Geophysical Research Abstracts
Vol. 15, EGU2013-5475, 2013
EGU General Assembly 2013

Application of Handles in the European Data Project EUDAT

Frank Toussaint (1), Martina Stockhause (1,2), Tobias Weigel (1,3), Heinke Höck (1), and Michael Lautenschlager (1)

(1) Deutsches Klimarechenzentrum, World Data Centre for Climate, Hamburg, Germany, (2) Max Planck Institute of Meteorology, Hamburg, Germany, (3) University of Hamburg, Hamburg, Germany

Increasing quantities of data lead more and more to automation of data handling. This needs to be closely linked to automated metadata handling, as well. In EUDAT (European Data Infrastructure), a European project for interdisciplinary, collaborative data infrastructures, Handles will serve as persistent identifiers (PIDs) to keep track of data and metadata, of predecessors and successors, and of datasets and their subsets. On an international level, the Research Data Alliance (RDA) aims to facilitate data-driven innovation by the development and adoption of, but not limited to, infrastructure, policy, practice, and standards for PIDs and other infrastructure features.

Many of a data object's metadata are kept close to the data. They mostly belong to the socalled use metadata like, e.g., descriptions of formats and coordinate systems. However, often there are separate data objects, containing more general metadata. These discovery metadata may comprise general information on content or data producer, references to papers or software, and so on. To keep the connection between data and metadata object, both can be linked by mutual pointers held in Handles. This requires, of course, that PIDs are assigned to both, data object and metadata object.

This concept strongly differs from the classical approach in which the data storage location is kept in the metadata as one attribute of many. In the former case, pairs of PIDs for data and metadata can support an infrastructure of services for data handling and processing in the full data lifecycle – an approach which is not only advantageous if the bulk of the data is stored on tapes.

In EUDAT processes of creation, movement, and deletion of data objects and their replicas will be tracked and guided by use of PIDs. In the project it was decided that all data objects handled need to be registered, i.e. they must have a PID. Wherever a PID is accompanied by storage information, this will be updated automatically via, e.g., a web service when the data is moved or replicated. Technically, the data replication in EUDAT will be based on IRODS, the Integrated Rule Oriented Data System.

One of the main advantages of using PIDs instead of classical MD is that it is easier to maintain up-to-date the metadata carried by a PID. Standardised web interfaces can help to log creation, transmission, replication, and deletion of data objects in the PID metadata. In any case, the structural information held with the PID (provenance, data to metadata relation, subsets) is kept and the pointing addresses will be changed accordingly.