Preserving State Government Digital Information
Core Legislative XML Schema Meeting

Minnesota Historical Society

Friday, June 6, 2008
Pillsbury Conference Room, Minnesota History Center
St. Paul, Minnesota

As part of its NDIIPP-sponsored project, Preserving State Government Digital Information*, the Minnesota Historical Society (MHS) hosted a meeting with its Minnesota partners to begin work on a core legislative XML schema.*

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

At the meeting's conclusion, these next steps were agreed upon:

1) *Invite meeting participants to Basecamp, the Web-based, project management application being used by MHS for this project.*

2) *Send out draft of meeting minutes for review by participants.*

3) *Begin a draft of core schema by combining Jolene Sather's data element list with Tim Orr's list (latter may be available by mid-July).*

4) *Convene a second meeting of this Core Legislative XML Schema group in August.*

5) *At a later date, seek ideas from larger audiences/professional organizations that perhaps we could make presentations to.*

6) *Disseminate for review the Library of Congress' Congressional Research Services' legal citation list, an example of which is linked here: <*http://thomas.loc.gov/cgi-bin/bdquery/z?d110:S1:*>.*

The following account of the meeting is paraphrased.

Bob Horton welcomed all the participants and gave a PowerPoint presentation overview of the project's efforts to date.

Michele: What is Illinois's involvement in the project to date? Minnesota, Illinois, and California are all pretty far along with XML.

Bob: Due to a snowstorm, we did have an initial, abridged meeting with them, where I talked them through my PowerPoint presentation via phone, then they conducted the rest of the meeting on their own. We'd like to meet with them again.

Dan: Will we have a list of document types that we want to support? At Thomson Reuters, we have a core model of paragraphs/lists already, but there is also a bigger value of structures that can be aggregated. What is the meaningful level where we want to aggregate data?

Bob: So far, the idea of structuring down to the paragraph level across document types is beyond our scope/capacity, but common elements across all sources are of primary interest. What is this, and where did it come from? What is the author, title, and subject information? [References his recent Temple University meeting that outlined different groups who're mashing up state, federal, and international legislative information with other resources (e.g., the Pennsylvania Policy Database Project mash-up of numerous resources—legislative content, newspaper articles, *Governing* magazine, legislative biographies) with one content code added to all resources, and access by key word searching <www.temple.edu/papolicy>.] We might take this as a starting point. I think we're interested in semantics, the meaning of the content, and the provenance, chain of custody.

Dan: What you're describing sounds like a framework for new tags, more a superclass of schema.

Bob: In previous meetings, there has been lots of interest in establishing context by referencing/linking numerous other resources (newspapers, mass media). We're talking about "universal" for a schema. [References Kansas wanting to link multiple types of content, both legislative and non-legislative.] XML is the lingua franca for digital data interchange.

Devan: There are possibly 2 levels of structure: 1) creating a crosswalk vocabulary for something like resource types (e.g., bills, measures); 2) on top of that, another level for indexing. Based on half a dozen personal experiences, common vocabularies are quite hard to agree that respective data is equivalent.

Michele: The National Conference of State Legislatures' (NCSL) previous attempt at this failed in great part due to each state's specifically evolved vocabularies/terms.

Dan: Thomson Reuters has also experienced something similar. Because each state does its own thing, Thomson Reuters is forced to impose its own schema.

Jolene: It sounds like you're talking about metadata schemes rather than actually marking up content itself.

Bob: [References MHS team member Charles Rodgers' efforts at documenting what types of state information resources beyond bills/acts are common across numerous states.] What can we do with all this material to harmonize it? Primary users are legislators and their staff. There are multiple metadata schemes that already exist across numerous professions, that perhaps what we're talking about is crosswalks, where XML allows for flexibility/adaptability.

Tim: The Revisor's Office has been kicking around ideas for new technology. We don't extend all technologies thoroughly enough. Are there technologies out there that allow for this cross-walking—RDF and OWL (Web ontology language), perhaps? Does anyone have experience with it?

Dan: It was a brilliant idea that XML allows different namespaces to be in the same XML document. We could create a namespace for our common vocabulary. We should also use linking standards such as XPath [XML Path Language; <http://en.wikipedia.org/wiki/Xpath>] and XPointer [a system for addressing components of XML based Internet media; <http://en.wikipedia.org/wiki/Xpointer>.

Devan: We could let everyone point to that namespace but have some required subset of terms.

Isaac: Do I need to access the other systems? Should we export our data and the historic contextual information to understand it? If you send me XML from one legislative session, can we access all of it, or is it parsed out in different relational databases? Can we do this with RDF/Owls?

