The Benefits of Cross Linking: The International Continental Scientific Drilling Program

Jens Klump

GeoForschungsZentrum Potsdam, Telegrafenberg, 14473 Potsdam, Germany
email: jens.klump@gfz-potsdam.de

Abstract

The International Continental Scientific Drilling Program (ICDP) addresses fundamental scientific problems of global importance as an element of geological and geophysical research programs. It seeks geological sites from around the world and involves the international community of scientists to optimize the results from drilling. Besides scientific and technical consulting to ICDP projects the ICDP Operational Support Group (OSG) offers data management in the field, and during the analysis and publication phases of scientific drilling project. Operational data management is based on the ICDP Drilling Information System (DIS), which allows data capture in the field and transfer to a central database at GeoForschungsZentrum Potsdam (GFZ).

Since the start of ICDP, data sharing has played an important part in ICDP projects and the OSG has facilitated dissemination of data within project groups and encourages the sharing of data to promote scientific progress. With the online Scientific Drilling Database (SDDB, http://www.scientificdrilling.org) ICDP and GFZ Potsdam created a platform for the dissemination of data from drilling projects. The thematic focus of SDDB is on data that were used as the basis of a publication. All data publications in SDDB are citeable and accessible through Digital Object Identifiers (DOI). In addition, SDDB maintains a record of links from its datasets to the literature, where the data are interpreted and will in future also reference the samples from which the data were derived. This network of data, literature and physical samples can be shared with other data providers and portals to enhance discovery of data through discipline specific catalogues. Novel added-value services may be developed by exploiting the semantic nature of this network.

Introduction

Projects in the International Scientific Continental Drilling Program (ICDP) produce large amounts of data. Since the start of ICDP, data sharing has played an important part in ICDP projects and the ICDP Operational Support Group has facilitated dissemination of data within project groups (Conze et al., 2007). Some of these data later became the basis of scientific publications, while others remained unpublished. However, in most cases the data sets themselves were not available outside of the respective projects. With the online Scientific Drilling Database (SDDB, http://www.scientificdrilling.org) ICDP and GeoForschungsZentrum Potsdam (GFZ) created a platform for the dissemination of data from drilling projects (Klump and Conze, 2007). The thematic focus of SDDB is on data that were used as the basis of a publication, but some published datasets are data publications in their own right.

To publish data requires that data are citeable. This means, a mechanism is needed that ensures that the location of the referenced data on the internet can be resolved at any time. In the past, this was a problematic issue because URLs are short-lived, many becoming invalid

after only a few months. Data publication on the internet therefore needs a system of reliable pointers to a web publication to make these publications citeable. To achieve this persistence of identifiers for their conventional publications many scientific publishers use Digital Object Identifiers (DOI) (Brase, 2004). GFZ Potsdam is a member of the project "Publication and Citation of Scientific Data" (STD-DOI), which is funded by the German Science Foundation. In this project the German National Library for Science and Technology (TIB Hannover), together with GFZ Potsdam, Alfred Wegener Institute (AWI) Bremerhaven, University of Bremen and the Max Planck Institute for Meteorology in Hamburg set up a system to assign DOIs to data publications.

Data management in scientific drilling projects

Data management in scientific drilling projects of the International Continental Scientific Drilling Program (ICDP) and the Integrated Ocean Drilling Program (IODP) pursues two main goals: on the one hand, to capture of drilling and scientific data as early as possible, and on the other hand, manage the long-term storage and dissemination of these data. The data capture in both ICDP projects and IODP-Mission Specific Platform (MSP) expeditions has three phases: fieldwork phase, analysis phase, and publication phase. Drilling, curation, logging, and basic scientific data are captured at the drill site during the fieldwork phase. This phase is followed by the analysis phase during which detailed measurements, descriptions, images and analytical data are captured within a laboratory setting. The data are subsequently transferred to the long-term data storage system. During the fieldwork and analysis phases data are, in most cases, only accessible by the scientists involved in the project.

