Preserving State Government Digital Information
All Partners Meeting Summary

Monday December 8, 2008
Cargill Commons, Minnesota History Center
St. Paul, Minnesota

Attendees

California Digital Library: *Stephen Abrams, Tricia Cruise, and Perry Willett*

California Legislative Council: *Bill Behnk and Linda Heatherly*

California State Library: *Kris Ogilvie*

Illinois State Library: *Andrew Bullen and Connie Frankenfeld*

Kansas Legislative Computer Services: *Terri Clark*

Kansas State Historical Society: *Matt Veatch*

Library of Congress*: Butch Lazorchak and Bill Lefurgy*

Minnesota Historical Society: *Nancy Hoffman, Bob Horton, Jennifer Jones, Carol Kussmann, Charles Rodgers, and Shawn Rounds*

Minnesota Legislative Reference Library: *Robbie LaFleur*

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

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

Tennessee Department of State: *Jami Awalt and Wayne Moore;*

Thomson Reuters*: Dan Dodge and Jolene Sather*

University of North Carolina*: Cal Lee*

Vermont State Archives*: Tanya Marshall and Scott Reilly*

Meeting Summary

Bob Horton from the Minnesota Historical Society gave two presentations to re-familiarize the group with the project, past meetings, current accomplishments, and future goals. He explained how the Minnesota Historical Society became involved with the government records National Digital Information Infrastructure and Preservation Program (NDIIPP) project and also mentioned additional NDIIPP projects being conducted by other states. A summary of the results from the site visits conducted in 2008 included main points of interest for each state. Common goals included the

authenticity of records, accessibility of records, preservation, and digitization.

To address some of the above issues, a next step in the project includes the building and testing of a system to capture and preserve legislative data. The recently developed model involves using an XML wrapper schema and a core set of metadata elements to mark up legislative data. This method limits the amount of additional work for the creators and users of information. The overall goal of the project is to create a framework that will be flexible enough for other states (current partner states – CA, IL, KS, MS, TN, VT – as well as states outside of the project) to use as a template to create their own method of capturing and preserving legislative records. It was stressed that each state is unique; and that they each have their own priorities, resources, audiences, and missions which is why a one size fits all solution is not feasible.

The afternoon session began with a presentation by Tim Orr from the Minnesota Revisor of Statues Office followed by a presentation by Daniel Dodge from Thomson Reuters. Both discussed the development of the XML Wrapper Schema for legislative records. Tim provided the background on the development while Dan discussed the technical details. There were many questions and comments surrounding the idea of the wrapper schema.

Bob Horton finished by discussing the work plan for 2009. If interested, site visits to each state would involve some sort of gap analysis to see how the wrapper schema could be used in other situations. The unique concerns and environments of the states involved in the project could provide the opportunity to test the wrapper with other data types or provide an opportunity to learn more about open government systems and methods of sharing different data types. Other activities for the next year include collaborating with other NDIIPP funded projects and continuing outreach and education about our NDIIPP project activities to the National Conference of State Legislatures (NCSL), the National Conference of Commissioners on Uniform State Laws (NCCUSL), and other possible user groups.

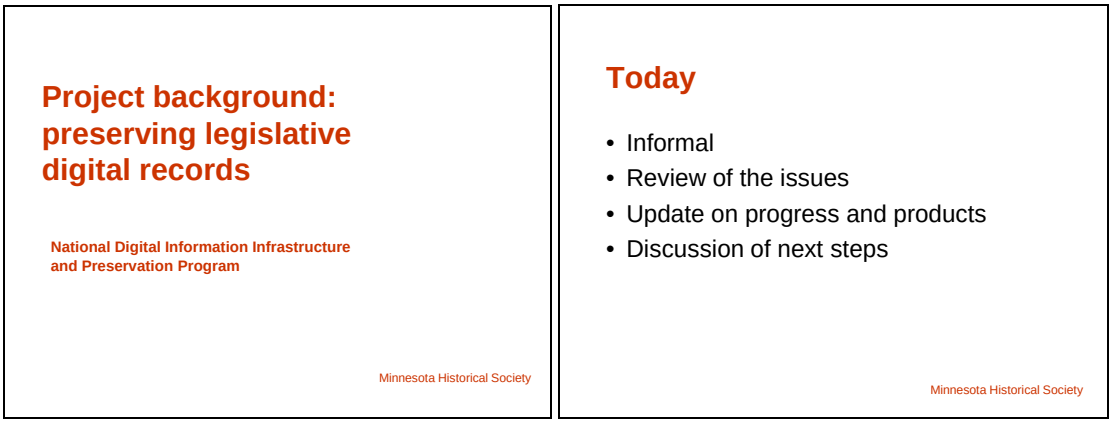
After further discussion about the goals of the gap analysis, the meeting concluded with the plan to follow up with each state individually to see what they were most interested in and what MHS could do for them in regards to this project.

Meeting Minutes and Detailed Notes about Presentations

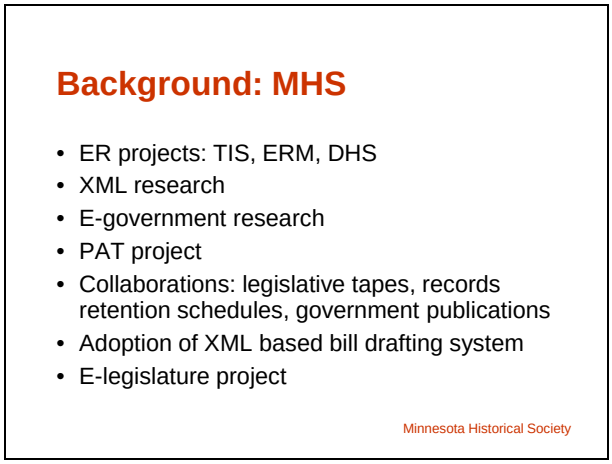
Below you will find the slides of the four presentations. Some slides are followed by text that provides additional information provided by the speaker. Questions, answers and discussion about the slides are enclosed in square brackets. PDFs of the individual Power Point presentations can be found on the All Partners Meeting webpage.

Project Background; Bob Horton, Minnesota Historical Society

Slides 1 and 2



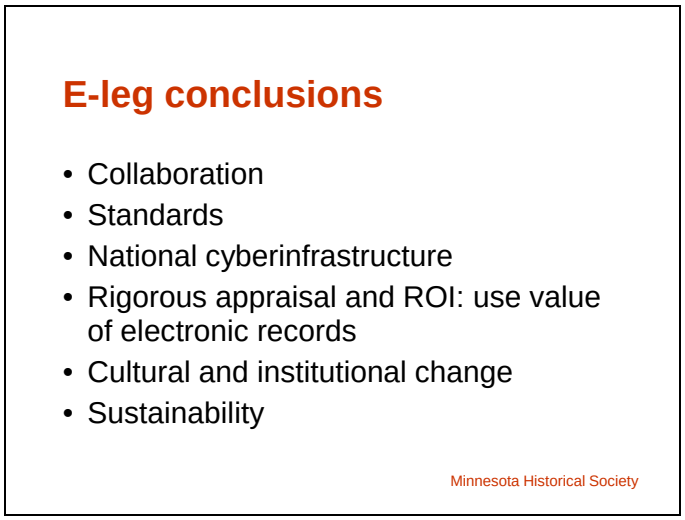
Slide 3



The Minnesota Historical Society (MHS) has been involved in a number of electronic records projects, including the Trustworthy System work and Preserving the Records of E-Legislature, both funded by the National Historical Publications and Record Commission (NHPRC). MHS has also developed a series of electronic records management guidelines, which have been adopted/adapted by other states, as well as a recordkeeping metadata standard for state government entities. Additionally, the State Archives has practical experience, having worked on contract for the Minnesota Department of Human Services and the Minnesota Bureau of Criminal Apprehension.

MHS previously worked with California on the E-Legislature project, as well as with the Minnesota Revior's Office and the Minnesota Legislative Reference Library. The E-Legislature project was possible because the Revisor's Office developed and implemented an XML-based bill drafting system (XTEND). This project set the stage for the business case for preserving and providing better access to Minnesota's legislative records.

Slide 4



Collaboration is essential; it allows institutions to share funds, technology, and intellectual strengths. Standards are needed especially during collaboration efforts so that all involved parties have the same basic framework to work from including common metadata and format types.

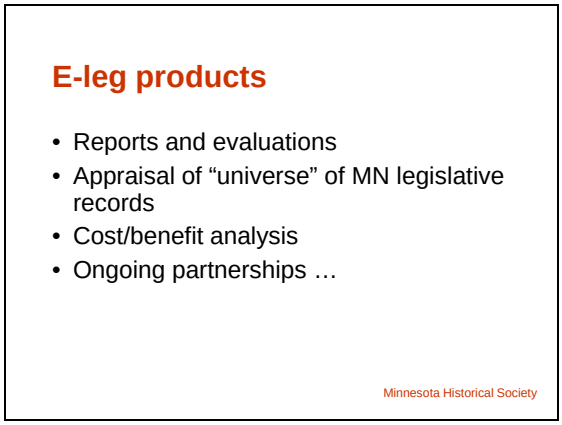
MHS also became aware of what others were doing on the national level, such as the Library of Congress (LOC).

You must know the value of your records, the use value of the electronic records. It is very important to do an appraisal on the records and program to see the return on investment. (Bob draws a graph of The Long Tail and indicates that it is really a small number of items that attract the greatest number of people. Many times we – archives - are collecting a lot of stuff that only a few people are interested in. When working with paper records, the storage overhead costs are minimal if the storage space exists, which is not true in the digital world. You can't manage information in the same way; the costs are too high for storage, hardware, software, staff time, etc. You don't want to be collecting a lot of stuff for a few people; you want the information collected to be useful to a lot of people to make it worthwhile.)

The business case for providing and preserving legislative records in digital format includes understanding that there are a wide variety of uses of government data that can be created for a little cost. This is one of the compelling reasons to invest in preserving digital legislative records. In addition to being the law to preserve and provide access to legislative information, the high use value increases the value of long-term preservation of the information.

