Geophysical Research Abstracts
Vol. 21, EGU2019-11774, 2019
EGU General Assembly 2019

Addressing source to delivery for accessible and sustainable open data: government agency realities for Linked Data; policies, tools and technologies

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

(1) Geoscience Australia, Canberra, Australia (william.francis@ga.gov.au), (2) CSIRO Land & Water, Dutton Park, Australia (nicholas.car@csiro.au)

To deliver open data, government agencies must deal with legacy processes, both social and technical, that contain barriers to openness. These barriers limit the true usability of open data - how it can be used over time and in multiple contexts - and are critical to address as governments seek to expose open data.

Linked Data (LD) has always been, at its core, about ensuring the FAIR Data Principles (Findable, Accessible, Interoperable, Reusable) by focusing on the identity and relationship of entities and exposing their context to consumers of data, even if these principles have only recently been named FAIR. A fundamental component of LD is that entities are identified by sustainable URI references called Persistent Identifiers (PIDs) which retain their utility over time despite system and organisation change.

This poster will show how Geoscience Australia (GA) is applying the use of LD & PIDS in a real world, production IT setting. Long running operational processes have been incrementally advanced to deliver data from relational databases as LD.

Policies, practices and tools have developed and applied to support these LD delivery. The key components are:

- **Data transformation tools**: reliant on a robust internal data schema, the Corporate Data Model, these tools export views of it as XML or CSV publicly which is then converted to RDF in another step
- **Overarching data model**: a Semantic Web ontology that outlines the types of entities delivered publicly by GA and their macro relations. To date, public entities are Datasets, Web Services, vocabulary terms and geological Samples, Sites Surveys and Stratigraphic Units. New objects will include images with multiple formats and resolutions
- **PID service**: an application that manages a series of PID redirection rules
- **PID governance policy**: the defined process to support the agency with its multiple teams and their different data sources to have consistent application of entity identification rules and ensure uniqueness across multiple systems in the same registers
- **pyLDAPI data service tools**: a Web API tool that can present LD endpoints for entities according to given ontologies

Cloud infrastructure as code (infracode): Provisioning of LD data holding RDF triple stores on the public cloud following agency best practice in delivering scalable solutions. The tools used are Apache's Jena/Fuseki triplestore and API deployed on Amazon Web Services (AWS) with scalability through AWS Elastic Load Balancer and Elastic File Store components. Further work will explore suitability of the new triple store on AWS Neptune.