Geophysical Research Abstracts
Vol. 19, EGU2017-10103, 2017
EGU General Assembly 2017

Combined use of semantics and metadata to manage Research Data Life Cycle in Environmental Sciences

Fernando Aguilar Gómez, Jesús Marco de Lucas, Esther Pertinez, and Aida Palacio
IFCA-CSIC, Santander, Spain (aguilarf@ifca.unican.es)

The use of metadata to contextualize datasets is quite extended in Earth System Sciences. There are some initiatives and available tools to help data managers to choose the best metadata standard that fit their use cases, like the DCC Metadata Directory (http://www.dcc.ac.uk/resources/metadata-standards).
In our use case, we have been gathering physical, chemical and biological data from a water reservoir since 2010. A well metadata definition is crucial not only to contextualize our own data but also to integrate datasets from other sources like satellites or meteorological agencies. That is why we have chosen EML (Ecological Metadata Language), which integrates many different elements to define a dataset, including the project context, instrumentation and parameters definition, and the software used to process, provide quality controls and include the publication details. Those metadata elements can contribute to help both human and machines to understand and process the dataset.

However, the use of metadata is not enough to fully support the data life cycle, from the Data Management Plan definition to the Publication and Re-use. To do so, we need to define not only metadata and attributes but also the relationships between them, so semantics are needed. Ontologies, being a knowledge representation, can contribute to define the elements of a research data life cycle, including DMP, datasets, software, etc. They also can define how the different elements are related between them and how they interact. The first advantage of developing an ontology of a knowledge domain is that they provide a common vocabulary hierarchy (i.e. a conceptual schema) that can be used and standardized by all the agents interested in the domain (either humans or machines). This way of using ontologies is one of the basis of the Semantic Web, where ontologies are set to play a key role in establishing a common terminology between agents.

To develop an ontology we are using a graphical tool Protégé, which is a graphical ontology-development tool that supports a rich knowledge model and it is open-source and freely available. To process and manage the ontology, we are using Semantic MediaWiki, which is able to process queries. Semantic MediaWiki is an extension of MediaWiki where we can do semantic search and export data in RDF.

Our final goal is integrating our data repository portal and semantic processing engine in order to have a complete system to manage the data life cycle stages and their relationships, including machine-actionable DMP solution, datasets and software management, computing resources for processing and analysis and publication features (DOI mint). This way we will be able to reproduce the full data life cycle chain warranting the FAIR+R principles.