In the publication phase of a scientific drilling project analytical data become available and their interpretations are published in the literature. The publication platform for public data from ICDP drilling projects is the Scientific Drilling Database (SDDB) (Klump and Conze, 2007). Some of these data can be imported directly from the DIS, but for most data ICDP has to rely on the authors to supply the data. Since most drilling project track their publications for reporting purposes the SDDB editorial staff has an overview of publications related to a specific project. The authors are then contacted and asked for their cooperation in the data publication process. At the present stage of the development of SDDB the right to upload data is still restricted to the SDDB operators. However, a prototype of the SDDB data upload assistant already exists and is being tested for usability. The aim is to make the SDDB data upload assistant as easy to use as possible so that authors can upload data themselves. Linking data to associated publications and to physical sample material is part of the editorial process.

Publication of scientific data is labour intensive and two points are crucial to make the system scaleable: (1) the publication process has to be automated as far as possible and duplication of work, e.g. metadata editing, has to be avoided, and (2) authors have to be motivated to participate in the data publication process. However, the process of scientific data publication is still hampered by structural deficits which will be discussed in the section below.

Data publication today

On 22 October 2003, a group of leading research institutions and research funding institutions published the ‘Berlin Declaration on Access to Knowledge in the Sciences and Humanities’ in order to “[…] promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider.” (Berlin Declaration, 2003). The Berlin Declaration has since been signed by 242 scientific bodies worldwide. The OECD Governments have also recognised the importance of

open access to knowledge. This new policy has been formulated in the 'Communiqué on Science, Technology and Innovation for the 21st Century' issued at the OECD ministerial meeting, 29-30 January 2004 (OECD, 2004). It has been followed by a recommendation on access to research data to the OECD Council (OECD, 2006). This 'soft legislation' has to be transferred into national legislation by the OECD member states.

Knowledge, as published through scientific literature, is the last step in a process originating from primary scientific data. These data are analysed, synthesised, interpreted, and the outcome of this process is published as a scientific article. The Berlin Declaration and the OECD Ministerial Communiqué look at the outcome of this process. Because scientific knowledge is ultimately derived from data, we wish to examine more closely the beginning of this process, the issues of data sharing and data publication.

Some organisations encourage scientists to share data freely and even make data sharing a part of their funding policy (e.g. NIH, 2003). In addition, cases of scientific misconduct in recent years have highlighted the importance of making scientific data available. As a consequence, the German Science Foundation, and other science organisations, adopted 'Recommendations for Good Scientific Practice' as part of their policy. They require that institutes archive data, which were used as a basis of a publication, on safe storage media for a minimum duration of ten years (DFG, 1998). Besides being a matter of common sense and good scientific conduct, thorough documentation of experiments also makes economic sense (Alexander et al., 2004).

