# Preserving State Government Digital Information
# Core Legislative XML Schema Meeting

## Minnesota Historical Society

Friday, September 12, 2008
Pillsbury Conference Room, Minnesota History Center
St. Paul, Minnesota

*Meeting participants included Dan Dodge and Jolene Sather (Thomson-Reuters), Tim Orr, and Michele Timmons (Minnesota Revisor of Statutes Office), Nancy Hoffman, Bob Horton, Jennifer Jones, Charles Rodgers, Shawn Rounds, and Christopher Welter (Minnesota Historical Society).*

***The following account is paraphrased.***

Bob gave a general introduction and review of the XML schema project to date.

Tim: Talked to developers about developing a simplified core schema. They said it wasn't feasible because it would break all the applications that rely on the schema structure. Tim constructed a core schema in spite of these comments and applied it to bill 1812 (see example). At this point, the problems the developers predicted became apparent. Based upon this, he came to the conclusion that a core schema would not be possible to implement for content capture but may have some use in framing a metadata structure for archiving purposes.

Bob: (referring to the metadata crosswalk document) We identified three metadata elements from your core schema: Title, ID, and Content. But how is the Metadata tag structured – is ID a part of it?

Dan: The Id and the Meta tags are children of the Metadata tag.

Tim: I was trying to create a wrapper for the Minnesota bill metadata by placing them within the Metadata tag as name-value pairs. That said there are many good existing metadata standards out there.

Bob: But this is how you do it now?

Tim: Yes, I copied the meta values out of the Minnesota schema and a pasted them into the example.

Bob: So these are specific to the bill drafting system. It would be interesting to add them to the crosswalk.

Shawn: Many of these are comparable to metadata elements in the Minnesota Recordkeeping Standard such as Date, Document type.

Dan: So, in a generic way, contributors would have a set of definitions for metadata elements that would allow humans to understand how the data exchange.

Tim: Some of our elements would drop out

Dan: Or would not be used

Bob: The Thomson Rueters and Library of Congress element sets are good to use for searching. It's important to keep in mind that we want to facilitate searching legislation across states.

Tim: Also, Minnesota has a lot of information in the bill status system that could be pulled out for archiving.

Dan: The bill status information would not be included in its entirety, but you could pull out the metadata to archive with the bill.

Jolene: Are the bill and the bill status information kept together now?

Tim: They are kept separate systems. Access to both is presumed. If a bill was archived elsewhere, some information that is record in the bill status system, but not in the bill itself would be missing. The metadata from the bill status would need to be added

Dan: Archiving would be a way of keeping this information together.

Jolene: (referring to the crosswalk) Where did the Thomson Rueters metadata elements come from?

Shawn: They cam from your suggestions at the 6/6 meeting.

Dan: We would want to go back and look at all of the metadata elements we use.

Jolene: These are important, but we have more. When creating a common interchange system, it's better to have creators send everything – unless it's proprietary – and the receivers can decide what to keep.

Tim: As Dan said, the bill with the status system metadata would be standard but also including the HTML, PDF, and XML – everything.

Dan: There are three important parts to this – a metadata dictionary for the wrapper elements, an HTML version for rendering so humans can put the legislation in a browser window, view it and do brute force searches by search engine, and an original XML instance. Minnesota, and every other state (or territory) would have its own namespace for their XML schema.

Bob: You mentioned brute force searches – the idea that a batch of all bills in one session could be made available.

Dan: The question is, at what level of granularity this information could be made available.

Bob: Use of a common schema for archived XML would be one approach.

Dan: Would this be applied at the document level, or for a batch of documents? If access were at the document level, each would need to have the original XML version, an HTML rendering, and the metadata.

Jolene: The common schema would server as the wrappers.

Dan: The HTML would have to be XHTML. We would have 50 namespaces for the XML documents. We could add references to the original namespaces so that anyone who wanted to could check it.

Jolene: As a receiver of material, we would want to know the metadata, the schema used, that the rendering was in XHTML, but would others care?

