Data centres and their role in publication and access to data

Jens Klump[1], Eleni Paliouras[2], Jan Brase[3], Michael Diepenbroek[4], Hannes Grobe[5], Heinke Höck[6], Michael Lautenschlager[6], Uwe Schindler[4] and Irina Sens[7]

[1] GeoForschungsZentrum Potsdam, Telegrafenberg A3
14473 Potsdam, Germany
jklump@gfz-potsdam.de
[2] World Data Centre for Remote Sensing of the Atmosphere, DLR-DFD, Postfach 11 16
82230 Wessling, Germany
eleni.paliouras@dlr.de
[3] Research Centre L3S, University of Hannover, Postfach 6080
30060 Hannover, Germany
brase@kbs.uni-hannover.de
[4] World Data Centre for Marine Environmental Sciences (WDC-MARE),
Zentrum für Marine Umweltwissenschaften (MARUM)
28359 Bremen, Germany
{mdiepenbroek, uschindler}@pangaea.de
[5] Alfred Wegener Institute for Polar and Marine Research (AWI), Am Alten Hafen 26
27568 Bremerhaven, Germany
hgrobe@pangaea.de
[6] World Data Centre Climate, M&D at Max-Planck-Institut für Meteorologie, Bundesstrasse 55
20146 Hamburg, Germany
{hoeck, lautenschlager}@dkrz.de
[7] German National Library of Science and Technology (TIB), Postfach 6080
30060 Hannover, Germany
irina.sens@tib.uni-hannover.de

Abstract. In 2003 the 'Berlin Declaration' was published as a guideline to policy makers to promote the Internet as a functional instrument for a global scientific knowledge base. Since knowledge is derived from data, the principles of the 'Berlin Declaration' should similarly apply to data. However, access to scientific data is hampered, in part, by structural deficits in the publication process. Thus, applying the 'Berlin Declaration' to data requires a publication system for data that goes beyond what is in place for 'traditional' media. The criteria of accessibility, persistent identification and long-term availability need to be met in order to comply with the declaration. The project 'Publication and citation of scientific primary data' (http://www.std-doi.de) has shown prototypically how these criteria can be met and soon a publication system for scientific data will be available to the scientific community. In the structure set up within the project, data centres act as registration agents encompassing all functions necessary for the publication of scientific data, pointing towards a new role for data centres in addition to their current archiving mandate.

1 Background

Since the advent of the 'Berlin Declaration' in 2003 [1], calling for promotion of the Internet as a functional instrument for a global scientific knowledge base, it has been used as a basis for more specific calls for action related to scientific data, as data is the foundation for most scientific knowledge. Notable among these is that which is contained in the 2004 International Council for Science (ICSU) report on 'Scientific Data and Information' of the CSRP Assessment Panel which called for "a long term strategic framework for scientific data and information (policies, practices and infrastructure)" [2].

One of the practices that is essential to foster when considering a strategic framework on scientific data is the publication of primary data. The almost complete inability to include scientific data along with articles which report the results and conclusions obtained using the data, usually due to the size of the data sets, has resulted in unnecessary duplication of research efforts as well as in difficulty in the verification of scientific results [3]. A system for publication that involves data centres, which are most often already responsible for evaluation, quality control and maintenance of large data sets, is a logical step. Such a system has been under development over the past several years in Germany and will be described in this paper.

2 ICSU World Data Centre System and the German World Data Centre Cluster for Earth System Science

There exist many excellent data centres worldwide specializing in the archiving of scientific data and data products. Those which are especially good excel in making the data in their archives easily accessible to users, usually via the Internet. The International Council for Science (ICSU) helped to pioneer the concept of data centres when it created its World Data Centre (WDC) System in 1957 to, at that time, archive and distribute data collected in support of International Geophysical Year activities. Since then, the ICSU WDC system has expanded to 52 WDCs in twelve countries and covers many scientific disciplines.

Part of the ICSU mandate is for the WDCs to, subject to their financial resources, accept data according to the data management plans of appropriate ICSU scientific programs or monitoring activities, and to store these data safely and in good condition. Additionally WDCs may enhance their holdings by seeking and collecting related data sets and/or by preparing higher-order data products such as indices and collated or condensed data sets. Additionally, the WDCs are expected to, and have pledged to, provide the resources required to perform these activities on a long-term basis [4], [5].