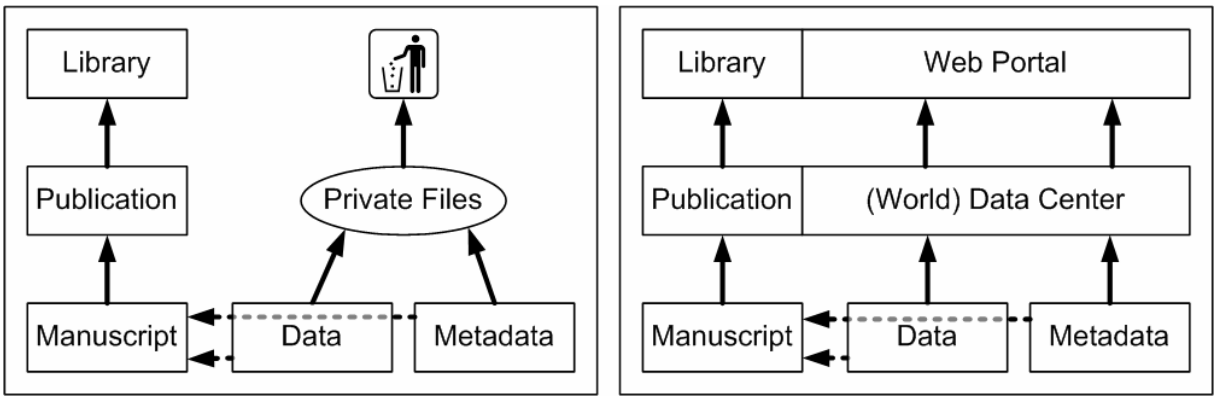

Figure 1. *(left) Schematic depiction of the flow of scientific information from research to published library resources as currently practiced (modified after Helly et al., 2003). (right) Potential approach based on data publication by data centres and content syndication to scientific web portals, which could also be library catalogues. Inter-linking publications and their underlying data will create new scientific products with added value. The dashed lines from data and metadata to the manuscript reflect the limited publication of these sources in our conventional scientific journals.*

Despite advances in the policies regarding access to data, only a very small proportion of the original research data are published in conventional scientific journals. Existing policies on data archiving notwithstanding, in today's practice data are primarily stored in private files, not in secure institutional repositories, and effectively are lost (Figure 1, left). This lack of access to scientific data is an obstacle to interdisciplinary and international research. It causes unnecessary duplication of research efforts, and the verification of results becomes difficult, if not impossible (Dittert et al., 2001). Large amounts of research funds are spent every year, while already existing data remain underutilised (Arzberger et al., 2004).

Publication and citation of scientific data

To improve access to data and to create incentives for scientists to make their data accessible, the German CODATA group initiated a project on publication and citation of scientific data which was funded by the German Science Foundation DFG for the periods 2003-2005 and 2006-2008 (STD-DOI, 2003). This project uses persistent identifiers (both DOI and URN) to identify datasets available in a digital format. The identifier is resolved to the valid location (URL) where the this dataset can be found. This approach meets one of the prerequisites for citeability of scientific data published online. In addition, the data publications are included into the catalogue of the German National Library of Science and Technology (TIB) (Brase, 2004).

