

Standards for the publication of scientific data by World Data Centres and the National Library of Science and Technology in Germany

Jan Brase

Research center L3S, University of Hannover

Michal Diepenbroek,

World Data Center for Marine Environmental Sciences (WDC-MARE), Bremen

Hannes Grobe

Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven

Heinke Höck

World Data Center Climate, Hamburg

Jens Klump

Geoforschungszentrum Potsdam

Michael Lautenschlager

World Data Center Climate, Hamburg

Uwe Schindler

Center for Marine Environmental Sciences (MARUM), University of Bremen

Irina Sens

German National Library of Science and Technology (TIB), Hannover

Background

In its 2004 report "Data and information", the *International Council for Science* (ICSU) strongly recommended a new strategic framework for scientific data and information.

On an initiative from a working group from the *Committee on Data for Science and Technology* (coData), the *German Research Foundation* (DFG) has started the project *Publication and Citation of Scientific Primary Data* as part of the program *Information-infrastructure of network-based scientific-cooperation and digital publication* in 2004.

Starting with the field of earth science the *German National Library of Science and Technology* (TIB) is now established as a registration agency for scientific primary data as a member of the International DOI Foundation (IDF).

Registration of scientific data

Primary data related to geoscientific, climate and environmental research is stored locally at those institutions which are responsible for its evaluation and maintenance. In addition to the local data provision, the TIB saves the URL where the data can be accessed including all bibliographic metadata. When data are registered, the TIB provides a DOI as a unique identifier.

Digital Object Identifier (DOI) is a system for identifying content objects in the digital environment. DOIs are names assigned to any entity for use on digital networks. They are used to provide current information, including where they (or information about them) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI will remain stable.

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.

Scientific data in the library catalogue

The scientific data is now accessible via the online library catalogue of the TIB (see fig. 1). The catalogue content is based on the application profile of the STD-DOI project for scientific data. The profile includes all metadata identified in the ISO 690-2 obligatory for the citing of electronic media, together with Dublin Core based standard metadata attributes.

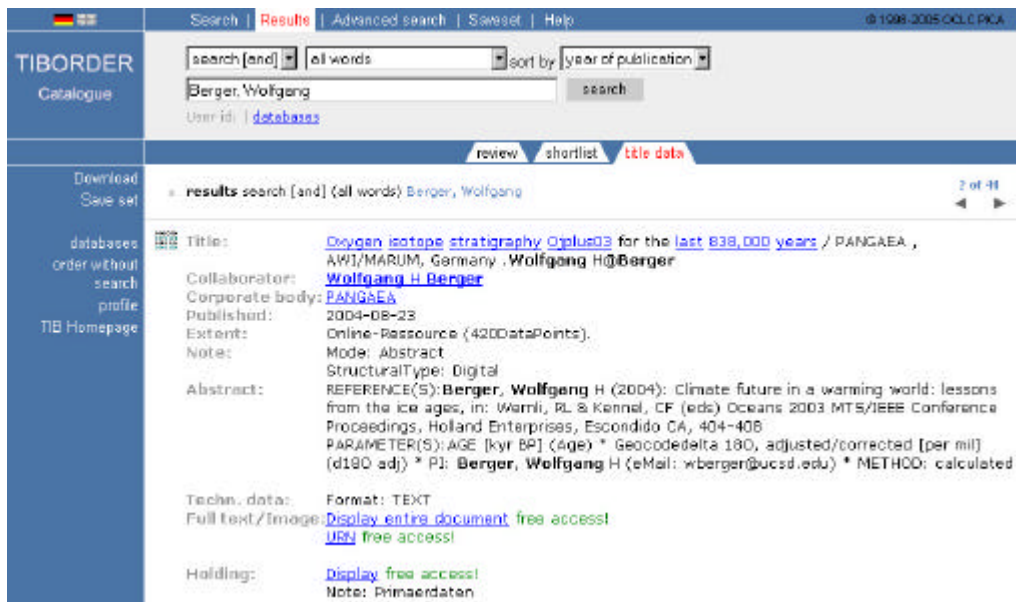


Fig. 1 A published dataset as a query result in the online catalogue of the TIB

The TIB offers an XML-based web service infrastructure that allows the data providers to include the registration and publication of scientific data into their infrastructure.

Standards set by the World Data Centers

At WDC-MARE, the webservice client is embedded into the metadata publishing workflow of the PANGAEA - Publishing Network for Geoscientific & Environmental data.

After inserting or updating a dataset in PANGAEA the import client queues background services which keep the XML metadata repository up to date (see fig. 2). The internal XML is stored as binary large object (blob) in a database table linked to the datasets. On top of this a

full text search engine (SYBASE EFTS) provides fast search access to the metadata. These XML blobs can be transformed into various other schemas with XSLT on the fly:

- ISO 19115
- OGC WebFeatures (for WFS)
- Dublin Core (for a OAI-PMH explain ! repository)
- another internal thumbnail format, which is also stored as blob for fast access by the PANGAEA search engine "PangaVista" \cite{pangavista}
- STD-DOI for the DOI registration of citable datasets

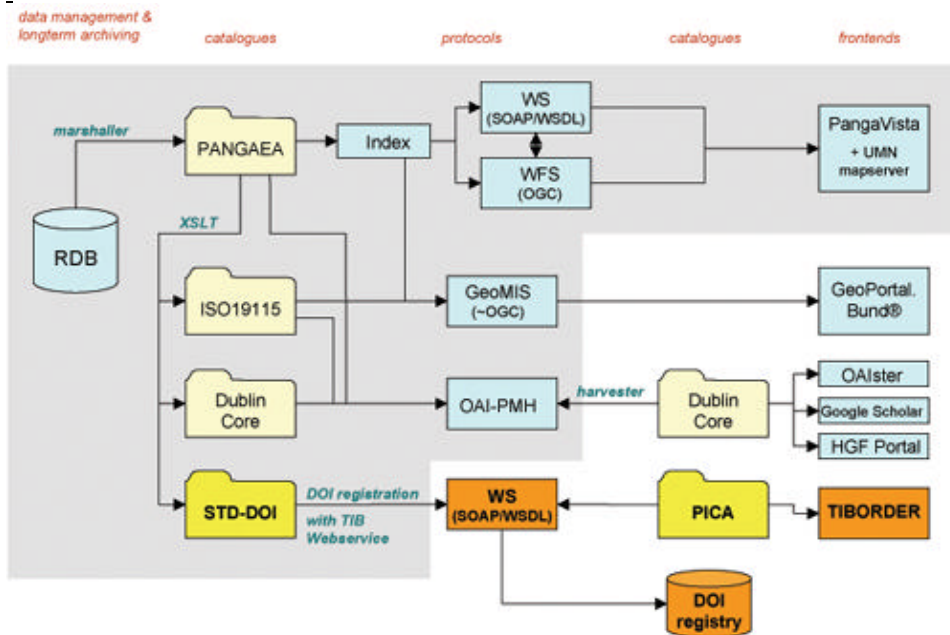


Fig. 2: PANGAEA middleware

Status

Registration has started for some fields of earth sciences, but will include other scientific disciplines in future.

We have registered 30 citable and 150,000 core datasets so far (March 2005), with an amount of expected 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 and is available to any data center worldwide