Devan: Each of the states would have to go through some translation step before providing content.

Dan: I've been reading about the Global Justice XML Data Model and Data Dictionary [a data reference model for the exchange of information within the justice and public safety communities <http://it.ojp.gov/topic.jsp?topic_id=43>] and security/authentication: with some fragments, we won't be able to get at original content, so we might have to point to the original content rather than encapsulate it.

Devan: [Acknowledges that the Global Justice model has matured.]

Dan: The fragment you may want, you can't get, so you point to it. You're not downloading information and adding it to your system, you're pointing to it instead of dealing with multiple copies.

Bob: Would that involve some permanent identifier? It requires controlled vocabulary. Would it also support some sort of API? [e.g., <http://www.followthemoney.org>]

Dan: I'm not familiar with that organization. We could create a SOAP/API. [SOAP is a protocol for exchanging XML-based messages over computer networks, normally using HTTP/HTTPS <http://en.wikipedia.org/wiki/Simple_Object_Access_Protocol>.]

Isaac: As for authentication, MV5 (a type of algorithm) could ensure the version you have is exact data.

Devan: Authentication is a very big issue, even within a legislature [he references one state where legislative research services are so efficient that legislators vote on early bill summaries rather than bill content itself; unnamed state used MV5 as a solution to this problem.] Preservation of links will be very important to maintain over time (i.e., PURLs).

Dan: Each domain can be assigned a context domain, and each domain can manage its herd of sheep.

Bob: Does this make sense for or have value to this project?

Devan: I was going to ask the same question. What is a state's motivation to participate in this project in the first place?

Bob: We must define that; it's going to be different for each state. There's a high expectation that all information can be found on the Web. At a minimum, evolving technology has changed basic research needs of numerous users (immediate needs), or pre-positioning to preserve and make access available to valid, valuable content. The states may not want to do it now, but the content is valuable, and participating in this project is a low-cost way to preserve it and use later.

Tim: The theory is that you can do anything with XML, so where do you start? If this project is able to devise a core schema that works among six states, that in and of itself has value, and perhaps the schema may also be extensible.

Devan: If this is primarily about research (e.g., Minnesota's gun laws vis-à-vis Kansas'), that's one issue. If it's about core schema that other states can use as a framework to build a new XML bill-drafting system, that's another issue.

Bob: I think we need to define our boundaries, and for purposes of this project, I think a schema used as a basis for a bill-drafting system is beyond our scope.

Tim: We have three or four good ideas so far.

Bob: Do you think what we have so far is feasible and good for Thomson Reuters?

Dan: Thomson Reuters would see value in a core schema across states.

Tim: I think our main ideas are context, indexing, and namespaces.

Dan: I would emphasize that we create a namespace for the core schema. This way, a good one can be adopted by other states, and when states want to exchange data, they can use that, or they can keep their own data separate when they want to. And they can keep things safe over time.

Isaac: For example, if someone uses Dublin Core terms, you have to use its namespace by necessity.

Dan: Under the system we use now, terms end up in the same namespace by brute force, and this isn't what the states intended.

Jolene: Is there a lot of need to look at information and issues across states? I can see that there's obvious need within a state.

Bob: Based on [Minnesota's Director of House Research] Patrick McCormack's comments [in a January 2008 meeting], yes, there's an interest across the board—though who's interested and what the issue is changes over time. The staff does research in other states based on references they hear about via word of mouth.

[Prompted by its Legislative Reference Librarians staff section, the National Conference of State Legislatures initiated and maintains a topical listing of legislative and statutory databases, compilations and state charts/maps <http://www.ncsl.org/programs/lis/lrl/50statetracking.htm> to facilitate searching across all fifty states.]

Jolene: I'm seeing two things here: 1) a common vocabulary for content sharing; 2) a new structure built on top of the vocabulary, common metadata elements to facilitate cross-state searching. Will the project address one or the other, or both of these?

Bob: I think a little bit of both is possible. Of the parties represented in this meeting, Thomson Reuters, XMaLpha, and the Revisor's Office all have experience with structuring content, while the State Archives has experience with providing access to numerous audiences.

Devan: XMaLpha consulted on the Human Genome Project, where the data elements had to been whittled down from way too huge a list to something like 10 or 12.

Bob: [draws diagram on paper; reaffirms Devan's point that the schema has to be very limited and extensible.]

Dan: So we have two goals: 1) Designing a system that has an API; 2) we design a format for data interchange that's transportable to others' systems. Once the metadata is defined, then people start figuring out tools to use that metadata.

Bob: [References GIS metadata that has 400+ elements, too ambitious a scale.]