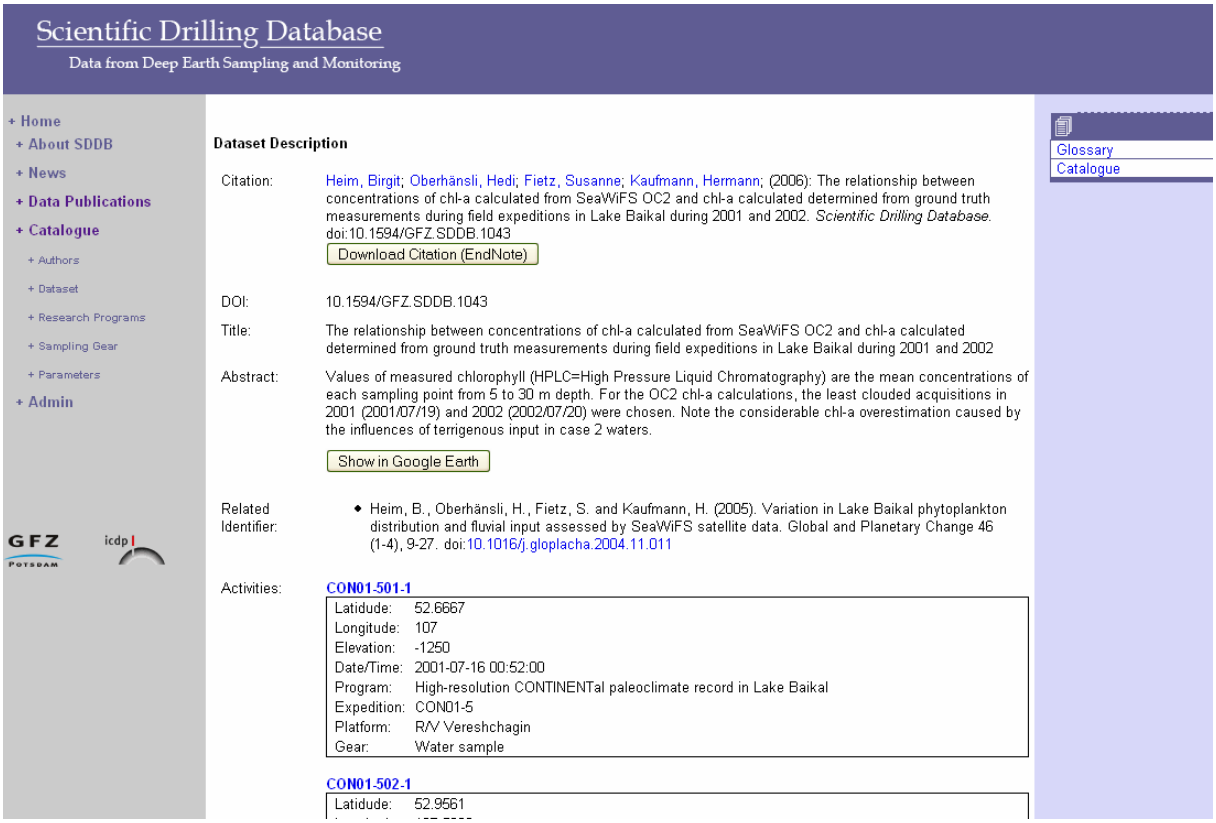

Figure 2: *Screenshot of a data citation in SDDB. Note the buttons for download of citation into a reference manager and for the visualization of sampling locations in Google Earth. The Digital Object Identifier of this dataset is doi:10.1594/GFZ.SDDB.1043*

In the project STD-DOI, the TIB acts as a registration agency for persistent identifiers. For every data publication, it requests a set of metadata to be incorporated into the library catalogue. The data sources are the participating World Data Centers in Germany, WDC-MARE (Bremen/Bremerhaven), WDC Climate (Hamburg), WDC-RSAT (Oberpfaffenhofen and the proposed WDC-TERRA (Potsdam). The data centres act as registration agents for scientific and technical data DOIs. These data centres are also responsible for technical quality control in their data domains, at the same time they also act as long-term archives. The project participants thus encompass all functions necessary for the publication of scientific data.

On May 1st 2005 the TIB became the world's first DOI registration agency for scientific primary data, working in cooperation with the World Data Center Climate (WDCC) at the Max Planck Institute for Meteorology Hamburg, GeoForschungsZentrum Potsdam, World

Data Center for Marine Environmental Sciences (WDC-MARE) at the Alfred Wegener Institute Bremerhaven and at the University of Bremen and technically advised by the Research Center L3S Hannover. Through this project, the foundations have been laid for a system of scientific data publication.

Information discovery through semantic linking

In many online scientific databases, the access to data is through some kind of search interface, which may consist of a simple field for the entry of keywords or may offer more elaborate search criteria. However, in most cases database users do not know the precise contents of a database, yet no database is so perfectly comprehensive to offer datasets matching any search query. Database users would therefore rather "browse" the contents like in a catalogue. In the design of SDDB we therefore took a very different approach to other databases and designed the graphical user interface to contain dynamically generated catalogue listings and cross-links, e.g. between datasets and sample material, datasets and authors, datasets and parameters, etc., to allow the database user to browse through the SDDB contents (Figure 3).

