Preserving State Government Digital Information
NDIIPP / Final All Partners Meeting Summary

Tuesday December 6 – Wednesday December 7, 2011
Minnesota History Center
St. Paul, Minnesota

Attendees

California Office of Legislative Counsel: *Bill Behnk and Mendora Servin*

Illinois State Library: *Andrew Bullen*

Kansas State Historical Society: *Patricia Michaelis and Matt Veatch*

Library of Congress: *Butch Lazorchak*

Minnesota Historical Society (MHS): *Shelby Edwards, Jennifer Jones, Carol Kussmann, Charles Rodgers, and Shawn Rounds*

Minnesota Legislative Reference Library: *Elizabeth Lincoln and Mike Schatz*

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

National Conference of State Legislatures: *Pam Greenberg*

Tennessee State Library and Archives (TSLA): *Cathi Carmack and Wayne Moore*

Tessella: *Mark Evans and Mike Thuman*

University of North Carolina: *Christopher (Cal) Lee*

Vermont State Archives and Records Administration*: Tanya Marshall*

Meeting Summary December 6, 2011

Welcome, Introductions and Orientation; Jennifer Jones, Minnesota Historical Society
[no notes]

State NDIIPP Evaluation Report; Christopher Lee, University of North Carolina
Presentation Summary

Description of the four state NDIIPP projects including:

1. Persistent Digital Archives and Library System (PeDALS) led by the Arizona State Library Archives and Public Records
2. A Model Technological and Social Architecture for the Preservation of State Government Digital Information (MTSA) led by the Minnesota Historical Society.
3. Geospatial Multistate Archive and Preservation Project (GeoMAPP) led by the North Carolina Center for Geographic Information and Analysis

4. Multi-state Preservation Partnership led by the Washington State Archives

Twenty six states are participating in one of the four state initiatives; ten are participating in two of the four projects, and 15 states are not participating in any of the projects.
The idea for completing project evaluation of all four state projects was initiated at the Best Practices Exchange in October 2010. The evaluation process was done with various methods including site visits, analysis of project deliverables and documentation, as well as monitoring project activities and announcements.

The main questions that were asked of participants during the evaluation focused on why they were interested in the projects, what related activities have they been involved in over time, how did the project fit into the mission and goals of their organizations, and how can the activities move forward. The evaluation itself answered questions such as which products and lessons from the project are most or least likely to be applicable to other states and why?

A general description of the *PeDALS project* was given including the project goals and products. These included developing a curatorial rationale to support an automated, integrated workflow to process digital collections; implementing digital stacks in an inexpensive storage network; and build a community of shared practice in which barriers were removed to keep costs low. Products include a repository system based on BizTalk and LOCKSS and the PeDALS email extractor.

Discussion
Was the issues of scale addressed in this project when working with the transmission of files? PeDALS did not test scale per se; MetaArchive is doing something similar at a larger scale.

The PeDALS project also developed an email extractor. It is not known if this has been adopted by others outside the project. There are a number of XML email approaches such as the one developed by the Smithsonian (Email Preservation Parser), but again the level of adoption is unknown.

A general description of the *A Model Technological and Social Architecture for the Preservation of State Government Digital Information (MTSA) project* was given. This project worked with state archives, libraries, and legislative communities, to explore access and preservation methods for digital legislative records. The project focused on education and produced guidance documents; testing access and preservation systems with different partnerships; and assisted other similar efforts such as KEEP.

The *KEEP (Kansas Enterprise Electronic Preservation) system* was developed with various funding sources including NDIIPP. KEEP is an enterprise wide preservation system for all branches of state government. In Kansas a bill was passed that authorized the State Archivist to set standards on the authenticity of electronic government records, certify systems for compliance with standards, and serves as the agent for authenticating records. The funding model includes fees for authenticating documents.

The *GeoMAPP project* focuses on geospatial content. The project team produced guidance on ingest and management of geospatial data as well as working on business planning resources. The *Multi-state Preservation Partnership* was built on the Washington State Archives digital archives environment to implement a regional repository for state and local digital information. They developed a component for submission (ArchiveThis!) as well as a component for ingest (Auto Todd). The code is written in C# and built on a Microsoft platform.

Discussion
Is there a concern about sustainability for when the system is built on a Microsoft platform? The underlying technology is Microsoft, but the funding is not. Much of the funding comes from record fees.

Project team members have also stated that the overlying code could be done in open source but this specifically was not tested as part of the project.

If this type of system was to be replicated, how would it be done? Would other states be able to use Washington's infrastructure? Could similar regional centers be developed?

During the evaluation process it became clear that many personnel and leadership changes had taken place over the duration of the projects. There were 20 changes in lead states alone. In total there were over 44 personnel and leadership across the four projects (numbers are probably higher).

Overall the major lessons and themes included:

- Building on strengths (past activities, local opportunities) [List of strengths from each lead partner was given]
- Building networks
- Ability to persist in face of dramatic disruptions
- Diversity of approaches from which others can learn is an asset
- Progress is often specific to content type
- Working on the flexibility of contractual arrangements

Specific observations on the MSTA project included the focus on collaboration, exploration of specific technical approaches, and creation of resources. This project took the least 'custodial' approach and worked on testing and sharing information.

General suggestions for other states based on the evaluation include:

- Progress comes through incremental advances
- Over the next few years, identify priorities of digital preservation
- Review what others have done, and 'follow' the ones that might help advance your priorities; share your experience
- Plan for sustainability in the face of continuous disruption
- Collaborate and look outward for digital preservation advances and resources

Other digital preservation activities and trends include umbrella initiatives such as NDSA, OpenPlanets, Digital Curation Centre, and Preservation and Archiving Special Interest Group

(PASIG). These organizations and others have also created software and support for digital preservation [examples on slide 19]. There are also specialized commercial vendors/systems available one of which MTSA tested. [slide 20].

Some other organizations have worked on audit and certification on what makes a good digital repository. In the risk management tool, Drambora, you can now pick specific functions that are applicable to your situation or may be a priority, rather than going through a complete system analysis. TRAC is another tool.

