Geophysical Research Abstracts
Vol. 17, EGU2015-7183, 2015
EGU General Assembly 2015

Technologies and practices for maintaining and publishing earth science vocabularies

Simon Cox (1), Jonathan Yu (1), Megan Williams (1), Fabrizio Giabardo (1), and Dominic Lowe (2)

(1) CSIRO, Land and Water, Melbourne, Australia (first.last@csiro.au), (2) Bureau of Meteorology, Canberra, Australia (d.lowe@bom.gov.au)

Shared vocabularies are a key element in geoscience data interoperability. Many organizations curate vocabularies, with most Geologic Surveys having a long history of development of lexicons and authority tables. However, their mode of publication is heterogeneous, ranging from PDFs and HTML web pages, spreadsheets and CSV, through various user-interfaces, and public and private APIs. Content maintenance ranges from tightly-governed and externally opaque, through various community processes, all the way to crowd-sourcing ('folksonomies'). Meanwhile, there is an increasing expectation of greater harmonization and vocabulary re-use, which create requirements for standardized content formalization and APIs, along with transparent content maintenance and versioning. We have been trialling a combination of processes and software dealing with vocabulary formalization, registration, search and linking.

We use the Simplified Knowledge Organization System (SKOS) to provide a generic interface to content. SKOS is an RDF technology for multi-lingual, hierarchical vocabularies, oriented around 'concepts' denoted by URIs, and thus consistent with Linked Open Data. SKOS may be mixed in with classes and properties from specialized ontologies which provide a more specific interface when required. We have developed a suite of practices and techniques for conversion of content from the source technologies and styles into SKOS, largely based on spreadsheet manipulation before RDF conversion, and SPARQL afterwards. The workflow for each vocabulary must be adapted to match the specific inputs.

In linked data applications, two requirements are paramount for user confidence: (i) the URI that denotes a vocabulary item is persistent, and should be dereferenceable indefinitely; (ii) the history and status of the resource denoted by a URI must be available. This is implemented by the Linked Data Registry (LDR), originally developed for the World Meteorological Organization and the UK Environment Agency, and now adapted and enhanced for deployment by CSIRO and the Australian Bureau of Meteorology. The LDR applies a standard content registration paradigm to RDF data, also including a delegation mode that enables a system to register (endorse) externally managed content. The locally managed RDF is exposed on a SPARQL endpoint. The registry implementation enables a flexible interaction pattern to support various specific content publication workflows, with the key feature of making the content externally accessible through a standard interface alongside its history, previous versions, and status.

SPARQL is the standard low-level API for RDF including SKOS. On top of this we have developed SISSvoc, a SKOS-based RESTful API. This has been used it to deploy a number of vocabularies on behalf of the IUGS, ICS, NERC, OGC, the Australian Government, and CSIRO projects. Applications like SISSvoc Search provide a simple search UI on top of one or more SISSvoc sources.

Together, these components provide a powerful and flexible system for providing earth science vocabularies for the community, consistent with semantic web and linked-data principles.