Geophysical Research Abstracts
Vol. 12, EGU2010-7577, 2010
EGU General Assembly 2010

Use of standard vocabulary services in validation of water resources data

Jonathan Yu, Simon Cox, and David Ratcliffe
(jonathan.yu@csiro.au)

Ontology repositories are increasingly being exposed through vocabulary and concept services. Primarily this is in support of resource discovery. Thesaurus functionality and even more sophisticated reasoning offers the possibility of overcoming the limitations of simple text-matching and tagging which is the basis of most search.

However, controlled vocabularies have other important roles in distributed systems: in particular in constraining content validity. A national water information system established by the Australian Bureau of Meterorology ('the Bureau') has deployed a system for ingestion of data from multiple providers. This uses a http interface onto separately maintained vocabulary services as part of the quality assurance chain.

With over 200 data providers potentially transferring data to the Bureau, a standard XML-based Water Data Transfer Format (WDTF) was developed for receipt of data into an integrated national water information system. The WDTF schema was built upon standards from the Open Geospatial Consortium (OGC). The structure and syntax specified by a W3C XML Schema is complemented by additional constraints described using Schematron. These implement important content requirements and business rules including:

• Restricted cardinality: where optional elements and attributes inherited from the base standards become mandatory in the application, or repeatable elements or attributes are limited to one or omitted. For example, the sampledFeature element from O&M is optional but is mandatory for a samplingPoint element in WDTF.

• Vocabulary checking: WDTF data use seventeen vocabularies or code lists derived from Regulations under the Commonwealth Water Act 2007. Examples of codelists are the Australian Water Regulations list, observed property vocabulary, and units of measures.

• Contextual constraints: in many places, the permissible value is dependent on the value of another field. For example, within observations the unit of measure must be commensurate with the observed property type

Validation of data submitted in WDTF uses a two-pass approach. First, syntax and structural validation is performed by standard XML Schema validation tools. Second, validation of contextual constraints and code list checking is performed using a hybrid method combining context-sensitive rule-based validation (allowing the rules to be expressed within a given context) and semantic vocabulary services.
Schematron allows rules to incorporate assertions of XPath expressions to access and constrain element content, therefore enabling contextual constraints. Schematron is also used to perform element cardinality checking.

The vocabularies or code lists are formalized in SKOS (Simple Knowledge Organization System), an RDF-based language. SKOS provides mechanisms to define concepts, associate them with (multi-lingual) labels or terms, and record thesaurus-like relationships between them. The vocabularies are managed in a RDF database or semantic triple store. Querying is implemented as a semantic vocabulary service, with an http-based API that allows queries to be issued from rules written in Schematron.

WDTF has required development and deployment of some ontologies whose scope is much more general than this application, in particular covering 'observed properties' and 'units of measure', which also have to be related to each other and consistent with the dimensional analysis.

Separation of the two validation passes reflects the separate governance and stability of the structural and content rules, and allows an organisation's business rules to be moved out of the XML schema definition and the XML schema to be reused by other businesses with their own specific rules.

With the general approach proven, harmonization opportunities with more generic services are being explored, such as the GEMET API for SKOS, developed by the European Environment Agency.

Acknowledgements:
The authors would like to thank the AUSCOPE team for their development and support provided of the vocabulary services.