PANGAEA
An ICSU World Data Center as a Networked Publication and Library System for Geoscientific Data

Michael Diepenbroek, Uwe Schindler
MARUM, Universitt Bremen, Leobener Str, D-28359 Bremen, Germany

Hannes Grobe
Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, D-27570 Bremerhaven, Germany

Keywords: Digital libraries, world data center, data publisher, data portals.

Abstract: Since 1992 PANGAEA® serves as an archive for all types of geoscientific and environmental data. From the beginning the PANGAEA group started initiatives and aimed at an organisation structure which – beyond the technical structure and operation of the system – would help to improve the quality and general availability of scientific data. Project data management is done since 1996. 2001 the ICSU World Data Center for Marine Environmental Sciences (WDC-MARE) was founded and since 2003 – together with other German WDC – the group was working on the development of data publications as a new publication type. To achieve interoperability with other data centers and portals the system was adapted to global information standards. PANGAEA® has implemented a number of community specific data portals. 2007 – under the coordination of the PANGAEA® group – an initiative for networking all WDC was started. On the long range ISCU supports plans to develop the WDC system into a global network of publishers and open access libraries for scientific data.

1 INTRODUCTION

Data centers were created with the motivation to assure the long-term availability of scientific data. The Geophysical Year 1957 had been the starting point for the foundation of the system of World Data Centers (WDC), a number of globally distributed data centers, which were supposed to archive and distribute the geophysical data produced in that and the following years. Since then, the WDC system, which is related to the International Council for Scientific Unions (ICSU), has been extended to more than 50 data centers covering all fields of geosciences. More recently, ICSU expects the system to go through a major revision process. The exponentially increasing data volumes and the development of the Internet led to many new data managing and archiving systems. One of them was the Publishing Network for Geoscientific and Environmental Data PANGAEA®[1] (Diepenbroek et al., 2002), implemented in 1992. In 2001 the PANGAEA group founded the World Data Center for Marine Environmental Sciences (WDC-MARE)[2].

From the beginning PANGAEA® was conceived as a system that could cope with a wide spectrum of observational data. The heterogeneity and dynamics of the geosciences (including biology) required a flexible system for the acquisition, processing and archiving of the various data.

Nevertheless, already in the first phase of implementation it became clear that an efficient technical system is a necessary prerequisite, but cannot solve the principal problems of data quality and availability. Following the principle of open access (ESF, 2000; President of the Max Planck Society, 2003; OECD, 2004) scientific primary data are – besides publications – the second important result that must be long-term available in a re-usable state. A few decades ago it was still usual to publish primary data directly within a publication. Due to increasing data volumes and the transition to electronic publishing this practice was left. Scientific publishers allow for storing pri-

[1] Publishing Network for Geoscientific and Environmental Data. http://www.pangaea.de

[2] World Data Center for Marine Environmental Sciences. http://www.wdc-mare.org

Diepenbroek M., Schindler U. and Grobe H. (2008).
PANGAEA - An ICSU World Data Center as a Networked Publication and Library System for Geoscientific Data.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 149-154
DOI: 10.5220/0001518401490154

mary data as electronic assets. Nevertheless, archiving is not compliant to any standards or unique structures and is excluded from peer review, hence, can also not be seen as a template for a general solution of the problem. In contrast, many data centers including a good part of the ICSU WDC, are well prepared in a technical sense, although, archiving mostly does not comply to global standards either. The separation of scientific publications from the underlying primary data can be seen as a severe structural problem in the empirical sciences. It hampers not only the evaluation of a publication but also re-usage of results.

There are no authorized and authenticated places for the long-term storage of scientific data, no cross-referencing between scientific publications and possibly archived data and no or only rudimentary networking between data centers. Needed are global library structures and systems for the publication of scientific data. In this context the ISCU WDC play an active role. The German WDC-Climate, WDC Remote Sensing, and WDC-MARE together with the GeoForschungsZentrum Potsdam and the Technical Library in Hannover have implemented a practical system for the publication of scientific data (Schindler et al., 2005). In this connection WDC-MARE with the information system PANGAEA® and its editorial system already can be seen as a reference for a publication and library system for scientific data. In addition, due to its interoperability, PANGAEA® is networked with various other data centers, libraries, portals, and services. In the following it will be described in more detail.

