Open Data on the Web position paper: Richard Light

Background

I have been in the cultural history domain all my working life, aiming to make museum, archive and library information more useful and interoperable. In the course of this work I have been active in CIDOC (the International Documentation Committee of ICOM), and currently chair its Documentation Standards Working Group. What “interoperability” looks like has changed over the years: we have moved from common database schemas through XML applications to object-oriented Conceptual Reference Models. Currently, we are talking about Linked Data Design Patterns for cultural history, and we have a newly-formed group which is trying to develop a unified approach for museums, archives and libraries.

Broadly speaking, I am frustrated with endless meetings at which the talk is of “strategy” and of our aspirations for our data. (A reason not to invite me, perhaps …) So instead I have tended to concentrate on building stuff, in two particular areas:

- Shared ontologies
- Tools to support “URLification” of data

Shared understanding = shared foreign keys

One striking feature of many Linked Data resources (e.g. the British Museum) is that they are essentially self-referential, i.e. they don’t contain URLs which point outside their own domain space. However, this is a chicken-and-egg problem, in that there are too few resources to which they *could* point.

I have been working both to provide a platform which will allow publication of “authorities”, and actually to publish suitable resources. The Modes Linked Data Framework enables any Modes user to declare which files are to be published as Linked Data, and provides a mechanism for filtering out non-public data and for generating a wide variety of physical formats.

Using the MDLF, I have published:

- All the standard “termlists” provided with the Modes software
- The complete works of Shakespeare (my contribution to the Will’s World Hack, December 2012)

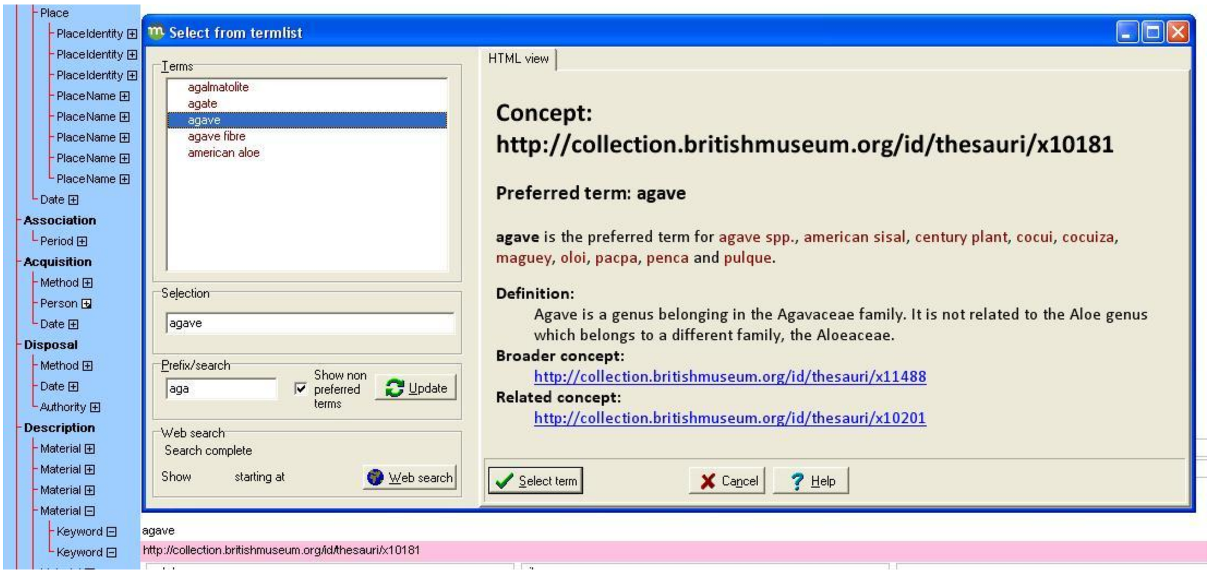
Publishing the Modes termlists will allow Modes users to include the URLs for concepts they have recorded using a controlled vocabulary, when they publish their own catalogues as Linked Data. It is not yet clear whether and how the Shakespeare resource will be used. Every speech Shakespeare wrote now has its own URL: indeed, so does every line and every word, but the focus of the XML resource on which I based my work is at the speech level.

The point about the Shakespeare experiment is that the Linked Data concept can be applied as comfortably to textual resources as it can to sets of assertions. It is obvious that giving a speech a LD URL will allow people to make RDF assertions about that speech, but it also provides a neat way of delivering the content of that speech, as and when it is required. The same approach can apply to articles and indeed books. I think that this aspect of the overall Linked Data picture – "non-non-information resources", if you will – has not been given enough attention. The Workshop has the topic "extracting human-readable "stories" from data", perhaps forgetting that we already have a wealth of stories in textual format, and simply need to engineer them into the overall framework.

Strings to URLs

It is a truism that approximately none of our existing data is ready to participate in the Linked Data world, because it is all stored as character strings or integer/number values. For cultural history resources, such as museum catalogue entries, I think that web services can make a difference by enabling the inclusion of Linked Data URLs as part of standard data entry work.

We have added the concept of a "web termlist" to our Modes software. This allows an arbitrary resource on the web to act as an authority file for controlling data values. The initial implementation linked to Geonames, but any web resource which supports HTTP requests and returns an XML response can be used as a web termlist. In particular, I have successfully linked up to two bog-standard SPARQL end-points (the British Museum materials thesaurus and the Ordnance Survey postcode resource).



Select from termlist

Terms: agalmatolite, agate, agave, agave fibre, american aloe

Selection: agave

Prefix/search: aga ☑ Show non preferred terms

Web search: Search complete

Concept: http://collection.britishmuseum.org/id/thesauri/x10181

Preferred term: agave

agave is the preferred term for agave spp., american sisal, century plant, cocui, cocuiza, maguey, oloi, pacpa, penca and pulque.

Definition:
Agave is a genus belonging in the Agavaceae family. It is not related to the Aloe genus which belongs to a different family, the Aloeaceae.

Broader concept:
http://collection.britishmuseum.org/id/thesauri/x11488

Related concept:
http://collection.britishmuseum.org/id/thesauri/x10201

agave

http://collection.britishmuseum.org/id/thesauri/x10181

Unless cultural history bodies start to use other peoples' URLs in their data, the idea of Linked Data URLs as tokens of shared understanding is going to be dead in the water.