Information exchanges and interoperability architecture [1]

Submitted by erinmr on Mon, 2014-04-21 11:44 Wednesday, July 9, 2014 - 15:15 to 16:45
Event: Summer Meeting 2014 [2]
Session Type: Breakout [3]
Collaboration Area: Documentation [4]
Abstract/Agenda:

One of the major impediments to data reuse and interoperability is obtaining documentation for the schemas and vocabularies uses in existing published data. An information exchange specification is a definition of an information model and an encoding scheme used to transmit content according to the information model between agents through a communication channel. Our specific use case is transmitting geoscience information between data provider servers and data consumer client applications. In a data network like EarthCube, registering structured descriptions of the interchange in a manner that allows information model components to be discovered reused in new exchange definitions could foster interoperability. Development of systems for registering models and encoding schemes that would enable machine-processable mappings to be constructed between them would lay the groundwork for progressively greater automation of the data integration process. This session is intended to introduce the ESIP community to some existing data sharing networks using information exchanges, and to discuss future architecture that could support use and reuse of exchange specifications.

Notes:

Steve will give a brief introduction and then transition to more of a discussion.

Based on work at the Arizona Geological Survey. And some of the infrastructure they have set up for that and how that can be used in interoperability.

Dealing with long tail, heterogeneous data sets. Largest file is 2 million wells, but mostly 100 records in excel sheets.

Information from various schemes and getting them to work together.

Problem space - lots of small lots of heterogenous data. And making it comprehensive. You end up with complex models and schemas that are too complicated for anyone to use. Especially when handing off to programmers.

Information models, and people only use part of them. So started with the geothermal system to make simple schemes. And there needs to be explanations of what is in the data, where columns in a spreadsheet are not easily understood. What the fields are, and what values mean and how it is set up. Written form of this. Also finding a way for this to be machine processable.

Need for unique identifiers. And a method for validation rules that machines can test. Working in XML - have schemas and schema rules that test for semantics. And can get more complex for custom validation.

US GIN and geothermal system. Been creating interoperability tiers. Tier 1- Documents, scanned well logs, pdfs etc. That people can use and read but are not machine processable. Tier two is structured data which can be accessed in files Tier three is more complex information which structured data with standard schemas and coding. It has conventions on how it is translated between computers. What is this data about and how is it structured?

Json, xml, csv, RDF, etc.

Content models for Tier 3 (see slides). This is based on the national Informaiton Exchange Model NEIM.gov and EPA environmental information exchange network exchangenetwork.net

For our community we need something more flexible and can be accessed more easily.

To support these information exchanges, they have a registry for content models. Which has the view people can use to understand what is there. It will also have a link to the service which is implementing this exchange so you can see an example of the data that comes out of that service.

Use GIS as well, which has some information that is standardize for display on a map. And just use one scheme for two different sets of data.

When data sets are uploaded, there is a way to validate. There is a web interface that pulls in CSV files in USGIN and is also accessible in an API. This includes telling which information exchange you are using, and it uses the constraints from that exchange to look at each row of the data to see if it fits.

From this they have seen similarities across models and then they can build an information model that includes all of these. They have properties that are common to all instances at a higher level.

Ecosystem of exchanges that have a low barrier of entry, that is documenting how you are delivering data. So if you want to register a process as an exchange would just require documenting the scheme, then you can map between these different content models for interoperability.

Builds up a framework for getting information online with little overhead. The marketplace will use the ones that are good and they can eventually start culling the ones that are not used.

They are using GitHub for new exchanges. There is little governance, but if there is enough coming in and it is needed, they can create governance, but until then it is not needed. They rarely get feedback during the review process now. So they are thinking about how this will scale out.

Development process - looking at existing model. and there is a balancing between peoples theoretical idea of what they need to know, and what is actually being captured and quantified.

Designing these as flat services with simple features. Providing data not building clients.

