections of bills from CA, IL, and MN, the Uploader, Search, Move Tool, Index Advisor, Check Sum, Stress Test Tool, Template and Dublin Core. Carol concluded her presentation by demonstrating a variety of ways to get data out of the database. This included saving a record to the open source citation tool, Zotero which uses HTML4 metadata tags. Details are not given here as the PowerPoint of the presentation includes detailed notes to follow along with.

  - Questions during demonstrations
    - Cal asked if eXist creates a transaction log and Dan McCreary explained that it can be set-up to generate who, what, where, when type audit trails.
    - Cal asked what would happen if indexing was run while a query was in progress. Carol said that the indexing would slow down the query, but not affect it in any other way.
    - Cal asked what the 'boost' values represent in the index advisor application. Carol told him them they are simply relative numbers.
    - Butch asked if it was possible to test whether or not the indexing was successful. Dan said that a formal test was not developed, but it could be checked using the stress test tool.
    - Richard Reed asked if a cross-collection search would require the creation of another index, but Dan told him it would not, it would just be a matter of changing the query code.
    - A discussion followed about how the wrapper metadata could be used to improve the search results ranking and presentation of the results list.
    - Andrew asked if we intended to aggregate all states data into one database, and Carol said each state could have its own group in a single database, or alternately, a federated search across several databases could be set-up.
    - Tanya asked if a single database was created, who would host it. Minnesota, LC, and NCSL would be among the possible choices for this.

        - Andy asked if the service could use a SOAP interface instead of WebDAV, which would eliminate the need for a common schema. Dan indicated that it would be possible to set-up the system in that way.
        - Richard asked about the advisability of packing the database with the oXygen XML editor. Carol explained that system could be made read-only for anyone but the administrator. There was further discussion about the security and trustworthiness of files in the database.

- Other Uses of eXist; Isaac Holmlund, Minnesota Office of the Revisor of Statutes
    - I am one of the people who make use of technology in Minnesota's legislative process. I will discuss what a native XML database can do for us. What I do is take the XML documents that are created and make them available for users in various formats. We have been looking for a way to simplify the way things are done.
    - Rube Goldberg, simple machines that are not so simple. There are many steps that need must happen and many points of failure.
    - Lots of work goes into publishing the statutes. At least one server must be devoted to this, it is then pulled out of rotation, we run prep scripts, create tables, pre-populate data, and then run publishing programs (which can take over 24 hours itself). We have information in various files and need to shred the XML to make it usable. This process takes three staff people three days to complete. It does not seem like this should take that long, how do we fix this?
    - The native XML databases use XQuery and can query XML documents natively. What this means is that shredding is no longer necessary.
    - I tried eXist myself and after I downloaded it to my system, all I had to do was drop and drag the files into eXist. This entire process took 10 minutes. It also gave me direct access to every XML node in my documents and I was able to write queries immediately. I had full control over what I could look at.
    - I then started to prototype the pages to make them look like our current webpage. I worked with statutes. On our current page there is a 'topics' section, this currently done physically by a person matching up the right data with the statutes, it is a document and not a database table. It took us a long time to figure out how to display a list of topics (about two weeks). Now that I could query the documents I wanted to see if I could reproduce the 'topics' section. I had it completed within 30 minutes. (Realistically it took 2 days… a day to develop and a day to test it, but it was up and running in 30 minutes.)
    - Another feature that we would like to put on our website is "referenced by". It is on the "to do" list, we know how we can do it but know that it will take at least a week of time for several people because it involves working with "many to many" relationships. But with eXist, I got it done with one line of code… I told the system to find the node for referenced statutes and it puts it in a list on the page.
    - By using a native XML database, the reduction in time for creating these features is tremendous.
    - I also thought it would be interesting to provide information on the bills that are currently amending a statute that is being looked at. People are not asking for this information, but because I can play with the data directly, I figured out how to make this possible.
    - The requirements of what we need to provide to people on the web are rapidly increasing and we need to meet those demands. To do this we need to be more flexible

and adaptable, and an XML native database is a good tool for that.

Wrap Up: Robert Horton, Minnesota Historical Society

- Next steps
  - Work more with the XML wrapper that we started
  - Create additional applications that work across boundaries
  - Gap analysis, how to take this and adapt to your environment

Conversation

- Andrew asked for more information about the wrapper. Tim Orr responded that we played around with the idea of an XML wrapper file that could take data that a legislature is already producing, wrap it with common metadata, and add it into eXist. If this is to be done, the metadata needs to be standardized and agree upon, but it provides a way to have consistent searches across multiple states or data types.
- Andrew asked if we looked at GILS?[4] He said that they spent a lot of time developing government information interoperability standards. The question that we need to answer is how can we compare different states on a basic level? Bob and Tim Orr talked about how the wrapper contains standard metadata and can be wrapped around information of different types, allowing us to compare apples and oranges. Bill commented that Tim Orr said the wrapper was 90 percent done and asked when others on the project will be able to use it. Tim Orr said it can be tested on their server now and that he will put links on Basecamp. It is the metadata that needs to be fine tuned and agreed upon.
- Cal remarked that it is important to make it easy for the data creators. If you put all of this information into one system, how do you get all the data to match. He asked if the California, Minnesota, and Illinois data was transformed. Dan McCreary responded by saying the data was not transformed, however the schemas were ignored because they did not match. It is possible to provide a tool that would allow each state to match up their schema with the wrapper metadata (by writing a path expression for each tag). Forms can also be set up as a way to get into the data.
- Bob said that there is no way that people will change how they are doing things now, but if you build in features that are seen as benefits it will get the ball rolling.
- Andrew stated that developing a wrapper that can move across state lines would be a very important tool to have.
- Tanya asked how to make information easier to exchange between state agencies. Bob talked about the Minnesota Campaign Finance and Public Disclosure Board and Shawn gave the example of an agency adopting a new electronic records management system. These situations demonstrate that records managers and archivists need to be ready to take advantage of opportunities, but can't force the issue.
- Shawn pointed out that education must be ongoing because every time a state has a new CIO,

[4] Global Information Locator Service; http://www.gils.net/about.html

priorities change. Bob mentioned that Minnesota has had nine CIOs in less than ten years and Cal pointed out that the average term of a state CIO is 26 months. Bob said that working with state legislatures is actually easier than working with the state CIO's office because the legislative staff is often more stable.

- Tim Orr announced that the next NALIT newsletter will have an article about the wrapper; it should be out by the end of January or beginning of February.