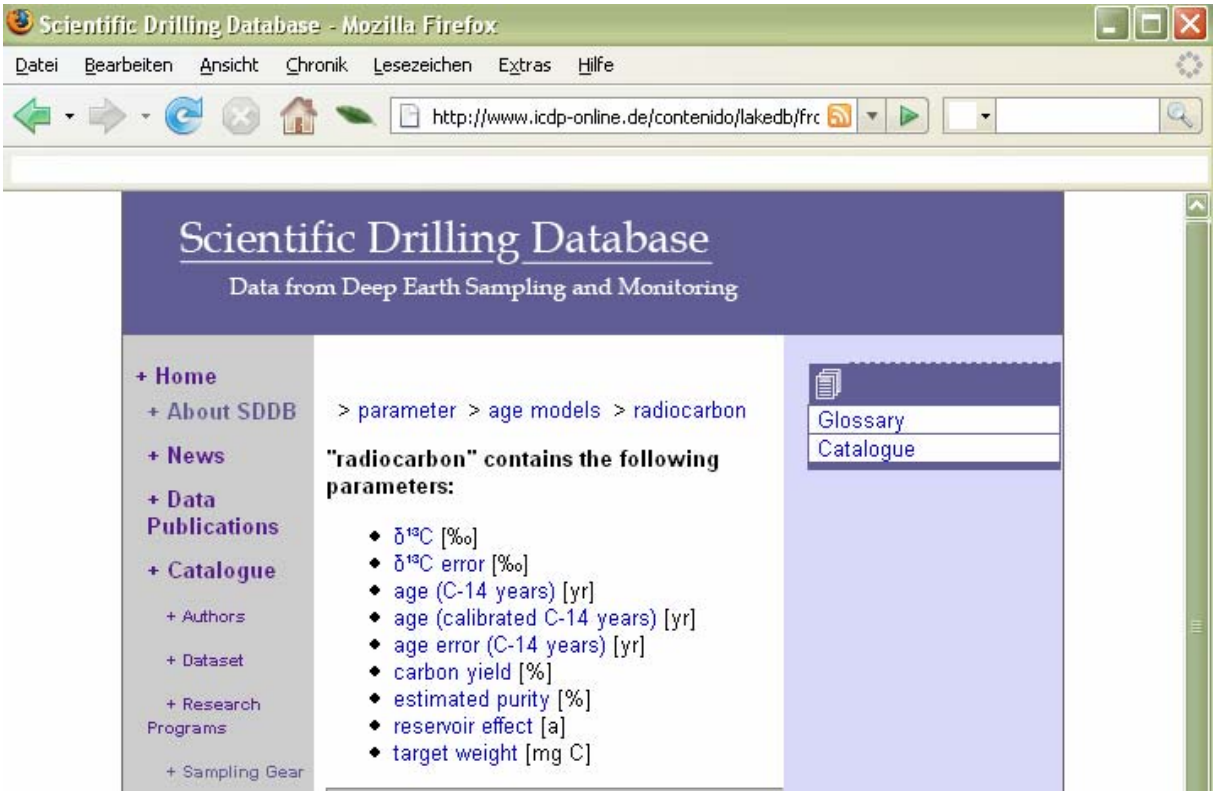

Figure 3: *Screenshot of SDDB showing a list of analytical parameters from the context of radiocarbon dating of fossil materials. All parameter names and their parent categories are active links that can be used to browse the contents of SDDB by analytical parameters. Other categories, like authors of data sets, or instruments used for sampling, can be browsed in a similar way.*

Any sample taken in the field is taken from a geographical and geological context. Not every database user will be familiar with the locality from where the sample was taken. Visualization of this sampling context helps to assess whether the offered datasets are useful to the particular question asked, or which subset of data to choose. At present this geographical visualization primarily shows the sampling positions in their geographical

context. Virtual globes, such as Google Earth, are useful and intuitive tools for geographical visualization (Butler, 2006; Lyon et al., 2006). To show the sampling locations in Google Earth, SDDB offers a kml-file for download with every dataset (Figure 2). The kml-file can be automatically imported as 'place marks' and viewed in Google Earth (Figure 4). The ‘place marks’ are interactive and act as links back to SDDB.

In a second step we plan to add more specific maps which will be displayed alongside the data or as separate maps in an online geographical information system. This information system will allow the dynamic generation of customised maps through a standardised Web Map Service. The elements of this service, i.e. the thematic map layers will are also published as data publications for re-use by other researchers (Heim et al., in press).