2 FROM DATA AQUISITION TO PUBLICATION

WDC-MARE / PANGAEA® is operated as a permanent facility by the Center for Marine Environmental Sciences (MARUM) of the University Bremen and the Alfred Wegener Institute for Polar and Marine Research (AWI) in Bremerhaven. 4 scientists are responsible for the organisation and development of the system. A team of 8-10 scientists take care of the data management services, which are supplied on an international level since 1996. Until 10/2007, PANGAEA® was and is partner in more than 60 European to international projects covering all fields of environmental sciences. The budget amounts approximately 1.2 Mio Euro per year for personnel, hard-, and software. Third party funds are about 70% of the total budget.

3 AQUISITION, QUALITY ASSURANCE, EDITORIAL AND ARCHIVING

The acquisition of scientific data is a time consuming problem. Based on own estimates only a few percent of the globally produced scientific data are generally available and even less is long-term archived in adequate data centers. Seldom, data are spontaneously handed out to a data center. For scientific institutions there is – since several years – an obligation for long-term storage of data. Likewise, many projects and programs are configured with corresponding constraints. Agreements in such contexts facilitate data acquisition, however, cannot solve the problem completely.

On the other hand data management as a funded component of scientific projects has proven rather efficient. For EU projects addressing environmental research data management is an important evaluation criterion. For projects like e.g. CARBOOCEAN[3],which aim at improved quantifications of CO_2 balances in the marine environment, a high availability of quality assured data is a necessary prerequisite for the success of the project. In general, large scale or complex scientific approaches in Global Change research are based on the results and data of many smaller projects.

The PANGAEA® group is supplying project data management since more than 10 years. This is the most important source for new data to be archived, mostly because of the proximity with the scientists. In addition, project data management considerably contributes to the operational costs of PANGAEA®. This creates capacities, which enables the group to realize also not funded projects as e.g. the final global harmonisation, archiving, and publication of data from the IGBP project Joint Global Ocean Flux Studies (JGOFS) (Sieger et al., 2005).

Quality assurance is an indispensable part of data management. Essential in this respect is not the data quality itself, but assessment of the data quality. Important are completeness und correctness of data descriptions (metadata) and compliance with existing content standards as ISO19115 (Kresse and Fadaie, 2004) or DIF[4]. At the minimum the metadata have to answer the question: Who has measured what, when, where, and how? In addition, PANGAEA® regularly checks the validity of used methods and whether the precision of data values corresponds with the meth-

[3]CARBOOCEAN. http://www.carboocean.org/

[4]Directory Interchange Format. http://gcmd.nasa.gov/User/difguide/

ods used. Outliers are identified and flagged. The data producer (principal investors or institution) take the responsibility for the actual quality of the data.

Editing and archiving of data sets varies with data types and data centers. Practically, there are neither archiving standards nor are there editorial systems, which could be generally used. Common is usage of relational data bases, which guarantees at least a certain consistency of metadata. At present, almost

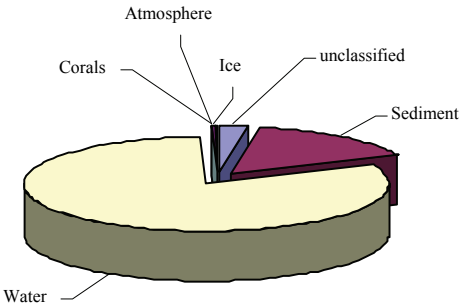

Figure 1: Contents of PANGAEA®. (9/2007): Data for ≈ 30 000 parameters (e.g.: sediment & ice profiles, seismic profiles, atmospheric profiles, ocean geochemistry, mineral distributions, geological maps, plankton & fish, sea floor pictures and films), data sets: 570 676, data items: 1 834 869 117.