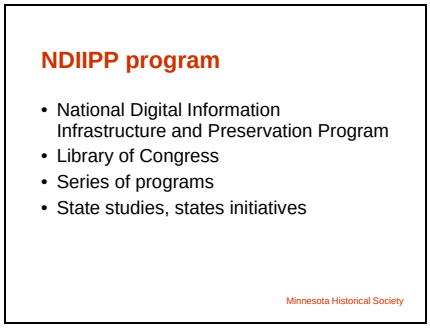
It will take cultural and institutional change to support digital archives. Here at MHS we still invest heavily on paper records, this will have to change if we want to sustain a digital archive.

Slide 5

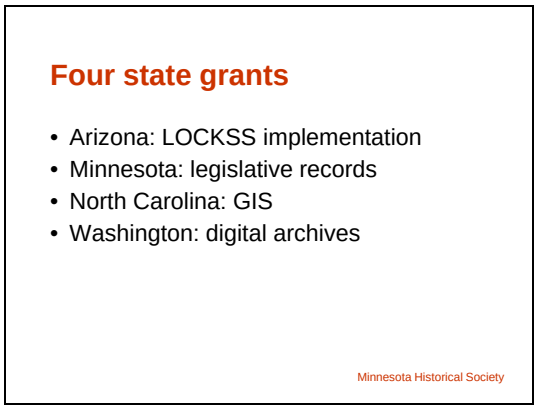


The E-Legislative project led to the current partnership with NDIIPP.

Slide 6

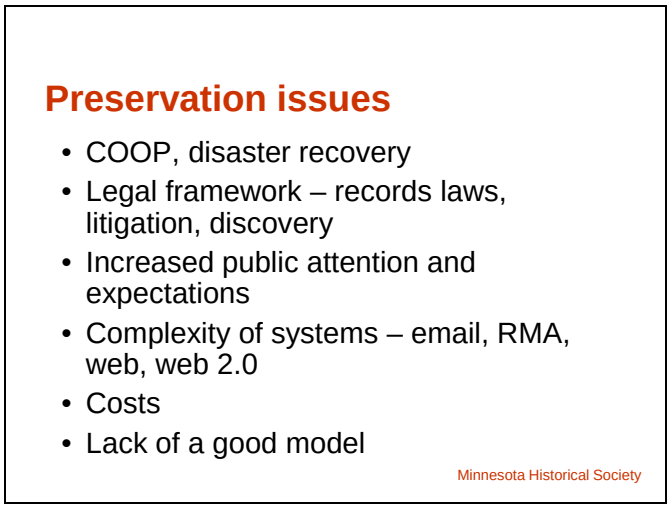


Slide 7



These four states are working with digital content with the intent to collaborate and create long-term applications that have not previously been studied. Arizona is looking at Lots of Copies Keeps Stuff Safe (LOCKSS) to determine if state archives can implement and maintain this type of system. We (Minnesota) are addressing legislative records. North Carolina is addressing GIS datasets, which we are particularly interested in following as there are possible parallels to our project. Washington has built a digital archive that currently serves the state government and for this project is extending its services to other states. Washington's project also includes a component that makes audio content searchable.

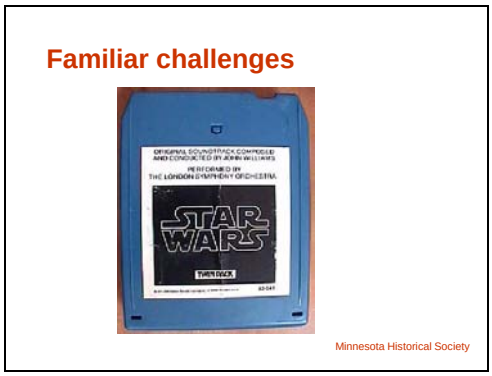
Slide 8



In addition to the legal requirements relating to the access of legislative records, people expect a certain ease of access to government information. They are used to finding things with a simple search or with just a few clicks of the mouse such as on YouTube or Flicker. Providing access is not as easy for government records. YouTube, for example, just needs to deliver video; it does not need to

link files of different types of datasets together. Government data is more complicated; different materials need to be integrated together for a complete package. Archives or government records must include emails, multiple versions of files, various applications and types of information. Currently we lack a good model for this type of preservation system even though many people are working on it. There seems to be bits and pieces of a model, but they are not very widely applicable.

Slide 9



Digital records face the same problems that archives currently face with other formats such as film, radio, TV, and audio files. All are subject to technological obsolescence and decay that continue to challenge archives. Once things are gone, they are gone. Digital files are no exception.

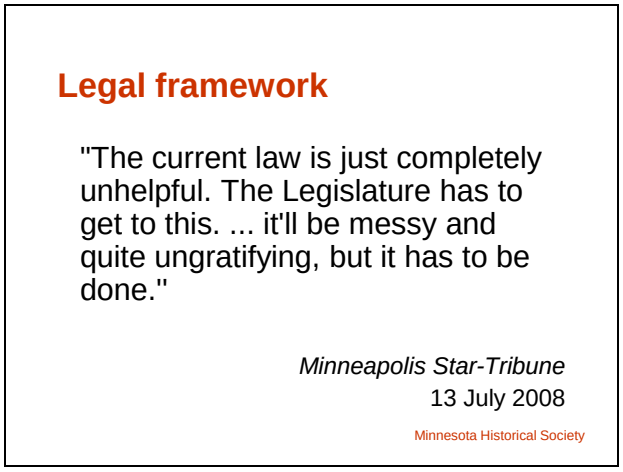
Slide 10

Must use a formula of migration, conversion, innovation, and collaboration as part of a preservation solution.

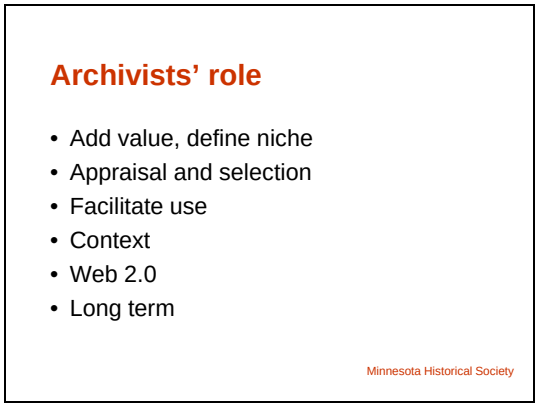
Slide 11

We must find appropriate solutions; there is not a one size fits all solution. Priorities and specific environments will determine the solution.
Solutions must be integrated into current routines to be useful; you don't want to completely change your routine, the costs will be too high.
Priorities must be set on what are the most important documents to capture. Selecting priorities forces you to review the use value of the records. [Andrew Bullen points out that you must also be able to catch lighting in a jar, to capture odd things and historically significant records that may not be part of your initial plan. You must be able to react to change and be ready to capture the needed information.]

Slide 12



Slide 13



As archivists we must define our role…

[Conversation between Wayne Moore and Bob about how the legal framework is often kept vague because it is so political. Bob states that politics and government reform will increase transparency, but political reasons to keep information close remain. Wayne comments that there is often no good incentive to change the record keeping laws. Why would political bodies want to change the laws to save more of the information about what they are doing?]

[Bob told a story about how one representative wanted to add a new law to the record that would allow operations to be shut down if they were not in compliance with the records management laws. They wanted to enforce the current laws, many of which organizations are not in compliance with. Under that proposed law almost everything could have been shut down. The law did not make it out of the legislative committee.]

[Connie Frankenfeld is working on trying to define what types of reports need to be kept and has created a list of required documents. She ends up getting the stuff she wants as well as a lot of unwanted stuff. Bob replies that it is often the case, which is one reason MHS focused on legislative records for this project. They do have a value, giving us the ability to address important issues without having to sort through unwanted items.]

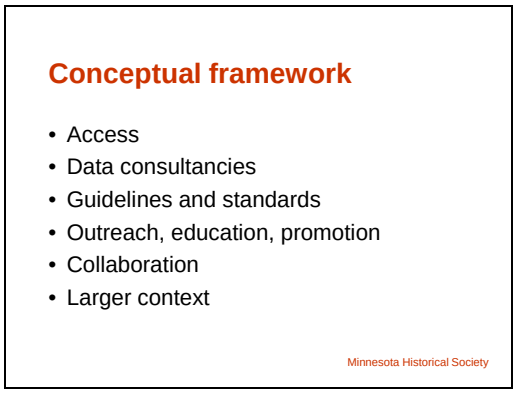
It is important to define your role, to find added value (something that helps you, and does not add to your cost). Find a niche that has not been covered, we think that facilitating use of the records will be key. Archives can link legislative material to other related materials, in a way that others can't do. There are many ways to use legislative data and we can't possibly think of all of them.

Web 2.0 will allow us to push out the content (legislative materials) so that others can use it in the way they want to. Our job will be to provide a standard data source that can be manipulated and used in conjunction with many other

applications. Others will create the products for their needs. Just look at all the ways GIS data is used. [Cal thinks that presentation and exposure will allow the public to come up with their own ways to use the data. Put it out there and they will use it.] Web 2.0 is cost effective because we are not building the applications to use the data, the people who manipulate the data are. Our goal would be to develop long term preservation of the data so people could continue to use it over time.

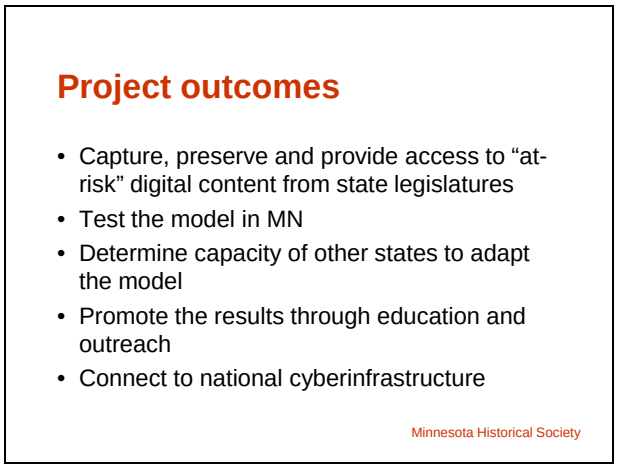
Bob talked about the paper; Government Data and the Invisible Hand… whose thesis is to create a simple accessible infrastructure that exposes data so that citizens can create the tools to use the data anyway they like. He also talked about Hack, Mash & Peer: Crowdsourcing Government Transparency.