> *Discussion*
> If TRAC is an all or nothing tool, and Drambora has allowed users to choose what they want to evaluate for themselves, what direction do you see this going in? Possibly going in both directions; larger institutions may find it necessary or important to be completely certified while others might want to be able to focus on what specifically applies to them based on priority.
>
> As a general model TRAC has been guiding KEEP. TRAC is the goal we are aiming for, this sets the bar high. As many parts of a system might be out of your control, all you can do sometimes is showing due diligence. For example, we need to make sure the vendors will meet these requirements as well. The requirements need to be passed along during system development.

Digital preservation is a dynamic landscape and we will never be able to level the ground, but we will always need to monitor the activities around us.

Authentication; UELMA Background, Minnesota Prototype, California White Paper

The Uniform Electronic Legal Materials Act (UELMA); Michele Timmons, Minnesota Office of the Revisor of Statutes

Presentation Summary

Michele became involved with the development of the Uniform Electronic Legal Materials Act in 2007 when the American Association of Law Librarians (AALL) conducted a state by state study on issues related to electronic legal materials. The AALL was asking what people were doing about preservation and authentication of digital legal materials. At the time, Minnesota was not actively pursuing these issues, but Michele had thought about these and other issues related to online security. In Minnesota, the executive branch decided to no longer publish the state register in print form, which started her thinking about these issues and how they might pertain to the legislative branch. AALL asked Michele to be a speaker at their 2007 summit meeting about authentication. As a Uniform Law Commissioner, Michele also addressed the Uniform Law Commission at the same summit about the same issues.

At the end of the summit, the efforts included a discussion of technologies, best practices, education, and a legal effort to establish a legal framework for the authentication of electronic

legal materials. Michele worked with AALL to write a proposal for a study committee and then on the drafting committee for a uniform act on authentication.

UELMA was approved on July 12, 2011 by the Uniform Law Commission. It established an outcome based framework for authentication, preservation and access to electronic legal materials. The idea of the act being outcomes based was important; the solution was not based on a certain technology that would need to be replaced. The main objective was to provide online legal material with the same level of trustworthiness traditionally provided in a law book.

UELMA requires that official electronic legal material be authenticated by providing a method to determine that it is unaltered; preserved in either print or electronic form; and accessible for use by the public on a permanent basis.

Examples of how other agencies are authenticating records:

U.S. Government Printing Office: One of the first agencies in the US to implement authentication. Uses the Adobe LifeCycle solution. The slide (4) shows a blue banner at the top that indicates the document uses certificates and is authenticated. The signature panel gives you additional information about the certificate. The blue eagle at the top left corner of the document is the watermark the GPO uses on official government documents.

State of Delaware: Uses a small subset of the Adobe LiveCycle solution. The certification process in Delaware is manual. If the number of certifications needed is low, this process may be more suitable for your situation. The slide (5) shows an example of a certificate that is NOT valid.

State of Ohio: The Supreme Court of Ohio no longer publishes court opinions in print. They also use the Adobe solution. This slide (6) shows that there is a problem with the signature, not the certificate.

State of Utah: Utah is not using Adobe, but uses hash values to show if a file is authentic or not. A hash value is created by running an algorithm on a file that creates a unique id based on the document content. Utah provides legal documents in html and compressed rtf; hash values are created for the rtf files using the MD5 hash. They provide a document that lists the MD5 hash values based on file name. To verify authenticity, users download the rtf file, and run it through a free hash reader, and then compare the values. If the values match, the document has not been changed, if they don't match, the document cannot be authenticated.

Minnesota Prototype

Minnesota looked into Adobe LifeCycle as the GPO and Deleware use it. Neither approach was a good fit. Started to look into building something based on what Utah was doing but making it a bit more user friendly.

Minnesota already uses secure server technology (https) to transmit files; this uses layers of encryption. Also in the URL field, there is a lock; this lock shows a certificate that states that yes this is the Minnesota Office of the Revisors official webpage.

Minnesota provides legislative information in both PDF and HTML. The main body of the website is in HTML so it can be navigated and searched easily. The PDF version which can be downloaded and shared will have a hash value associated with it. If you want to authenticate a PDF document, you will be able to do this on the Revisors website.

From the Revisors website, click the authenticate button at the top of the page; browse to the file on your computer that you want to authenticate; click authenticate. A message will appear in green if the document is authentic, and red if it is not. The hash values in the prototype are being created using SHA1. The hash code for the document is also provided, so the document can be independently verified if someone wishes to do so.

Minnesota is thinking of moving to SHA2 for the authentication algorithm, based on discussions that took place at another NDIIPP meeting regarding California's white paper on authentication. Feedback from a group of Minnesota law librarians suggested adding more detailed documentation for the authentication process on the website. More user testing will be helpful.

Discussion

Is this a record in itself? Will you keep track of people verifying documents? This is probably not necessary and we don't plan on saving this data. As documents are sent to different people, they can continually be validated as needed.

Kansas is also thinking of something like this. They would however need to save the authentication record. The key is being able to go back to the repository and verify the document.

Discussed what might need to happen if you move from one hash value to another. You can re-run all the hash values, and match the old with the new.

How do you organize the data behind the scenes? If you use file names, you might run into problems. People change file names all the time. The hash values themselves do not take the title into account, so the title can change and the value will remain the same; however if your backend database is tied to these file names there may be problems. There needs to be a key that links the hash to the records. How will the computer find the files? Minnesota retains the hash values in an Oracle database behind a firewall and expects to someday have to rerun hashes; does not expect varying file names to affect this process.

What is the legal effect of authentication?
If a record is authenticated, it is presumed to be an accurate copy of the legal material. Presumption applies in another state that has adopted UELMA. The party contesting authentication has the burden of proving by a preponderance of the evidence that the legal material is not authentic.

What is required if preservation is done electronically?
Preservation must ensure the integrity of the record, provide for back-up and disaster recovery and ensure the continuing usability of the material.

The preservation options for digital material grid provides a starting point on what can be done based on good better and best practices for particular preservation issues. The Minnesota Digital Library created a preservation options matrix based on available preservation systems.

What materials are covered by UELMA?
The state constitution, session laws, state code or statutes, and state agency rules that has or had effect of law are considered mandatory, whether or not they are in effect. Others are permissive; state agency decisions, reported decisions of specified courts, state court rules, and any other category of legal material.

Who must implement the requirements of the act?
For each type of legal material, the state must name a state agency or official as the 'official publisher'. The official publisher has the responsibility to authenticate, preserve, and provide access to that specific material.