600 000 data sets with nearly 2 billion observations (numerical, text, or binary data items) are available. The data are related to about 30 000 different measurement types (parameter), more than 10 000 principal investigators (PI), about 6 000 scientific publications, and more than 300 000 sample locations. The yearly increase is more than 10% of the total inventory (see figure 1).

In PANGAEA® data and metadata are systematically recorded through an editorial system. The system contributes significantly to the efficiency of data curation. For smaller data centers with relatively specialized data contents such a system might be dispensable. PANGAEA®, however, was conceived as a large scale system to handle various types of data.

On the server side the challenge of managing the heterogeneous and dynamic data of environmental and geosciences was met through a flexible data model, which reflects the information processing steps in the earth science fields and can handle any related analytical data. The basic technical structure corresponds to three tiered client/server architecture with a number of clients and middleware components controlling the information flow and quality. A relational database management system (RDBMS) is used for information storage. Physical backups are regularly stored in different locations, thus protecting the data inventory from loss. Figure 2 shows the simplified setup of PANGAEA®. Mass data, like geophysical data or binary objects, as e.g. pictures and films, are stored on hard disk arrays from where they eventually migrate into related tape silos. All data are replicated on a frequent base into a data warehouse (Sybase IQ). This enables high-performance retrievals of any space/time or keyword constrained section of the data inventory. The compiled metadata are part of the search results. The web-based clients include a simple search engine and an IQ interface which will be productive by the end of 2007.

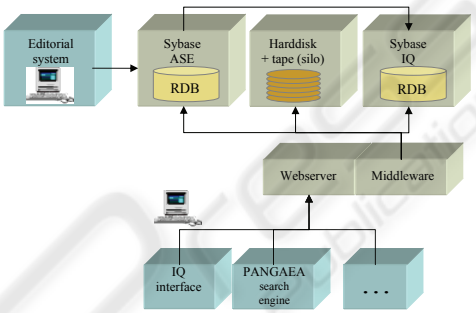

Figure 2: Technical setup of PANGAEA®.

4 DATA PUBLICATION

Within the last three years the PANGAEA® group together with the WDC-RSAT, WDC-Climate, and the German Technical Library (TIB) has developed and prototypically implemented a concept for the publication of scientific data (Schindler et al., 2005; Klump et al., 2006). The project – funded by the German Science Foundation (DFG) – investigated general requirements for this new publication type:

- The formal structure of the publication, that is, which describing elements are mandatory, which are optional, how should they be configured, which data formats and standards are useful?
- The granularity of data sets to be published.
- The development of "peer review" like procedures for quality assurance.
- The requirement for data centers with respect to long-term archiving and persistent referencing of archived data, e.g. through "Digital Object Identifier" (DOI). Certification of data centers is – besides own experiences – essentially based on the OAIS reference model (NASA Consultative Committee for Space Data Systems, 2002) and results from the German BMBF project NESTOR[5].

[5]NESTOR. http://www.langzeitarchivierung.de/

The results were used as guidelines in the participating facilities to adapt organisational and technical structures, in particular to develop editorial schemes for the import and curation of data. Such schemes were prototypically realized in all data centers. They are, however, more or less – depending on the facility – integrated into the technical environment and the scientific process. In this respect the above mentioned problem of granularity is crucial. A principal problem is that data centers traditionally treat their data archives as a continuously extendible and updatable data space which does not allow for a subdivision into static data entities. With data publishing, however, persistence and version control of data entities are needed. So far, the WDC in the data publication project have agreed on a simple model which differentiates between archived or accessible data entities and citable data entities: Archived data entities can be citable or may be comprised to citable data entities in a second step. Citable data entities represent the interface between data archive and scientific literature. They allow for cross-referencing data publications and traditional scientific publications.

