Geophysical Research Abstracts
Vol. 16, EGU2014-9710, 2014
EGU General Assembly 2014

Creating preservation metadata from XML-metadata profiles

Damian Ulbricht (1), Roland Bertelmann (1), Petra Gebauer (2), Tim Hasler (3), Jens Klump (1), Ingo Kirchner (2), Wolfgang Peters-Kottig (3), Nora Mettig (2), and Beate Rusch (3)

(1) GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany, (2) Institute for Meteorology, Free University Berlin, Berlin, Germany, (3) Zuse Institute, Berlin, Germany

Registration of dataset DOIs at DataCite makes research data citable and comes with the obligation to keep data accessible in the future. In addition, many universities and research institutions measure data that is unique and not repeatable like the data produced by an observational network and they want to keep these data for future generations. In consequence, such data should be ingested in preservation systems, that automatically care for file format changes. Open source preservation software that is developed along the definitions of the ISO OAIS reference model is available but during ingest of data and metadata there are still problems to be solved. File format validation is difficult, because format validators are not only remarkably slow - due to variety in file formats different validators return conflicting identification profiles for identical data. These conflicts are hard to resolve. Preservation systems have a deficit in the support of custom metadata. Furthermore, data producers are sometimes not aware that quality metadata is a key issue for the re-use of data.

In the project EWIG an university institute and a research institute work together with Zuse-Institute Berlin, that is acting as an infrastructure facility, to generate exemplary workflows for research data into OAIS compliant archives with emphasis on the geosciences. The Institute for Meteorology provides timeseries data from an urban monitoring network whereas GFZ Potsdam delivers file based data from research projects. To identify problems in existing preservation workflows the technical work is complemented by interviews with data practitioners. Policies for handling data and metadata are developed. Furthermore, university teaching material is created to raise the future scientists awareness of research data management.

As a testbed for ingest workflows the digital preservation system Archivematica [1] is used. During the ingest process metadata is generated that is compliant to the Metadata Encoding and Transmission Standard (METS). To find datasets in future portals and to make use of this data in own scientific work, proper selection of discovery metadata and application metadata is very important. Some XML-metadata profiles are not suitable for preservation, because version changes are very fast and make it nearly impossible to automate the migration. For other XML-metadata profiles schema definitions are changed after publication of the profile or the schema definitions become inaccessible, which might cause problems during validation of the metadata inside the preservation system [2]. Some metadata profiles are not used widely enough and might not even exist in the future. Eventually, discovery and application metadata have to be embedded into the mdWrap-subtree of the METS-XML.

[1] http://www.archivematica.org
[2] http://dx.doi.org/10.2218/ijdc.v7i1.215