What electronic legal material is official?
If the legal material defined in the act is published only electronically, it must be designated official and must meet the requirements of the act to authenticate, preserve, and provide access to the material. If there is both a print and electronic version, the electronic version may be designated as official but the requirements to authenticate, preserve and provide access must be met.

When does the act apply?
Prospectively to official electronic legal material first published on or after the effective date of the act as chosen by each state.

Are there issues not addressed by UELMA?
The relationship between the official state publisher and a commercial publisher left to contract law. Copyright laws are unaffected as well as rules of evidence.

How will the act provide guidance as technology standards for electronic legal material continue to evolve?
The official publishers required to consider most recent standards regarding authentication, preservation and security, and public access. Consider the harmony with methods and technologies used by official publishers in this state and other states that have adopted the act. The outcomes-based requirements allow flexibility to respond to emerging standards.

After 2011 ULC Annual Meeting, UELMA was styled by Uniform Law Commission in September, 2011; the chair and reporter finalized comments in October, 2011; and UELMA will be presented to American Bar Association in February, 2012. UELMA is ready for 2012 enactments. The Act is "targeted" by Uniform Law Commission and AALL will be instrumental in enactment efforts.

California and Authentication; Mendora Servin; California Legislative Counsel
Presentation Summary

California has been involved in various aspects of the NDIIPP project. One of these has been working with historical statutes from 1849-1988. These records have been digitized and used the project's XML wrapper, as well as California's XML format (CAML) for drafting, and will be made available on the pubic website. Last month after hearing that Minnesota uses a secure website for legislative records, California hopes to follow suit.

This project has given us a strong commitment to archiving our documents. In the past, it was thought that archiving should be the responsibility of someone else, the state archives, or the publishing group. But the idea kept coming around that as creators of the electronic records and the fact that there was a law that said the records need to be available, the importance of preservation kept bubbling up. The Office of the Legislative Counsel in California also wants to make authentication part of their daily workflow.

The question of what is archived or should be archived started because of their involvement with this project. Part of this was also driven by the new drafting system, and that people no longer have to manually put documents together – the printing company no longer typesets.

In California, there is currently no official publisher, and nobody is responsible for the long-term curation of the records. The Legislative Counsel is working at getting UELMA enacted so they can take responsibility for the electronic records they create. This is a huge change in our philosophy.

California has two requirements for authentication; the ability to verify that the document is actually from the source that it claims to come from and that the document has not been altered since it left its source. (Authenticity of origin and document integrity).

California will be looking at authenticating chaptered bills (statutes), chaptered resolutions, constitutional amendments, the state constitution, and state codes. While thinking about taking on this responsibility, the questions of methods and cost were brought up.

California is working on a white paper that addresses both of these issues. The goal of the white paper is to explore some of the options, associate costs, and determine what the best option is for California. The paper covers California's current legal landscape, authentication requirements, descriptions of basic methods, commercial offerings, and six typical configurations with their associated prices.

The Legislative Counsel currently doesn't archive or authenticate anything on the web site. Originally planned on only having 4 years of data on the web site but because of requests from the public we stopped removing previous sessions of bills off the web site. We currently have Bill Information back to 1993 on the web site. This is the only archive and public access we have and is not planned archiving but archiving by default and is not where we want to be. A team of Lawyers and technical staff are now planning a long term archiving process and they will also address the authentication process.

Slides 7 – 10 provide a visual of authentication building blocks and processes. Building blocks include hash codes, encryption, certificates, digital signatures, certificate authorities, and public key infrastructure. [The white paper will contain detailed information on these.] Slides also visually show the verification of a hash code, the creation of a signed document, and a PKI signed document.

> A certificate authority is a group that has the authority to say that a document is authentic. You have to make the decision, do you trust that authority. This can be done by self signing methods or by using a higher authority.
>
> PDF is a standard, not an Adobe product. Other vendors can use the open standard based around PDFs to create signed documents. In some cases, all you need to do is open the document and the authentication result is displayed. The method is embedded within the document.
>
> PKI is the highest level of a signed document. A company such as VeriSign or InTrust issues a Certificate Authority to an agency, giving them the right to create and authenticate. Adobe products look at the hierarchy of certificates and because they approve VeriSign, they trust the certificates given out by VeriSign, so these documents are also automatically verified. (slide 10)

The components of a solution include file type considerations, certificates, signing software, and validation software. You must understand what types of files you want to authenticate (xml, rtf, html, pdf…). California has XML files that use a metadata wrapper and also PDF versions for archiving; the question remains if they will work with other document formats such as the rtf and html that are online for long-term preservation.

Slide 12 shows six typical configurations of authentication methods. These rank the level of security, the volume of records, appropriate formats, initial costs, and annual costs.

The next steps are to make the white paper available, to choose and implement a strategy, enact UELMA, develop an archiving solution, and make records available to the public.

> *Discussion*
> There was talk about certificates in detail; specifically about how the GPO has their own certificate authority, where they issue certificates to other agencies. This idea is very interesting; if there is a way to have the government or a government agency that will be

in existence 'forever' there will be less risk of relying on an outside company as an authority.

California plans on authenticating down to the code section of the law; with a high volume the question about self-signing, or paying for a certificate authority is still up in the air. California is considering using XML with Java for digital signatures or possibly the Multi-Doc Type configuration. California has access to IT developers which is what the initial cost depends on.

[For more information on the California White Paper and the authentication methods, please see the presentation by Bradlee Chang on these Authentication methods. http://www.mnhs.org/preserve/records/legislativerecords/authentication.htm]

XML Standards for Archiving Legislative Records; Dan Dodge, Thomson Reuters
Presentation Summary

It would be nice to have a schema that everyone could use, but others have tried to do that with limited success. The real goal was to archive material, so we worked on developing a schema to wrap any type of legislative material; however the project team focused on bill data. With this, there were no transformations to write, no loss of semantic markup, and no mapping to common structures. The wrapper was a way to transport data (xml, html rendering, and binary attachments).

The basic document structure (slide 6) shows the file (the outer box) with the major sections of metadata, XML source, HTML rendition, and the encoded attachments (encoded as Base64 or other plain-text format).

Discussion
Are you worried about the Base64 encoding changing? It could, but that goes along with making the decision to work with a standard technology.

