Preserving State Government Digital Information XML Working Group Meeting Summary

March 6, 2009
Pillsbury Conference Room, Minnesota History Center St. Paul, Minnesota

Attendees: *Dan Dodge (Thomson Reuters); Isaac Holmlund and Tim Orr (Minnesota Revisors Office); Ward Johnson and Dan McCreary (Synatctica); Nancy Hoffman, Bob Horton, Jennifer Jones, Carol Kussmann, Charlie Rodgers, and Shawn Rounds (Minnesota Historical Society)*

Summary

The discussion focused on how to account for schemas within the wrapped bills. A schema registry was identified as a workable approach and an additional metadata tag, catalog ID, will be added to the wrapper to allow for schema references in such a registry in the future, though it was deemed unnecessary for the proof of concept. The registry will likely become an important element as the project gets larger and data is added from more sources. A decision was also made, based upon current best practices, to annotate the schema metadata to include the element definitions. Methods for creating unique document identifiers were also considered. Stable URIs offer many advantages and would be well suited to this purpose, but implementation will need further examination.

Discussion

This discussion is paraphrased.

Introduction by Bob Horton:

Bob began by saying that this meeting would further discussions and help answer questions that were brought up about the first draft of the XML wrapper that was posted on Basecamp. He hoped the group would reach a consensus on the wrapper and get it ready for testing with CA and MN data. He also thought the group could explore how the wrapper might work with the eXist database. He indicated that we would like to share this information with our partners and the National Council of State legislatures (NCSL).

Ward Johnson and Dan McCreary were introduced to Dan Dodge and a discussion followed summarizing the proposed use of the eXist database in our pilot project, including how the database and wrapper might work together since Dan Dodge had not met Ward and Dan or participated in any previous discussions about use of the database.

Dan Dodge then presented an overview of the wrapper document structure using the slide from his PowerPoint presentation given during the December 8th All Partners meeting, indicating that we would like to get each jurisdiction (state, county, city) to use native XML. He pointed out that the wrapper can also capture the content in various formats (PDF, Word, and other arbitrary binary files), but for purposes of this discussion, we wanted to get into the details of the wrapper schema. It was decided to leave parsing of the native XML out of the picture for the time being and just focus on the wrapper and parsing a standard metadata set. Dan Dodge noted that some of the metadata elements are required, while providing optional metadata would add value to the data set.

Dan McCreary commented that the wrapper had a flexible and well structured design.

Dan Dodge explained that the wrapper schema does not use strict validation, but this issue will probably be discussed later. He said the current lax schema was done with XML Spy, which helped when figuring out how the structure would work. He went over printed copies of the files he had posted with the first draft of the XML wrapper. He noted that some of these files were from the Revisors Office, and others are the examples of wrapped bill files.

Discussion began about how get the native XML schema from the provider to the receiver of the data, as it is currently not part of the wrapper. In particular, how would the native schema be associated with a large batch of files to avoid having a copy of the schema sent with each individual file?

Dan Dodge suggested adding a metadata field that could be the schema identifier. Then when files are sent, whether it's just one, a few, or thousands of files sent at a time, the schema is automatically linked to all files.

Dan McCreary said the he has done this using a schema registry, which is a table that includes a schema and a variety of descriptors, and optional name spaces. An associated name space with a URI allows you to keep track of the versions of schemas. He indicated that one problem with schemas is that they get changed without people telling you. Then, when you discover a broken application, you need to go get the new schema. These registries can be as simple as a spreadsheet to begin with, or something more elaborate, however, a registry would be valuable.

Dan Dodge asked if this registry would be centralized.

Dan McCreary said, yes, the registry would be controlled in a central location. It would have role-based access. A schema administrator would verify the information and schema. For quality-control reasons, information is generally pulled by the central administrator, rather than pushed by the data provider. For example, the administrator would want to verify that all 10,000 documents work and the paths don't break. He added that there are federal guidelines that define exactly what it means to break an existing schema.

Bob Horton asked if someone would need to host and maintain the registry or if there was an agency that would do it.

Dan McCreary said it would need to be hosted and explained that a registry is basically an XML file that lists schemas, versions of schemas and the dates they came in. When a new version is created, quality control can be performed.

Tim Orr suggested that a prerequisite for the use of this wrapper might have to be the ability to pull associated schemas as needed.

