Geophysical Research Abstracts
Vol. 18, EGU2016-13253, 2016
EGU General Assembly 2016

panMetaDocs, eSciDoc, and DOIDB – an infrastructure for the curation and publication of file-based datasets for 'GFZ Data Services'

Damian Ulbricht (1), Kirsten Elger (1), Roland Bertelmann (1), and Jens Klump (2)

(1) GFZ German Research Centre for Geosciences, Potsdam, Germany, (2) Commonwealth Scientific and Industrial Research Organisation, Mineral Resources Flagship, Kensington, Australia

With the foundation of DataCite in 2009 and the technical infrastructure installed in the last six years it has become very easy to create citable dataset DOIs. Nowadays, dataset DOIs are increasingly accepted and required by journals in reference lists of manuscripts. In addition, DataCite provides usage statistics [1] of assigned DOIs and offers a public search API to make research data count. By linking related information to the data, they become more useful for future generations of scientists. For this purpose, several identifier systems, as ISBN for books, ISSN for journals, DOI for articles or related data, Orcid for authors, and IGSN for physical samples can be attached to DOIs using the DataCite metadata schema [2].
While these are good preconditions to publish data, free and open solutions that help with the curation of data, the publication of research data, and the assignment of DOIs in one software seem to be rare.
At GFZ Potsdam we built a modular software stack that is made of several free and open software solutions and we established 'GFZ Data Services'. 'GFZ Data Services' provides storage, a metadata editor for publication and a facility to moderate minted DOIs. All software solutions are connected through web APIs, which makes it possible to reuse and integrate established software.
Core component of 'GFZ Data Services' is an eSciDoc [3] middleware that is used as central storage, and has been designed along the OAIS reference model for digital preservation. Thus, data are stored in self-contained packages that are made of binary file-based data and XML-based metadata. The eSciDoc infrastructure provides access control to data and it is able to handle half-open datasets, which is useful in embargo situations when a subset of the research data are released after an adequate period.
The data exchange platform panMetaDocs [4] makes use of eSciDoc's REST API to upload file-based data into eSciDoc and uses a metadata editor [5] to annotate the files with metadata. The metadata editor has a user-friendly interface with nominal lists, extensive explanations, and an interactive mapping tool to provide assistance to scientists describing the data. It is possible to deposit metadata templates to fill certain fields with default values. The metadata editor generates metadata in the schemas ISO19139, NASA GCMD DIF, and DataCite and could be extended for other schemas.
panMetaDocs is able to mint dataset DOIs through DOIDB, which is our component to moderate dataset DOIs issued through 'GFZ Data Services'. DOIDB accepts metadata in the schemas ISO19139, DIF, and DataCite. In addition, DOIDB provides an OAI-PMH interface to disseminate all deposited metadata to data portals.
The presentation of datasets on DOI landing pages is done though XSLT stylesheet transformation of the XML-based metadata. The landing pages have been designed to meet needs of scientists. We are able to render the metadata to different layouts. Furthermore, additional information about datasets and publications is assembled into the webpage by querying public databases on the internet.
The work presented here will focus on technical details of the software stack.

[1] http://stats.datacite.org
[2] http://www.dlib.org/dlib/january11/starr/01starr.html
[3] http://www.escidoc.org
[4] http://panmetadocs.sf.net
[5] http://github.com/ulbricht