The core schema metadata elements include an identifier, title, type, jurisdiction, agent, date, session, description, subject, relation, governor action, management history, and rights. Of these 13 elements, only seven are required. [See the XML Schema working group page of the NDIIPP project[1].] The ID for example, is the unique identifier, but it only needs to be unique within a state, as the state codes would differentiate the two 'collections'. Examples of the metadata are also given in slides 9 and 10.

The requirements for the source XML include preserving the original text content and original markup. Example on slide 12. The xml.source tag is at the beginning and ending of the [CDATA] section. This tells the programs not to parse the data but to treat it as text. This allows the markup to be passed through – including special characters.

[11] http://www.mnhs.org/preserve/records/legislativerecords/xml1.htm

The requirements for HTML include allowing an optional HTML rendition to be included (HTML / XHTML) and data with and without named entities. Example on slide 14. This allows the information to be previewed easily. The 'html.rendition' tags are at the beginning and end of this section.

The requirements for the attachments include allowing encoded binary files to be included, and that the format types are extensible to include almost anything including Microsoft Word, PDF, compressed zip files, RTF, Images. The attachments themselves look like a jumbled mess of characters, because the need to be un-encoded. Example on slide 16. Slide 17 shows a file after it has been run through a Base64 encoder making the file readable.

The proof of concept in Minnesota was to create a prototype using the XML wrapper. The Minnesota Revisor's office made the standard available on a website. To call up a certain bill, users must enter parameters into the web address to select a particular year, session, legislative body, bill number, engrossment, and formats. A zip file is then generated that included the object as well as two hash code files (md5 and sha1) that could be used for authentication. This can be seen in slide 19.

The California Legislative Counsel also tested the wrapper using chapter bills. These included the metadata and XML source (using the caml namespace). Upon initial review, gaps within the metadata values were identified. Metadata definitions were them modified to make the metadata schema more extensible to other states.

[Sample files from Minnesota and California can be seen on the XML Working Group webpage.[2]]

The next steps would be to complete additional features for proof-of-concept. These include having a secure interface for aggregation, creating a centralized repository, and being able to authenticate/verify transmitted data. Having other states test these would be beneficial

Some of the possibilities include creating a toolkit based on the website interface the Minnesota Revisors office created; explain how to do this and why it might be beneficial to them. Create a toolkit for verifying the contents of an archival package. Create a legislative content XML schema template to make it easier for states to adopt.

> *Discussion*
> As stated earlier it might be useful to include a digital signature into the metadata. It will be helpful to think about this and see where and how it would fit.
>
> *Is this a PHP program that is run on the website [MN prototype]?* The program creates the XML file on the fly, and runs the MD5 and SHA1 algorithms and creates a file for each. These three files are then zipped together and sent to the requester.

[2] http://www.mnhs.org/preserve/records/legislativerecords/xml1.htm

How will you accommodate other states or new changes? Will the schema continue to be modified? The working group can continue to meet and make changes as suggested by users.

It might be helpful to look at PREMIS for information on the digital signature. The semantics are captured, but it is not known if it has been used in practice.

Preservation Systems; Carol Kussmann, Minnesota Historical Society
Presentation Summary

The goal of the project has been to gather and share information about access and preservation of digital materials. This past year has focused on preservation systems. We started by exploring different preservation methods, soon to be documented in a white paper. After learning a bit about different preservation systems and evaluating the options, you will need to determine what method is the best fit for your current and future environment and goals. During the project we were able to test web archiving and two micro-services based contract service repositories. We looked at CDL/UC3's Web Archiving Service (WAS) and Merritt (repository) and Tessella's preservation repository, the Safety Deposit Box.

Web Archiving

Before you decide to archive web content, it is important to evaluate what it is you are looking to capture. Ask the questions, what do you want to capture? Why? Is this material only available for capture through a website or is there another method that would satisfy your requirements?

You must think about the investment (time, staffing, and money), technical responsibilities (skills, equipment, technical components of crawling – blocked content, opting out), and administrative details (long-term access, scoping and test crawling, analyzing content, de-duplication, deletion of content, navigation and search-ability, public or private archive, costs)

We tested the Web Archiving Service. We created sites/collections, captured site/s, viewed the captures, and chose not to provide public access. Creating the collections, or group of websites to crawl was only part of the process. A large amount of time was spent evaluating the crawls to determine if the captures included too much or too little content. The entire process was documented in a white paper on web archiving using WAS as a case study.[3]

Contract Service Repositories

We also had the opportunity to test two different contract service repositories. One of these was Merritt, developed by CDL, and the other was the Safety Deposit Box (SDB), created by Tessella.

[3] http://www.mnhs.org/preserve/records/legislativerecords/WebArchiving.htm

As part of the partnership, CDL allowed us to test the repository in return for feedback on how the system worked for non-university content; our focus was on government records.

Working with Tessella allowed us to test a multi-tenancy instance of a contract service repository (shared system resources with access to only our content). Representatives from Illinois, Tennessee, Minnesota, and Vermont all took part in this process. Each state focused on their own interests; the resulting reports are on the project website.[4]

The main goal for testing both of these repositories was to test major repository functions: ingest, storage, management of data, preservation methods, authentication methods, and exporting data. The records that we were using were permanent public records. We tested various formats and content types, in large and small quantities.

Archive Organization
You must first understand how an archive is going to be organized.

Merritt defines a collection as a group of objects. An object can be made up of one or more files.

SDB defines a collection as one or more deliverable units. A deliverable unit is made up of one or more files. Deliverable units can be a parent to other deliverable units. Deliverable units can also have multiple manifestations.
When viewing the contents of the archive, Merritt shows the ten most recent entries of a collection, while SDB uses a folder structure for navigating to a record of interest.

Ingest
Both systems create a Submission Information Packet (SIP) that includes the object/s and system created metadata. With the main process we used with Merritt, Merritt created the SIP behind the scenes. With SDB, users create the SIP and upload it into the system.

With Merritt, you create a collection within the archive and then upload objects into the collection. Uploading objects can be done manually or as part of an automatic process. The system can handle both single file and batch uploads. Batch uploads work more effectively as part of an automatic process which we were never able to do because of firewalls and other system issues.

With SDB, the SIP can be used to create a collection and upload content at the same time, or to just add content into an existing collection. The process we used was manual but it could have also been automatic. Content could be made up of single files, a folder of multiple files or an entire folder structure; each could easily be uploaded in both the manual and automatic methods.

