8-1-2014

Harmonization Of Vocabularies For Water Data

Simon J.D. Cox

Jonathan Yu

Bruce A. Simons

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

HARMONIZATION OF VOCABULARIES FOR WATER DATA

S J D COX, J YU & B A SIMONS

Environmental Information Systems, CSIRO, Highett, Australia

Observational data encodes values of properties associated with a feature of interest, estimated by a specified procedure. For water the properties are physical parameters like level, volume, flow and pressure, and concentrations and counts of chemicals, substances and organisms. Water property vocabularies have been assembled at project, agency and jurisdictional level. Organizations such as EPA, USGS, CEH, GA and BoM maintain vocabularies for internal use, and may make them available externally as text files. BODC and MMI have harvested many water vocabularies alongside others of interest in their domain, formalized the content using SKOS, and published them through web interfaces. Scope is highly variable both within and between vocabularies. Individual items may conflate multiple concerns (e.g. property, instrument, statistical procedure, units). There is significant duplication between vocabularies.

Semantic web technologies provide the opportunity both to publish vocabularies more effectively, and achieve harmonization to support greater interoperability between datasets.

- Models for vocabulary items (property, substance/taxon, process, unit-of-measure, etc) may be formalized OWL ontologies, supporting semantic relations between items in related vocabularies;
- By specializing the ontology elements from SKOS concepts and properties, diverse vocabularies may be published through a common interface;
- Properties from standard vocabularies (e.g. OWL, SKOS, PROV-O and VAEM) support mappings between vocabularies having a similar scope
- Existing items from various sources may be assembled into new virtual vocabularies

However, there are a number of challenges:

- use of standard properties such as sameAs/exactMatch/equivalentClass require reasoning support;
- items have been conceptualised as both classes and individuals, complicating the mapping mechanics;
- re-use of items across vocabularies may conflict with expectations concerning URI patterns;
- versioning complicates cross-references and re-use.

This presentation will discuss ways to harness semantic web technologies to publish harmonized vocabularies, and will summarise how many of the challenges may be addressed.

INTRODUCTION

The publication of vocabularies of water observation parameters, and related classifiers, has been made significantly easier with the development of semantic web technologies, particularly RDF, SKOS and OWL [1] [2] [3], [4], [5]. A number of services have been established that host and publish vocabularies relevant to applications in the earth and environmental sciences, and which use semantic web technologies for formalization and publication. Some patterns for these are now emerging, as well as some questions.

Of particular note are the following systems which include vocabularies of relevance to the environmental sciences, including hydrology, and are explicitly targeted for use by multiple organizations and initiatives:

1. SWEET Ontology from JPL/NASA [6] - a set of interlinked ontologies, that covers a large fraction of the natural sciences;
2. OBO Foundry ontologies, developed by the biomedical community [7], in which ChEBI provides a set of >30,000 chemical species [8], [9].
3. NERC Vocabulary Service (NVS) [10], [11] hosted by the British Oceanographic Data Centre on behalf of the UK Natural Environment Research Council, and has grown out of a collection of more than 50 vocabularies developed in, or commonly used by, the oceanography and marine science communities.
4. The Marine Metadata Initiative [12] has collected a similar, and in some cases overlapping set of vocabularies to NVS, converted to SKOS, and published vocabulary items using unique URIs in their namespace;
5. Environmental Thesaurus Server [13] from City University of New York has done another similar harvest of existing vocabularies, and provides a service to access them. However, individual items are accessed by URIs that depend on the current software implementation, so are unlikely to be persistent long term;
6. The World Meteorological Organization has recently developed the WMO Codes Registry [14] for publication of codes from various WMO standards, and also registering some external vocabularies. For the latter group the vocabulary items retain the original URIs, with a separate registry record in the WMO domain;
7. The GeoSciML community, under the auspices of the IUGS, publish more than 50 controlled vocabularies used in geology and mineral resource data, including multiple versions of the geologic timescale [15]. The vocabularies are published primarily as SKOS, though in some cases there is an alternative view using a specific model;
8. The Australian National Environmental Information Infrastructure (NEII) is publishing vocabularies in the domain environment.data.gov.au. This includes an observable property vocabulary for water data [16];
9. The Open Geospatial Consortium manages ‘definitions’ in many standards, which are formalized and published using SKOS [17]. An additional service delivers coordinate reference systems, formalized in XML using the GML Coordinate Reference System schema (e.g. http://www.opengis.net/def/crs/EPSG/0/4326);
10. GEMET Thesaurus of environmental information terms [18] from the European Environment Agency;
11. INSPIRE registry [19] includes vocabularies needed for the INSPIRE spatial data infrastructure.

