Geophysical Research Abstracts
Vol. 15, EGU2013-6894, 2013
EGU General Assembly 2013

Multi-facetted Metadata - Describing datasets with different metadata schemas at the same time

Damian Ulbricht, Jens Klump, and Roland Bertelmann
Deutsches GeoForschungsZentrum GFZ, Potsdam, Germany ({ulbricht, jens.klump, rab}@gfz-potsdam.de)

Inspired by the wish to re-use research data a lot of work is done to bring data systems of the earth sciences together. Discovery metadata is disseminated to data portals to allow building of customized indexes of catalogued dataset items. Data that were once acquired in the context of a scientific project are open for reappraisal and can now be used by scientists that were not part of the original research team. To make data re-use easier, measurement methods and measurement parameters must be documented in an application metadata schema and described in a written publication. Linking datasets to publications – as DataCite [1] does – requires again a specific metadata schema and every new use context of the measured data may require yet another metadata schema sharing only a subset of information with the meta information already present.

To cope with the problem of metadata schema diversity in our common data repository at GFZ Potsdam we established a solution to store file-based research data and describe these with an arbitrary number of metadata schemas. Core component of the data repository is an eSciDoc infrastructure that provides versioned container objects, called eSciDoc [2] “items”. The eSciDoc content model allows assigning files to “items” and adding any number of metadata records to these “items”. The eSciDoc items can be submitted, revised, and finally published, which makes the data and metadata available through the internet worldwide. GFZ Potsdam uses eSciDoc to support its scientific publishing workflow, including mechanisms for data review in peer review processes by providing temporary web links for external reviewers that do not have credentials to access the data.

Based on the eSciDoc API, panMetaDocs [3] provides a web portal for data management in research projects. PanMetaDocs, which is based on panMetaWorks [4], is a PHP based web application that allows to describe data with any XML-based schema. It uses the eSciDoc infrastructures REST-interface to store versioned dataset files and metadata in a XML-format. The software is able to administrate more than one eSciDoc metadata record per item and thus allows the description of a dataset according to its context. The metadata fields can be filled with static or dynamic content to reduce the number of fields that require manual entries to a minimum and, at the same time, make use of contextual information available in a project setting. Access rights can be adjusted to set visibility of datasets to the required degree of openness. Metadata from separate instances of panMetaDocs can be syndicated to portals through RSS and OAI-PMH interfaces.

The application architecture presented here allows storing file-based datasets and describe these datasets with any number of metadata schemas, depending on the intended use case. Data and metadata are stored in the same entity (eSciDoc items) and are managed by a software tool through the eSciDoc REST interface – in this case the application is panMetaDocs. Other software may re-use the produced items and modify the appropriate metadata records by accessing the web API of the eSciDoc data infrastructure. For presentation of the datasets in a web browser we are not bound to panMetaDocs. This is done by stylesheet transformation of the eSciDoc-item.

[1] http://www.datacite.org
[2] http://www.escidoc.org , eSciDoc, FIZ Karlruhe, Germany
[3] http://panmetadocs.sf.net , panMetaDocs, GFZ Potsdam, Germany
[4] http://metaworks.pangaea.de , panMetaWorks, Dr. R. Huber, MARUM, Univ. Bremen, Germany