Slide 14



We must facilitate access to the records. MHS would act as a data consultant with the ability to evaluate data, and then see if there is a way to enhance the data to make it more valuable. We must continue to look at the larger context of the technological world and keep track of what else is going on.

Slide 15



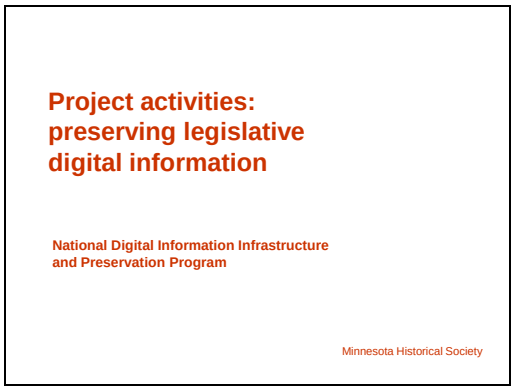
Questions…

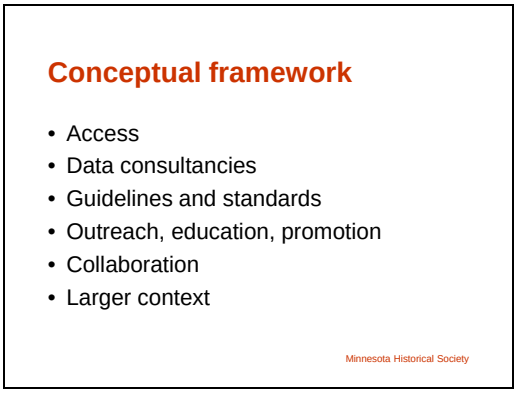
Question from Andrew, response from Bob: What about relational databases? How are you getting the information into or out of a relational database? We have not worked on gathering data from a relational database into the XML system, so it remains to be seen. We are currently ingesting data in separate pieces.

Comment from Tricia, response from Bob: When looking at slide 13 (archivist role), you have not included appraisal and selection, which would be part of the archivists role, correct? Yes, it should be in there as well. [The slide has been modified to include this.]

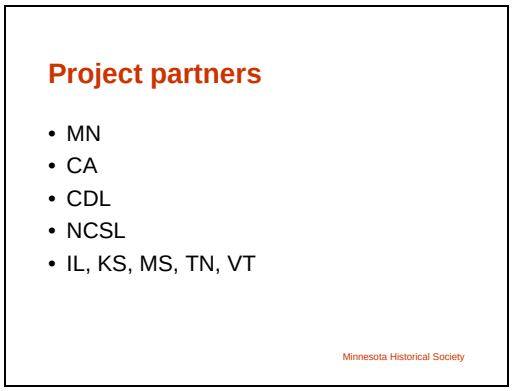
Project Activities; Bob Horton, Minnesota Historical Society

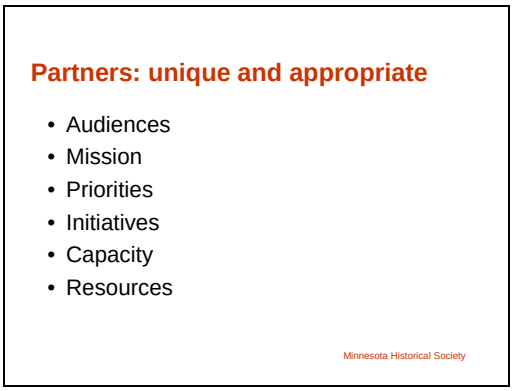
Slides 1 and 2



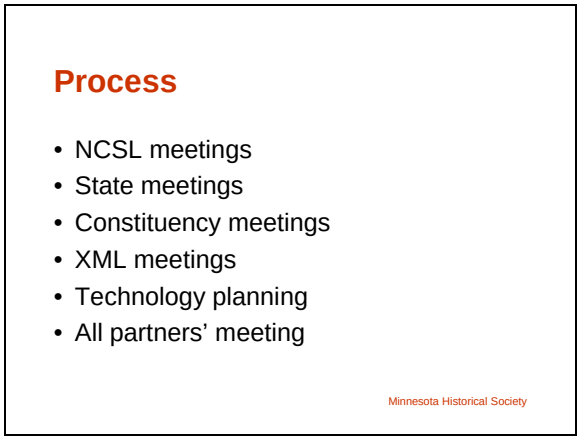


Slides 3 and 4



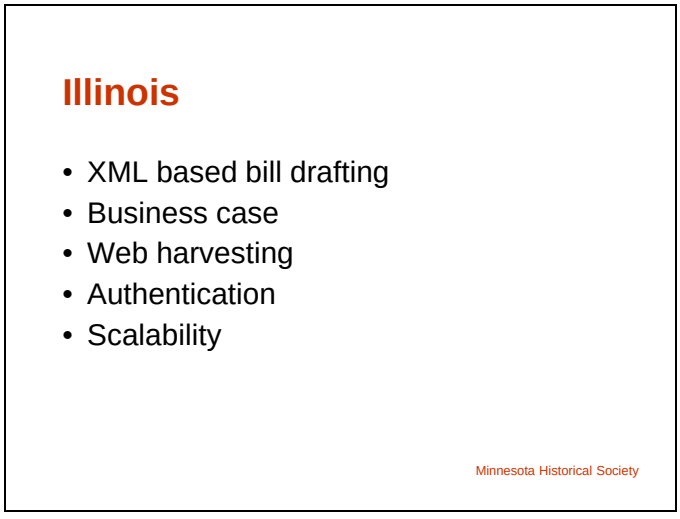


Slide 5



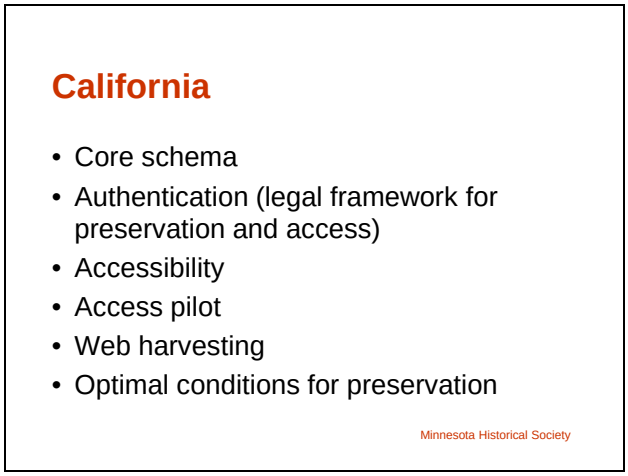
We have attended meetings to meet legislative IT people and held XML meetings here. (Propylon is working with KS and OR.) We researched what technologies are out there… Today's meeting is for feedback and to help define our next steps. I see us duplicating these steps in 2009, but moving towards implementation.

Slide 6



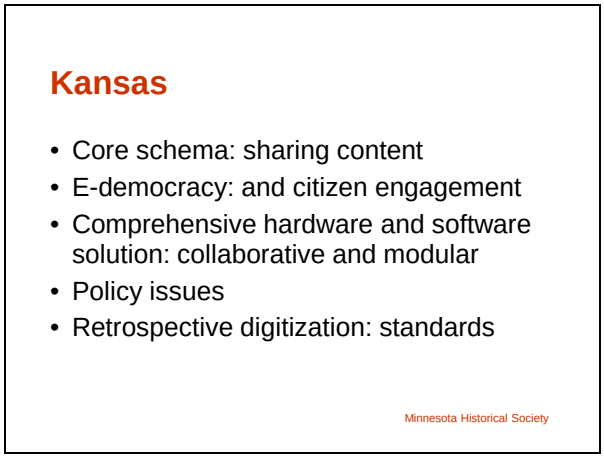
Illinois uses an XML bill drafting system. They are interested in business case scenario, web harvesting of government publications, authentication (chain of custody), and scalability. [Connie from Illinois recently asked the head of legislative information systems where the authoritative version of a document was and he could not answer; highlighting the importance of needing a way to authenticate records.] [Michele is looking to see if a state law is needed for authentication and preservation of legal materials online.] (A draft of the authentication paper written will be posted on Basecamp and on the project website soon.)

Slide 7



California suggested that if people are using XML, and others are looking at using XML, we could make recommendations to make it easier to others. This idea has evolved into the core metadata for the XML schema. The goal of the core schema is to be a support for interchange of information. In addition, preservation is not a one size fits all process. It might be possible to define optimal conditions for preservation in various situations that can make it easier for people to do what they can now instead of having to solve preservation problems in a time of crisis. [The goal is to complete a preservation grid with a narrative that will function as such a tool.]

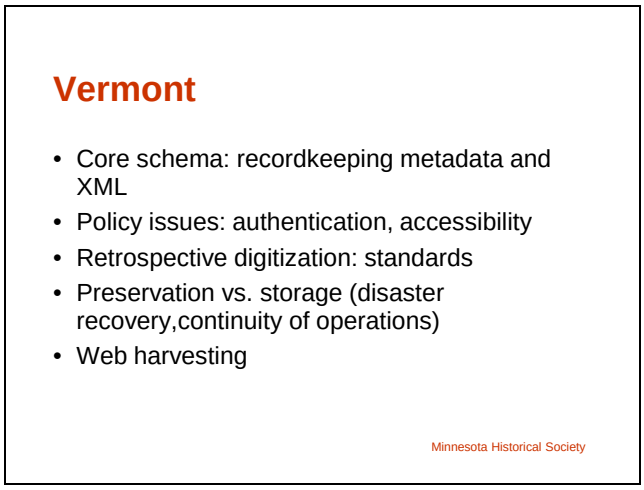
Slide 8



Kansas was also interested in a core schema. They have a very strong E-Democracy vision that makes the legislature open to citizens. They are currently using an architecture that allows citizens to be part of committee hearings. The process of allowing citizens to participate in legislative sessions mandates that many institutions work together including public libraries, university systems, and

the government. The system they developed uses a modular approach to hardware and software implementation so items can be rolled out component by component, increasing flexibility when needed. Other issues that are important to Kansas include authentication, preservation, retrospective digitization, standards, and policy statements. [Andrew asks if the steps for the E-Democracy plan are laid out, and Terri confirms.] [Bob feels that retrospective digitization is a bit outside the scope of the project of working with born digital materials but understands the importance of being able to interpret the past information in conjunction with the born digital materials of today.]

