Geophysical Research Abstracts
Vol. 21, EGU2019-11615-2, 2019
EGU General Assembly 2019

Solving Real-World Interoperability Problems with Sustainable Semantic Web Technologies

Alex Ip (1), William Francis (1), Nicholas Car (2), and David Lescinsky (1)

(1) Geoscience Australia, Symonston, ACT, Australia (Alex.Ip@ga.gov.au), (2) CSIRO (Land & Water), Dutton Park, Queensland, Australia (Nicholas.Car@csiro.au)

Linked Data (LD) and related Semantic Web technologies have promised much but have proven difficult to deploy at scale. Despite this, LD remains one of the most viable means of delivering interoperable data across differing technology stacks and organisational boundaries. Geoscience Australia (GA) is an active participant in the Australian Government Linked Data Working Group, and is committed to delivering its extensive data holdings according to the FAIR data principles (Findable, Accessible, Interoperable, Reusable) using LD.

Instead of attempting a single, massive LD mega-deployment, GA has a strategy to meet specific operational needs by incrementally publishing subsets of its data holdings as LD. 2019 will be a pivotal year in GA's LD journey, with several projects underway using common architecture and components.

Entities which are being published by GA as LD include:

- **Vocabularies** - GA is moving to manage all of its codelists in a central triple-store as vocabularies with mappings between terms according to the SKOS ontology. This will be accessible via SPARQL endpoints, and used to populate database lookup tables and codelists in GA's metadata catalogue.
- **Samples** - GA's extensive geological sample metadata database is being published via an LD API with samples identified by International GeoSample Numbers.
- **Sites** - Metadata for GA's sampling sites is already presented as LD. These site URIs are being linked to samples and other resources such as site photographs.
- **Surveys** - Authoritative metadata for GA's geophysical surveys is being presented externally as LD. The API doing this is already used to synchronise metadata between GA's internal database, ISO19115 XML metadata records, and ACDD metadata encapsulated in netCDF datasets.
- **Stratigraphic Units** - GA's stratigraphic units and the relationships between them will be published as LD in 2019 to support machine-to-machine querying of geological knowledge. A primary driver of this is the knowledge management subsystem of the Loop3D subsurface modelling project.

GA and CSIRO are collaborating in the LocI "Location Identifiers" project, in which LD is used to publish authoritative spatial datasets and relationships between their data items across government. GA also works with Australian state government agencies to improve interoperability, including sample metadata alignment with Geological Survey of Victoria, and sites modelling alignment with the Geological Survey of Queensland.

Entities managed in GA's internal data repositories are dynamically presented as Linked Data using the pyLDAPI Python package currently being developed by CSIRO. pyLDAPI provides a flexible but lightweight mechanism for delivering LD responses, including human-friendly landing pages and machine-readable RDF. pyLDAPI can source data from relational databases, triple-stores and other custom APIs. Each LD web resource is accessed via a Persistent Identifier (PID) URI, which is referenced in triple-stores to express relationships between entities. GA is initially deploying operational triple-stores using Apache Jena/Fuseki hosted in the AWS public cloud.

Challenges in delivering LD have included poorly normalised relational schemas, incomplete data, and the need to maintain consistent and scalable PID schemes. LD is not a panacea for interoperability, nor is it easy to implement well. It does, however, provide a future-proof mechanism for making data FAIR.