XML Wrapper Discussion
January 8, 2010

Attendees: Tim Orr, Isaac Holmlund, Michele Timmons (MN Revisors Office - RO); Dan Dodge (Thomson Reuters); Dan McCreary (Syntactica); and Robert Horton, Jennifer Jones, Charles Rodgers, Nancy Hoffman, and Carol Kussmann (Minnesota Historical Society - MHS)

Purpose: To discuss the wrapper prototype that was developed, address any concerns and determine the next steps for this portion of the NDIIPP project.

The following account is summarized.

Tim Orr described the background behind the code for the wrapper prototype. It was written in PHP but can be coded in many different languages. The XML that is returned is well formed and valid. The 'General Usage' section of the wrapper documentation handout explains the prototype. There are six URL line parameters that need to be edited to retrieve a wrapped bill. This method was chosen because it was the easiest method for the prototype; it can be changed if necessary. Dan McCreary stated that this method is very convenient and is quite standard.

Tim ran through an example. He cut and pasted the sample URL into browser, pressed enter, and was asked if wanted to open or save a file. It was saved as a zip file. Unzipping and opening the folder showed three files (the xml file and two different checksums –sha1/md5). Tim then opened the file in Mozilla to show the format including a section for metadata, XML, HTML, and encoded PDF.

Since the original rendition of the prototype, the XML is now available and needs to be updated. Tim has verified that the 'md5' and 'sha1' hashes are working as intended. (Hash values verify that the transmission was a good one and that files were unaltered in the process, which is especially important with the bills that are hundreds of pages long.)

Discussion on authenticity: The RO uses a secure server https and also has a digital certificate that is used on all RO web pages. Michele stated that eventually the hash values could also play a role. It was discussed that hash values could be used to verify that the original document that was downloaded from the RO site was the same as another version if there was a question of authenticity – a double check. The goal of the RO would be to be able to verify documents electronically; something that currently must be done by hand. Making this an automated process is a goal. Dan McCreary said that there are ways to do this using document IDs and having programs match hash marks inside a system and via a RESTful web service.

Bob asked Dan Dodge if these items are of interest to Thomson Reuters. Dan answered that being able to authenticate that the data we receive is that data that we were supposed to receive is very important to Thomson.

Tim talked about how at first the PDFs were not encoded in the wrapper but now they are. Dan McCreary agreed that it is a much better practice to encode the documents, because it is so much more reliable to a reader. If you use a binary format and wrap it in CData section, you can't tell,

and it fails; but if you encode it in base64 it can be reliably un-encoded. The file looks big, but it compresses well.

Isaac asked about encoding images. Dan McCreary said he does not recommend encoding them. He recommends using the new standards for Sprites[1] where all the images from a website are gathered and then CSS is used to point to the image.

Tim then asked about the tag label format in the metadata section. He points out that metadata section tags all start with 'meta.' and wondered if there was a utility to this, or if it was left over from previous discussions. Dan Dodge says that it was a convention that Thomson Reuters uses and he just followed it (their naming convention is two characters 'md' which allowed element names to be repeated without being confusing), it does not have to stay that way.

It was discussed that including 'meta' in the tag label could be good for the reader so they automatically know that they are in the metadata section however there is other contextual data that tells you this so it is not really needed, Dan McCreary agreed. Bob stated that it would be OK to remove the 'meta'. The group agreed.

Dan McCreary asked about the '.block' naming scheme that was used and stated that it does not follow the standards Dan McCreary is familiar with. Dan Dodge confirmed that it was similar to a Thomson Reuters convention. Dan McCreary suggests that maybe the '.block' does not need to be here either however, something is need to show each metadata section. A discussion followed and issues of cardinality, using plural tags, and creating different names. In the end, it was decided that '.block' does not need to be removed at this time.

Dan McCreary talked a little about the federal guidelines for metadata standards and standards on documentation called NIEM (the National Information Exchange Model)[2]. Once this information is packaged together, then it can become an artifact other agencies could use. Currently there is a federal metadata standard for statues, but not bills. Some terms may match, such as 'description.' 'Session.description' for example may be able to be mapped to this standard.

Isaac stated that it is worth looking at and Dan Dodge asked if it would be useful to have someone evaluate NEIM. Bob asked Nancy and Carol to look at the current standards. He also mentioned that we could talk with staff at the Congressional Research Service and the staff at the Library of Congress who administer the THOMAS system.[3] Dan McCreary said if we wanted to propose a federal standard, then adding the opinions of other peers will be important. Dan Dodge said that this would be useful for Thomson Reuters as well.

---

[1] "Some Web designers make sprites for small images or icons to reduce the number of requests the browser makes to the server. CSS is used to select the parts of the composite image to display at different points in the page. If a page has 10 1kB images, they can be combined into one 10kB image, downloaded with a single HTTP request, and then positioned with CSS. Reducing the number of HTTP requests can make a Web page load much faster. http://en.wikipedia.org/wiki/Sprite_(computer_graphics) [accessed 1/11/2010]

[2] http://www.niem.gov/ [accessed 1/11/2010]

[3] http://thomas.loc.gov/[accessed 1/11/2010]

Dan Dodge stated that documentation will be a vital part of the metadata element definitions.

