Publishing datasets with eSciDoc and panMetaDocs

D. Ulbricht, J. Klump, and R. Bertelmann

German Research Centre for Geosciences, Potsdam, Germany (damian.ulbricht@gfz-potsdam.de)

Currently serveral research institutions worldwide undertake considerable efforts to have their scientific datasets published and to syndicate them to data portals as extensively described objects identified by a persistent identifier. This is done to foster the reuse of data, to make scientific work more transparent, and to create a citable entity that can be referenced unambigously in written publications.

GFZ Potsdam established a publishing workflow for file based research datasets. Key software components are an eSciDoc infrastructure [1] and multiple instances of the data curation tool panMetaDocs [2]. The eSciDoc repository holds data objects and their associated metadata in container objects, called eSciDoc items. A key metadata element in this context is the publication status of the referenced data set.

PanMetaDocs, which is based on PanMetaWorks [3], is a PHP based web application that allows to describe data with any XML-based metadata schema. 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 make use of contextual information in a project setting. Access rights can be applied to set visibility of datasets to other project members and allow collaboration on and notifying about datasets (RSS) and interaction with the internal messaging system, that was inherited from panMetaWorks. When a dataset is to be published, panMetaDocs allows to change the publication status of the eSciDoc item from status „private" to „submitted" and prepare the dataset for verification by an external reviewer. After quality checks, the item publication status can be changed to „published". This makes the data and metadata available through the internet worldwide.

PanMetaDocs is developed as an eSciDoc application. It is an easy to use graphical user interface to eSci-Doc items, their data and metadata. It is also an application supporting a DOI publication agent during the process of publishing scientific datasets as electronic data supplements to research papers.

Publication of research manuscripts has an already well established workflow that shares junctures with other processes and involves several parties in the process of dataset publication. Activities of the author, the reviewer, the print publisher and the data publisher have to be coordinated into a common data publication workflow. The case of data publication at GFZ Potsdam displays some specifics, e.g. the DOIDB webservice. The DOIDB is a proxy service at GFZ for the DataCite [4] DOI registration and its metadata store. DOIDB provides a local summary of the dataset DOIs registered through GFZ as a publication agent. An additional use case for the DOIDB is its function to enrich the datacite metadata with additional custom attributes, like a geographic reference in a DIF record. These attributes are at the moment not available in the datacite metadata schema but would be valuable elements for the compilation of data catalogues in the earth sciences and for dissemination of catalogue data via OAI-PMH.

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