Within Germany, there exist three ICSU WDCs and another which is under application with ICSU. These centres, to be described next, have joined together as a Cluster for Earth System Science, and as such have endeavoured to join forces in

advocating the archiving/preservation of scientific data and data products through various means.

2.1 WDCC

The WDC for Climate (WDCC) is maintained by the Model and Data (M&D) group hosted at the Max-Planck-Institute for Meteorology and is realized in cooperation with the German Climate Computing Centre (DKRZ). WDCC aims at collecting, scrutinizing, and disseminating data related to climate change on all time scales. Emphasis is on data products from climate modelling and related observational data. The WDC focuses on geo-referenced data using the operational CERA data and information system.

2.2 WDC-MARE

The WDC for Marine Environmental Sciences (WDC-MARE) is maintained by the Centre for Marine Environmental Sciences (MARUM) at the University of Bremen and the Alfred Wegener Institute for Polar and Marine Research (AWI). WDC-MARE aims to collect, scrutinize, and disseminate data related to global change in the fields of environmental oceanography, marine geology, paleoceanography, and marine biology. It focuses on georeferenced data using the information system PANGAEA.

2.3 WDC-RSAT

The WDC for Remote Sensing of the Atmosphere (WDC-RSAT) is hosted by the German Remote Sensing Data Centre (DFD) of the German Aerospace Centre (DLR) and aims at the provision of data and information on atmospheric trace gases, clouds, and the Earth's surface which are primarily gathered from satellite-based sensors. Higher level data and information products are also generated from the data through assimilation into numerical models of the atmosphere and of its interaction with the biosphere. The available data is always being updated as sensors, missions and assimilation techniques are added or improved.

2.4 WDC-TERRA

The planned WDC of the Lithosphere will be operated by the Data Centre of the GeoForschungsZentrum Potsdam. The scope of WDC-TERRA will be Earth's gravity field and gravity field models, geomagnetism, atmospheric sounding by GPS radio occultation, superconducting gravimetry, seismology (GEOFON and other sources), aerogravimetry, lithosphere soundings/seismics, magnetotellurics, and scientific

continental drilling. The WDC facilities will be integrated into the local framework of lithosphere observation and modelling. The planned WDC will cooperate closely with the WDC cluster for Earth System Science and other thematically related WDCs.

Some of the joint activities of these data centres will be described in the next section. However, it should first be stated here that, in addition to the ICSU WDC system, other large organisations, such as the World Meteorological Organisation (WMO), support or encourage large data centres. Although the remainder of this text will primarily refer to ICSU WDCs, all cases and models considered here could as easily be applied to non-ICSU data centres.

3 German Initiatives for Improved Data Access and Publication

Over the past several years, members of the German Earth System Science WDC Cluster, sometimes separately, sometimes together, have been involved in a number of innovative initiatives primarily at the German national level. Several of these will now be briefly described in order to demonstrate the evolution of the concepts involved in the most recent push to involve WDCs more centrally in the push for realising the goal of data publication.

3.1 STD-DOI: Publication and Citation of Scientific Primary Data

This project, started in 2004 and still ongoing, has focused many of the various efforts in Germany and has made possible the most direct steps towards a system for registration and publication of scientific data [6]. As a result of this project 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. The primary stakeholders, as well as steps for carrying out the data publication process developed and included in the project, are schematically represented in Fig.1 and Fig. 2 below. STD-DOI is being funded by the German Research Foundation and was undertaken to support an initiative from a working group of the ICSU Committee on Data for Science and Technology (CODATA).