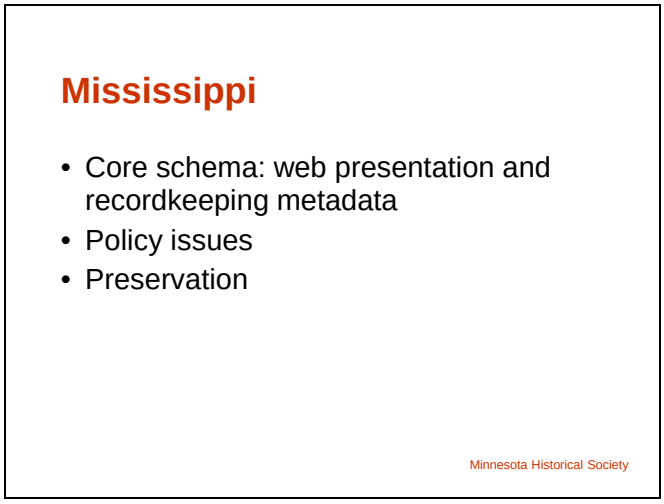
Slide 9



The concerns of Vermont mirror the other states as they also feel that a core schema, authenticity, accessibility, retrospective digitization, preservation, and web harvesting are important.

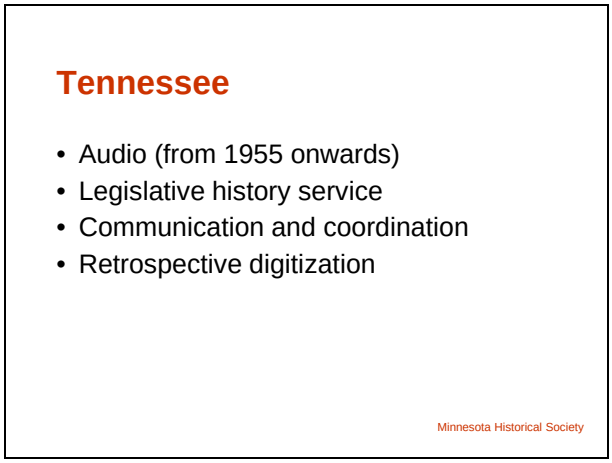
[Tanya explains that the state enterprise is conducted by the executive branch who states the paper copy is legal document, but the acts are done with LexisNexis which uses the digital copy as the legal copy. This highlights the importance of the need for authenticity.]

Slide 10



During the first visit it was said that nobody uses XML, but later it was determined XML is used during the transition between the proprietary bill drafting system and getting the information online. Mississippi was also concerned with the same policy issues and preservation issues as the other states.

Slide 11



Tennessee has a large investment in the audio format since 1955. They have recently moved into digital audio. The audio is of committee meetings and floor sessions.

[Wayne asks if the state archives or state legislature supports Kansas's efforts and is the web audio archived. Terri responds that they are not currently archiving the web streaming but would like to after the project components are in place.]

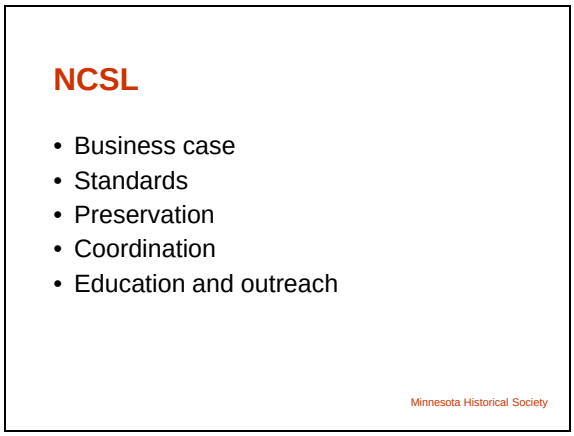
[Andrew asks about autograph recordings… Wayne replies that obsolete formats will be migrated and backed up.]

[Connie asks Wayne if the audio is available online, which it is not. The digital audio has only been recorded for six months. Connie also wants to know about accessibility. Are the audio records transcribed for the deaf, Wayne says eventually they will be. Connie states that the Illinois legislature wants their audio tapes destroyed because they don't want to transcribe them. Bob explains the project Washington is working on creates text from audio, making the information searchable. CSPAN and AZ are developing similar tools. Eight states want to send their audio to Washington and LOC was surprised that this was so popular.]

[Vermont records committee meetings and hearings, and has stopped creating transcripts of the recording. Vermont has also stopped taking minutes of these meetings, leaving only the recordings as the only documentation of the activities.]

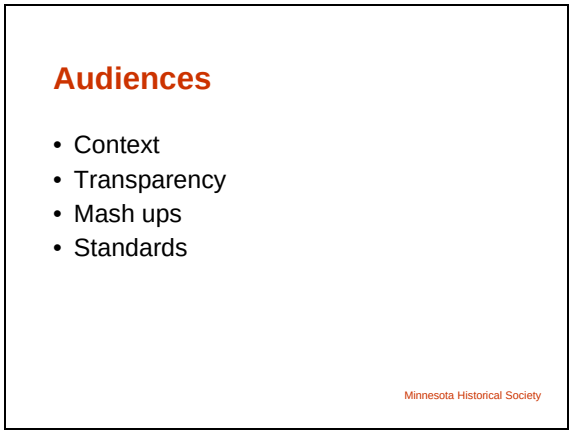
[Bill B. asks if the Tennessee recordings are used very much. Yes, there is a high level of use by attorneys, legislature, and media. It takes nine archivists and assistants to index the tapes during the legislative sessions. Bob states that the use is a reflection of how accessible the archive has made them.]

Slide 12



Attended a few NCSL meetings and we want to go again this spring. The legislative staff and IT staff are interested in the business case. These meetings provided opportunity for education and outreach while gaining knowledge on standards, preservation and coordination of efforts.

Slide 13



Our audiences are people who use legislative content and they themselves help facilitate the use value of the information. We heard that the context of the data is what increases its overall value. People have different needs for the data, things that we could not imagine. This transparency of data/content allows others to use it for mashups. Standards would ensure it would be supported by many people.

Slide 14

Authenticity is also not a one size fits all solution, but we need to determine how to adapt procedures in a cost effective manner. I am sitting in on meetings with Michele and the Uniform Law Commission (ULC – formally NCCUSL) that address such issues. The solution may vary, but identifying options is valuable. Following the current rules of determining authenticity the assumption is that the records are authentic unless proven otherwise. However, for some records it might be worthwhile to invest in a system to ensure authenticity (for highly sensitive records).

Preservation efforts now will ease the preservation efforts needed down the line, especially for legislative IT staff.

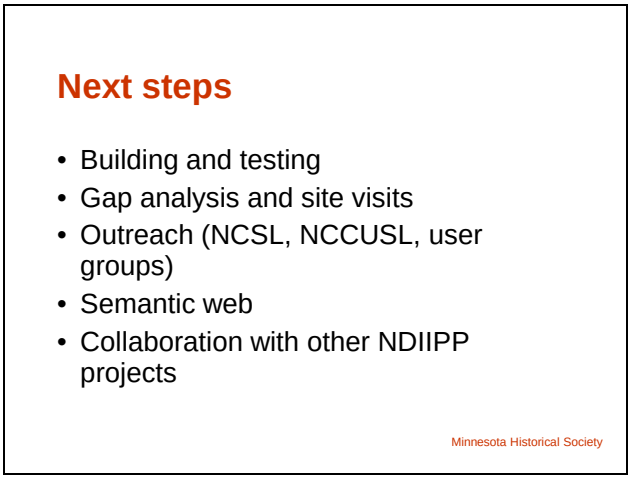
Context includes a recommendation of digitization standards.

Access. We would like to test and develop an access site. We will need to address questions on how to provide batch access.

Technology. Work with the idea of harvesting government web pages. CDL is using Heritrix.

Sustainability. The partnering with the records creators (Revisor's office) and records users (Thomson Reuters) during the development of the XML core schema and wrapper schema helps support the idea that different people want this type of data for various reasons.

Slide 15



Building and testing of the XML Wrapper in both MN and CA

Semantic web: I view XML as a stepping stone to these more complex systems.…

Questions and Answers…

Tricia: For the idea of batching content, where does content come from? Bob: You can choose to expose all of the bill data on the web or ftp site or just the metadata. You can control the dataset.

States can choose what data they want to expose and at what level, there is not a one size fits all solution. This is another area for the gap analysis. Different conclusions are ok.

Terri said that legislatures want transparency in principle but they also don't want others to make a profit on open government data. Bob stated that we don't want to preclude anyone profiting from value added services. One way to avoid this problem might be to make only the older data available.

Bill L. asks about the sense of the effectiveness of the business case, do you see being able to make a successful argument here and in other states? Bob responds that we are driving costs down as much as we can. For example with authenticity it is cheaper to provide the appropriate level based on the sensitivity of the document rather than running high level authenticity measures on every document regardless of its overall importance. We are also using standards for preservation now to help reduce costs later. The business case is good to drive costs down while minimizing adding new tasks to current routines. The interest in preservation of legislative materials has been identified as most are already saving materials for disaster recovery purposes. Our success will depend on relocation of our resources.

Cal asks about storage networks and states that NC has a storage area network (SAN) at George Mason. Will MHS be a repository? Bob confirms that MHS also has a SAN but we are also looking at the Washington and AZ NDIIPP projects to see if there are ways that MHS can participate in their preservation initiatives.

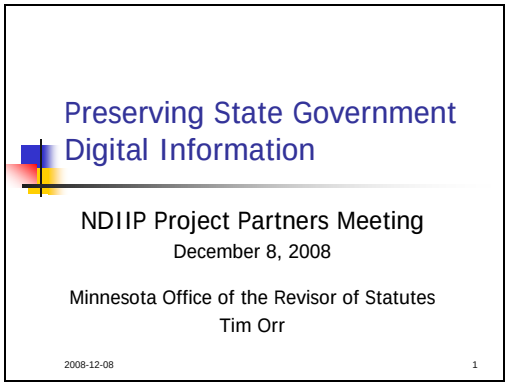
Has MHS inquired about the storage available through the Internet Archives? No, we have not talked to them about storage, only web harvesting.