It is important to understand the details of available ingest processes and how they could be integrated into your current workflows and systems.

[4] http://www.mnhs.org/preserve/records/legislativerecords/Tessella.htm

Metadata
Both systems use metadata generated by the SIPs and the users.

User generated metadata in Merritt includes four metadata elements (author, title, date, and ID); SDB allows users to add any amount of metadata that may or may not conforms to any schema. With SDB, metadata can also be added after ingest to any entity.

Metadata becomes searchable. Not requiring specific metadata lowers the barrier for participation or entry into an archive while accepting any and all metadata adds flexibility to a system and increase the ability to find relevant content.

Search and Navigation
After content is in the archive, you must be able to find it again. Both systems use keyword searching on metadata fields.

Merritt can only find items based on system generated metadata (unique ID numbers) and the optional four additional metadata fields. If you didn't enter in any of this metadata, you will have to know the unique ID number to find a specific record; otherwise you are limited to navigating through files 10 at a time, in the order that they were ingested. You also are limited to searching one collection at a time.

SDB also searches system and user generated metadata, but it also indexes content of many document types (PDF, Word, XML..). The entire archive is searched. Searches can be narrowed down by using metadata filed filters. You are also able to navigate through the archive using a folder structure to view archive contents.

Retrieval / Export
Not only is it important to understand and know how to get content in, but you will need to get content out for various reasons. You may have a need for single file retrieval or access to a group of files for disaster recovery purposes.

Merritt allows for single file retrieval, retrieval of a complete object (multiple version), and a single version of an object.

The system we tested with SDB allowed for retrieval of single files and an entire deliverable unit. The system however could also be set up to export an entire collection or entire archive.

Both systems use fixity checks to verify the authenticity of packages.

Manage
You must be able to manage your content in an archive. Some of the main topics include reporting, editing content, storage, access, documentation, preservation, and user interfaces (described below).

Reporting: You want to be able to know what you have in your archive at any given time. Reports are one way to do this. Merritt did not allow users to run reports, but were capable of doing so on their end. SDB provides users with standard out-of-the-box reports that address issues of format types, number of documents, etc. SDB also allows users to create their own reports to address specific needs and requirements.

Editing content: Merritt was designed to be a repository for research and scholarly materials – materials that were expected to be updated. To upload a new version, the entire original submission needed to be re-submitted. This was easy if it was one document, but if it was part of a batch, you would need to go and find all the documents, and resubmit the entire package (only the new material would be saved). SDB does not work with versioning the same way, but allows users to add to existing collections or deliverable units at anytime.

Storage: Storage for Merritt is at UC3. Storage for SDB can be anything (local servers, cloud storage, etc....) but for the pilot it was hosted by SDB in the UK.

Access: Both systems use log in permissions for access. Merritt's system uses persistent URLs to track objects. SDB can be designed to provide 'public' access with other systems if desired. Public access was not tested in either system during the project.

Documentation: Background information and user guides are available online and within the Merritt system. Information is available as text and sometimes video tutorials. These can be accessed by anyone. SDB provides users with system and user guides. User guides were created for the pilot that addressed specific needs. The need for additional documentation was one of the points discussed as part of feedback given to Tessella. User guides are very valuable and should be very detailed.

Preservation
Many 'preservation' repositories are basically storage repositories. How do you actually preserve the content? How do you make sure what you put in stays there (remains unchanged) or what you get out is identical to what you put in?

Most repositories perform fixity checks to verify content has remained unchanged at the bit level. Fixity algorithms are run on documents upon ingest and can be re-run to make sure the file remains the same over time. Both Merritt and SDB run fixity checks.

In addition to performing fixity checks, SDB allows users to migrate from one format to another, i.e. Word to PDF, Tiff to Jpg… These transformations can be done to assist with either preservation or presentation. When a transformation is made, both the original and new manifestation are kept and tracked in the system.

The SDB system also includes a technical registry that can hold preservation policy information. These include risk scores, migration pathways, and other information about formats. Users can set risk scores and create preservation plans to address at risk formats.

User Interface
The user interface is also important to explore. (These comments are specifically about the interface when browsing the archive.)

On the home page, Merritt displays an object's primary ID and provided metadata for each object in a list of 10, based on the most recent files ingested. SDB uses a folder structure to navigate through collections, deliverable units, and files.

When looking at individual objects, Merritt shows the metadata and version information about a file while SDB shows information on the technical properties, the title, format information, fixity information, and a thumbnail of the file.

Lessons Learned
You need to know what you want to preserve and why? How you want to use the records now and in the future.

You must understand the elements/features of the system and how they fit into current routines. System features include, documentation, access and ownership issues, the user interface, how to deposit material, how the system can integrate with others, how to manage records, how to access files, how to preserve the files, and what the associated costs are.

It is also very important to understand the level of technical skills and experience that are required by users. What capacity do you have for learning new skills? How much technical knowledge is available, IT department time, time to learn new skills??

Safety Deposit Box Pilot Project
During the testing of SDB, feedback was given on both a general and detailed level. When possible, changes were made during the pilot; other changes have been and will be incorporated into future releases of SDB.

Mark will give a quick demonstration of the ingest process using a cloud storage system rather than the hosted method we used during the pilot as well as discuss some other changes made to the system based on our suggestions.

Background and Demonstration of SDB; Mark Evans, Tessella

Tessella was interested in understanding the state level requirements and possible uses for a preservation system. During the testing, stresses were put on the system that had never been put on the system before. This made the pilot project a valuable experience for Tessella. The feedback gathered during the pilot project has already influenced the technology roadmap of SDB.

One of the biggest challenges from the perspective of transferring and ingesting content during the pilot was working around the firewalls and security systems from having to use a server hosted in the UK. FTP was used to navigate through these, which was a slow process.

Tessella has started to look at a cloud storage system. The cloud offers advantages in that it can grow elastically with the content. The processing power in the cloud also adds flexibility.

[Demonstration of a SIP creation and upload using the cloud. This process showed participants how much quicker this method was than the FTP used during the pilot. Also discussed the development of a .warc file viewer.]

Tessella may consider a cloud pilot activity to gain a more complete understanding of what the service offerings should be. It is anticipated that a complete service offering is preferred alternate to states having to manage their own infrastructure.