Legacy data are a further problem. For WDC-MARE / PANGAEA® a significant effort is needed to replenish the whole data inventory in a way to get citable data sets in the end. Each data set needs consultation with the original PI(s) or further scientists from the corresponding research field and eventually manual changes on the metadata. The current work led to first trials of a "peer review" for scientific data.

All data sets are annotated with a Digital Object Identifier (DOI) and are registered at the DOI registry for scientific data at the Technical Library in Hannover (TIB), which have a corresponding contract with the International DOI Foundation (IDF)[6]. Citable data sets are subsequently recorded in the library catalogue of the TIB[7]. Both, DOI registration and migration into the library catalogue, are automated routines.

Since more than 10 years PANGAEA® uses a client/server system for the import of new data and the curational works. The system minimizes the manual work for the data curators and can be used globally. The development of the system into an editorial system is an iterative process in which system managers, data curators, and partners of the data publication project are participating. Besides numerous adaptations it was necessary to include a chronological sequence into the editorial process. Newly imported data sets are not registered immediately but with a time-lag of 28 days, allowing for further changes on or replacement of data sets. After expiry of the time-limit data sets are registered and might be flagged as citable. Except for some minor metadata elements data sets are subsequently static. The DOI registration was harmoniously integrated into the existing infrastructure.

Overall, the necessary conversion to a publication system has been completed. The editorial effort – except for the increased communication – has stayed about the same. This is an important aspect with respect to the running operational costs. Examples of citable data sets are e.g.: doi:10.1594/PANGAEA.472287, doi:10.1594/PANGAEA.472492, doi:10.1594/PANGAEA.370797

5 STANDARDS, NETWORKING AND PORTALS

Networking of data producers, archives, and consumers, compliant to Global Spatial Data Infrastructures[8] (Nebert, 2004) is a necessary prerequisite for geospatial one stop shops and large scale data compilations. It is a vision, more recently described in the 10-year implementation plan of a Global Earth Observing System of Systems (GEOSS) (Battrick, 2005) of the Group on Earth Observations (GEO)[9] which – on the ministerial level – the first time supplies an efficient framework for networking geospatial service suppliers and users. Due to lacking resources, however, GEOSS is highly dependant on existing capacities and activities. Therefore, on the meeting of the WDC directors in 2007 it was decided to start an initiative for networking WDC. This is not only a useful contribution to GEOSS but can also be seen as a first step towards the creation of a global network of data libraries. The WDC system so far is a unique consortium of data centers that supply free and unrestricted online access on their data holdings. The WDC networking initiative is coordinated by PANGAEA®. During the last 5 years the group has worked systematically on the networking capabilities of the PANGAEA® system and by now supplies a variety of different, generally available services, all conform to global geospatial standards (ISO, OGC und W3C). The metadata for each data set are 'marshalled' from the relational database into a XML blob. The corresponding scheme is proprietary. It comprises all in-