All of these systems expose vocabularies using RDF-based semantic web technologies. This enables and should encourage re-use, and recording of explicit dependencies, mappings,

and cross-references, using links based on the URIs for vocabularies and vocabulary items. However, there is significant variation in the ways that the vocabularies are formalized and structured, the approach used to link between related vocabularies, the URI patterns used to denote items and services, and the web interfaces to these. In addition, the organizational arrangements are quite heterogeneous. Some vocabularies are backed by organizations that might be expected to be able to manage long-term persistence (European Commission, NERC, NASA, WMO, OGC) though the organizational stability is not necessarily translated into a commitment to a particular maintenance regime. Other systems appear to have emerged from a local project, with less confidence about long term reliability.

Common approaches to harmonization are under development. A number of technical and business concerns must be taken into account. In this short paper we will enumerate some of the key issues, and note that there are many viable approaches and best practice is yet to emerge.

FORMALIZATION

Most vocabularies listed above are formalized using the Simple Knowledge Organization System (SKOS) [3] with all items modelled as individuals of the class skos:Concept. In contrast, SWEET [6] and OBO [8], [9] model some or all concepts as OWL Classes [4], [5].

Some vocabularies take a hybrid approach, with vocabulary items being primarily instances of classes from an ontology or model that is tailored to the application, but also aligned with SKOS (e.g. [15], [16]), so individuals in the vocabulary are simultaneously members of both the application specific class and the class of SKOS concepts. This means that the vocabularies can be accessed using a generic vocabulary API based on SKOS [20], but the graph describing each concept that is delivered to the user has rich semantics relevant to the domain.

MAPPINGS AND OTHER CROSS-REFERENCES

Properties are provided in various well known vocabularies to be used to assert equivalence and other mapping relationships that may be used to link terms in different vocabularies. SKOS provides exactMatch, closeMatch, narrowMatch, broadMatch for concepts [3], and OWL provides sameAs for individuals, equivalentClass for classes and equivalentProperty for properties [4]. Other relations are provided in Dublin Core [21], PROV-O [22], VoID [23], VOAF [24].

Apart from matching items of similar scope, concept definitions may refer to concepts in related vocabularies. For example, entries in the Observable Properties vocabulary that define concentrations of various species link include references to both internal and external vocabularies, using properties defined in the OP ontology [16].

Linking between vocabularies defined in SKOS and vocabularies defined primarily as OWL classes can be problematic if the goal is to support automated reasoning. When used in systems based on formal semantics, there are also questions about using SKOS at all because of its less formal semantics. RDFS and OWL2 ‘punning’ permit arbitrary links [2], [4], but does not solve the challenge in reasoning applications.

COLLECTIONS

There are multiple options for containers to collect a related set of terms into a ‘vocabulary’. SKOS provides either Concept Scheme or Collection [3], with different behaviours. OWL provides Ontology [4] though there is no explicit way to indicate membership. Other standard RDF vocabularies, such as VoID [23] provide further alternatives.

SKOS Collection appears to provide useful functionality along with the flexibility to use them nested. Because collection membership is defined through skos:member properties on the container element, new collections may be composed by re-mixing existing content.

URI PATTERNS

In principle URIs are a separate concern, orthogonal to structure and relationships within the resources that they denote. For the larger vocabularies there is a tendency towards simple structure and opaque tokens [7], [9]. In practice, however, many providers see some advantage in memorable URIs, with patterns on the path that imply

- ownership of vocabulary items by an authority;
- membership in a container vocabulary or collection;
- hierarchical relationships within vocabularies.

A common and generally useful pattern is where the URI for a concept can have the last element trimmed to make the URI for the primary collection that contains or 'owns' the concept (e.g. http://environment.data.gov.au/water/quality/def/property/ is a collection containing http://environment.data.gov.au/water/quality/def/property/cadmium_concentration and others).