> *Discussion*
> It might be useful to incorporate into the user interface an indication of the associated costs of performing an action such as ingest and preservation (cloud services). Tessella still has to complete the exercise of gaining a complete understanding of the costs of providing a "Preservation As A Service" offering.

Illinois Update; Andrew Bullen, Illinois State Library

The Illinois update described the system Illinois state agencies to permanently archive electronically published state documents into the system. The system is called EDI, Electronic Documents of Illinois (http://www.ediillinois.org/).

Illinois state agencies have logins and passwords to the system. They use a cataloging utility to associate metadata with deposited items. (One person can have login permissions for multiple agencies if desired.)

After the metadata record is created, it is placed in the database system which is pulled and extracted to the permanent archive on a nightly basis. The tagged data is then placed into EDI and waits for documents to be associated with it. Documents corresponding to the descriptive metadata are either manually uploaded or collected from a user-supplied URL.

The system then builds an XML wrapper around the submission package, which includes management data. Md5 and SHA-1 fixity checks are also done on the files. The XML record can then be translated into a web page which is human readable allowing access to the publication. As it was originally created, the search facility of EDI was not being used very well – search engines could only read the metadata and not the data in the deposited documents.

When users deposit items into the system, IL State Library staff have a two step acceptance process for the ultimate permanent deposit of the items and their associated metadata. At the

time of first acceptance, ISL staff can edit metadata, approve it, or send it back to the depositors for further clarification. (The system also uses calculated bigram and trigram natural language algorithms to see if the content was previously uploaded, and to assist in fixity checking.) The second acceptance is the final acceptance after which it is accepted for permanent deposit.

The system stores files in three different systems; XML, an HTML proxy, and a collection of the documents themselves. As previously stated the original method was not user friendly and it was hard for users to find documents in the system. ISL staff decided to convert the records into ContentDM form. (ContentDM was previously in use at the Library, at http://www.idaillinois.org/.) The archived records are now converted into ContentDM form. All records are or will be available through the Illinois Digital Archives (ContentDM). As 87% of the documents deposited in EDI are PDFs, ContentDM works well at retrieving them, since it has a robust PDF/OCR facility. We can perform OCR on the PDFs, exposing the content of the PDFs so we can use keyword searching functionality. There are currently 86,174 documents deposited in EDI, averaging 5.3 pages per document.

Future directions arrived at after participation with the Minnesota NDIIP project, include:

A.) Retrospective copy and conversion of deposited PDFs to PDF-A format; this archive copy will be kept along with the original document to ensure that all deposited items can always be accessed. Likewise, all images will have a JPEG2000 copy made, and all MS-Word documents will be copied and converted to PDF-As.

B.) Creation of an independent (from the system) fixity database, strengthening the audit mechanisms of the deposited documents.

C.) Batch conversion of deposited XML records at the time of creation into ContentDM acceptable upload files.

D.) Failover system. While the EDI system is backed up, the system does not have proper failover capabilities.

Kansas Update; Matt Veatch, Kansas Historical Society

There was a discussion of Kansas record laws and record governance. (slides 3-7)

Before KEEP (1996-2008), Kansas was involved in many digital education and training activities, partnerships, system design, and tin rusted digital repository development. (slides 8-17) KSPACe (2004) was a good exercise and useful, but lacked forward momentum and was not going to be the foundation for the new project.

KEEP (Kansas Enterprise Electronic Preservation) is a trusted digital repository for Kansas government records with long-term value. They system goals include using standards and best

practices, open source tools, and providing access to authentic digital records. Partners include the legislature, the judicial and executive branches, the information network of Kansas, the Library of Congress, and other business partners. This was being done in the middle of IT consolidation. There was an effort to keep multiple agencies from creating their own repositories and to develop financial sustainability.

Policy framework for KEEP was developed in September 2010 with a prototype in June 2011; estimated production is in spring and summer 2012. Progress has been incremental. The prototype has everything in it, but it will be deployed incrementally.

The KLISS to KEEP connector was developed to assist in automated transfer of legislative records to the KEEP repository using RESTian APIs for a variety of record types. This connector also allows snapshots of the KLISS repository to be reconstructed at any revision to reveal point-in-time instances.

The lessons learned include it is essential to create a flexible and pragmatic business case; you must develop and maintain partnerships; have project champions; be able to influence system design; have financial sustainability; and have resources for education and training.

Vermont Update; Tanya Marshall, Vermont State Archives and Records Administration

The Vermont State Archives was formally established in 2003, and was given records management authority in 2008. In 2006, the archives starting collaborating with the State's CIO's office and has been working on creating a fundamental culture change on how records are managed. State IT staffs had focused largely on hardware and software, but now they are being asked to know and do more things, such as records management

Currently the focus is on people, processes, and systems and how these important pieces need to be resourced and pulled together. One thing that the Vermont State Archives has noticed is that agencies do not want to lose access to their digital records; therefore the Archives' continues to focus on Enterprise (statewide) systems and the integration of records management functionality into these systems, including archival management. There are few, if any, electronic records management systems; but the mentality is if you build it they will come. Records management and digital preservation should, over time, become a systematic government activity, thus reducing dependencies on personal champions.

Discussion

There was general discussion on funding models currently applied in various states to support the management of records, especially digital record preservation.
Vermont currently has an internal service fund to support centralized IT services and systems that is based on the number of positions in each business unit (http://dii.vermont.gov/DII_Divisions/CIO/allocation). Kansas reported that theirs is transaction or needs-based.

Meeting Summary December 7, 2011

Minnesota Update; Shelby Edwards, Carol Kussmann, and Shawn Rounds, Minnesota Historical Society; Mike Schatz, Minnesota Legislative Reference Library

Resource Center, Carol Kussmann
Minnesota will be producing a resource center as a final product for the project. This will be called the Center for Archival Resources On Legislatures (CAROL) and will be focusing on four main areas surrounding digital records – foundations, access, preservation, and authenticity.

Foundation includes things like record appraisal, retention schedules, legal requirements, the use of standards and best practices as well information about record inventories and business cases.

Inventory Project; Shelby Edwards, Mike Schatz
The Minnesota Historical Society worked with the Minnesota Legislative Reference Library (LRL) on a record inventory project answering questions about records that require long-term care in Minnesota. The end result demonstrated the diversity of legislative records and possible challenges associated with long-term care.