Dan: It depends upon the ultimate use of the information – in some cases metadata alone might be enough. Thomson would care about all three pieces for publishing and archiving, and would have to do some customization.

Tim: An agreement on standard metadata elements will probably be good enough for 90% of the potential users.

Dan: Yes, we could find some existing metadata standards and extend them, as needed.

Jolene: What would be the burden on the states – to create the XHTML and populate the metadata elements?

Tim: The only real requirement would be the creation of the metadata.

Shawn: And the metadata could be extracted automatically from the system.

Bob: What are our next steps?

Jolene: At our last meeting, we talked about what people are doing with legislative data and why, but could we review that?

Bob: I met with a group of researchers in Pennsylvania. They were interested in the legislative records, but also a wide range of contextual information such as biographies and newspaper articles. They used a simple subject code system to bring together related materials, but other researchers may want on-the-fly search across metadata fields. MHS recently met with a technology consultant, Zepheira [http://www.zepheira.com/] that has developed a spreadsheet-like application for combining data.

Nancy: They have developed an application that allows heterogeneous data sets to be combined in any way the user finds meaningful while automatically generating triples. This exposes the data to semantic applications that can facilitate more complex searches.

Bob: Regardless of the approach we decide to take, metadata standards will facilitate our work. Using a metadata wrapper for legislative records will need to take into consideration the fact that many states do not use XML. Mississippi, for example, has some of their data in PDF format.

Tim: Many states publish their legislation in PDF format.

Dan: PDF can be a useful way of sending a receiving data.

Tim: XML would have all of the metadata

Dan: Could binary data be placed inside a PDF?

Tim: Minnesota encodes for it (?)

Dan: You could hide RTF inside legal binary file formats and coding systems (?)

Jolene: This would just be for states with PDF format (?)

Dan: There are other ways to send original data, and Word documents can be sent as easily as PDFs.

Jolene: But some states wouldn't (?)

Dan: We could have namespaces for Word and PDF documents so that it would not be left up to the receiver to sort out.

Jolene: The document schema would be the core metadata elements. XHTML would be optional. Senders would add a wrapper, if they can. The package would include the original format, with some choices – it could be XML or it could be PDF or Word that had a standard wrapper.

Bob: Breakage at receiver's end (?)

Jolene: So we want to let the states do whatever they need to do for their own use, but set standards for data interchange.

Bob: We would set the metadata standard.

Dan: Metadata can do for data exchange what the Euro has done for economic exchange in Europe, provide a single, common standard.

Bob: We need to make sure that it is a standard derived from what the states are already doing.

Dan: A wrapper would improve the current workflow at Thomson when we receive a PDF.

Bob: It would help you to know if you have gotten everything.

Dan: It would help by providing metadata they we current get by opening up the document and having someone find and enter the key information, such as title.

Jolene: Could we identify required metadata, but allow each state to expand upon it to suite their needs?

Bob: Sure, more metadata would be OK. You just need to make sure you get everything. The information has to be comparable across file types. The ability to track provenance is particularly important for verifying authenticity. A content and/or subject field and title would be important.

Dan: We take the domain of facts about legislation and use it to derive the core metadata.

Tim: The NCSL has done something like this for their bill status system. Thomson must have some ideas.

Dan: The XML schema enforces or validates the required metadata. This is done with a DTD. A uniquely named tag allows enforcement, but name-value pairs do not. […?]

Jolene: How do we define the core elements? We've talked about looking at the NCSL and Thomson for their standards, but what about our original discussion. What do the people who will use this data want?

Bob: We have identified a small number of key elements that we can test. Subject and a full text search might turn out to be the only thing we need to use. Librarians, on the other hand, specify a large number of metadata elements and get into a lot of detail. Finding something that would address 90% of the users needs is do-able. Each user would also have unique needs that we cannot address. For example, Lawyers doing electronic

discovery would need name and term searches. They could pull out results and put them into their own system for further analysis. The idea is to give access at a basic level and let people do what they want with the data. So, what metadata elements do we think should be included in the core set? Tim has mentioned title.

