Geophysical Research Abstracts
Vol. 19, EGU2017-11151-1, 2017
EGU General Assembly 2017

NetCDF4/HDF5 and Linked Data in the Real World – Enriching Geoscientific Metadata without Bloat

Alex Ip (1), Nicholas Car (1), Kelsey Druken (2), Yvette Poudjom-Djomani (1), Stirling Butcher (1), Ben Evans (2), and Lesley Wyborn (2)

(1) Geoscience Australia, Symonston, ACT, Australia, (2) Australian National Computational Infrastructure, Acton, ACT, Australia

NetCDF4 has become the dominant generic format for many forms of geoscientific data, leveraging (and constraining) the versatile HDF5 container format, while providing metadata conventions for interoperability. However, the encapsulation of detailed metadata within each file can lead to metadata "bloat", and difficulty in maintaining consistency where metadata is replicated to multiple locations. Complex conceptual relationships are also difficult to represent in simple key-value netCDF metadata. Linked Data provides a practical mechanism to address these issues by associating the netCDF files and their internal variables with complex metadata stored in Semantic Web vocabularies and ontologies, while complying with and complementing existing metadata conventions.

One of the stated objectives of the netCDF4/HDF5 formats is that they should be self-describing: containing metadata sufficient for cataloguing and using the data. However, this objective can be regarded as only partially-met where details of conventions and definitions are maintained externally to the data files. For example, one of the most widely used netCDF community standards, the Climate and Forecasting (CF) Metadata Convention, maintains standard vocabularies for a broad range of disciplines across the geosciences, but this metadata is currently neither readily discoverable nor machine-readable.

We have previously implemented useful Linked Data and netCDF tooling (ncskos) that associates netCDF files, and individual variables within those files, with concepts in vocabularies formulated using the Simple Knowledge Organization System (SKOS) ontology. NetCDF files contain Uniform Resource Identifier (URI) links to terms represented as SKOS Concepts, rather than plain-text representations of those terms, so we can use simple, standardised web queries to collect and use rich metadata for the terms from any Linked Data-presented SKOS vocabulary.

Geoscience Australia (GA) manages a large volume of diverse geoscientific data, much of which is being translated from proprietary formats to netCDF at NCI Australia. This data is made available through the NCI National Environmental Research Data Interoperability Platform (NERDIP) for programmatic access and interdisciplinary analysis. The netCDF files contain both scientific data variables (e.g. gravity, magnetic or radiometric values), but also domain-specific operational values (e.g. specific instrument parameters) best described fully in formal vocabularies. Our ncskos codebase provides access to multiple stores of detailed external metadata in a standardised fashion.

Geophysical datasets are generated from a "survey" event, and GA maintains corporate databases of all surveys and their associated metadata. It is impractical to replicate the full source survey metadata into each netCDF dataset so, instead, we link the netCDF files to survey metadata using public Linked Data URIs. These URIs link to Survey class objects which we model as a subclass of Activity objects as defined by the PROV Ontology, and we provide URI resolution for them via a custom Linked Data API which draws current survey metadata from GA's in-house databases.

We have demonstrated that Linked Data is a practical way to associate netCDF data with detailed, external metadata. This allows us to ensure that catalogued metadata is kept consistent with metadata points-of-truth, and we can infer complex conceptual relationships not possible with netCDF key-value attributes alone.