Geophysical Research Abstracts
Vol. 15, EGU2013-8371, 2013
EGU General Assembly 2013

Concepts for PID Collections and Application to the Data Life Cycle

Tobias Weigel (1,2), Martina Stockhause (1,3), Frank Toussaint (1), Stephan Kindermann (1), and Michael Lautenschlager (1)

(1) Deutsches Klimarechenzentrum (DKRZ), Hamburg, Germany (weigel@dkrz.de), (2) University of Hamburg, Germany, (3) Max Planck Institute for Meteorology, Hamburg, Germany

Persistent Identifiers (PIDs) are here understood as worldwide unique IDs which can be submitted to a global resolution system to retrieve an arbitrary number of typed key-value pairs and referenced Digital Objects. Such lightweight PIDs intentionally differ from DOIs in that they do not mandate quality control nor do they guarantee persistence of referenced objects. Lately, the concepts of PIDs and Digital Objects have become the focus of international collaborative work, formed under the umbrella of the newly founded Research Data Alliance (RDA).

Using PIDs at a broad operational perspective for research data will require a number of structural elements to keep hold of relations between different objects and efficiently manage large amounts of PIDs. Management tasks are relevant for example in the EUDAT project and the DKRZ long-term archive. PID collections that are fully encoded and persistent within the PID framework form one promising approach to not only facilitate automatic data management but also enable a large number of downstream use cases.

In this poster, we present the fundamental concepts of Persistent Identifiers (PID) collections and provide examples taken from the generic data life cycle that illustrate how collections may be used. While the abstract concept is clear, we argue that there is no single best implementation of a collection for operational purposes. We rather use exemplary points in the data lifecycle to show how each detail use case requires a different trade-off between flexibility and computational complexity of the particular collection implementation. One key implementation of collections is formed through small tuples which can be instantiated in massive amounts to hold data and metadata objects together. This is particularly important for Earth science data management tasks. Another implementation form targets more individual, user-customized collections which may have a much more diverse number of elements and sacrifice low complexity in favor of increased flexibility.

With multiple possible implementations at hand, interoperability problems ensue. These can be dealt with at the level of broad collaboration and standardization e.g. through the RDA and by providing migration pathways between implementations. Migration should however be avoided in the first place because one major objective of PIDs is that they require as low a maintenance effort as possible.

Establishing information structures like collections for intermediate results in the data lifecycle ultimately contributes to a basic layer of provenance information, potentially covering the whole cycle from early computational results to dissemination and long-term archival. This basic, intentionally limited layer should be enriched by more sophisticated solutions. The relevant trade-off in this respect deals with how these structures are largely kept ignorant of upper layer functionality while nonetheless the necessary elements at the lower level are sufficient for managing scientific data. Collections are in this respect only one piece of a larger provenance graph, represented by persistently encoded relations between individual PIDs.