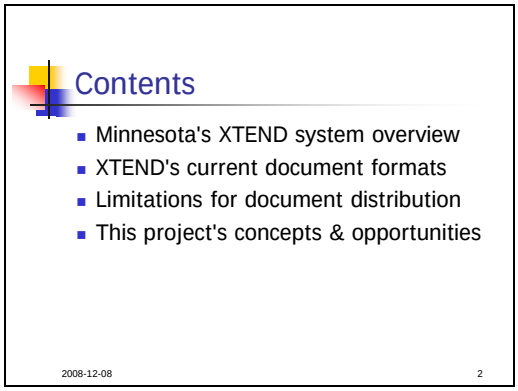
Andrew asks if anyone is spidering state websites. Yes, Tennessee uses Archive-It. The CDL is using an Archive-It like tool (Heritrix?). They recently crawled all state web pages in CA, and captured 1.9 terabytes of data. This process will be repeated every quarter. The CDL developed a tool to compare the new crawl with the previous one. Only the changes will be kept from the new crawl. Will you provide access to the crawled material? We are participating in another NDIIPP grant that will assist us in determining how to provide access to the crawled material. Cal said that NC has crawled state pages also. Shawn states that MHS has tried Archive-It, but found that it had too many limitations. For example, it was difficult to get to the pages that we were interested in, and it could not capture pages behind the state portal. It also could not capture dynamically created content. Cal said that they have to ignore blocks on public sites (robot text no follows). Connie has also run into spiders being blocked, but has found that a phone call can often provide access to the page they are interested in. Wayne uses these phone calls as an opportunity for outreach to agencies. There was general agreement that pages can be crawled despite no follow tags if a light touch is used. Bob said that he thought that web harvesting was a blunt instrument. Connie and Wayne said that they use a targeted approach.

Lunch break…

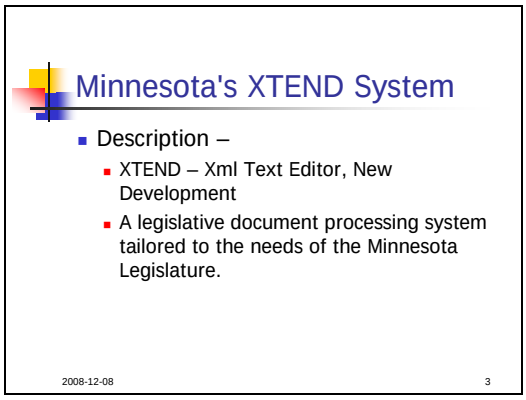
Preserving State Government Digital Information; Tim Orr, Minnesota Office of the Revisor

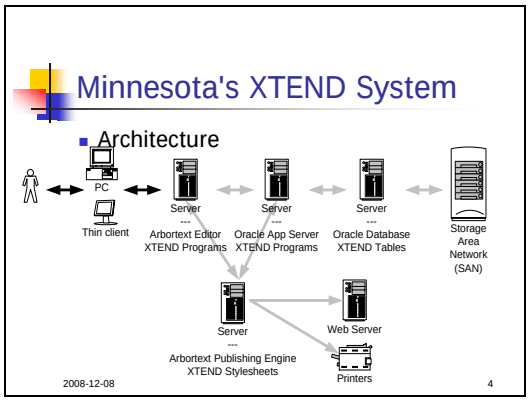
Slide 1 and 2



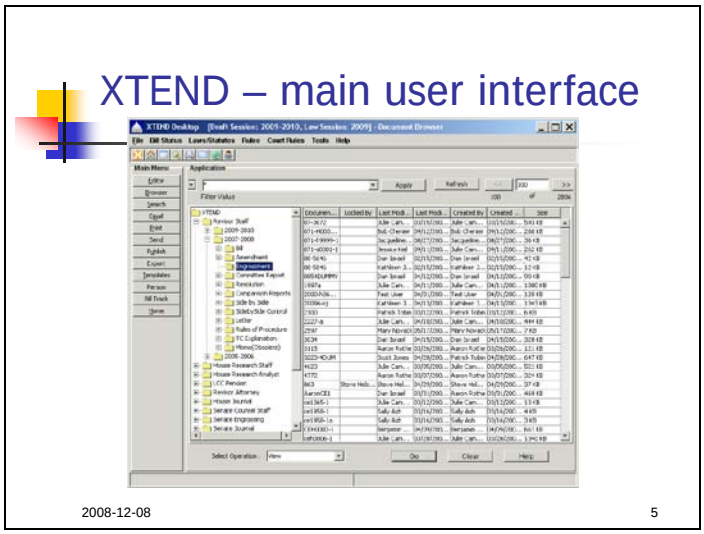


Slide 3 and 4



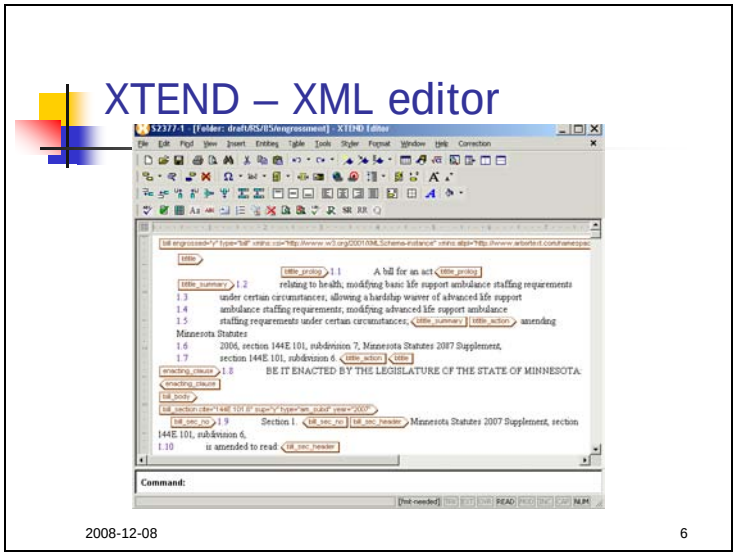


Slide 5

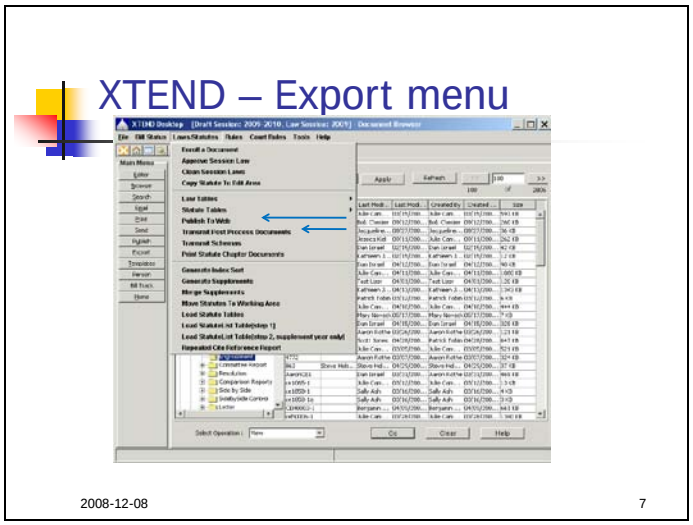


The main user interface is used to create a new document or edit an existing document. The users are the Revisor's office, the House and Senate attorneys and support staff. 2006 was the first year it was used in session and was primarily used by support staff, two years later more and more attorneys are using it.

Slide 6

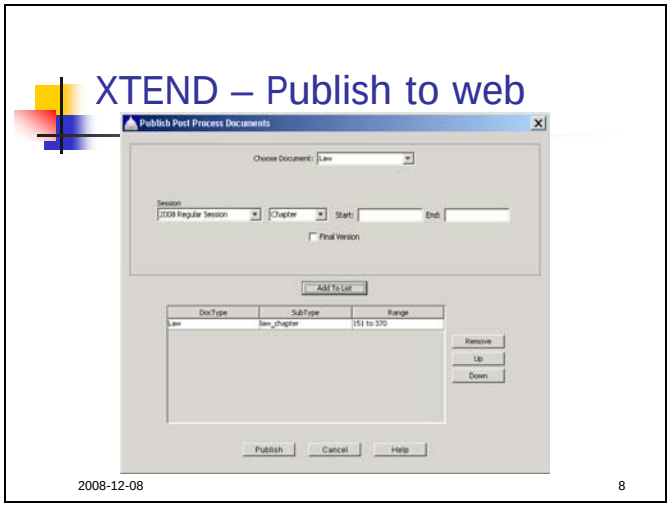


Slide 7



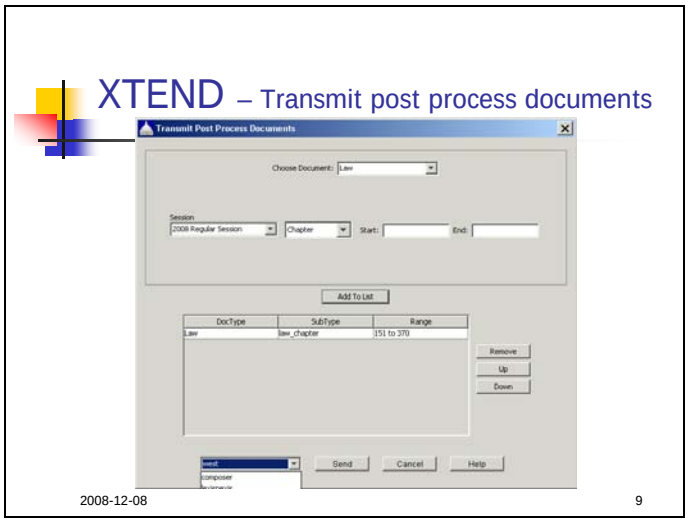
The export menu allows you to convert XML to html and PDF, to transmit docs to Thomson Reuters, LexisNexis, and others… This requires little new software development.

Slide 8



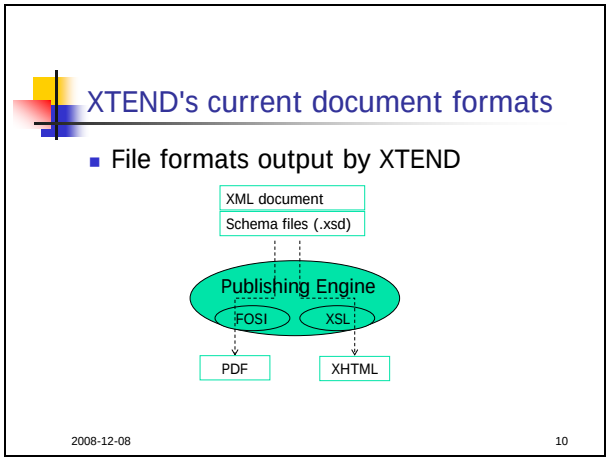
To publish to the web, fill in a few fields, hit a button and the information is sent.

