Workshop Report: Vocabularies and the potential for linkage (VPLink)

1. Summary

The aim of the half-day Workshop was to bring participants up to speed with current approaches to vocabulary linkage, and empower them to apply these methods to their own projects and developments. It began with four brief presentations highlighting different aspects or approaches to linkage. A fifth presentation took the form of a case study around practical implementation of these techniques, in the case of the ESCO taxonomy. The workshop then split into two groups of about 20 participants each, focusing respectively on Mapping/alignment of Vocabularies and on Linked Data. At the end of the morning the groups reunited to present their conclusions. Slides of all the presentations are already on the workshop website, and audio recordings will be added when ready.

2. Presentations

Linked Open Vocabularies – the vision and the reality
Pierre-Yves Vandenbussche, Fujitsu Ireland
Principal developer behind the LOV project, our speaker contrasted the approach of Linked Open Data, in which the focus is on linking instances, with that of Linked Open Vocabularies, which facilitates data linkage and enables multiple vocabularies to be exploited in ontology design. With 444 vocabularies listed so far, the LOV "ecosystem" relies on VOAF, a vocabulary specification providing elements allowing the description of vocabularies (RDFS vocabularies or OWL ontologies) used in the Linked Data Cloud. Pierre-Yves described the linkage types available (see LOV Home : http://lov.okfn.org) and pointed to the high maintenance cost of links in a distributed system. Is this approach an opportunity for Everybody, or just for the committed ontology geeks?

Schema.org, SKOS and Wikidata
Dan Brickley, Google
The development of schema.org stems from collaboration amongst search engines (sponsored by Google, Microsoft, Yahoo! and Yandex) to make pages and documents easier to understand. Their initiative directly contradicts the popular myth "Google pays no attention to your metadata". As Dan explained, access to structured data can help the search engines enhance the user's experience of a particular site – but webmasters do need to engineer their markup actively to get the benefits. They do not necessarily have to publish their own vocabularies, however, if they link up with externally maintained schemas.

Despite the consortium's efforts to guide prospective users, there is some doubt how many webmasters have scaled the learning curve. Dan posed some questions for those who back KOSs (aka value vocabularies):

- How far to go with type/sub-type modelling before moving to a less formal notion of taxonomy, code-list?

- How can we help vocabulary maintainers expose taxonomy data using Web standard formats? (RDFa? CSV? JSON-LD?)
- Can we use Wikipedia/Wikidata as a linking hub?
- How can we make all this super-simple for webmasters?

Metadata vocabulary alignment: opportunities and challenges
Gordon Dunsire
Gordon's talk pointed to the proliferation of standards and schemas, new and old. Steering a path amongst them can be fraught with perils, as an element with the same name in two or more schemas can have subtly different shades of meaning, liable to knock a mapped query off track. The scope of a specific element or concept can evolve over time, invalidating the mappings previously established for it. Thus mappings can be unreliable, version control becomes complicated as well as essential, and provenance an ever present issue.

Applying ISO25964 to thesaurus mapping and other forms of linkage
Stella Dextre Clarke
As well as replacing ISO 2788 and ISO 5964 (the international standards for monolingual and multilingual thesauri), ISO 25964 is the standard for setting up mappings and other forms of linkage between vocabularies. It guides the types of mapping applicable, with examples to illustrate good practice. It is well aligned with SKOS, the W3C standard for publishing vocabularies and mappings between them. A correspondence table and derived RDF schema between the data models of SKOS and ISO 25964 is available 24/7 without charge or password control.

Challenges for building a European labour market taxonomy: ESCO
Johan De Smedt and Agis Papantoniou, TenForce
The ESCO (European Skills, Competences and Occupations) taxonomy has a total of over 200,000 terms in the 22 languages of the European commission. They are organized in three main hierarchies or "pillars" corresponding to Skills/Competences, Occupations and Qualifications respectively, with internal linkages between pillars as well as relationships to and from terms and concepts in external vocabularies such as ISCO (ILO Occupation codes) and NACE (Eurostat economic activity sector coding). This was fertile ground for illustrating the challenges highlighted in Gordon Dunsire's talk! The complexity of the interlinkages also highlighted the importance of an excellent understanding, not just of the vocabulary content and usage context, but also of the standards/technology environment, including ISO 25964, SKOS, RDF, JSON, OWL and others. Mapping has to be done with care, and on such a scale it is a team effort.
Johan ended by listing the following problems and conclusions:

- The ISO 25964 OWL schema does not allow for top level array organization. This could easily be remedied by allowing skos:ConceptScheme to be in the domain of iso-thes:subordinateArray.
- In the case of Concept groups holding hierarchies, neither ISO 25964 nor SKOS provides for properties to pick out the top members of a group. Although this could be inferred, a "topMember" list would be very efficient.
- skos:memberList (rdf:List) is good for sorting concepts and arrays, because it avoids the problems associated with using a sort key (e.g. in cases of polyhierarchy).
- When mapping concepts in different skos:ConceptScheme, problems arise using rdfs:subPropertyOf (between skos:mappingRelation (skos:broadMatch, skos:narrowMatch) properties and the respective skos:semanticRelation (skos:broader, skos:narrower) properties) because it sets up a dependency between hierarchies in the different concept schemes.
  Proposal: Make the mentioned property hierarchy not a required part of SKOS or SKOS-XL, but give them as a possible SKOS extension.