Tim Orr talked about identical tags (description) that are in different metadata blocks. This needs to be changed or identifiers/qualifiers need to be added.

Overall we want to make it easy for others to extract the data.

Tim sees most of the future work on polishing up the metadata.

Tim asks if there are suggestions on how we actually call this program, currently it is URL based. Dan McCreary likes this; it is short and easy to use. You could use SOAP[4] but for smaller operations using SOAP is much harder and users that want to do mashups prefer the URL based RESTful web service.

Tim expressed concern that 6 parameters would be too many in the URL line and would make it too difficult to employ a RESTful web service, but Dan McCreary indicated this would not be a problem.

Talk about how the parameters are accessed (6 of them). Currently you must first search the RO website and find the parameters. It would be nice if there was a webpage that allowed for navigation or provided pick lists to help users find what they are looking for. Having a way to browse through the collection to get to the bills and having a link to click on would be a nice feature (like Data.gov).

Dan Dodge comments that Thomson Reuters is using more and more RESTful interfaces over SOAP.

(Bob asked when the article for NALIT that Tim and Isaac wrote about the wrapper would be out – in about a week or two. Bob states that we are working on a publication with NCSL on digital preservation and offers the Revisors Office a chance to review it and see if there is a place where information about the wrapper could fit as another avenue to help promote it.)

Discussion on Schemas

Tim asks about how to distribute the schema for the XML. It was said before that each legislature should have separate page where the schemas are located. Is this still the same thought?

Dan McCreary suggests having a path to the central schema page and using three levels of revision (major, minor, and revisions). Major changes should only happen once a year, minor

[4] "SOAP, originally defined as Simple Object Access Protocol, is a protocol specification for exchanging structured information in the implementation of Web Services in computer networks. It relies on Extensible Markup Language (XML) as its message format, and usually relies on other Application Layer protocols (most notably Remote Procedure Call (RPC) and HTTP) for message negotiation and transmission." http://en.wikipedia.org/wiki/SOAP [accessed 1/11/2010]

can be monthly, and the revisions are for fixing bugs. You need to make sure that the documents are synchronized with the schemas, or the results will break.

Tim asks if you have a schema that includes other schemas, and if any one schema changes, do you need to apply the changes to all documents (for all schemas)? There was then a discussion on when to change schemas and how to update them; the important thing to think about is what did you change and will it break a path link.

Isaac talks about the idea that schemas evolution is complicated and everyone has to be doing it right in order for it to work. We can build a system, but many states may not be doing it right, or have different abilities of how they can provide their schemas (for example MN is not public, it is private). We may need to consider other options for states to send us their schemas.

Dan Dodge, as a receiver of the data, would also like to get the schema for the file to copy to his own system. It would be nice if each jurisdiction would make it possible to download their schemas. Dan McCreary says that XML databases allow you to download the schemas, but then you are not going to the web to validate the schemas every time.

Isaac says that there are many different ways to do this, but we want a schema to have a namespace with a public URL.

Dan McCreary talked about walking people through different states. For example, we will help you in the first version and teach you (as there is not always someone in an organization who has done this before or knows about schemas and best practices). We need to start with education, simple best practices and move people up from stage 1 to stage 5. First, states just need to produce XML files. The second step is to ensure schemas are used. The third step is to use a versioned schema. The fourth step to produce schema versions with URIs and the final step is ensuring that those URIs are URLs … working up to a goal.

Isaac stated that we need to determine how much we can ask for from the states. In general, it must be easy. Dan Dodge suggested that maybe at the very least states would have to provide us with a person to call who can email us the schema. A personal contact and someone to work with can be very useful. Isaac agreed that is that is acceptable, for this process and this project.

Tim asked Dan Dodge if Thomson Reuters has an application that compiles all the schemas into one. Dan Dodge said that yes, it is done with a XSLT transformation using 'include' statements.

Carol asked about retrieving multiple documents rather than individual documents. It was discussed and determined that it could be set up using an asterisk for one of the variables that would pull a group of results. (This would not be possible on the current server due to load and stress on the system.) Dan McCreary said that if you are able to provide a pick list of metadata, to help users select the items they need, it will help to curb abuse; otherwise the hits can get out of hand. Atom[5] can be used to provide access to metadata.

Next Steps

[5] http://en.wikipedia.org/wiki/AtomPub

1. Drop ‘meta.’ on all tags.
2. Add xml to the prototype now that it is available.
3. Reissue the schema with schema revisions, so they are documented.
4. Address the two tags that are replicated (Tim and Dan Dodge). Make sure the MHS metadata documents also get changed.
5. Nancy and Carol will look at the federal standards (NIEM)
6. Send NCSL paper to RO
7. Determine what states may be able to help test the wrapper Dan Dodge suggests KS, Bob suggests CA, and Nancy suggests IL.