However, caution should be applied in embedding semantics, such as a semantic hierarchy, in the concept URI. Assigning the URI implies fixing semantics, which may be re-evaluated later. And a URI path can only support a single hierarchy, while RDF, SKOS and OWL are based on a set theoretic model in which poly-hierarchies are quite natural. In practice, the URI stem for a concept URI usually indicates the *original* collection of which it was a member–the 'maintenance collection'–without any limitation on its participation in additional collections.

CLONE, OR LEAVE ALONE

Given the increasing availability of relevant vocabularies formalized using the semantic web technologies and published using URIs, the question arises about how to incorporate an existing vocabulary within a new service, and existing terms in a new vocabulary. Many of the 'first generation' of semantic vocabularies formalized the content into RDF for the first time, so the concept URIs were necessarily new, and therfore in a namespace controlled by the publisher, rather than by the originator.

However, now we have a corpus of vocabularies published, there is the opportunity to re-use elements in-place. Re-use may be of whole vocabularies, of parts of existing vocabularies in new collections, and by cross referencing between individual items. For example, the water-quality vocabularies published in NEII link to ChEBI for definitions of chemicals [25][16]. In contrast, the vocabularies in the NERC service are all cached locally, forming a closed system [11]. The latter approach overcomes issues of reliability of source services, and there are also significant performance advantages when reasoning over more than one vocabulary if they are held in the same system: federated SPARQL is notoriously slow. But synchronization with the source must be managed, and individual entries are not tied to the originals.

VERSIONING

Some vocabularies include a version indicator in the URI for each concept (e.g. SWEET [6], OGC definitions [17]). This is now understood to be usually unnecessary and probably unwise. The reasoning is partly theoretical, partly pragmatic. The theoretical argument against versioning concept URIs is that a concept is an abstract thing. What we *know about it* may change, but that implies that the description (RDF graph) about the concept should be

versioned, while the concept itself is a node at the centre of the graph whose identity is stable. The pragmatic concern is how to manage references to concepts over time if the identifier is versioned, particularly if the descriptions of related concepts do not include any provenance information. For example all concepts get the URI updated in every new version of SWEET. So although the final element in the URI may be the same or similar, there is no formal indication as to whether

```
http://sweet.jpl.nasa.gov/1.1/time.owl#PLEISTOCENE
```

means the same as

```
http://sweet.jpl.nasa.gov/2.0/timeGeologic.owl#Pleistocene
```

or

```
http://sweet.jpl.nasa.gov/2.2/stateTimeGeologic.owl#Pleistocene .
```

A standard RDF processing system would assume they do not.

It is more likely that a Collection need be versioned, particularly if its membership changes. However, if a URI pattern has been adopted which makes the collection URI the stem of the member concept URIs (see above), then a version number cannot appear even here since if we want the concept URI to be stable, and its stem to denote the primary collection of which it is a member, then the collection URI cannot be versioned.

CONCLUSIONS

The emergence of semantic web technologies supports a significant convergence in standardizing the encoding and delivery of technical vocabularies, as needed for water observation data. In particular SKOS and OWL provide formalizations for vocabulary structure, and SPARQL a standard low-level API. However, there is great diversity of practice within this, particularly around the construction of collections, URI patterns, the deployment of closed or open systems, and the design of ontologies to support domain specific vocabularies. Diverse patterns are seen even in vocabularies developed by the same technical team, because of differing context and expectations [26]. A number of relevant vocabularies are now deployed and maintained at stable web addresses, so we can expect increasing collaboration and harmonization between different providers, and the developed of some shared practices.

ACKNOWLEDGEMENTS

Many thanks to Laurent Lefort and David Ratcliffe for explaining some of the finer points of OWL, and to Adam Leadbetter, John Graybeal, Roy Lowry and Jeremy Tandy for numerous discussions on the finer points of vocabulary standardization.

REFERENCES

[1] R. Cyganiak, D. Wood, and M. Lanthaler, “RDF 1.1 Concepts and Abstract Syntax.” World Wide Web Consortium, 2014.

[2] D. Brickley and R. V Guha, “RDF Schema 1.1.” World Wide Web Consortium, 2014.

[3] A. Miles and S. Bechhofer, “SKOS Simple Knowledge Organization System Reference.” World Wide Web Consortium, 2009.