Many of the records of concern are not part of official record laws, such as committee meeting minutes, audio tapes, audio files, etc. Many offices took care of these things themselves, and created records in whatever format they wanted. With digital files, the LRL was not guaranteed to get copies of these records as easily as they did with paper records. The first goal was to create the database (described below) to assist with storing retention information about record types. The next step is to use forms to allow people to add more data about the materials they are creating/collecting. LRL can't keep track of it alone. This is a proactive approach to driving a conversation about creating a preservation plan to better collect and preserve digital materials.

When creating the inventory, ten attributes were used to define and describe the chosen records. These included record title, record dates, record creator, format/s, access mechanism (how does a user access the records), retention schedule (yes or no), responsibility (who is responsible for the long-term care of the records), purpose/use (how and why are the records used), authenticity, and general concerns (other related information, a catch all such as storage size, how long do you want to be able to access these, more information on file format).

The inventory was created in Microsoft Access as a flat database table. After initial research and input, the database was shared with LRL who was able to add to and expand the inventory to include more specific information about digital file formats, especially about the digital audio and video formats. LRL will modify this tool and create a multifunctional web tool for others to use to learn about retention schedules as well as be able to find legislative records.

Creating a record inventory is a process and data will constantly change; share this information and it will become more meaningful and useful to others.

Business Case; Shawn Rounds
Understanding your environment and knowing your goals will be helpful in creating a business case for digital preservation. The business case itself can be a useful tool to help gain support and resources for your goals.

A white paper was produced that includes a list of essential elements of a business case, discusses some of the issues to consider when developing a business case for digital preservation, and provides resources for additional information.

Access, Carol Kussmann
Over the years we have tested and explored various access options and tools including, eXist (XML native database) and the preservation repositories (WAS Merritt, and SDB) that can also be used as access points. We also worked with Sunlight who created a Minnesota OpenGovernment web page and a mobile app using Minnesota legislative data. Results on all of these are provided on our website and will also be in the resource center.

In addition, we addressed web accessibility issues in terms of ADA compliance with white papers and information resources such as the W3C. The W3C not only addresses ADA compliance issues but web design, xml technology, eGovernment, and more.

As we focus on education about how to make government data more accessible to the public as well as to those who may want to reuse it; we created, with input from the Sunlight Foundation, a document on best practices for opening up government data.

Sunlight Projects, Carol Kussmann
In addition to working with Sunlight to create a document with best practices for opening up public data, they created two products based on Minnesota data. They created an Open Government site with MN data and are also developing an App with MN legislative information from their Open State data.

The MN Open Government page is based on the Federal OpenCongress site. The site takes Minnesota data from the 2009-2010, and 2011-2012 sessions and puts it all in one place. This site blends information on legislators, bills, news sites, and campaign contributions in one location. It is intended to allow the public to learn, track, and share information on legislative activities. Five other states have similar pages.

Sunlight also is working on an App that will allow Developing an App based on TexLege. The author of TexLege used Open States data (API, Bulk downloads), and through collaboration, the

underlying code could be used to support all of the states that Open State supports. The Minnesota app will be the first one released. Open States data includes bill actions, bill sponsors, bills, legislator roles, legislator votes, bill votes, and committees, legislators… in JSON or CSV files or APIs. The app will look something like the screen shots (slides 25-26). Users will have the ability to pick a state that Open States has data for, view data on Legislators, Committees, Maps, and news and more.

Preservation, Carol Kussmann
We discussed the preservation systems that were explored yesterday. Documentation of the testing for the preservation systems is available online and will also be available in the resource center.

In addition, there will be links to the previously created preservation options grid which provides suggestions based on the issues of policy framework, storage media, file format, security requirements, documentation, staff skills, and time; this can be a guide for issues to consider when trying to determine the level of preservation actions that your records require.

Other materials will also be produced that address the various preservation repository options and more.

During testing, the Minnesota Digital Library (MDL) also looked at preservation repository options for their needs as well as evaluate others based on a criteria matrix. Results are available on their website.

The Minnesota Digital Library created a matrix based on evaluation criteria specific to their needs when exploring repositories. Information on this report can be found on their website. MDL used about 50 different criteria. Criteria revolved around issues related to being able to ingest original records and associated metadata. In other words, how easy it was to get materials in and out in their original formats? How materials were were stored? What was the reporting functionality? Could retention schedule information can be automatically applied? How good was the partnership publicly and privately?

> *Discussion*
> *At the Authentication meeting in Washington DC there was talk about creating a white paper surrounding the issues of preservation similar to the white paper California is developing on authentication. Would it be possible to create a paper about the methods of preservation?* This would be a difficult task, as the systems that offer digital preservation cannot be equally compared. It might be possible to explain different types of preservation, but not compare solutions.
>
> During this project the business drivers for preservation were also discussed, but it is not easy to sell preservation, but rather access or some other business driver. Some of the

business drivers for preservation include e-discovery and lawsuits, public access, transparency and accountability, and disaster recovery.

In Minnesota, the e-discovery expenses are more of a cost of business. Sometimes it just takes one big scandal to bring records issues to light. The 35W bridge collapse started a huge discussion regarding record keeping practice in MN. Often it takes a law suit to bring light on questions such as record access, etc.

Authentication, Carol Kussmann
As discussed yesterday, authentication will be a large part of electronic records management moving forward. You must be able to prove that your records are authentic. UELMA will assist with moving this forward. Ideally, the research done by Minnesota and California will aid others in making forward thinking decisions for preservation plans.

All of these authentication resources are or will be available on our website and resource center.

Other Activities, Carol Kussmann
Other activities that were discussed yesterday include the metadata wrapper and schema, KEEP, the NDIIPP evaluation report, and our general education activities. As stated earlier, the resource center will be the final product that pulls all of the information together.

Discussion
Minnesota is looking for review and comments on the schema wrapper. We are still interested in people's feedback. California liked the wrapper, there were a few initial issues with metadata descriptions, but overall it was easy to use.

National Conference of State Legislatures (NCSL); *Pam Greenberg, NCSL*
NCSL has been a partner on this project from the beginning. As general background on NCSL, we serve over 7000 legislators and 35,000 legislative staff and have offices in Washington DC and Denver. The mission of NCSL is to improve the quality and effectiveness of state legislatures, promote policy innovation, and ensure state legislatures and strong cohesive voice in the federal system.

