Pittsburgh Linked Data Notes

These notes are a rough account of discussion and major topics at the 3 Linked Data Special Sessions at Dublin Core 2010, in Pittsburgh, PA, October 20-22, 2010. Sessions one and two were held back to back on Thursday afternoon, 21 October, from 2-3:30 and from 4-5:30. Attendance at these two sessions numbered roughly 100. Session 3, on Friday 22 October, 4-5:30pm, was opposite the second half of the joint DCMI Architecture Forum and W3C Library Linked Data Incubator Group session, and was more sparsely attended, with approximately 15 participants.

Session I: Domain Models

The session began with a short introduction to data models, presented by Corey Harper, and was followed by discussion.

The first point of discussion was about the FRBR model, and more broadly, about where one finds a model. A review of the DBPedia ontology's relationship to Wikipedia's "Info Box Templates" was seen as informative here.

It was considered that a data model should be based on purpose, which FRBR is (Find, Identify, Select, Obtain).

This led into a discussion of whether these verbs accurately reflect the research process. It was pointed out that "Find" is actually more akin to "Explore", where one is linear and the other is more non-linear.

It was noted that Metadata provides information and context for these non-linear processes.

Other use cases were discussed that skew the model:

- Digitization on demand changes find & obtain
- Use cases from XC: PIM, Collaborate, Categorize
- Linking out to other things is missing from the FRBR User Module

Library Data may be to broad a domain anyway. It includes:

- Published and purchased resources
- Licensed Resources
- Digitized objects
- Special collections and archives
- Cultural heritage objects

Perhaps this implies there is more than one data model.

Discussion ensued about the need for flexibility to change the model after and during experimentation - not needing to think of everything up front. Conclusion seemed to be that modeling is hard, but it doesn't need to be perfect up-front. Lightweight is good enough to bootstrap a community.

The idea of "agile modeling", that the modeling is iterative, and the data is iterative, though this raised the question about how to manage change sets and whether data should be marked as "beta" and consumers should be encouraged to use at own risk and not to cache.

A review of the relationship between

Session II: Vocabulary Selection & Development

The vocabularies session was opened with a short presentation from Karen Coyle, which closed by asking the following three questions:

- Is it better to reuse a property, or create a new one?
- How does one find existing vocabularies?
- How does one evaluate existing vocabularies

Following on the "is FRBR sufficient" conversation, there was a conversation about the scope and scale of the RDA vocabularies, and what constitutes essential or important metadata. RDA has 311 properties and 314 relationships, and they are all essential and important to someone. The reality is that you're needs may not be met "exactly" by an existing vocabulary and you will need to create your own.

This does, though, raise the question of how to manage relationships across vocabularies, given the size and scale. How to know what terms match up with what? Same for classes? Are FoaF and FRBR Person the same?

The question of searching for vocabularies was left largely open, though there was some agreement that Metadata Registries as well as effective vocabulary description had a role to play here.

Evaluation also proved a complex question. Among the factors to look at include who maintains the vocabulary, how it is maintained, and the goals of its development community.

It was noted that vocabularies presupposed the existence of a lot of specific "things" (resources) to describe, and those resources may be rooted in a particular worldview, making reuse difficult. This ties back to a previously discussed notion of building both domain models and vocabularies toward a principle of "least ontological commitment", which is to say limiting the extent to which your world-view is reflected in your ontology.

The idea was also raised of aggregating multiple views of the same data that adhere to multiple worldviews. Like pushing multiple different citation formats from our ILS systems. This approach is being developed in VIAF, which represents the persons in its authority file as FoaF Persons, as SKOS Concepts and as FRBR Persons, and these could in turn be validated according to more than one validation schema (whether a DSP or an OWL ontology, or something else).

This is a very expensive approach, though, and may not be right for all data sources. For some data, it makes more sense to ask yourself the question, "What would other communities most want in my data?" and focus on that. Again, this can be iterative and can change over time.

The session again seemed to come down to, let's build something now, stop counting angels on pins, and know that we can iteratively improve it as we go along. And again, the same question as before came up about how to not inconvenience or mislead data consumers with regularly changing data.

Session III: Breakout Sessions & Follow-up

The final session was very small, and very informal. Attendees were more people that were newer to this space, and a bit overwhelmed by the discussions in Sessions I & II, and wanted a sort of debrief.

People were looking for spaces to continue these conversations. ALA still seems like a good, though less international and cross-domain, option.

Additionally, attendees wanted a place to bring questions. To bring a draft data model and flesh it out collaboratively. To discuss and get help choosing &/or developing vocabularies. This need echoed Mike Bergman's keynote that morning, in which he suggested that DCMI is exactly the forum for taking on this role.

A developer from the Public Broadcasting System gave a presentation on initial data modeling work he's doing for that community, which led to some very interesting discussion.

Notes by:
Corey A Harper
Metadata Services Librarian
New York University Libraries