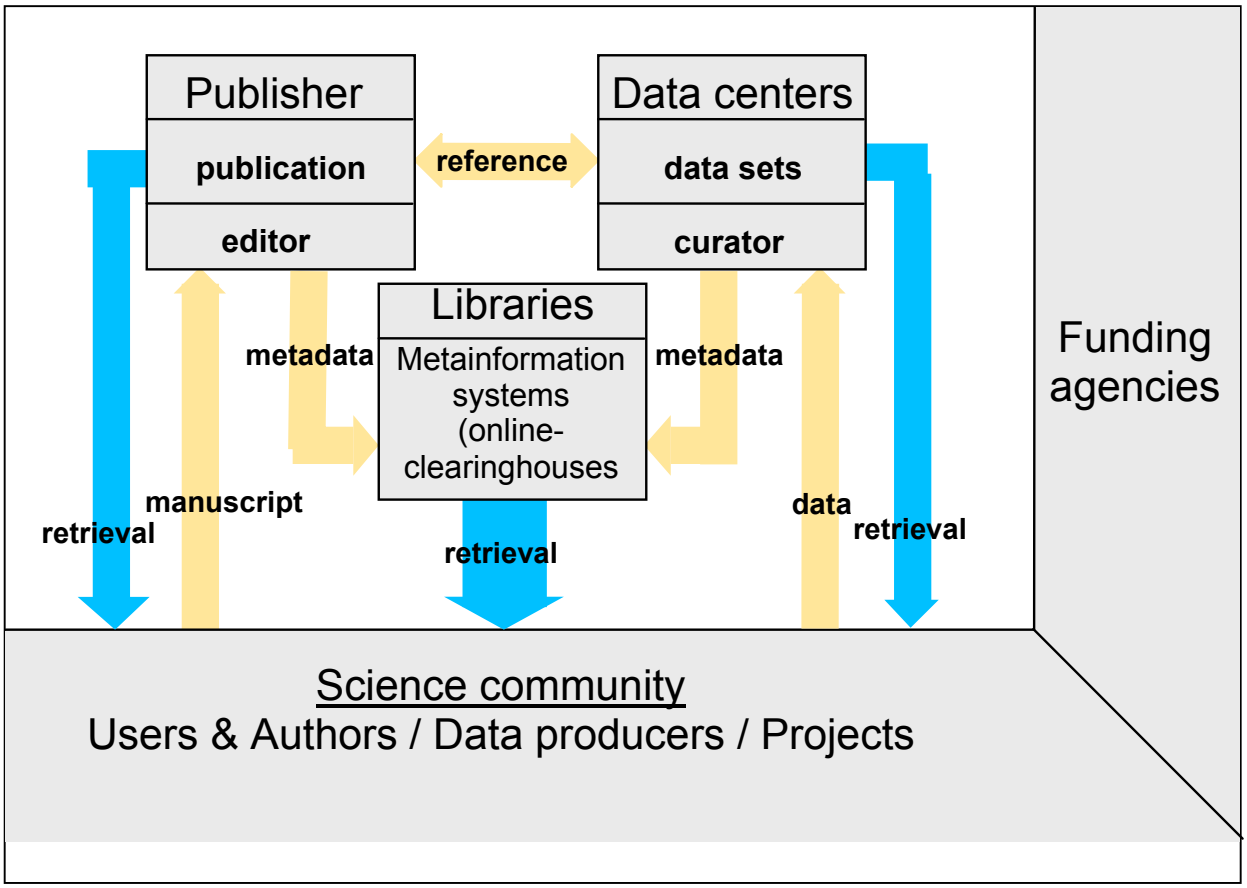

Fig. 1. Roles played by the science community, publishers, libraries, and data centres in the data publication process. With the exception of publishers, all of the stakeholders are represented and active in the STD-DOI project

3.2 C3-GRID: Collaborative Climate Community Data Processing Grid

This project, started in September 2005, aims to develop a highly productive grid-based environment for the German earth system science community in order to foster effective scientific analysis of the vast amounts of data resulting from modelling and observation programmes. This environment will be built on available Grid technology but will require within the project the development of a new generation system. The system will be composed of a meta-directory structure for the consistent description of data leading to

effective access and use of data archived in data centres such as ICSU WDCs, as well as to efficient use of distributed processing capability. C3-GRID is funded by the German Federal Ministry for Education of Research (BMBF).

3.3 CeGIM

CeGIM was a pilot project conducted by the German ICSU WDC Cluster for Earth System Science aimed at building a centre of excellence in data and information management in the geological sciences. This management would support the entire life cycle of scientific data, from capture and storage of primary data, to dealing with complex integrated types of problems. Publication of data was seen as a part of the life-cycle of scientific data and as a part of the project, a study of both the information needs and search strategies of scientists was conducted.

3.4 AG EUDIM

AG EUDIM is a working group of the Helmholtz Association of German National Research Centres that focuses on data and information management in the earth and environmental sciences. As such, AG EUDIM coordinates the data and information management strategies among all of the related Helmholtz Research Centres. Its concepts and expertise have earned the group a leading role in promoting access to data and the development of an eScience infrastructure for the geosciences. The group has been appointed by the German Science Foundation to coordinate these developments on a national level.