Shawn: We can start a list now, but anyone can send us more ideas later.

Dan: The next version could be for review and trial.

Bob: We should take Devon's suggestion and try to wrap something.

Dan: We could define 12-13 buckets for common metadata: Identifiers, Dates, Subjects, Authors / Contributors or more generally People – Institutions would be Agents, Descriptions, Title, Keywords, and Jurisdiction.

Bob: How do we specify a unique identifier?

Dan: Two we find useful are: the Identifier assigned by the originator, and the one assigned by Thomson.

Bob: What is the identifier string on the example document we have? (Bob and/or Tim find and read the bill id number from a printout on the meeting table).

Dan: we would use that number plus the state postal code, in this case "MN."

Jolene: We are essentially adding jurisdiction information.

Tim: Do we also need to add session information, if the identification numbers used are only unique within a given session?

Bob: We could add the year to the id number also.

Dan: The original id number is the authoritative piece of information, but the jurisdiction and session fragments are also useful for searching and sorting purposes. But as a receiver of data, we don't want to have to concatenate three fields to make a new id field.

Bob: And it would mean repeating data in separate metadata fields.

Michelle: We could leave the id field blank and say that it is to be determined.

Dan: Yes, we would need to make sure it is unique within the system we are using.

Shawn: We can say that the receiver will assign it.

Dan: The original identifier is not for us to dictate.

[Shawn – Michelle – Jolene discuss]

Bob: Does anyone have suggestions about how to handle the various fragments of the identifier?

Jolene: The sender shouldn't care about that part.

Bob: [writing on the board] So we would have the "original id" assigned by the sender and entered as- is, the "receiver's id" would not be in the schema. "Jurisdiction" would be a separate element, not part of an id number. We could provide examples of the format.

Jolene: We should have a set of rule for how to populate the metadata elements and a set of examples also.

Shawn: We could have elements like "session" and "jurisdiction" as core elements, but we could also have higher-level buckets such as "people" with sub-elements such as "sponsor" and "contributor."

Michelle: A bill can have many authors.

Bob: And researchers are particularly interested in sponsors and finding their connections to other interests.

Jolene: Under "contributors" we could allow one or more. The same would be true for "sponsors," and "authors."

Dan: A sponsor might be an organization and an author might be a person. Are these required?

Bob: Not yet.

Shawn: We could use "agent" as the higher-level bucket for all these roles and make it required without specifying particular roles.

Jolene: What about the "title" element?

Michelle: Minnesota's bill titles are very long.

Dan: There can be formal titles and popular titles. The "title" element many need a sub-element for "other titles."

Bob: Let's consider "bill summary."

Tim: Minnesota provides a short description of a bill that would serve as an appropriate summary.

Dan: How would this relate to a “description” element?

Jolene: Would “description” be a bucket or a flat element?

Tim: It could have two levels.

Shawn: If we start with a tree, we can always flatten it and decide what is required later.

Dan: Too deep a tree will look daunting or confusing.

Jolene: What would “subject” cover?

Bob: This could go under “description.” “Summary and “title” could have a lot of the information found in the “subject” element.

Shawn: “Description” and “summary” could be optional, but “subject” could be required.

Dan: And Thomson does require at least one “id.”

Bob: The “description” bucket could have many sub-types.

Dan: If the model changes or is unknown, the description can always be searched. Dates of all varieties can go into one “date” bucket.

Shawn: If you want to develop a crosswalk with other metadata schemas, “subject” and “title” are kept separate in most systems.

Jolene: How would we define “document type?”

Dan: I don’t think we could come up with a finite list of values.

Tim: Maybe we can. Legal terminology is pretty consistent.

Dan: Except Louisiana. Because the rest are grounded in English Common Law, they are probably the same.

Jolene: What would “keywords” look like?

