

Use of Persistent Identifiers in the Publication and Citation of Scientific Data

Jens Klump (1), Jan Brase (2), Michael Diepenbroek (3), Hannes Grobe (4), Beate Hildenbrandt (5), Heinke Höck (6), Michael Lautenschlager (6), and Irina Sens (2)

- (1) German Research Centre for Geoscience (GFZ), Potsdam, Germany
- (2) National Library of Science and Technology (TIB), Hannover, Germany
- (3) MARUM, University of Bremen, Bremen, Germany
- (4) Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven, Germany
- (5) DLR German Remote Sensing Data Center, Oberpfaffenhofen, Germany
- (6) Max Planck Institute for Meteorology, Hamburg, Germany

In the last decade the primary data, research is based on has become a third pillar of scientific work alongside with theoretical reasoning and experiment. Greatly increased computing power and storage, together with web services and other electronic resources have facilitated a quantum leap in new research based on the analysis of great amounts of data. However, traditional scientific communication only slowly changes to new media other than an emulation of paper. This leaves many data inaccessible and, in the long run exposes valuable data to the risk of loss.

Most important to the availability of data is a valid citation. This means that all fields mandatory for a bibliographic citation are included. In addition a mechanism is needed that ensures that the location of the referenced data on the Internet can be resolved on a long-term. Just using URLs by doing "data management on web servers" does not help at all because it is short-lived, mostly becoming invalid after just a few months. Data publication on the Internet therefore needs a system of reliable pointers to each digital object as integral part of the citation. To achieve this persistence of identifiers for their conventional publications many scientific publishers use Digital Object Identifiers (DOI). The identifier is resolved through the handle system to the valid location (URL) where the dataset can be found. This approach meets one of the prerequisites for citeability of scientific data published online. In addition, the valid bibliographic citation can be included in the catalogues of Libraries.

To improve access to data and to create incentives for scientists to make their data accessible, some german data centers initiated a project on publication and citation of scientific data.

The project "Publication and Citation of Scientific Data" (STD-DOI) was funded by the German Science Foundation (DFG) between 2003 and 2008. In STD-DOI the German National Library for Science and Technology (TIB Hannover), together with the German Research Centre for Geoscience (GFZ Potsdam), the Alfred Wegener Institute for Polar and Marine Research (AWI) Bremerhaven, the University of Bremen, the Max Planck Institute for Meteorology in Hamburg, and the DLR German Remote Sensing Data Center set up the first system to assign DOIs to data sets and finally to its publications.

The STD-DOI system for data publication is now used by eight data publication agents. Data publication through specific agents addresses specific user communities and cater for their requirements in the data publication process. The registration process between TIB and the publication agents is based on a SOAP web service.

This presentation will show the organisational and technical aspects of the data publication process through the STD-DOI project and give examples of a successful workflow towards established data citations in the earth sciences.