THE PUBLICATION OF SCIENTIFIC DATA BY WORLD DATA CENTERS AND THE NATIONAL LIBRARY OF SCIENCE AND TECHNOLOGY IN GERMANY

J. Brase [1*] ***and U. Schindler*** [2]

[*1] *Research center L3S, University of Hannover, Hanover, Germany*
Email: brase@l3s.de
[2] *World Data Center for Marine Environmental Sciences (WDC-MARE), MARUM, University of Bremen, Bremen, Germany*
Email: schindler@wdc-mare.org

ABSTRACT

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).

1 REGISTRATION OF SCIENTIFIC DATA

Primary data related to geoscientific, climate, and environmental research is stored locally at those institutions which are responsible for their 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. Any scientist working with this data is now able to cite the data in his work by its DOI. By this, scientific primary data is not exclusively understood as part of a scientific publication but has its own identity. If a scientist reads a publication where registered data is used, he might be interested in analysing the data under different aspects. He can now cite the data in his own publications using its DOI, referring to the uniqueness and separate identity of the original data. Since academic regard is often measured in so-called "citation-indexes" which count the number of citations of a scientist's work, collecting data can therefore be accomplished as an important part of academic work. Because of 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, for example to refer to the data inside a publication. Only citable datasets, usually collections of or publications from core datasets, are included in the catalogue.

2 THE DIGITAL OBJECT IDENTIFIER

To register the data, the TIB awards it with a DOI as a unique identifier. In May 200, the TIB became an official DOI Registration Agency. A DOI consist of two parts: a prefix and a suffix. For scientific data, a DOI looks like this:

10.1594/WDCC/IPCC_EH4_OPYC_SRES_B2_MM

10.1594 is the prefix and identifies that this DOI belongs to a scientific data set, registered at the TIB; WDCC stands for the respective research institute (World Data Center for Climate, in this case), followed by the internal name of the data record at the WDCC.

A DOI can be resolved in every web browser worldwide, using the *Handle system* from the *Cooperation for National Research Initiatives (CNRI)*. A Handle server, for example, is installed at the webpage of the *International DOI Foundation (IDF)*. Resolving of this DOI is therefore possible by using the URL:

http://dx.doi.org/10.1594/WDCC/IPCC_EH4_OPYC_SRES_B2_MM

Furthermore, it is possible to install a free plug-in into the Internet Explorer to resolve this DOI by typing it into the address bar of the browser. The DOI registration at the TIB works on a cost-recovery basis, enabling a persistent registration of scientific results for less than half a dollar.

3 SCIENTIFIC DATA IN THE LIBRARY CATALOGUE

Scientific data is now accessible via the online library catalogue of the TIB (see fig. 1). The catalogue data for the 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. A detailed analysis of the metadata used can be found in Brase (2004).