[4] J. Bao, E. F. Kendall, D. L. McGuiness, and P. F. Patel-Schneider, “OWL 2 Web Ontology Language Quick Reference Guide (Second Edition).” 2012.

[5] W3C OWL Working Group, “OWL 2 Web Ontology Language Document Overview (Second Edition).” World Wide Web Consortium, 2012.

[6] R. G. Raskin and M. J. Pan, “Knowledge representation in the semantic web for Earth and environmental terminology (SWEET),” *Comput. Geosci.*, vol. 31, no. 9, pp. 1119–1125, Nov. 2005.
[7] “The Open Biological and Biomedical Ontologies.” [Online]. Available: http://www.obofoundry.org/. [Accessed: 27-Mar-2014].
[8] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, “ChEBI: a database and ontology for chemical entities of biological interest.,” *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D344–50, Jan. 2008.
[9] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, “The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013.,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D456–63, Jan. 2013.
[10] “NERC Vocabulary Server version 2.0 (NVS2.0) at BODC.” [Online]. Available: http://vocab.nerc.ac.uk/. [Accessed: 13-Feb-2014].
[11] A. Leadbetter, R. Lowry, and D. O. Clements, “The NERC Vocabulary Server: Version 2.0,” in *Geophysical Research Abstracts 14, Proceedings EGU General Assembly 2012v*, 2012.
[12] Marine Metadata Interoperability, “MMI Ontology Registry and Repository.” [Online]. Available: http://mmisw.org/orr/#b. [Accessed: 24-Mar-2014].
[13] P. Ji, “Environmental Thesaurus Server,” 2013. [Online]. Available: http://edscvs.ccny.cuny.edu/. [Accessed: 27-Mar-2014].
[14] World Meteorological Organization, “WMO Codes Registry.” [Online]. Available: http://codes.wmo.int/. [Accessed: 28-Mar-2014].
[15] S. J. D. Cox and S. M. Richard, “A geologic timescale ontology and service,” *Earth Sci. Informatics*, 2014.
[16] S. J. D. Cox, B. A. Simons, and J. Yu, “A harmonised vocabulary for water quality,” in *Proceedings, 11th International Conference on Hydroinformatics - HIC 2014*, 2014.
[17] S. J. D. Cox, “OGC Definitions Service.” [Online]. Available: http://def.seegrid.csiro.au/sissvoc/ogc-def/resource?uri=http://www.opengis.net/def/. [Accessed: 27-Mar-2014].
[18] EIONET, “GEMET Thesaurus,” 2012. [Online]. Available: http://www.eionet.europa.eu/gemet/. [Accessed: 27-Mar-2014].
[19] European Commission - Joint Research Centre, “INSPIRE registry.” [Online]. Available: http://inspire.ec.europa.eu/registry/. [Accessed: 28-Mar-2014].
[20] S. J. D. Cox, J. Yu, and T. Rankine, “A Linked Data API for SKOS vocabularies.,” *Semant. Web J.*, vol. submitted, 2014.
[21] “Dublin Core Metadata Element Set, Version 1.1.” [Online]. Available: http://dublincore.org/documents/dces/. [Accessed: 30-Mar-2014].
[22] T. Lebo, S. Sahoo, and D. McGuinness, “PROV-O: The PROV Ontology.” World Wide Web Consortium, 2013.
[23] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao, “Describing Linked Datasets with the VoID Vocabulary.” W3C, 2011.
[24] L. Rozat and P.-Y. Vandenbussche, “Vocabulary of a Friend (VOAF),” 2012. [Online]. Available: http://lov.okfn.org/vocab/voaf/v2.0/. [Accessed: 30-Mar-2014].
[25] S. J. D. Cox, B. A. Simons, and J. Yu, “eReefs vocabularies - substances and taxa.” [Online]. Available: http://sissvoc.ereefs.info/sissvoc/ereefs/resource?uri=http://environment.data.gov.au/water/quality/def/object/. [Accessed: 30-Mar-2014].
[26] S. J. D. Cox, “Best practice in formalizing a SKOS vocabulary,” 2014. [Online]. Available: https://www.seegrid.csiro.au/wiki/Siss/VocabularyFormalizationInSKOS. [Accessed: 20-Mar-2014].