Dan McCreary explained that schema gathering could be initiated with an email message to do so. It could also indicate files have been placed in one spot on a public website so the information can be pulled at will. Schemas and URIs should comply with W3C standards for naming and design. Having a local cache of schema information and reliable data without having to go across the web to find them is ideal because it improves response time and performance.

Dan Dodge agreed that because we are archiving data, we would want to have the schema locally, and ideally would not want to transverse the web looking for this.

Dan McCreary indicated there are pros and cons for keeping a local copy of the schema versus link references, but if you just keep links to the schema namespace with archived data, the schemas could break when you try to bring the data out of the archive.

Tim Orr agreed that, for archival purposes, we would not want to rely on a website that could change. So how do you archive the schemas themselves?

Dan McCreary said that he had seen the health industry use checksums to verify valid transfer of files. Regarding the method of transfer, he recommended that, in general, the files should be written to write-once media for transfer and schemas kept with files would also need to be verified.

Bob Horton observed that the registry host must be equipped to archive documents as well as schemas and the Minnesota Historical Society (MHS) State Archives could serve as a registry archive.

Dan McCreary reiterated that archiving new documents in a native XML database like eXist is simply a matter of dragging and dropping the files into the database folder Bob Horton understood that the schema is identified in a document and the schema and document should be kept together.

Dan Dodge said for his documents (at West), he would pull the schema from the official source, such as the MHS registry and replicate it.

Dan Dodge said this sounds like an SGML catalog.

Dan McCreary replied that most of these registries actually use the name 'catalog' when describing themselves.

Dan Dodge said we could embed the catalog identifier in the wrapper schema in order to associate the native XML document with its provider's schema.

Dan McCreary explained there are federal guidelines to follow when creating version numbers for schema changes. It uses a three part number, such as, V.1.1 (with the first number referring to major and the second number to minor revisions).

Bob Horton suggested that we would then need to add an entity to the metadata elements – such as catalog identifier – and it should be required, with the idea that the data can't be trusted until we have an agreed upon catalog identifier.

Dan McCreary said we needed two things – if the source has a schema and a namespace, we would want to store that, and we would want to store the catalog identifier.

Bob Horton observed that, looking at the structure, the wrapper schema sits over schema that governs the source document and we want to identify the schema for the XML source.

Dan Dodge pointed out that the current version of the wrapper completely ignores the source document's schema. He said if we want to associate the schema as a metadata element this puts responsibility on the receiver of information to go to the catalog, find the appropriate schema, and enter that information into the wrapper metadata. The metadata wrapper could also capture the actual schema and namespace references used in the document.

Tim Orr asked what would happen, if we decided to move ahead with this idea of a schema registry, what would we do with the files the Revisors Office plans to send in June as a catalog will probably have not been created yet?

Dan Dodge said the information will have to be hard coded at first for the proof of concept. He suggested that we could include a placeholder for the registry information in the wrapper metadata for now, and then build the registry at a later date.

A discussion followed about whether or not there was currently a central spot for government schemas. Dan McCreary and Shawn said there was not one, but all agreed when Charlie suggested that the state Office of Enterprise Technology would be a good place to collect this type of data.

Tim Orr asked if it might also be helpful to add another tag to define the document type. He gave the example that they currently have bills which have associated schemas in various versions, but statutes may also have versions and need multiple schema versions, which could be different from the bill schemas.

Dan McCreary said that if there are a limited number of document types being entered into the database, it may be good to use a document type code, three letters for example, and add to it as needed.

Dan Dodge asked if a list of codes already existed.

Dan McCreary didn't know, but if you are working with a large number of documents, they could be evaluated to see how many different types of documents existed. For a small number of documents, the types could be easily identified and codes created, but first you need to get the data and do an analysis on the queries.

Isaac Holmlund asked about the states that do not have legislation in XML format. He said that we hoped to have several states provide us with information, however many do not have any documents in XML.

Bob Horton affirmed that Isaac was correct, only a few states are currently using XML bill drafting systems, which is why we planned to test first with documents from California and Minnesota.

Isaac Holmlund observed that we had been talking about requirements for states that have XML, but he wondered if there were any standards, ISO for example, for other document types.

Dan McCreary mentioned some of the standards that he knew of that might be applicable.

Isaac Holmlund recapped by saying he understood the registry would be an XML schema catalog that would indicate what versions of XML schemas have been used but since not all states are using XML, this cannot be required as previously suggested.