3. Discussion Groups

Group 1: Mapping/alignment of vocabularies
Group Leader: Gordon Dunsire
Rapporteur: Leonard Will

Gordon led the discussion by distinguishing between three main types of vocabularies:

- Element sets (schemas, ontologies)
- Datasets (lists, databases)
- Value vocabularies (KOS).

Once again he stressed the risk of misdirection if mappings are established in a careless way, leading to propagation of errors. The question was raised of whether mapping is worthwhile, if it carries such risks as well as maintenance overhead. On the basis of the considerable experience of ZBW Leibniz Information Centre for Economics in establishing good quality mappings between vocabularies in overlapping fields, Joachim Neubert pointed out that their information retrieval results are considerably enhanced by using the mappings between concepts to extend searches with all the synonyms gleaned in this way. Part of their success may be attributed to the ability of an end-user to detect and reject any irrelevant results obtained. It was noted that indiscriminate addition of synonyms would be more risky in a context where computer inferences are applied without human oversight or mediation. Others pointed out the value of concept definitions or scope notes to resolve doubts.

The development of FRBR and ISAD (G) are leading to improved practice in many sectors. In the cultural heritage sector considerable efforts have been applied to development of the CIDOC CRM (Conceptual Reference Model), which provides definitions and a formal structure for describing concepts and relationships used in cultural heritage documentation. This common, extensible semantic framework is intended to support mappings to any cultural heritage information, and provide a common language for domain experts and implementers to formulate requirements for information systems. While the direct beneficiaries should be users of museums, libraries and archives, there is some hope that other sectors could follow the CRM as a guide to good practice in conceptual modelling. Meanwhile, release of the Getty Research Institute's *Art and Architecture Thesaurus* (AAT) as linked open data should help us all see the potential of these techniques.

Some specific recommendations for good practice include:

- Make sure URIs are kept stable across vocabulary versions, where concepts are unchanged. This may apply even if the preferred term changes.
- Watch out for terms with identical form denoting different concepts in different vocabularies; small subtle differences (such as different cultural expectations when language barriers are crossed) can have large unforeseen implications.
- Within *one vocabulary*, skos:broader, skos:narrower and skos:related are the main semantic relations available, corresponding to BT, NT and RT in a thesaurus. Hierarchical relationships within a single vocabulary should be reliable because these relationships serve in part to define the scope of a concept, as intended by the vocabulary creator. Hierarchical mapping

relationships between different vocabularies are open to some doubt, as they are established by comparison of concepts that may be defined differently. The applicable relationships for mapping between concepts in *different* vocabularies are skos:broadMatch, skos:narrowMatch and skos:relatedMatch. The properties skos:closeMatch and skos:exactMatch are available too in the context of mapping, and they correspond to the tags ~EQ and =EQ using the conventions set out in ISO 25964.

- When vocabularies (including metadata schemas) are updated, any external mappings to or from concepts that have changed need to be checked and revised.
- In KOSs, it is important to give each concept a URI; in datasets, the same applies to each instance. In both cases this enables Linked Data applications, as well as helping generally with identification of the item. In the case of datasets that are published independently but happen to describe copies or editions of the same item in local collections, this can lead to multiple URIs for that item.

Finally this Group voiced concern about the need for learning materials and training opportunities for all the people involved in vocabulary linkage. This includes the experts who establish the semantic links between concepts, as well as the developers and implementers of ontologies. A lot of people need to extend their skills if vocabulary linkage is to enjoy widespread, effective adoption.

Group 2: Linked Data
Group Leader: Dan Brickley
Rapporter: Judi Vernau

“By publishing our vocabularies on the Web we can enable the linkage of vast and diverse bodies of knowledge. So what stops us all doing it, now?”

In Dan Brickley’s view, general purpose generic "triple/graph browsers" lead to an unconvincing experience for navigating, browsing and using RDF data. “You don't want the exact same user interface for navigating detailed genome data as for a bibliographic dataset, just because both are shared as RDF,” he said, but “Creating an entirely new user experience / interface layer for every single dataset also feels like it's missing an opportunity for re-use.” Instead he proposed a “middle level”, possibly around the level of DC/FOAF/SKOS, with events, people, places, works (documents, images etc.), timelines and topics, that is appropriate for a rich shared user interface, and which would work with data merged from hundreds or thousands of fairly diverse linked data collections. This type of development would be unlikely to materialize without explicit planning.

The group went on to list the following issues in summary:

- Ontological models differ. Which to use, and why?
- Lack of experience and detailed understanding of the tools and techniques
- Lack of registries
- Lack of good metadata on potential resources, and URIs may not be persistent
- Data quality often inadequate, perhaps arising from poor information governance
- When you publish a dataset, it is hard to predict how/where it will be used, or even find out whether and where it has been used.

- Not enough mid-level tools.
- Implementation may be harder than we think.

What we need to overcome these problems:

- More registries of vocabularies and other datasets
- Good metadata (including ownership, rights, scope, frequency and more)
- Good success stories to publicize
- More "middle-level" tools for data linkage
- More guides, tutorials, workshops and other opportunities for learning how to do it.

4. Conclusions

As Chairman, Gordon Dunsire concluded by expressing confidence that vocabulary linkage techniques will soon become widespread, as the tools we need are developed and made readily available. This report will be circulated and posted on the ISKO UK website, and transmitted to the organizers of the forthcoming NKOS workshop to be held in London on 11-12 September, with the main theme of *Mapping between Linked Data vocabularies*. Among the conclusions are not just the points raised in discussion, but also the questions raised by speakers, as reported above.

S G Dextre Clarke, 9 July 2014