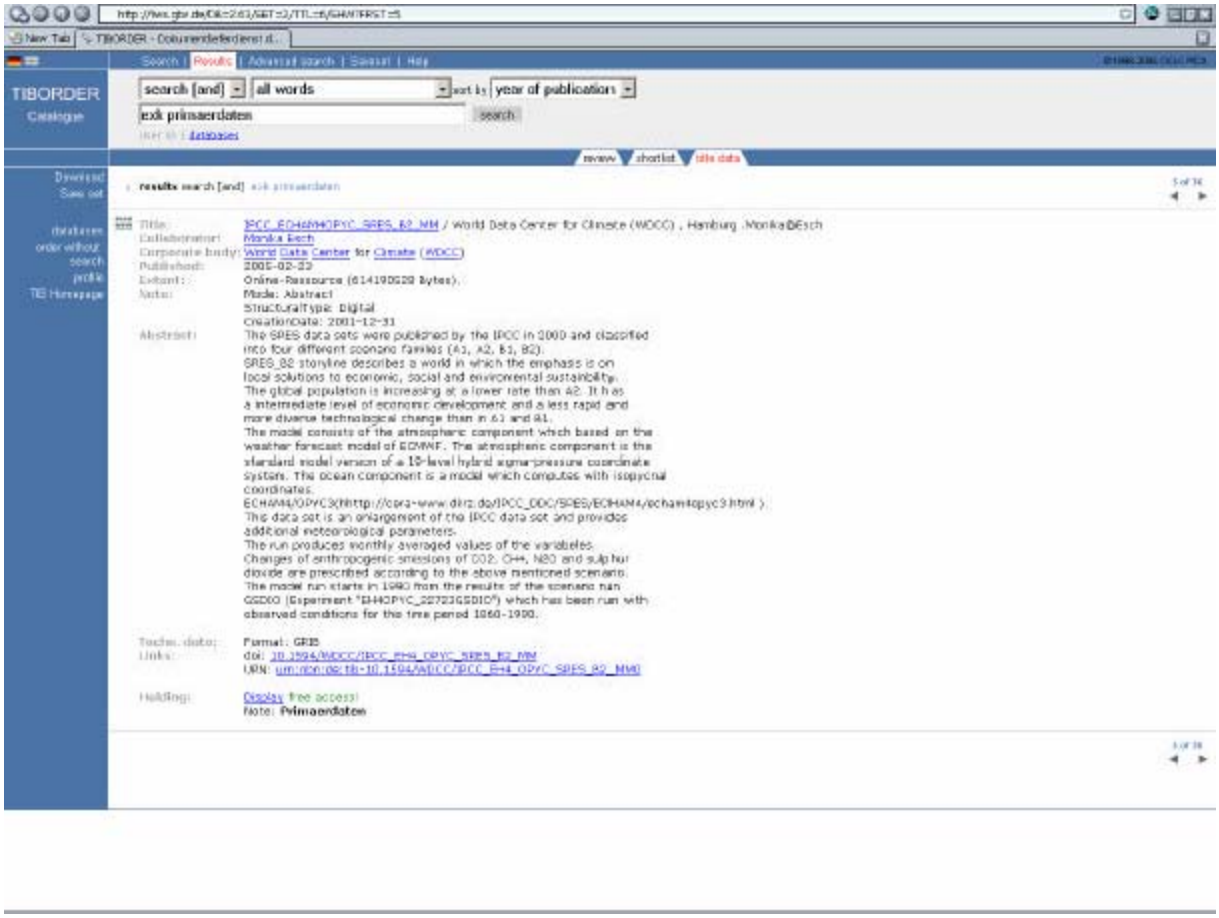

Fig. 1 A published dataset as a query result in the online catalogue of the TIB

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.

4 AN EXAMPLE OF WORKFLOW AT THE WORLD DATA CENTERS

At WDC-MARE, the web service 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). First, these background services marshal the metadata into an internal XML schema. This schema reflects the PANGAEA database structures and is optimized for simple marshalling of database records and transformation into other formats. With this software, the underlying database structure can be easily mapped to a given XML schema.
Because of the relational database structure, a change in one relational item can lead to a change in several XML files. Database update triggers fill the background services queue with changes for the related tables. This keeps the "flat" XML table in synchronization with the relational data. The internal XML is stored as a binary large object (blob) in a database table linked to the datasets. The full text search engine, however, provides fast search access to the metadata. These XML blobs can be transformed into various other schemas with XSLT on the fly.

.