Dan McCreary agreed giving the example of HTML documents that don't have any meaningful XML markup. For those documents he suggested we may want to scrape the HTML, and then write extractors that collect the metadata for the wrapper.

Isaac Holmlund continued, saying that tracking versions of schemas by using the version numbering sequence would require states to change existing systems or develop new systems for tracking versioning of their schemas. He pointed out this would place another requirement on the states, to conform to a standard numbering system.

Dan McCreary agreed, but added that using a federally recognize standards and formats improve access and when you begin to share data, you see how internal practices begin to filter down and affect others.

Bob Horton pointed out that this helps to show that there are good reasons to follow best practices.

Tim Orr said that they currently don't number schema versions this way, but he thought it was a good goal to work towards.

Dan McCreary added you begin to see that changing one thing may break ten feeds further downstream; with more collaboration, these implications become magnified and following best practices becomes more important.

Tim Orr indicated that this was valuable recommendation for his office to institute.

Bob Horton said that most states are not yet in a position where they need to worry about XML schema standards, so if we can prove our concept, these things can be built into workflows and it may be easier for those states to adopt them if they move to XML in the future.

Dan McCreary said that, in general, most people don't use best practices when establishing schemas.

Bob Horton said the right schema has to be matched with the right file.

Isaac Holmlund indicated that there are a number of ways to deal with schema changes. The one they use now is to simply convert old documents to the new schema. When documents are archived, as we have discussed, the document creators will need to number the schema versions and remember to send copies of the new ones to the archive.

Tim Orr asked even if they change the schema so it no longer matches that for bills that were sent out for the year 2000, the content of the bill files has not changed, so why would they need to use schema versioning?

Bob Horton and Isaac Holmlund both answered that the Revisors Office would not have to associate schema versions with files, but they do need to keep track of the changes.

Dan Dodge pointed out that we would still want to keep an archival record of what schema was used.

Tim Orr asked how we would transmit schemas – one idea is to bury it in the wrapper and another is to send it separately.

Bob Horton said that he thought we were now talking about sending it separately. *(Originally we were thinking to include it in the wrapper.)* Including it in the wrapper would be too verbose.

Isaac Holmlund thought that would be great because the responsibility would fall on the receiver of the information to get the schema however, most people do not have theirs available online.

Dan McCreary mentioned a set of standards for transmitting files with published document standards used by the criminal justice system that he would be happy to explain further if we wanted learn more about it.

Bob Horton asked Dan where the best practices he cited were coming from. He noted that Dan had mentioned the criminal justice department a number of times.

Dan McCreary replied by saying the criminal justice department has a good understanding of the relationships of schemas between databases. He recommended consulting the Naming and Design Rules[1] for XML Schemas. He said there is a group in Washington DC that goes though case studies of federal and state agencies and use W3C standards to help find ways to transfer data between states. The document is based upon the idea that consumers and producers need to use XML to communicate precise semantics. Dan said the key is to adopt just enough standards to effectively accomplish your goals. .

Tim Orr asked if we even want to include transmission of schema as part of this project.

Bob Horton indicated that the pilot will just look at Minnesota and California so we probably would not have to include it yet.

Tim Orr expressed his concern that because of the in-depth discussion we already had about complicated schema transmission issues; we could be getting into scope creep.

[1] http://xml.coverpages.org/ndr.html

Isaac Holmlund thought that if we ignore it now, the schema transmission issue might get left behind.

Bob Horton thought we could provide education about best practices for schema transmission, but any given state would not use them unless there were compelling internal reasons to do so.

Dan Dodge remarked that without the schemas, he felt the value of the XML data would go down.

Dan McCreary thought perhaps we could talk about transmission concerns more specifically later.

Tim Orr stated that the wrapper could help transmit content to a receiver, but he wondered if it could or should transmit the schema.

Dan Dodge said that as long as there is a way to keep the schema associated with the files this is not necessary. It is not a priority with our proof of concept, but I think that this is important for the governance of schemas.

Bob Horton mentioned that we could make our system extensible so we can address this issue later.

Isaac Holmlund said that we may want to make schemas optional, and let the issue be decided by whether or not if the state is providing public access to their schemas online. Developing a complicated system may not be the best way to go.

Dan McCreary notes that such websites are time stamped so it is always possible to see if the schema has changed since the last time you gathered data.

Dan Dodge agreed that this approach could work.