Slide 9

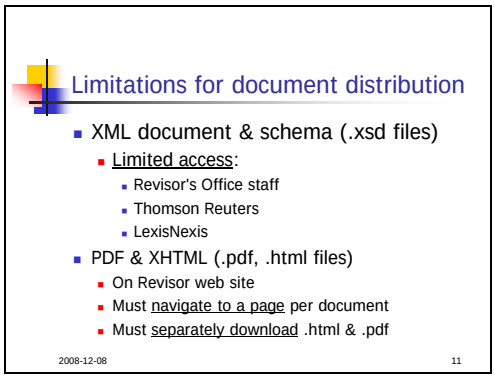


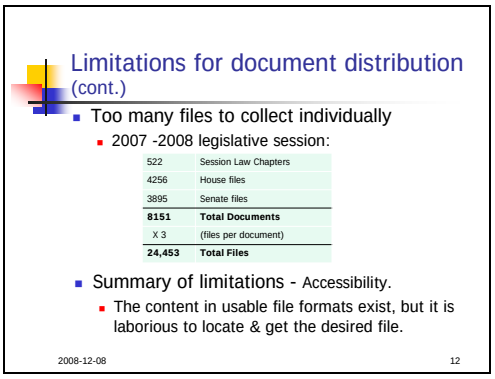
To transmit data to offsite recipients, fill in a few fields and information is sent.

Slide 10



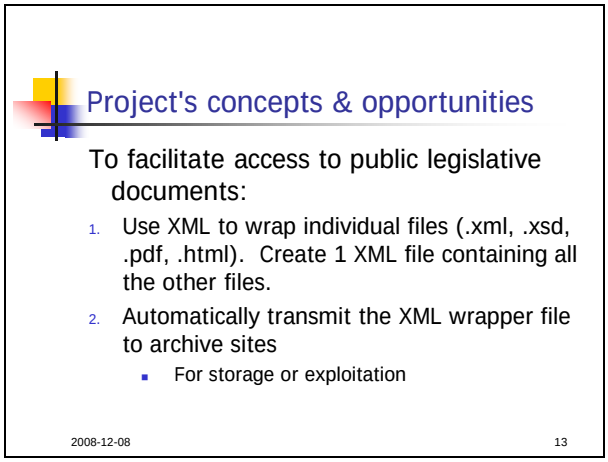
Slides 11 and 12





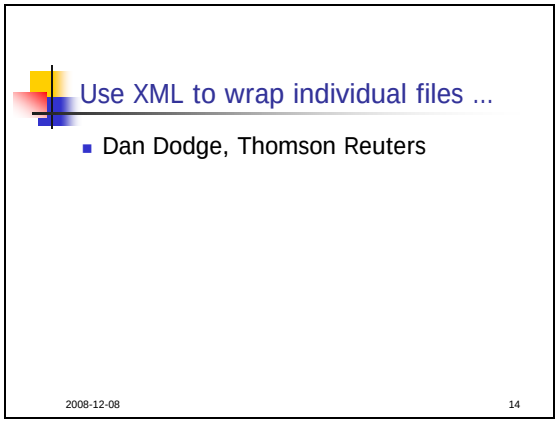
The file formats created may not be accessible to the public and there is no mechanism to make it so. The PDF version is on the web, but it is hard to get to, it takes many clicks to get down to the information you might be looking for. Another limitation is that there is not an automated way to collect the information or files. Access is the main limitation of the current system.

Slide 13



Creating an XML wrapper has a low impact on current tasks as it uses the formats already in existence. This can be formatted to various file types which increases the flexibility of the wrapper.

Slide 14



Questions….

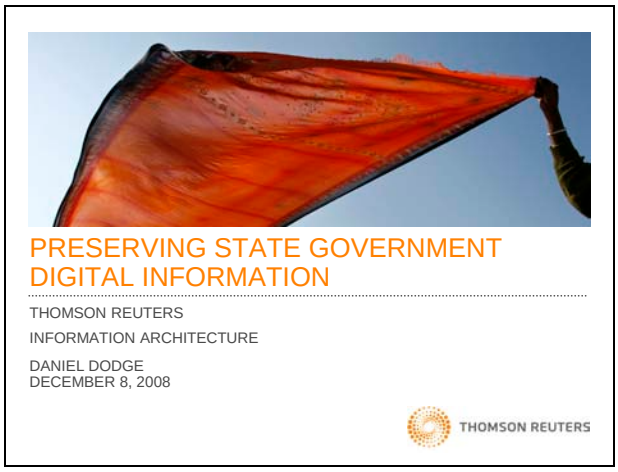
Tricia asked if people just want the XML files. Tim responded that it is hard to say because the XML is not readily available. He thinks that people would like the XML because of the flexibility of the format. However, currently there is not a way to direct people to such a harvesting site.

Andrew wondered how the Revisor's office chose to use XTEND and licensed software over open source software and a web interface. Tim responded that they knew they did not want to build a system themselves, they tested Arbortext and it seemed to do what they wanted, and the company itself was well supported. The Revisor of Statutes Office has been using a thin client since the 1990s and it is in a system that works well for them. ("A thin client does most of its processing on a central server with as little hardware and software as possible at the user's location."[1])The new bill drafting system is also compatible with Citrix which was important. It was a simple install; they only had to install it on the three projection servers which allow the other computers to talk to each other.

[1] http://en.wikipedia.org/wiki/Thin_client

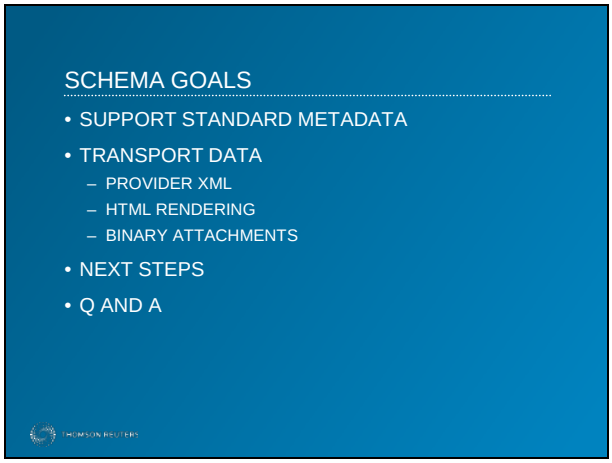
Preserving State Government Digital Information - XML Schema; Daniel Dodge, Thomson Reuters

Slide 1



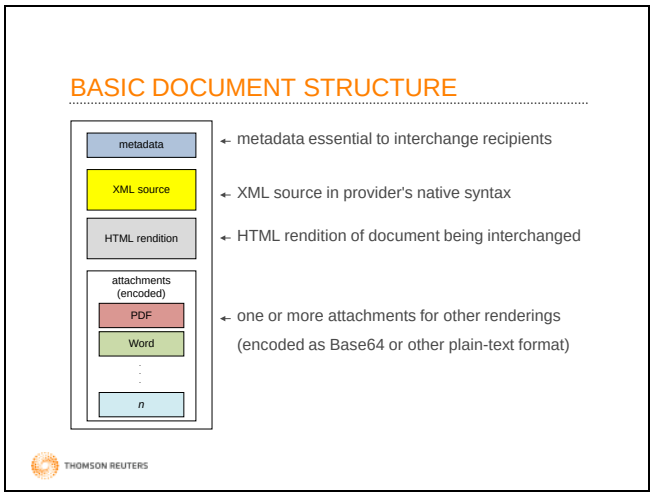
Introduction… Keep in mind that this is not a full proof of concept yet. We are still developing and testing this.

Slide 2



After some discussion it was concluded that metadata was an important way to access information. In addition, we wanted to design a method that could transport a variety of data packets.

Slide 3

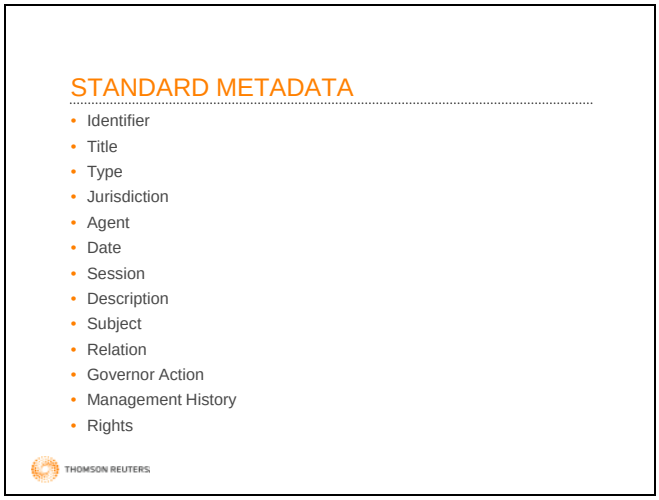


The metadata is the only part of the schema that is standardized. If this is going to work we need to have everyone's buy-in.

The purpose of this schema is to wrap native XML syntax as well as HTML and XHTML or other optional file formats (such as PDF or Word) together is a data package.

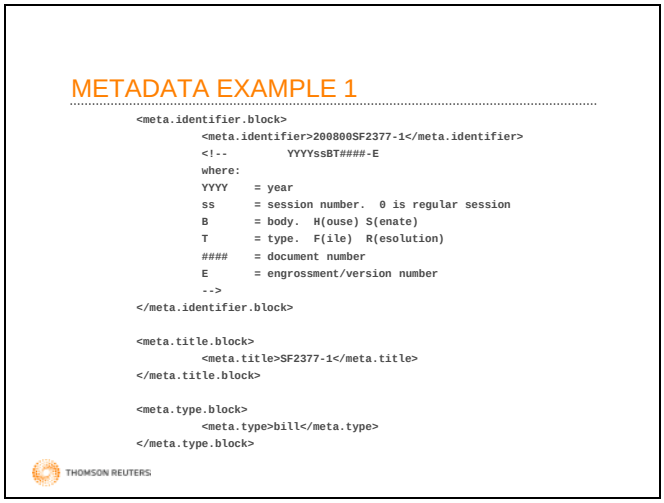
[Are the HTML formats inside the wrapper? Yes and others could also be included.]