NCSL provides many services to its members, including information requests, creating and distributing publications, consultation, and educational opportunities. There are twelve legislative staff organizations, four of which shared many of the same goals of this MN NDIIPP project as well as two of the twelve standing committees.

Slides 12-24 highlight the joint activities between NCSL and the MN NDIIPP project through conferences, meetings, and publications.

NCSL has seen a rise in interest over preservation and authentication issues during the project; expected next steps include tracking related legislation technology initiatives, tracking UELMA bill introductions, and working with others on possible future sessions at NCSL and other organizations such as NDSA and others about authentication issues.

Discussion
What is the way to get bills published? The NCSL does not lobby in support of specific state legislation or model legislation, but makes information available. UELMA is not partisan so it will be easier to provide information to interested people.

What has been the experience of other model laws and their success? The Uniform Electronic Transactions Act was passed by many states quickly; UELMA will probably be slower because there are cost implications associated with adoption.

Library of Congress, NDIIPP; *Butch Lazorchak, Library of Congress*
Our job is to collect for the future, as well as preserve the past. NDIIPP has helped create national networks and partnerships that reach across traditional industry boundaries.

The Minnesota project in particular has been an iterative process, where we learned by doing. There were general ideas that flowed as the landscape changed. Being flexible has been an important element to the success of this project. The next steps are to determine how to hand off the work that has been done.

Looking at all of the projects there have been many areas of concentration and many tools and services have been developed supporting these. (slides 8-15).

The Library of Congress has been involved in other initiatives and projects. These include capturing the Twitter archive, focusing on personal digital archiving, and moving forward with the National Digital Stewardship Alliance (NDSA). Outreach is a large part of spreading the word. This is done through blogs, newsletters, Twitter, Facebook, videos, podcasts, book festivals, brochures, and other events.

[This leads into the final wrap up and conversation.]

Wrap Up; *All*
Open Discussion

The Memento Project draws on web archiving technology. Memento uses a server to house an index of what content various web archiving services have captured. This content is then displayed as a result.

In answering the question of what is next; it would be nice if work could be done on a data interchange format so that federated searches would be better utilized.

It might be useful to concentrate more on storage solutions, spending time on exploring DuraCloud or a federal version of DuraCloud. Signing contracts is a challenge, makes it difficult when working with vendors. Sometimes the time it takes to go through all the channels takes multiple years. It is also a challenge to overcome the idea that the cloud is just one big security risk, and being able to see the benefits. Butch asks what can DuraCloud do for you? Having yearly contracts would be beneficial, as well as possibly create some sort of network for state government data needs.

How useful do you think the wrapper schema is or can be? It is a good start. There are questions about how you define formats, if web services would be used, SOAP protocol… There is a communication layer and then the data interchange format. What works for legislative records may not work with other permanent state documents like the Illinois State Library collects.

Long-term preservation planning is important. You can't just put something on a shelf and forget about it as you could with paper records. You can't just forget about the data you create or collect. Information about creating preservation plans would be helpful for many.

It would be nice if there was 'someone' out there who could inform others on when a format or standard was getting old or out of date and then provide a tool to migrate old formats to a better one. If each organization has to figure this out on their own there will be a lot of duplication of efforts and a lack of standards across agencies. PRONOM and the Library of Congress's Sustainability sites are first steps at this but people are looking for more information on what to do, not just information about formats in general. People want a policy to adapt (a vetted policy) rather than trying to figure it out for themselves.

It would be interesting to have a pilot project that was able to test migrations on a large scale and document what happened, whether data was lost or retained. GeoMAPP moved things from place to place and documented the changes; perhaps something similar could be done with testing migrations. The results may be surprising. Many migration tools are being used blindly. There have been instances where there has been a loss of authenticity with migration tools. A tool registry that defines the tools and provides a critique may also be useful.

The Library of Congress can't make recommendations, but if an organization created a list that includes value judgments, others would be able to point to it. In this day and age, it will be difficult to get an organization to commit to making value judgments. One way around this might be to use something like Yelp – a more social network of information; crowd sourcing. Or perhaps create something on a Wikipedia like platform that could include a registry of transformation tools.

There has been a lot of discussion on what to do with records after they are created and are ready to be archived. People are interested in ways to communicate with record creators at the point of record creation, to streamline the process of archiving records. How do you get people to use common data standards before the records come to the archives? Information on formats, metadata, and exchange models all need to be shared.

In the past, we heard more about replicating data as part of a preservation method. Is this something that needs to be looked at in more detail? Many preservation repositories do this as part of their policies, and could be considered another option.

Sometimes moving data 'offsite' or out-of-state may not be possible due to legal constraints. What then? This is something else to think about. Using an outside vendor versus housing materials somewhere in your own state sometimes comes down to a policy issue.

Solutions must also fit within your own technical abilities.

Preserving electronic data is our responsibility and we do have to pay for it.

There are some concerns about how to keep things upgraded and accessible. We would like to avoid the cycle of 'upgrade' costs. If companies go out of business, we don't want to have to keep starting over.

It is a challenge to try to insert new costs into budgets that are constantly being cut. We need to figure out new ways to fund these activities. It is not something that can be done overnight.

Continued discussion on the cost of paper versus digital materials. You fund space (paper) vs. servers. You pay people to move the paper around and find certain records. You may lose some supporting staff with electronic records. People have always expected to have to pay for paper storage costs, how do you get people to expect to pay for digital storage? Perhaps we might be able to appeal to electronic collection storage as an essential function of government, so it remains a shared burden. Keep in mind that it is not always a cost comparison – storage may be cheap, but access isn't.

It is not just the cost of storage that you need to invest in, you need to make the commitment to keep up with the digital records to keep them viable. Digital records require care and feeding. This responsibility should not be a requirement for ever agency; you don't want to reinvent the wheel, you want to pool resources.

How do you preserve the legacy of public libraries? How do you make the library be a place where people can publish local things? Can you make a toolkit for libraries on how to do this?

COSA has conducted a study on electronic records and plans on developing a grant for education on related topics. Archivists are the main focus, but they can also find good partners within state government. The level of attention provided to digital records preservation is below average across the board. What does this mean for the future of electronic record preservation?