XML Wrapper Discussion
August 20, 2010

Attendees: Tim Orr, and Isaac Holmlund (MN Revisors Office - RO); Dan Dodge (Thomson Reuters); and Robert Horton, Jennifer Jones, Carol Kussmann, and Shawn Rounds (Minnesota Historical Society - MHS)

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

The following account is summarized.

Background and General Discussion

Bob gave a background on the wrapper and its development. Originally the project found people to be interested in a core schema that could be used as a basis for XML bill drafting systems; some basic set of entities that could be used to describe legislative documents. During development of this core schema, people saw the possibility of wrapping content with a core schema that could lend itself to the standardization of objects, and possibly even authentication. A prototype was developed and minimally tested.

The goal of this meeting is to compare notes, see where we are, and determine if what we have is what we want to share or if there is additional or further work to do. Bob sees the possibility of three products coming out of this work: 1) A Core Schema to identify and recommend, 2) Use of a wrapper with authentication, and 3) Use of a wrapper for transfer of batch content that is automated.

Comments on Changes to Wrapper

Tim says that the changes Dan made to the makes the schema easy to read, it looks much cleaner. It would be useful to know if end users find it useful, what file formats would they like to see included in the wrapper? If an end user downloads a file, what would they do with it next? Is there enough information there to be useful? Who could test this? CDL?

Other questions that end users (who?) could help answer is if the metadata intuitive enough? Does it need to change?

The Wrapper as an Authentication Tool

There are many ways to authenticate files; people are beginning to think that digital signatures may be overkill for every situation. Could the wrapper be used to authenticate objects from a record creator?

The wrapper as currently designed could be used to authenticate information passing from the record creator to an 'archive'. The XML format may not be very user friendly to a general user, data needs to be extracted; the package needs to be unwrapped and dissected to reach the desired file format.

If there was some sort of desktop utility that could unwrap the files, general users might find it more useful. (Something like WinZip.)

Authentication becomes more complex as data is passed from hand to hand and reused in various ways. There is only so much authentication that you can be responsible for. For example, if you make your data available for consumption (XML or other formats), you can authenticate the transfer (making sure the data you sent is the data received), but you (as the record creator) can't ensure to anyone that the data the user is making use of is the same authenticated data, it may have been changed. The responsibility for authentication must get passed down the line, from one user to another. There is no way to authenticate what users do with the data/information.

The Revisors Office currently provides a secure connection (https) to let people know that nothing has happened in a transfer as you downloaded a document. But once that document gets moved from place to place, there is no way to show that it is still authentic. To improve this the RO is looking at ways to seal downloaded documents, and has found two basic strategies: 1) verify the bits or 2) verify the language itself. Verifying the bits involves comparing that the bits match with hash tags/codes and give two entities ways to validate a document by comparing results. If you choose to verify the language, you are saying that the words and text themselves are valid. This is much harder to do. For example if you provide XML to someone and they reuse the information on another website, and someone else downloads information from this site and posts it elsewhere, anyone who then finds this information would be able to come back to you, the original record creator, and have the language validated (bit for bit the documents have changed and document formats may have changed but the language itself remains the same).

So maybe the wrapper, as currently developed, would be useful for arching authentication rather than with end user authentication. The wrapper can authenticate the transfer of files from one location to another as well as add metadata to the files.

[The Revisors office has been looking at outside vendors for authentication options and these companies are focusing on authenticating files by comparing the bits. Being able to authenticate a file regardless of file type (at the language level) seems almost impossible at this time.]

The Wrapper as an Aggregator

The wrapper could be used as an aggregator for data from multiple states. It would allow them to choose what format they are able to provide (XML, PDF, Word…). Users could pull what they want. It's the metadata that can be used to aggregate the information.

Finding What You Want

When the project first started, the push or pull relationship of data transfer was discussed and it was decided that users could pull the information when they wanted it, but how do they get to subsets of information. For the archives, the easiest would be to receive everything in a single package at the end of a session, but what needs to be in this package?

It may be useful to have a way to determine what files you want. The archive wants everything, but other users may only want information relating to education or budgets. How do they determine what bills they need to download? There is no way to find the files you want easily without knowing what it is you want. Might there be a way to allow users to explore data via metadata or date ranges. A service that indicates which bills fit a query would allow users to easily find the bills they would be interested in. Some users may want only new bills from the last time they collected information or bills on a certain topic or a certain format. A list of URLs could be called, which could be used to gather the data. Next step would be a download all button, so each individual URL does not need to be selected one at a time.

Multiple customers have multiple needs. Let them decide what they want.

Individual Bills or Collection of Bills

The wrapper currently wraps each bill file individually. The wrapper does not wrap associated bills (engrossments of a bill) together. A second application could do this. The metadata can be used to associate related bills. The wrapper as built needs to wrap each individual bill as it associates metadata not contained in the bill with the bill. A 'super wrapper' could wrap all the related parts if necessary.

Next Steps

Tim and Isaac can write code to return a list of URLS available for download and if possible work on a way to use the metadata to refine choices.

Want to be able to test the process of requesting files, wrapping the files, receiving the files and validating the files.

Can CDL also test this wrapper? Can they comment on the wrapper? Does it help them? How do they feel about the transfer and portability of the package?

The Revisors Office will work on creating code to show what files are available for use and an interface to assist with automation of use of the wrapper. Deadline is the end of October.

Other Issues

Isaac wants the project to keep an eye on National Information Exchange Model (NIEM), to make sure that our schema is close enough or along the same lines as similar NIEM standards to ensure an easy transition in the future if necessary. We want to avoid the possibility of needing to hand code items in the future. [Looking at NIEM, the only government related area is the Justice Domain which has a lot of metadata elements that follow ANSI standards when possible (ie. State codes). The metadata elements reference statutes, but they relate to things moving through the courts, not about the documents themselves.[1]]

[1] The NIEM tools site allows you to search for metadata elements and schemas. The Justice Domain may have the most elements that are similar to the wrapper, however the context and use is different. Information can be found here: http://niem.gtri.gatech.edu/niemtools/home.iepd