3.5 AG Open Access

AG Open Access is a working group of the libraries at Helmholtz Research Centres to promote the idea of Open Access to scientific knowledge among its member institutes. The Helmholtz Association of National Research Centres is a signatory of the 'Berlin Declaration'. In June 2005 the Assembly of the Helmholtz Association of National Research Centres adopted a proposal by the working group on how to promote Open Access among its member institutions. The adopted policy and work programme includes the publication of primary scientific data through Helmholtz data centres with the members of AG EUDIM in a leading role.

4 The Role of Data Centres

As illustrated in Fig. 1, data centres play an important part in the publication process for scientific data in a generic sense, especially in providing the science community with a link to the libraries. Within STD-DOI, to date, WDC-MARE, WDCC, and GFZ (proposed to host WDC-TERRA) are the participating as data centres. In addition to their baseline tasks as data centres, namely evaluation, review, storage and management of the scientific data in the archive, they have become responsible for assisting data "authors" by adding a new responsibility, namely that of data agent. In that role, the data centres register the data set on behalf of the authors at the TIB where, according to ISO standards, an electronic citation is issued along with a Catalogue DOI and an XML file with the relevant bibliographic metadata. The TIB saves these files which point to the URL at the data centre. If referenced, the data will be referred to by its related DOI (Fig. 2). For a more detailed analysis of the workflow at the TIB and the data centres, please refer to [4].

Due to the large anticipated number of datasets that will need to be registered, it has been decided to distinguish between *citable datasets* on the collection level and *core datasets* on the item level. Core datasets will receive their identifiers, but their metadata will not be included in the library catalogue. Up to August 2005, the STD-DOI project had registered 40 citable and 240,000 core datasets with an expectation that 500,000 datasets will be registered by the TIB by the end of 2005.

The project STD-DOI will be extended to include WDC-RSAT, as represented in the schematic in Fig. 2. Although well established and successfully delivering operational and historical data related to remote sensing of the atmosphere to many users, WDC-RSAT has, at this time, only a simple method for cataloguing its holdings. If a user knows what he/she is looking for, they can easily find it. However, in order to make the data holdings of WDC-RSAT more accessible and more widely known and used, the holdings must be published. In the coming year, WDC-RSAT will embark on this endeavour. Through the process, it is expected that some kind of model can be developed that may help other data centres as they become publishing agents.

Through the increasing inclusion of ICSU WDCs and other data providers, more scientific disciplines will be represented in primary data publication.