Slide 4

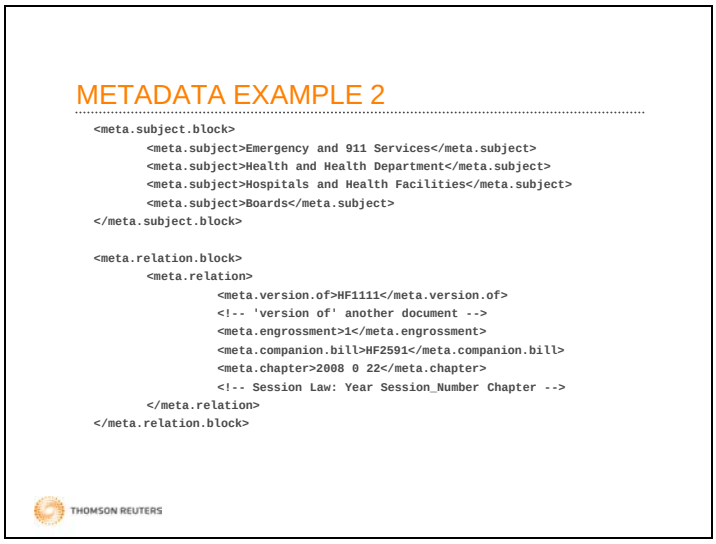


It was determined that the identifier is the most important element and that it must be unique. [Comments about how it will be unique if the information is shared with other states. Will state abbreviations need to be used? Will other metadata fields help create the identifier? If so, you add layers of complexity to the schema and metadata requirements. The XML working group also considered concatenating the original identifier with a state field to create a unique identifier.]

Slide 5

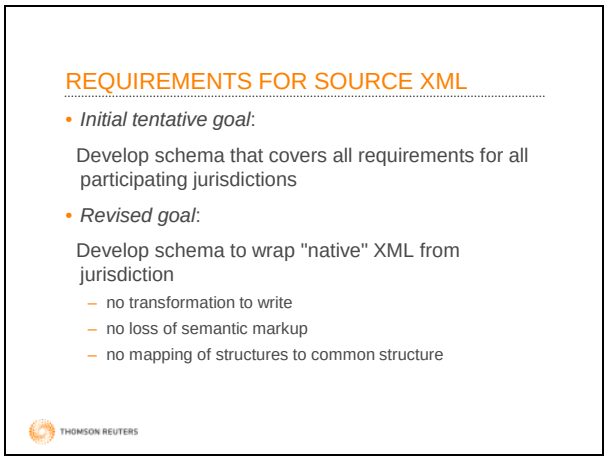


Slide 6



This example shows what is out there… If I want to query other states, can I find stuff from other states…? [Connie asks if a controlled language is used in this system. Bob and Dan respond no. Bob continues with the idea of trying to build off what is available. Others suggest that keywords or indicating what controlled vocabulary is being used per record set would be useful but not required. A question is also asked about the repeatability on the name element. The name element can be repeated as can other elements. A separate draft version of the metadata requirements includes these details. How can you support relations between other documents? With the relationship element.]

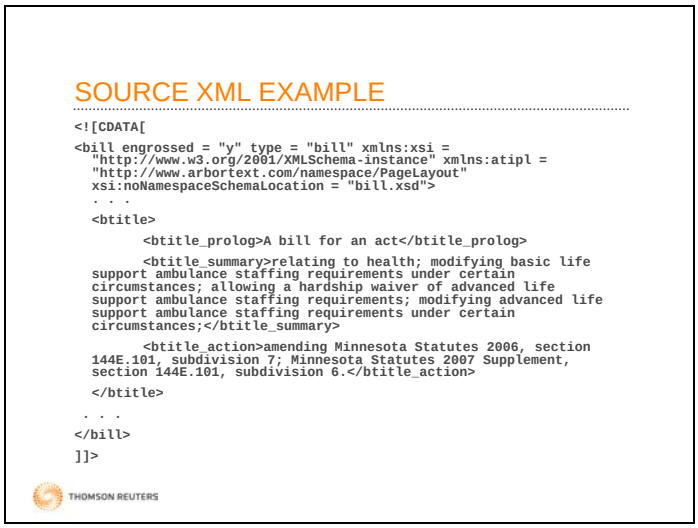
Slide 7



As previously discussed, you can't create a one size fits all solution. You do need to preserve what was created by the legislature and from the archival perspective,

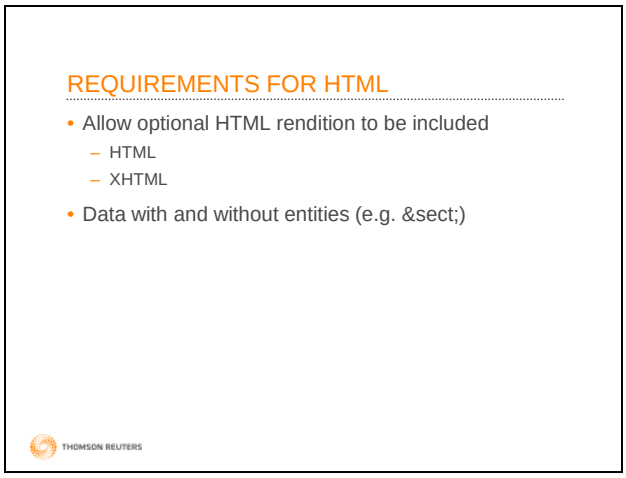
you want the original creation; you don't want modified information. By using a wrapper, you can preserve the document in it is original form. The wrapper creates a package that is interchangeable. (The package is the original XML and other file formats with this XML wrapper.)

Slide 8



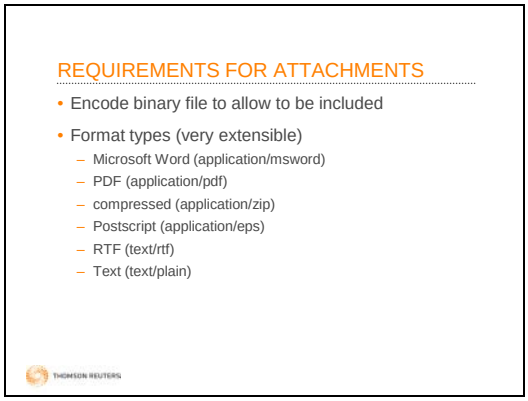
How can this be done? The CDATA section is where the original bill information is located. If you have a piece of data with its own schema and tags that are valuable, you can wrap the schema location into the file and import it all to the desired location. It is important to preserve the schema along with the data at the time of transfer.

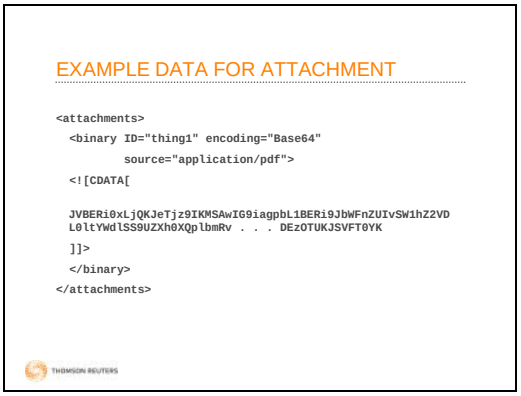
Slide 9

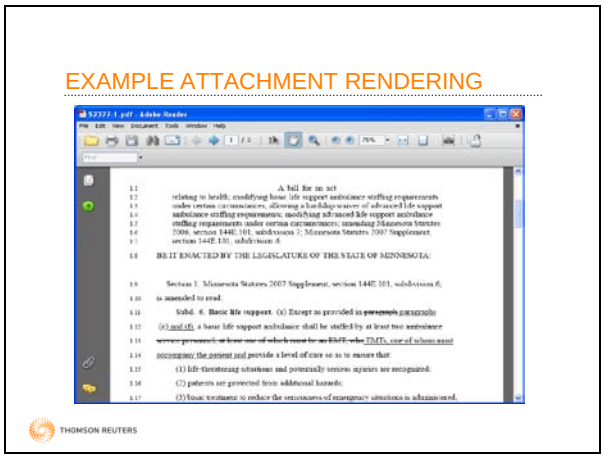


[Do you need to attach software type and version information to open the other file formats? No, but it can be specified in the XML and only high value formats are chosen here.]

Slides 10, 11, and 12

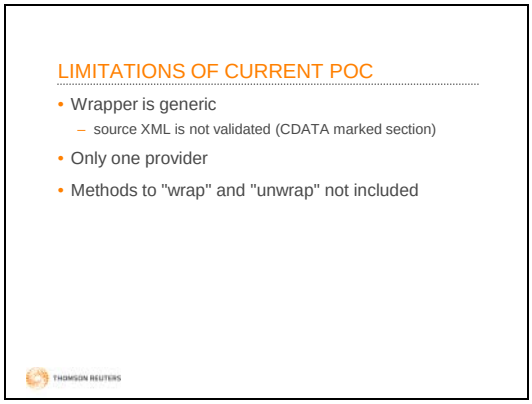






The requirements for attachments are included in wrapper schema; the schema does not try to convert items. [Question about encoded blobs. Dan states that he took a PDF file from Tim. More questions and discussion about documents. The format of additional files can be any standard mime type. Andrew states that it seems like a lot of information is being left out such as information on the version type, hardware, and software being used. Bob asks whose problem is it, the sender or the receiver, believes that it is the receiver or the one who wants the data because it is not possible to support all the issues that may come up. Dan adds that you could add an optional metadata block to include this type of info (version, programs). Bob agrees that the model should be adaptable or extensible. Dan states that if you want to specify a lot of different file types you will eventually run out of space. Could you use PDF-A which embeds the information in itself? Yes, it would be preferable to have a self-standing PDF. Could the average file size get too big? Andrew then asks if there would be a possibility to page the document to reduce file size. Butch asks if these packages are much larger than just an XML file. Dan confirms that yes they are larger, it took a minute to open the example file in Notepad, which is what he used as a test of file size. Tim states that you don't have to process the entire XML file, you could just grab the PDF. Andrew asks about multiple binary instances which Tim responded will be proportional to the bill that is being worked with, which will change the size of the file.]