[6]International DOI Foundation. http://www.doi.org

[7]TIBORDER. http://www.tib-hannover.de/en/

[8]Global Spatial Data Infrastructures (GSDI). http://www.gsdi.org/

[9]Group on Earth Observations. http://www.earthobservations.org/

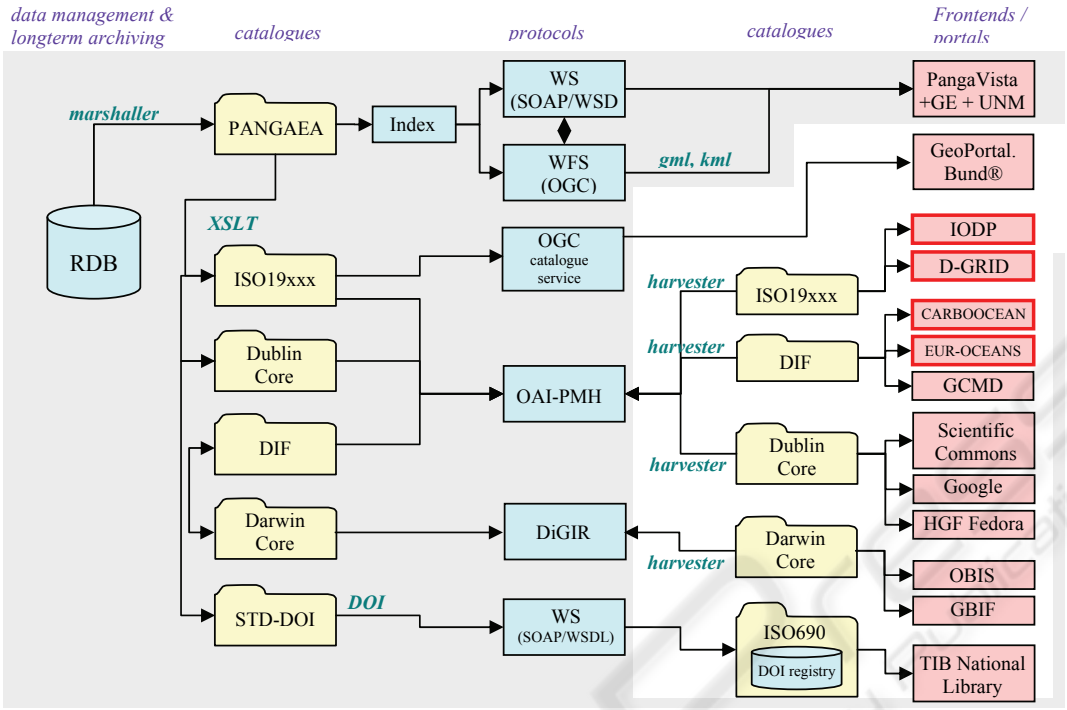

Figure 3: Metadata infrastructure for PANGAEA®. Grey shaded parts belong into the domain of PANGAEA®. Portals with a red outline were implemented by the PANGAEA®. group.

formation needed for mapping (per XSLT) the metadata on to the various content standards as ISO19115 or the Directory Interchange Format (DIF), Important protocols are the OGC Catalogue Service (CS-W)[10] and the Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH) (Van de Sompel et al., 2004). The latter is relatively simple to be implemented and is widely used in the library world. An overview supplies Figure 3. Because of the dynamics of IT developments PANGAEA® deliberately builds on an internal architecture that can cope with different or new standards.

In addition, the PANGAEA® group has implemented a number of community and project specific metadata portals. The portal framework is generic and based on the components harvester, indexer with search engine (Apache Lucene[11]) and corresponding API (Schindler and Diepenbroek, 2008)[12]. Examples are the portal for the International Ocean Drilling Program[13] and for the EU projects EUR-OCEANS[14] and CARBOOCEAN[15]. A precondition for these portals is that participants not only supply metadata catalogues, but also enable direct and open access to the corresponding data entities.

An even higher level of networking was reached with the participation in the German Community GRID C3[16]. In this project PANGAEA® supplies its portal framework and contributes to the data GRID with observational data served by the data warehouse. Nevertheless, GRID projects are still restricted to special data types and workflows. For general and simple to be implemented architectures more development is needed. A special problem with heterogeneous data as supplied by PANGAEA® is the availability of standardized vocabularies for the control of applications. Corresponding concepts are supplied by ISO19109 and ISO19110 (Kresse and Fadaie, 2004). Practical progress can be expected through the European ini-

[10]OGC Catalogue Service. http://www.opengeospatial.org/standards/cat

[11]Apache Lucene (Hatcher and Gospodnetic, 2004). http:// lucene.apache.org/java/docs/

[12]PANGAEA Framework for Metadata Portals (panFMP). http://www.panFMP.org/

[13]Scientific Earth Drilling Information Service - SEDIS (Miville et al., 2006). http://sedis.wdc-mare.org/

[14]EUR-OCEANS data portal. http://dataportal.eur-oceans.eu/

[15]CARBOOCEAN data portal. http://dataportal.carboocean.org/

[16]Collaborative Climate Community Data and Processing Grid. http://www.c3grid.de/

tiative INSPIRE[17] (The European Parliament, 2007). This, however, must be regarded as a long-term task.