Dan: Once metadata is defined, and everyone knows the structure, then others start to use the data to solve problems in the public domain.

Bob: We could test this very quickly.

Isaac: Think of this as a nationwide state database; years from now (using the Semantic Web), a user can search for a specific issue legislated across 3 states, then also search for video from committee hearings for that issue. [For an explanation of the Semantic Web, see W3C's site <http://www.w3.org/2001/sw/>.]

Bob: Yes, but the first step would be setting this up on a state level with bills and acts and linking to contextual items. Information from other states is contingent upon their participation, after this project's timetable. MHS has to carve out a sustainable niche, for example by positioning itself on the preservation level (i.e., as a trusted repository), whereas Thomson Reuters excels at providing a value-added service.

Jolene: Based on Isaac's comments, I just came up with 6 data elements: date / jurisdiction / content classification / document type / sub-topics / author.

Devan: RDF can have subsets sitting on top of [?].

Jolene: I have little experience with RDF, so how facile or difficult might it be?

Devan: It would require some work. There are ways to generate RDFs automatically.

Tim: The core schema would make that easier.

Devan: There's enough expertise in this room to create the RDFs automatically.

Isaac: RDF, I think, will be a critical component. Without it, more interpretation of XML content by both people and automated computer systems will be required.

Devan: One other strong advantage to Isaac's suggestions: you're talking about an industry standard that may appeal to numerous entities.

Bob: Both the NCSL and Library of Congress can promote our project's outcomes.

Dan: With three parts to an interchange document (being metadata, content, and relationships) all being owned by the primary (or source) document, it would be easy to use standards like RDF and XPath/XPointer to pull documents.

Bob: Sounds similar, perhaps, to METS [the Library of Congress's Metadata Encoding and Transmission Standard <http://www.loc.gov/standards/mets/>].

Dan: XLink [XML linking language; <http://www.w3.org/TR/xlink/>] defines the location in the target document where the data is. It's an industry standard.

Tim: In preparation for this meeting, I invited the Revisor's Office Information Systems staff to comment on the XML schema being used for bills and what changes they would make, if given the chance. [Hand-out includes 6 suggestions to clean up the code and/or make it more efficient; see Appendix A.] Moreover, the IS staff has suggested roughly 12-15 possible tags for the project's schema [Tim said he likely could share these tags by mid-July].

Devan: So a next step would be to come up with a schema called "common vocabulary."

Dan: [Agrees with point 1 of Tim's IS report using XHTML tags.] We should do things that have bigger meanings (i.e., broader definitions).

Upon summarizing what next steps would be taken, Bob concluded the meeting at this point.

Appendix A

To: MHS's NDIIPP Project Team
Subject: Current experience with XML schemas
From: Tim Orr, Minnesota Revisor's Office
Date: June 6, 2008

Observations by Minnesota Revisor's Office IS staff on XML schemas for bills

1. Use XHTML tags when possible.
   XHTML tags applicable to bills are:
   <p>
   <del> <ins>
   <sub> <sup>

2. Try to use generic tag names. Then use 1 schema in many document types.
   This: <section>
   Not: <bill_section> <amendment_section> <statute_section>

3. Create a document level tag that includes the 1, generic schema.
   Example:

```
<document type="bill">
        <section>
                <p> This is paragraph number 1. </p>
                <p> This is paragraph number 2. </p>
                <p> This is paragraph number 2. </p>
        </section>
</document>

<document type="law">
        <section>
                <p> This is paragraph number 1. </p>
                <p> This is paragraph number 2. </p>
                <p> This is paragraph number 2. </p>
        </section>
</document>
```

   The benefits of this approach are:
   a. The 1 generic schema can be used for multiple document types. The number and complexity of schemas is reduced.
   b. The document type="" allows style sheets to display paragraphs in "bill" differently than in "law".
   c. Software (e.g., XPath code) becomes more consistent and manageable because the tag names are constant across multiple document types.

4. Opinion – Using attributes within a tag is preferable to many unique tags.

This:	<ref type="statute">	<ref type="law">	<ref type="adminrule">
Not::	<statute_ref>	<law_ref>	<adminrule_ref>

5. In addition to capturing content, tags and attributes are used for these purposes
   a. publishing to web (i.e XSLT programming to generate .html)
   b. publishing to paper (i.e., FOSI, XSL-FO programming to generate paper)
   c. software functions (i.e., MN XTEND function to auto generate the bill title)
   d. generate reports & QC (e.g., List all bills amending Statute chapter 287.)

6. Use features of XSD language to :

a. include one schema into another schema
b. extend or override the definition of a tag