Slide 13

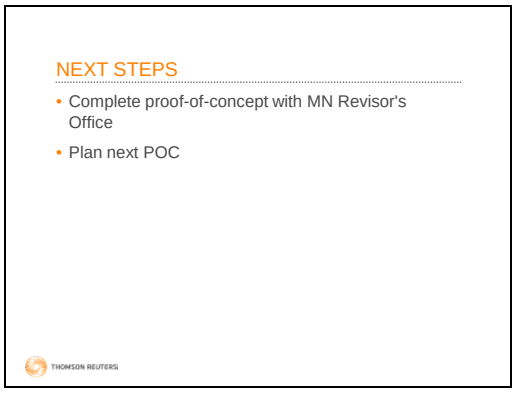


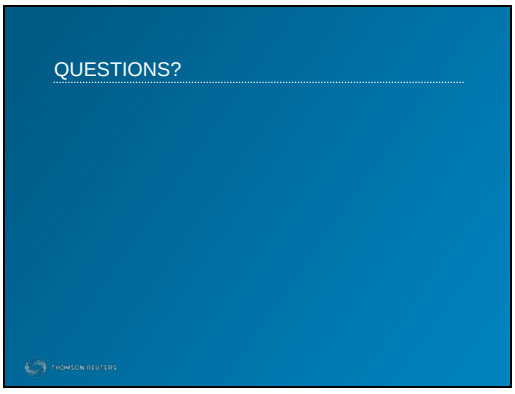
The wrapper is totally generic in order to wrap the native XML/HTML, and it only needs to be well formed; it does not need to be validated. The schema must be imported with the XML file. It would be good to verify that everything is included and perform a periodic re-audit to ensure that nothing was lost.

[Andrew states that we would want to validate for quality of information. Dan replies, true but in order to validate the XML file you currently need to un-wrap the bill. It would be nice to make this easier.]

We are currently only working with Minnesota legislative data, but will need to work with others to develop a way to easily wrap and unwrap files and validate the schemas.

Slide 14 and 15





Questions…

There were many questions that are summarized below…

Who creates the metadata in the schema if it is not created/generated by mapping from the native schema? Whoever is creating the wrapper would need to add the standardized metadata which is much easier than adjusting the original file schema.

Kansas was interested to see if this worked and if metadata adoption could be implemented in the design stage of their project. This would make the capturing of data instantaneous and allow all the related information wrapped up at one time reducing costs down the road. Bob states that it is great if you can implement this at the time of development, however most states have a set process, and the wrapper allows them to keep their own methods while still providing a standard set of information to others without breaking their systems.

Another benefit for Thomson Reuters is that if this system can create connections to other documents, Thomson Reuters could spend less time cataloging the data they currently receive from the Revisor's Office. The metadata would be used to connect other documents to each other. The responsibility of implementation of the schema and metadata does fall onto the states. However the core set of metadata should not be too much extra work for most states because it is a minimal core set of metadata. Minnesota might be able to find out a way to extract this metadata from original XML bill schemas.

Would it make sense to have a local and a national unique identifier? This was briefly discussed with the conclusion that the group could talk to someone at CDL who has done a lot of research on identifiers.

It may also be possible to store information about the file type, digital signatures, creation date, etc… in metadata that could assist with authenticity.

Illinois and California stated that they would like to test the model.

General Discussion

Bob informs the group that the XML Working group will continue to test the XML wrapper and are looking for additional sources of data, and if you are interested let us know. MHS will also continue to write white papers and create guidelines based on the topics you are interested in. If you have other specific interests please let us know and we can work on that with you. Current information for project members can be found on Basecamp (https://mnhscollections.seework.com/login). A general overview of the

project and related resources can be found on the project's website (http://www.mnhs.org/preserve/records/legislativerecords/).

Discussion ensued on other projects including comments from Bill L. about the California Archives working with Washington's NDIIPP project. Kris stated that the California State Library was going to be an observer on that project but budget cuts prevented it, so they are only involved in this project and are observing the AZ project.

Bill B. talked about the lawsuit over access to legislative data in California. Maplight wants to be able to link monetary contributions to the votes on bills. People want the government to put the data in an easily accessible place. The data is currently in paper form so it takes a lot of work to come up with the same ideas that if done electronically would take little time. The concept is information vs. knowledge and the public perception of what it is being provided.

Next Steps; Bob Horton

NDIIPP Brochure…

Bob hands out a proof of the NDIIPP brochure and discusses how it can be adapted to fit their institutions. (By changing the logo and adding their name on the front). The audience for the brochure is seen to be people in archives and government. It will be a useful handout at conferences as it will give attendees a general background on the project and information on who to contact for more information. The hope is to raise the profile of the project.

Discussion about Site Visits…

Bob hopes that the next site visits could include a gap analysis on using the XML wrapper in some context for each state. The visits are also an opportunity to talk to collaborative partners in the legislature and IT sector of each state. In order for this to happen, we need to have closer contact with each other. We would need to work together and spend time evaluating the situation before the site visit. The goal will be to produce a 'cheat sheet' that could be used as a work plan in the future. If this does not work or sound like what you are looking for, we can also use the site visit to brief you on what we have been working on, the outcomes with other states, and on various aspects of the project.

Outreach…

We will continue to do outreach with partners like the Society of American Archivists (SAA), the American Law Librarians Association and others to promote the project. Others are welcome to assist in this effort. If you have any suggestions for conferences or would like us to do a presentation about the project for your local partners we would be happy to do so.

Open Discussion…

Members of NCLS stated that it would be important to get the IT staff together as well as the archives staff… such as through NALIT.

Wayne asks about gap analysis for Tennessee, which uses little XML. Bob suggests that Tennessee has digital audio and video content which could be used as additional formats that would test the flexibility of the XML wrapper. This would be a proof of concept and others would be interested in the compatibility of digital audio files in general. A cost benefit analysis could also be done. Wayne states that the Tennessee legislature is also doing parallel recording and the records are not preserved. Bob confirms that a preservation layer could be added to this project. Kansas might be able to share information on their e-democracy project with Tennessee and the rest of the group if interested.

Bob suggests that a gap analysis with Kansas might be from them to us, not us to them. Kansas representatives state that archiving is not a big part of their current design plan, but believes that it should have a higher value. The current NDIIPP project might be able to stress the importance of preservation in their E-Democracy plan.

Cal offers the resource of student labor for the project. Students could help with technical issues such as dealing with quirky formats and would be available for a semester at a time.

When thinking about scheduling the next site visit, keep in mind upcoming Legislative Sessions, which are followed by the summer months and into the beginning of another semester and into 2010. Let Minnesota know what works for you and what you would like us to do with/for you, and we can schedule a date.

The grant ends in late 2009 which is contingent on the gap analysis. It is hoped to do the pilot test in 2009.

Library of Congress would like to have all four (AZ, CA, MN, WA) of the NDIIPP project partners come together to give a report in late 2009 or early 2010. This would also include people from outside of the project.

Matt feels that the gap analysis is important… this is a gap that really matters. The business case application is what attracted him to this project.

Cal has applied for a grant relating to policy and information management… IMLS

Cal wonders about long-term preservation? How much interest is there concentrating on preservation? Andrew feels that legislative data should be preserved. They are public records which are supposed to be around for all to see forever. Cal wonders about how to build a repository for such information. Bob feels that this is where collaborations provide a cost effective method, for example the CDL is currently building a repository.

Bill B. brings up the GIS database clearinghouse created by David Erickson and wonders is this is a model we could follow. Studying this model would allow us to expand our ideas. Butch states that he sees similar points between our project and the current GIS project being conducted by NC and feels that there is an opportunity to share information. North Carolina has been looking for an interchangeable format to use with GIS data and an XML wrapper might be of interest to them.

Discussion about the preservation grid…

> Shawn states that the purpose of the preservation grid is to highlight the reasons you care about preservation and what you can get in return from your efforts. The goal is to create a narrative that explains why and how to best preserve legislative data. Connie stated that the Center for Research Libraries recently evaluated their repository (the Illinois State Library) which was valuable in that it let them see what they needed to work on. Many of the things Connie is now working on can be found in the best column in the grid. Connie wonders how much information will surround the proposed narrative, for example the grid states that it is best to have a migration plan – but what does this mean? Will a sample migration plan be included? What is involved, what do you need to consider in a migration plan? Where do I go to look?

Authentication Paper…

> Michele views the paper as a valuable resource for her, but wonders how the authentication paper fits into the project. Is it authentication of the data being sent? Bob replies that if we use the Trustworthy Information Systems as a model, the data in a trustworthy system by definition is authentic; however transactions between systems present authenticity problems. Perhaps this is where a preservation copy could be used to check the authenticity after several transactions. Andrew states that check sums are ok, but wonders if it constitutes legal authentication. Bob assures him that it does.

Tanya states that it would be nice to have some sort of a check list as a way to rate your state to determine where you are on the preservation grid. This is a way to show many different people who look at things in different ways where the institution stands, it gets people on the same page and allows for a more productive conversation. Shawn said that a similar tool was used in the TIS handbook and it proved to be useful. Bill L. also likes the idea and feels that such a tool would be valuable. The TIS handbook can be found on the Minnesota State Archives website, but it needs to be updated. TIS was inspired by COBIT (a systems security), a method that IT people use to verify information. TIS creates guidelines to follow for a more granular level. The TIS guidelines assist with creating methods to ensure secure content, ERMS goes into more specifics, and the Record Keeping Metadata Standard created by MHS creates a common set of metadata elements based on Dublin Core, which we are now mapping to XML.

Cal mentions the UK's DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) and how it is a tool that can be used to rate your repository.

Wrap Up; Bob Horton

Bob confirmed the work plan would include a visit to each state; the planning to begin in early 2009 and visits to take place at the host states' convenience.

The meeting was positive and it helped to have 'outsiders' come and talk about the importance of preservation and access to other partners of the individual states. Some wondered if we have heard enough from the other states and what they want from the project. In response, fleshing out what the states want is a major goal of the next site visits. Others commented that the goals of the project are much clearer than they were at the first site visit. A comment was made about the NCSL common bill drafting schema and how it was problematic because each state has unique requirements. MHS looked at the NCSL project; one of the reasons we decided to try to create a generic wrapper is that it would allow a standard set of metadata to be added to existing unique XML schemas of individual states.