6 CONCLUSIONS

With its long-term and secured archiving structure, the highly efficient editorial system, and the extensive interoperability with other data centers and portals, PANGAEA® has developed into an exemplary publication and library system for scientific data. The approach for publication of scientific data developed within the German WDC consortium and realized within PANGAEA®, is way beyond the usual interlinking of scientific publications with related data as e.g. practiced within the Human Genome Community. It allows for self-contained data publications. Each data publication is provided with a meaningful citation and a persistent identifier (DOI) und thus enables reliable references. The citability gives a strong motivation for scientists to publish their data. It is a bottom-up approach which on the long range will improve data quality and availability.

The concept met with wide response from data producers. Nevertheless, it might take years for this new publication type to be generally accepted. First talks with ISI Thompson have indicated that data publications might be recognized for the citation index. The reference systems, developed within the German WDC, need to be extrapolated. With the networking initiative of ICSU WDC a first step is done in the direction of a global library consortium for scientific data. Such a network would be trans-disciplinary and has the advantage that all data are available without any restriction according to the open access rules. However, a sustainable framework is needed on the one hand to guarantee long-term availability of scientific data and on the other hand to foster the work in the data centers in the direction of standards for processing, archiving, and publication of data as well as interoperability of data centers. The revision of ICSU WDC will support such a framework. Nevertheless, long-term operation requires further safeguarding through national or international contracts. A memorandum of understanding could be a good starting point.

REFERENCES

Battrick, B. (2005). *Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan Reference Document*. ESA Publications Division.

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., and Wefer, G. (2002). PANGAEA–an information system for environmental sciences. *Computers & Geosciences*, 28(10):1201–1210.

ESF (2000). *Good scientific practice in research and scholarship*.

Hatcher, E. and Gospodnetic, O. (2004). *Lucene in Action*. Manning Publications.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Hck, H., Lautenschlager, M., Schindler, U., Sens, I., and Wchter, J. (2006). Data publication in the open access initiative. *Data Science Journal*, 5:79–83.

Kresse, W. and Fadaie, K. (2004). *ISO Standards for Geographic Information*. Springer, Heidelberg.

Miville, B., Soeding, E., and Larsen, H. C. (2006). Scientific Earth Drilling Information Service for the Integrated Ocean Drilling Program. *Geophysical Research Abstracts*, 8:05486.

NASA Consultative Committee for Space Data Systems (2002). Reference Model for an Open Archival Information System (OAIS).

Nebert, D. D., editor (2004). *The SDI Cookbook, Version 2.0*. Global Spatial Data Infrastructure Association, Technical Working Group Chair.

OECD (2004). Science, Technology and Innovation for the 21st Century. In *Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique*.

President of the Max Planck Society (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*.

Schindler, U., Brase, J., and Diepenbroek, M. (2005). Webservices Infrastructure for the Registration of Scientific Primary Data. In Rauber, A., Christodoulakis, S., and Tjoa, A. M., editors, *Research and Advanced Technology for Digital Libraries*, volume 3652 of *Lecture Notes in Computer Science*, pages 128–138. Springer.

Schindler, U. and Diepenbroek, M. (2008). Generic XML-based Framework for Metadata Portals. *Computers & Geosciences*. Submitted.

Sieger, R., Grobe, H., Diepenbroek, M., Schindler, U., and Schlitzer, R., editors (2005). *International Collection of JGOFS – Volume 2: Integrated Data Sets (1989-2003)*. Number 0003 in WDC-MARE Reports. WDC-MARE.

The European Parliament (2007). *DIRECTIVE 2007/.../EC Of the European Parliament and of The Council of establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*. Directive not yet officially released.

Van de Sompel, H., Nelson, M., Lagoze, C., and Warner, S. (2004). Resource Harvesting within the OAI-PMH Framework. *D-Lib Magazine*, 10(12).

[17]INSPIRE. http://www.ec-gis.org/inspire/