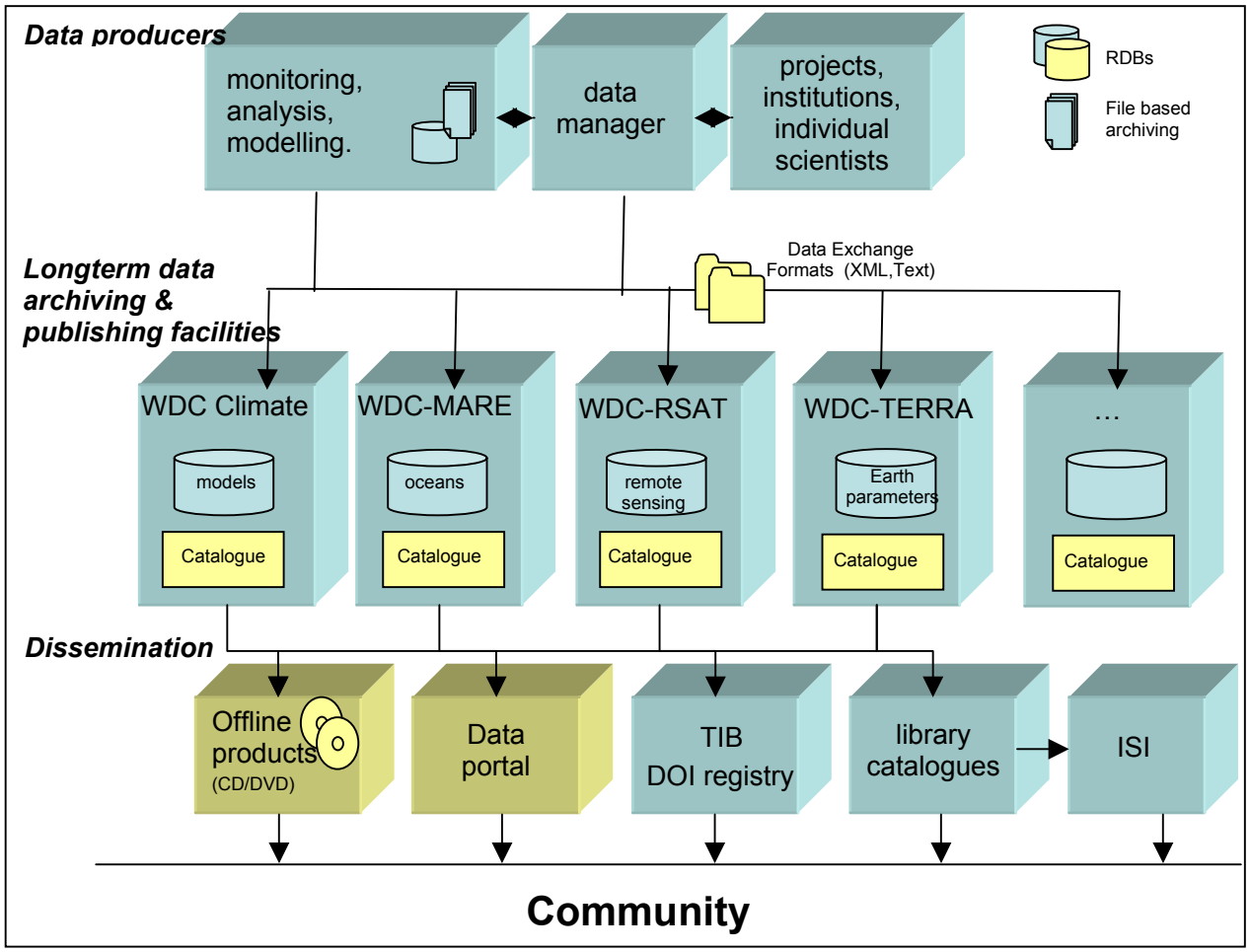

Fig. 2. Initial model for publishing data in Germany utilizing the ICSU WDC Earth System Science Cluster

5 Conclusions

Data centres are a natural partner of libraries in moving ahead with the, some may argue overdue, step forward in publication of data. First, they encompass significant holdings of primary data and information products and as such provide good targets for getting as

much data published as quickly and efficiently as possible. Second, they often offer insight and connections to unique scientific disciplines and communities, and as such, can help to broker solutions to challenges that may be unique to a particular part of the scientific community. Third, data centres often have interactions with each other, such as the German ICSU WDC Cluster for Earth System Science, and through these interactions can find solutions to challenges more quickly since they can each rely on the experiences of the others. Finally, and most importantly, data centres can play a vital role in the move towards publication of primary data since they already have the mandate, officially or unofficially, to assure the long term availability of their data holdings – one of the key elements of data publication.

In order to assure the continued progress in the area of data publication, there are of course several issues that still require attention. One is the establishment of a publishing platform that is equivalent to a journal. The Helmholtz Association of German National Research Centres is preparing an Open Access data journal for earth and environmental data, which will be associated with an Open Access journal of articles in this field. Additionally, the German Science Foundation has begun to cooperate with the Helmholtz Association of National Research Centres to carry this concept of into the geosciences community by expanding its membership beyond Helmholtz institutions and to include universities and other research entities.

References and Further Recommendations

1. http://www.zim.mpg.de/openaccess-berlin/berlinedeclaration.html
2. International Council for Science (2004) ICSU Report of the CSPR Assessment Panel on Scientific Data and Information
3. Dittert, N., Diepenbroek, M., and Grobe, H. (2001) Scientific data must be made available to all *Nature,* 414, 393.
4. Brase, J., Schindler, U. and Diepenbroek, M. (2005) Webservice infrastructure for the registration of scientific primary data. 9th European Conference on digital libraries (ECDL 2005), September 2005 Vienna, Austria
5. http://www.icsu.org
6. http://plato.wdcb.rssi.ru/wdc/guide/wdcguide.html.
7. Lautenschlager, M. and Sens, I (2003) Konzept zur Zitierfähigkeit wissenschaftlicher Primärdaten. *Information* 54, 463-466.
8. Arzberger, P. et al. (2004) Promoting Access to Public Research Data for Scientific, Economic, and Social Development Data Science Journal 3: 135-152. http://journals.eecs.qub.ac.uk/codata/Journal/contents/3_04/3_04pdfs/DS377.pdf
9. Maurer, S., Firestone, R. and Scriver, C. (2000) Science's neglected legacy - Large, sophisticated databases cannot be left to chance and improvisation Nature 405. http://dx.doi.org/10.1038/35012169