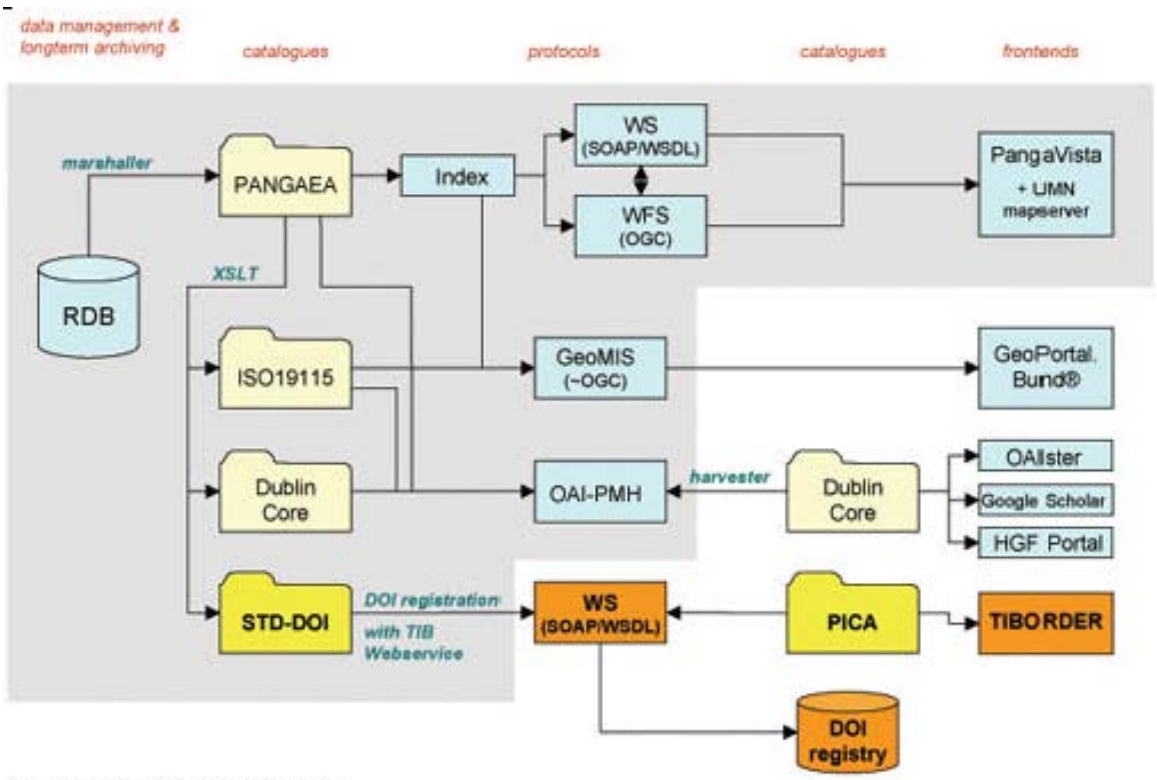

Fig. 2: PANGAEA middleware

For DOI registration, another background service registers all new or updated data-sets with the status "published" after a lead time of 30 days at the TIB. The lead time helps prevent inadvertent registration of datasets. During this time other data curators can look after the data and metadata and make changes which reset the lead time to 30 days. After registering the core datasets, the data curator can group them into a *collection dataset* (e.g. all data of a project or all data linked to a single publication) and give them a separate *citable DOI*. For that, the assigned metadata gets transformed by XSLT to the STD-DOI schema from the internal XML file. Nevertheless, it is also possible to choose a single core dataset and make it citable. Because of this workflow, the registering of citable PANGAEA datasets is always an upgrade of a previous core dataset (single data file or collection) to a citable one by adding metadata at the TIB.

5 STATUS

Registration has started for some fields of earth science but will include other scientific disciplines in the future. We have registered 40 citable and 240,000 core datasets so far (August 2005), with an expected number of 500,000 datasets to be registered by the TIB by the end of 2005. The registration of primary data will be widened to other science fields in 2006 and will be available to any data center worldwide. First discussions were held about cooperation with the *European Radiology Congress (ECR)*, the *European Academy for Allergology and Clinical Immunology (EAACI)*, the *Danish Research Database (DEF)*, and the *International Union of Crystallographers (IUCr)* to begin next year. Further projects have started to analyse the DOI registration of other kinds of scientific content such as Learning Objects, Simulations, or Pictures. For further information about how to cooperate with us, please visit the project's webpage (http://www.std-doi.de) or contact the author.

6 ACKNOWLEDGEMENTS

Everything described in this paper is the result of a joint cooperation. Apart from the authors, the following people are working for the project *Publication and Citation of Scientific Primary Data:*

- Michael Diepenbroek and Heinke Höck (World Data Center for Marine Environmental Sciences, MARUM University of Bremen, Germany),
- Hannes Grobe (Alfred-Wegener-Institut für Polar und Meeresforschung, Bremerhaven, Germany),
- Jens Klump (GeoForschungsZentrum Potsdam, Germany),
- Michael Lautenschlager (World Data Center for Marine Environmental Sciences, MARUM University of Bremen, Germany),
- Irina Sens (German National Library of Science and Technology, Hannover, Germany).

7 REFERENCES

Brase, J. (2004) Using digital library techniques - Registration of scientific primary data in (LNCS 3232) *Research and advanced technology for digital libraries*, Heidelberg, Germany: Springer