Geophysical Research Abstracts
Vol. 16, EGU2014-3739, 2014
EGU General Assembly 2014

Persistent Identifiers in Earth science data management environments

Tobias Weigel (1,2), Martina Stockhause (1,3), and Michael Lautenschlager (1)

(1) German Climate Computing Center, Hamburg, Germany (weigel@dkrz.de), (2) Universität Hamburg, Hamburg, Germany, (3) Max Planck Institute for Meteorology, Hamburg, Germany

Globally resolvable Persistent Identifiers (PIDs) that carry additional context information (which can be any form of metadata) are increasingly used by data management infrastructures for fundamental tasks. The notion of a Persistent Identifier is originally an abstract concept that aims to provide identifiers that are quality-controlled and maintained beyond the life time of the original issuer, for example through the use of redirection mechanisms. Popular implementations of the PID concept are for example the Handle System and the DOI System based on it. These systems also move beyond the simple identification concept by providing facilities that can hold additional context information. Not only in the Earth sciences, data managers are increasingly attracted to PIDs because of the opportunities these facilities provide; however, long-term viable principles and mechanisms for efficient organization of PIDs and context information are not yet available or well established. In this respect, promising techniques are to type the information that is associated with PIDs and to construct actionable collections of PIDs.
There are two main drivers for extended PID usage: Earth science data management middleware use cases and applications geared towards scientific end-users. Motivating scenarios from data management include hierarchical data and metadata management, consistent data tracking and improvements in the accountability of processes. If PIDs are consistently assigned to data objects, context information can be carried over to subsequent data life cycle stages much easier. This can also ease data migration from one major curation domain to another, e.g. from early dissemination within research communities to formal publication and long-term archival stages, and it can help to document processes across technical and organizational boundaries.
For scientific end users, application scenarios include for example more personalized data citation and improvements in the amount of context available for unfamiliar datasets. We can see how Earth system model data is spatially and temporally transformed to better fit the differing scenarios relevant in consecutive life cycle stages. At the end, users often want to cite and use distinct subsets of data which are disseminated through e-science infrastructures. If actionable collections of fine-granular PIDs are available, much more precise citation and use can be supported. This can also help to establish interoperable input and output references for processing tasks during intermediate life cycle stages.
The current working draft API of the Research Data Alliance's working group on PID Information Types combined with more elaborate collection mechanisms can provide the necessary foundations and tools to enable wide-spread use of PIDs for data life cycle management and user applications. This contribution will highlight some of the available mechanisms and existing efforts with particular focus on applications for institutional data management and e-science infrastructures such as the Earth System Grid Federation.