Each exchange would be particular implementing model. Which means you would know many of the things to design a client without many issues.

Between client and server? OGC tunneling protocols. and also more common to use HTTP method (RESTful).

Usage - software developer can look to see how to get data out of an exchange and share it with servers. For the data consumer it is more transparent. The responses will be for services you can pull. For data providers, you get a template or framework. For brokers, it can provide a mapping object to create the transformer between systems.

Breaking things up in to smaller chunks that are still useful in a framework for interoperability. But you can still customize it for what you need. Each chunk is identified by a URI. That will be a reference point. And can put links between the chunks to make more complex data structures. using a linked data approach if you want to collect more information.

You build up a set of reusable property definitions that can be mapped to other schemes to automate information exchanges.

Architecture components - registry for exchange, descriptions, apis, dereferencing the identifiers for exchange and providing validation for trust. It works, it has been tested.

Last piece is a registry for how these materials are mapped, with RDF triples and some kind of API that lets things be mapped between the different schemes.

As far as USGIN is set up, the validation and registry is operational, but not yet the registry of properties and mapping between schemas.

To the group - your ideas for exchanges? What APIs do you use?

To build up a shared information model that can be more easily integrated to what is already there.

Question - are you planning to test that this can work for two heterogenous data sets?

Steve, I think we are. We have 45 subcontractors with their own data. And each come with their well logs in a database and ask, how do I get them in to this exchange format? And the bottom line is data integration is going to involve some transformation and mapping. That can be done on the client side. You can do it on the server side (which they are trying to do in the geothermal system). Giving them the work. And finally data brokering which requires someone else to make the decisions. But the work has to be somewhere. And wants to create patterns that are already defined and makes this process easier.

Question - either individual changes their data to model the other set, or asks the owner of the collection to change it for them to work in that system?

Denise - that is exactly what we are dealing with now. Modifying things to work with different types of data.

Steve - we are using GitHub for publishing the data, and publishing the schemas as well. and they collect issues there, and if one emerges for a model, they can make it evolvable. And these changes are versioned as well. Things are documented and one is not guessing.

Question - UDDI? have you heard of it as an API or service registry component.

Steve - has not seen an implementation that is scalable. Gave an example of a group who has done some exchanges between darwin core and another xml format, of various materials at museums and they create a global cache in UDDI.

Doug - Taxonomy data, which might be analogous with deep sea drilling data. But a few different, maybe less requirements given the commercial use of Steve’s data. They have lithologies up too.

http://schemas.usgin.org/models/ [5]

Steve gave example of how with deep sea drilling the data is associated with the boat whereas continental wells are not associated with the drill rig or driller. And is a different site than say TSR. This would be a new well header that has different information like ship, leg, lat/lon. And some other differences like top and bottom of well location.

Denise talked about her process for translating materials into this system.

Further discussion was held, but not captured in the notes.

Session Leads:	**Name:** Steve Richard [6] **Organization(s):** Arizona Geologic Survey [7] **Email:** steve.richard@azgs.az.gov [8]
Notes takers:	**Name:** Sarah Ramdeen [9] **Organization(s):** School of Information and Library Science UNC-CH [10] **Email:** ramdeen@email.unc.edu [11]

Accepted:

Source URL: https://commons.esipfed.org/node/2410

Links
[1] https://commons.esipfed.org/node/2410
[2] https://commons.esipfed.org/2014SummerMeeting
[3] https://commons.esipfed.org/session-type/breakout
[4] https://commons.esipfed.org/collaboration-area/documentation
[5] http://schemas.usgin.org/models/
[6] https://commons.esipfed.org/node/1398
[7] https://commons.esipfed.org/taxonomy/term/312
[8] mailto:steve.richard@azgs.az.gov
[9] https://commons.esipfed.org/node/557
[10] https://commons.esipfed.org/taxonomy/term/373
[11] mailto:ramdeen@email.unc.edu