Isaac Holmlund pointed out that it gave the provider a reason to use good system architecture, because it offers a reason for having schema online. He thought this seemed like a better method.

Tim Orr reviewed the wrapper components we had agreed upon as: the metadata block, and an optional URL tag for the schema.

Dan McCreary added that we also discussed including a catalog ID that points to the URL in the schema catalog.

Dan Dodge suggested that we could create two optional fields in our metadata elements.

Bob Horton thought that the cataloging ID could be filled out by the repository later. He did not see this field as necessarily being populated by the sender of data.

Isaac Holmlund said that if senders were using a URI in the XML source, it should be self describing.

Dan Dodge worried that a URI is not necessarily a URL, but we could specify that it must be dereferencible.

Dan McCreary indicated that providing dereferencible URIs is a good practice.

Dan Dodge agreed that the URIs should be accessible on the web.

Isaac Holmlund wondered if we are trying to figure out how every state can version schemas and send them. If our solution takes an extra transaction to send the schemas, he thought no one would do this, and the optional metadata tags would become useless.

Bob Horton said that if we think of the schema registry catalog and creating the catalog ID number as an archival function then he didn't think it would be necessary for the sender to provide the an ID number, they would only need to provide the schema, or URI to the schema.

Tim Orr speculated that there may be instances where providers do not want to keep their schema on a public website, for example, when the Revisors Office makes changes during a session. He would want to keep these schema changes internal until they were ready to publish at the end of the session.

Isaac Holmlund thought that public data in XML should have a public XML schema.

Tim Orr said making the schema public should be part of the publishing process.

Isaac Holmlund commented that in the end it should not matter whether the schema change is part of an internal or external process.

Dan Dodge pointed out that the XML source if in a CDATA section and it would be harder to get schema information from a CDATA section than from the wrapper metadata.

Isaac Holmlund suggested that the receiver could parse the CDATA section.

Dan McCreary said the best practices he knows of say that CDATA should only be used when you can't guarantee the well-formedness of the XML data inside.

Dan Dodge suggested that the receiver could look in the native XML for the schema references.

Isaac Holmlund asked if we could require providers to submit XML that it is well-formed.

Dan Dodge agreed it was possible, but thought that it raised questions that should probably be addressed in another discussion.

Bob Horton summarized by stating that we have decided to include but not require the new metadata elements discussed.

Dan Dodge agreed that they should be kept optional.

Isaac Holmlund disagreed saying that he didn't think we should have them at all, but if they are optional, it would not matter.

Dan McCreary asked if we had definitions for the metadata elements. Dan was given a paper copy of the metadata element definitions and crosswalk written up as a result of discussion at a previous wrapper development meeting. Shawn Rounds briefly went over the material with Dan.

Dan McCreary commented that strong schema designs include metadata annotations and the metadata element definitions must be precise, concise, distinct, noncircular and unencumbered by business rules. He referred to the data element definitions in Wikipedia.[2] He understood that we had written the definitions, but had not put them in the schema.

Dan Dodge said this was correct, but that he would put them into the wrapper schema.

Dan McCreary pointed out that once the definitions were written, the metadata could be registered, and you can query for the definitions.

Dan Dodge observed there is a big payoff if you have good metadata, because it allows cross cataloging of data.

Dan McCreary said, for example if there are ten elements and they are applicable to other states and the definitions are set, the only thing you have to do is get the data in the database and draw the maps, and you have a basis for canonical data.

Isaac Holmlund pointed out that this amounted to schema normalization.

[2] http://en.wikipedia.org/wiki/Data_element_definition

Dan McCreary said there were also ISO standards on how to write good metadata definitions.

Dan Dodge said that we don't really care about the native XML so much as having good metadata.

Dan McCreary agreed that the socializations of standards was key especially so that we can map between systems. This can be done using XPath expressions. It would be possible to use a 'publish once and subscribe' model of loosely coupled data.

Dan Dodge wanted to talk about metadata information that is not in the bill documents.

Isaac Holmlund said that Minnesota keeps some of the metadata elements in a separate database, so we would need rules about how to get that information.

Dan Dodge said that it sounds like for the proof of concept, we need the annotated schema, which will socialize the vocabulary so California and Minnesota can talk to each other about the data.

Dan McCreary noted that with a schema, it is in the leaf elements that you care about the definitions. The precision there will determine the validity of the standards.