Dan: May be found in the “description” or “subjects.” (?)

Tim: Minnesota bills have “topics.”

Bob: Tennessee has staff assign keywords and topics to audio and video recordings.

Shawn: We could refer to related material in a “relationships” element.

Dan: Thomson keeps track of this in our metadata.

Shawn: This is a common way to reference versions.

Dan: The version can also be indicated by the structure of the id number.

Tim: It might be nice to break out a “version” element.

Dan: And any changes in the id.

Tim: Yes, that might be useful for an analysis.

Jolene: “Effective dates” might be useful in regard to version.

Michelle: Minnesota has “enacted dates” noted for its bills.

Jolene: That could indicate another version of the document.

Tim & Michelle: Resolutions and bill engrossments go through many versions. The final version is considered a new document and gets its own id.

Dan & Shawn: The “relationships” element could show how all versions are connected.

Jolene: Would the new instance have a new “document type” such as from an act to a law?

Dan: This could be part of the process (?)

Michelle: Act would be a better term to use. There are acts, statutes, and codified laws.

Jolene: Would his include statutes? (?)

Tim: This could be extended later from bills to statutes.

Jolene: What is the relationship between a bill and an act?

Dan: Historical?

Shawn: They are different versions.

Dan: The “relationships” element could indicate relationships to other state laws.

Bob: This would be the case when a state has enacted a law written by the uniform laws commission. [http://www.nccusl.org/Update/]

Jolene: This could also be used to show related press coverage.

Tim: And the relationship to the last engrossment.

Jolene: But this would not be required?

Michelle: A person would have to create this information.

Bob: Or MHS staff could do it.

Dan: For this to work easily, the ids should be consistent, preferable in the same domain. Pointing to a California law could be difficult.

Michelle: Minnesota also records an act's effective date.

Shawn: The "date" element could have a number of sub-elements.

Bob: There is a lot of information out there to capture, but it has to be the same to be useful.

Shawn: The metadata documentation would define elements, sub-elements, standards, and include examples.

Bob: We need to remember that validation is an important function.

Dan: It's always important to start with work that has already been done.

Bob: We mapped the Minnesota Recordkeeping Metadata elements to the Dublin Core set. To recap: we have decided that the metadata will go into an XML wrapper around instances of legislative data in its original form, whether that is XML, PDF, or Word, as well as an XHTML version. We may have to define acceptable formats to support transmission or extraction of information, preservation, search and validation.

Dan: Discovery, comparability, and provenance – some pointer back to the original source – will be important.

Bob: Repurposing includes publishing and discovery.

Dan: From the transmission point of view, the information might not be there, but it could be prepared on request.

Bob: Many states use XML for at least some of their data. To review: next steps will be to work out the crosswalk details, and have everyone review them. We will define the required elements to minimize the burden on senders by finding the metadata already being captured.

Jolene: So the state submitting legislative data would provide the metadata.

Bob: Yes, 13 elements, more or less.

Tim: The Minnesota Revisor's Office would do this.

Shawn: We will post the meeting minutes, the metadata crosswalk, and check the way the legislative information is organized on the NCSL website.

Tim: Devon could probably write the XML wrapper for the metadata. I think this would not be difficult for him.

Dan: We would want to test the draft version of the wrapper.

Tim: The Revisor's Office could participate in testing the wrapper sometime after the start of the next legislative session

Bob: MHS will be bringing all of the state here on December 8th. The Revisor's Office plans to attend.

Dan: Jolene and I would like to attend also.

Bob: The attendees will be here from Sunday night through Monday night. Representatives will come from the NCSL, the Library of Congress, and the partner states.

Michelle: The meeting agenda will include an item on authentication. I have the handout I used at the Portland presentation and will be getting input from the Uniform Laws Commission.

Bob: Great. We should have version three of the authentication paper ready at that time also.

Tim: We will send our metadata elements list to you.

Shawn: We will include it in the next version of the crosswalk.