Geophysical Research Abstracts
Vol. 15, EGU2013-4254-1, 2013
EGU General Assembly 2013

Concept for Future Data Services at the Long-Term Archive of WDCC combining DOIs with common PIDs

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

(1) WDC Climate / DKRZ, Hamburg, Germany (stockhause@dkrz.de), (2) Max Planck Institute for Meteorology (MPI-M), Hamburg, Germany, (3) University of Hamburg, Germany

The World Data Center for Climate (WDCC) hosted at the German Climate Computing Center (DKRZ) maintains a long-term archive (LTA) of climate model data as well as observational data. WDCC distinguishes between two types of LTA data:

1. **Structured data:** Data output of an instrument or of a climate model run consists of numerous, highly structured individual datasets in a uniform format. Part of these data is also published on an ESGF (Earth System Grid Federation) data node. Detailed metadata is available allowing for fine-grained user-defined data access.
2. **Unstructured data:** LTA data of finished scientific projects are in general unstructured and consist of datasets of different formats, different sizes, and different contents. For these data compact metadata is available as content information.

The structured data is suitable for WDCC's DataCite DOI process, the project data only in exceptional cases. The DOI process includes a thorough quality control process of technical as well as scientific aspects by the publication agent and the data creator. DOIs are assigned to data collections appropriate to be cited in scientific publications, like a simulation run. The data collection is defined in agreement with the data creator.

At the moment there is no possibility to identify and cite individual datasets within this DOI data collection analogous to the citation of chapters in a book. Also missing is a compact citation regulation for a user-specified collection of data.

WDCC therefore complements its existing LTA/DOI concept by Persistent Identifier (PID) assignment to datasets using Handles. In addition to data identification for internal and external use, the concept of PIDs allows to define relations among PIDs. Such structural information is stored as key-value pair directly in the handles. Thus, relations provide basic provenance or lineage information, even if part of the data like intermediate results are lost. WDCC intends to use additional PIDs on metadata entities with a relation to the data PID(s). These add background information on the data creation process (e.g. descriptions of experiment, model, model set-up, and platform for the model run etc.) to the data. These pieces of additional information increase the re-usability of the archived model data, significantly. Other valuable additional information for scientific collaboration could be added by the same mechanism, like quality information and annotations.

Apart from relations among data and metadata entities, PIDs on collections are advantageous for model data:

- Collections allow for persistent references to single datasets or subsets of data assigned a DOI,
- Data objects and additional information objects can be consistently connected via relations (provenance, creation, quality information for data),
- PID Collections enable identification and citation of any groups of data, e.g. unstructured project data or all files belonging to a time-series of a variable or all datasets used for a scientific publication.

WDCC/DKRZ participates in the recently founded Research Data Alliance (RDA) initiative.

References:
DataCite: http://datacite.org

DKRZ: http://www.dkrz.de
ESGF: http://www.esgf.org
RDA: http://rd-alliance.org
WDCC: http://www.wdc-climate.de