Isaac Holmlund wondered if we should have some resolution about receiving schemas. He thought the reality was that transmitting schemas becomes unwieldy and that perhaps we need to rely on the ideal situation that there are URIs with valid URLs.

Bob Horton acknowledged that there are some difficulties with this, but he didn't think we need to resolve them for our proof of concept. He said we should start by building the capacity to include schema IDs in the metadata, but for our test, we can pull it off the website, or fill in the blanks. He suggested we can outline possible options, and may build in the functionality in the future.

Bob went on to begin outlining the next steps for the wrapper project. He indicated that Dan Dodge will pull in the metadata definitions after Shawn Rounds has verified that the current definitions of the metadata elements are correct. We will then transfer data from two states (California and Minnesota) and try mapping to the schema we have.

Tim Orr asked if the wrapper schema was basically done.

Dan Dodge said yes, apart from adding the Catalog ID element.

Tim Orr said that the Revisors Office would wrap and ship the files to MHS.

Bob Horton said we would then put the files into the eXist database and start working with Dan McCreary on education.

Tim mentioned that currently the Revisors Office staff use a FTP based system to push the data to MHS.

Dan Dodge asked about the scale involved – how many files?

Tim Orr said the he and his staff would like to include all of the bills in the 2009 session, currently around 2000, and probably about 3000-4000 by the end of the session.

Bob Horton reviewed the sequence of events, saying that Shawn would work on the metadata, and then the wrapper would go to Tim and Isaac. He indicated that he would talk to Dan McCreary about a contract for the work with eXist.

Talk about the time frame for future meetings. It was agreed that before April would be best because May is a bad month for the Revisors Office. Tim said he would hope to have the bills ready to go by the end of May. We would want to schedule the eXist training in April. Dan Dodge and Jolene Sather were invited to participate and Dan indicated they would be interested. Bob suggested we talk with the Legislative Reference Library to identify contextual materials we could put into the database.

Dan Dodge asked when we would get the California data.

Bob Horton said we are currently discussing that, but they are on hold right now unknown because of their budget issues.

Tim Orr wondered if his next question might be a topic for another discussion, but he asked what we wanted to do about unique identifiers for files names. He pointed out that when files come in from various states this might become an issue.

Dan Dodge thought that it would be the responsibility of each state to make sure their own files have unique names.

Dan McCreary added that in the eXist database we can then add a two letter ISO state code on at the beginning to ensure unique file names across the database.

Bob Horton suggested adding the year as well, since some states may reuse file names from year to year.

Tim and Shawn asked if this would look like a “MN2009” prefix, for example.

Dan McCreary said that the semantic approach would be to ensure each unique document file had its own stable URI bookmark.

Shawn Rounds asked Tim if their file names were currently based on a tree structure.

Tim Orr told her that was correct.

Shawn Rounds said that, as an archival function, we would probably want to keep an original file outside of the eXist database, so we would have to either keep your file structure or allow the files to repeat names in some way.

Isaac Holmlund said that if they package sessions together and sent the package that would keep the files together, but if they didn't send it at the end of a session, file names could end up being duplicated.

Tim Orr pointed out special sessions could cause problems in this scenario.

Bob Horton thought that if files were sent to the archives on specified dates, the archives would keep a copy as it came in, and put another into eXist, which can have a concatenated, unique name for the file.

Tim Orr suggested the unique name could start with a unique URL.

Dan McCreary agreed. He said the process of shortening a URL can create document identifiers, CURIEs.[3] These can help with searching and filling out form – a variety of things.

Bob Horton said that this kind of stable reference could play a role in establishing chain of custody and authentication, if URNs can become PURLs.

Tim Orr observed that if transmitting data becomes a pull rather than a push, like we are doing now, we will need a unique identifier before that happens.

Isaac Holmlund noted that the Revisors Office would have to create the URI rather than having the archives do it.

Dan Dodge thought that this could be developed into RDF identifiers, but was not necessary to do now.

Dan McCreary asserted that if you reference a version of the truth, you should not change it.

Bob Horton pointed out again that this process would create a chain of custody that could allow for certification and access.

[3] http://www.w3.org/TR/curie/

Tim Orr and Isaac Holmlund said that the statues that already have a unique URLs, so they thought this wouldn't be too hard to create these for the bills.

Dan McCreary said it was never too early to start thinking about creating URIs.

Meeting adjourned.