Citability of scientific primary data - New ways for access techniques

Jan Brase[1]
Michael Diepenbroek[3], Hannes Grobe[4], Heinke Höck[5], Jens Klump[6], Michael Lautenschlager[5], Uwe Schindler[3] and Irina Sens[2]

[1] Research center L3S
University of Hannover
Expo Plaza 1
30539 Hannover
brase@l3s.de

[2] German national library of science and technology
Hannover

[3] World Data Center for Marine Environmental Sciences
MARUM, University of Bremen

[4] Alfred-Wegener-Institut für Polar- und Meeresforschung, Bremerhaven

[5] World Data Center Climate
Max-Planck-Institut für Meteorologie, Hamburg

[6] GeoForschungsZentrum Potsdam

1 Motivation

Registration of scientific primary data, to make these data citable as a unique piece of work and not only as part of a publication, is becoming a more and more important issue. In its 2004 report of the CSPR assessment panel, the *International Council for Science* (ICSU) strongly recommended a new strategic framework for scientific data and information. In the context of the project "Publication and Citation of Scientific Primary Data" founded by the German research foundation (DFG) the German national library of science and technology (TIB) has become the first registration agency worldwide for scientific primary data.

2 Scenario

The data are still stored at the author's local data center or research institution, where the responsibility for valuating and maintaining of the data still lies. In addition to the local data preparation the research institutions transmit the URL where the data can be accessed to the TIB, together with a XML-file containing all relevant bibliographical metadata.

The TIB stores this information about the primary data and awards the primary data with a *Digital object identifier* (DOI) as unique identifier for registration. Any scientist working with this data is now able to cite these data in his work by its DOI. A citation will for example look like: *Miller, 2003: Temperature in Hannover for*

2003.[doi: 10.1594/WDCC/W_Han_2003_MMB_2]
By this, scientific primary data is not exclusively understood as part of a scientific publication, but has its own identity. All bibliographical information about the data is now accessible through the online library catalogue of the TIB. The entry is displayed with all relevant metadata and its identifier as a link to access the dataset itself (see fig. 1).

Due to the expected large amount of datasets that need to be registered, we have decided to distinguish between *citable datasets* on the collection level and *core datasets* on the item level. Core datasets receive their identifiers, but their metadata is not included in the library catalogue. The DOI guarantees the accessibility of this data to refer it inside a publication for example. Only citable datasets, usually collections of, or publications from core dataset will be included in the catalogue.

3 Realization

For this task, we have developed a web service infrastructure at the TIB, which is able to be accessed from the middleware of any involved research institution. A core dataset can be registered in about 4 seconds, including error handling.

Citable datasets, usually collections of, or publications from core dataset will be included in the catalogue. To register citable datasets, we have built a web interface on top of the web service (see fig. 2) This interface receives XML files from the data providers, and starts the registration process:

- The DOI is registered via a java based transmission to the DOI foundation.
- The metadata has to be transformed to PICA format and uploaded on a ftp server at the central library database.

Registration has started for the field of earth science, but will include other scientific disciplines in future. We have registered 30 citable and 150,000 core datasets so far (March 2005). We expect an amount of approximately 500,000 datasets to be registered by the TIB until the end of 2005. The registration of primary data will be widened to other science fields in 2006.

The possibility of citing primary data as a unique piece of work and not only a part of a publication opens new frontiers to the publication of scientific work itself and to the work of the TIB. For a medium amount of time, the availability of high-class data respectively content can be assured and may therefore significantly contribute to the success of "eScience".

Fig. 1. A dataset as a query result in the library catalogue

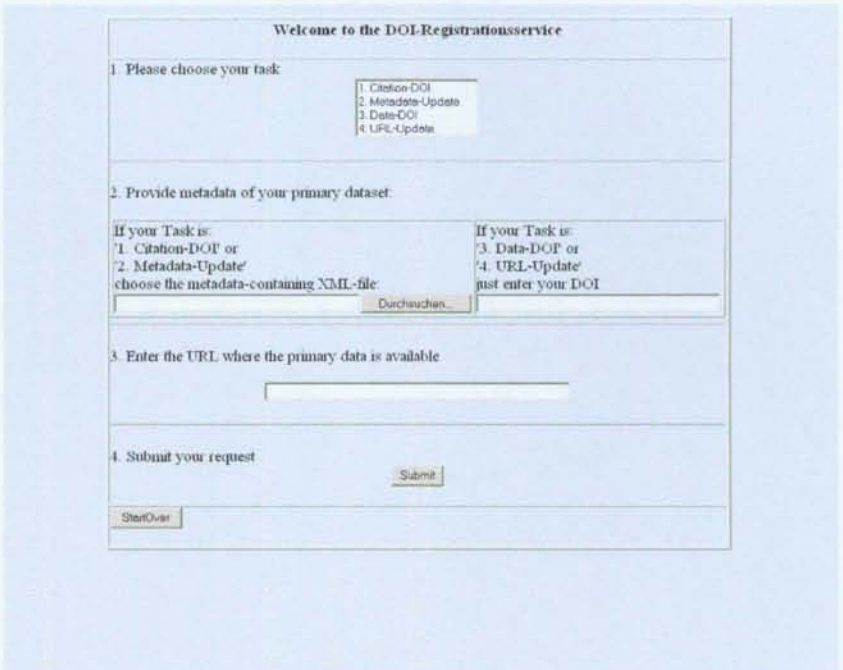

Fig. 2. Screenshot of the web interface to register datasets