Geophysical Research Abstracts
Vol. 18, EGU2016-15650, 2016
EGU General Assembly 2016

Using a Simple Knowledge Organization System to facilitate Catalogue and Search for the ESA CCI Open Data Portal

Antony Wilson (1), Victoria Bennett (2), Steve Donegan (3), Martin Juckes (3), Philip Kershaw (2), Ruth Petrie (3), Ag Stephens (3), and Alison Waterfall (3)

(1) Science and Technology Facilities Council, Scientific Computing, Didcot, United Kingdom, (2) STFC Rutherford Appleton Laboratory, NCEO/Centre for Environmental Data Archival, Didcot, United Kingdom, (3) STFC Rutherford Appleton Laboratory, NCAS/Centre for Environmental Data Archival, Didcot, United Kingdom

The ESA Climate Change Initiative (CCI) is a €5m programme that runs from 2009-2016, with a goal to provide stable, long-term, satellite-based essential climate variable (ECV) data products for climate modellers and researchers. As part of the CCI, ESA have funded the Open Data Portal project to establish a central repository to bring together the data from these multiple sources and make it available in a consistent way, in order to maximise its dissemination amongst the international user community.

Search capabilities are a critical component to attaining this goal. To this end, the project is providing dataset-level metadata in the form of ISO 19115 records served via a standard OGC CSW interface. In addition, the Open Data Portal is re-using the search system from the Earth System Grid Federation (ESGF), successfully applied to support CMIP5 (5th Coupled Model Intercomparison Project) and obs4MIPs. This uses a tightly defined controlled vocabulary of metadata terms, the DRS (The Data Reference Syntax) which encompass different aspects of the data. This system hs facilitated the construction of a powerful faceted search interface to enable users to discover data at the individual file level of granularity through ESGF's web portal frontend.

The use of a consistent set of model experiments for CMIP5 allowed the definition of a uniform DRS for all model data served from ESGF. For CCI however, there are thirteen ECVs, each of which is derived from multiple sources and different science communities resulting in highly heterogeneous metadata. An analysis has been undertaken of the concepts in use, with the aim to produce a CCI DRS which could be provide a single authoritative source for cataloguing and searching the CCI data for the Open Data Portal. The use of SKOS (Simple Knowledge Organization System) and OWL (Web Ontology Language) to represent the DRS are a natural fit and provide controlled vocabularies as well as a way to represent relationships between similar terms used in different ECVs.

An iterative approach has been adopted for the model development working closely with domain experts and drawing on practical experience working with content in the input datasets. Tooling has been developed to enable the definition of vocabulary terms via a simple spreadsheet format which can then be automatically converted into Turtle notation and uploaded to the CCI DRS vocabulary service. With a baseline model established, work is underway to develop an ingestion pipeline to import validated search metadata into the ESGF and OGC CSW search services. In addition to the search terms indexed into the ESGF search system, ISO 19115 records will also be similarly tagged during this process with search terms from the data model. In this way it will be possible to construct a faceted search user interface for the Portal which can yield linked search results for data both at the file and dataset level granularity. It is hoped that this will also provide a rich range of content for third-party organisations wishing to incorporate access to CCI data in their own applications and services.