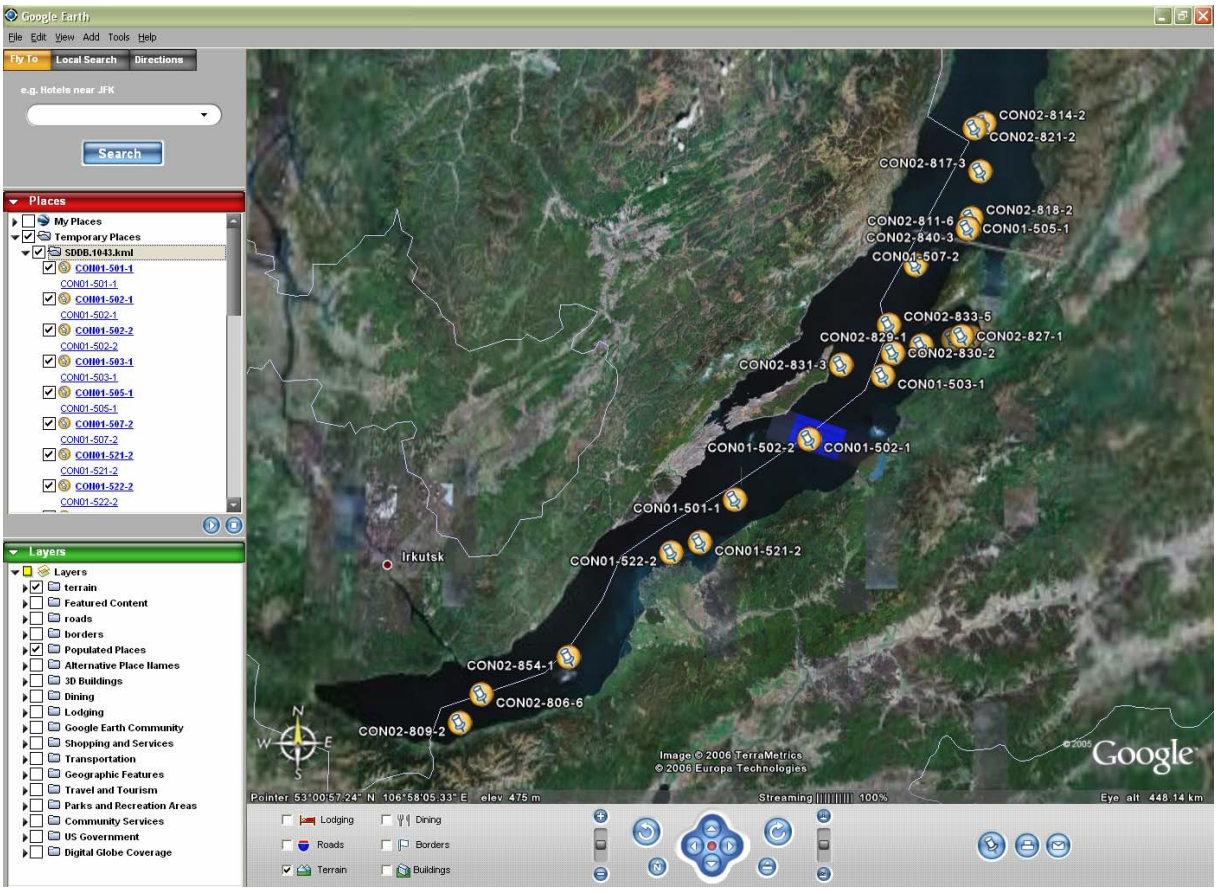

Figure 4: *Sampling locations of doi:10.1594/GFZ.SDDB.1043 by (Heim et al., 2006) displayed in the Google Earth virtual globe. The place marks at the sampling locations link back to an online description of the fieldwork activities at the respective locations as recorded in SDDB.*

Starting in 2008, ICDP will also assign identifiers to physical samples taken in the field or in the laboratory to allow unambiguous identification of sample materials and associated data. This identifier is called an International Geosample Number (IGSN) and based, like DOIs, on the Handle Service®. The service for registration of IGSN will be run in co-operation with the System for Earth Sample Registration (SESAR, http://www.geosamples.org) at Lamont Doherty Earth Observatory of Columbia University (New York, NY) (Lehnert et al., 2006).

Ideally, literature should already reference the materials used and the data derived from these. Since this is not yet done, repositories publishing data and tracking sample material record the literature based on these data and samples in their databases. In the case of the STD-DOI project, its metadata profile includes identifiers of related material, e.g. literature interpreting the data, related datasets, or samples from which the data were derived. These metadata can then be used to create ontologies interlinking literature, data and samples.

The challenging task ahead is that of interlinking literature, data and samples with as little editorial work as possible. Keeping the amount of work small is essential to allow the indexing of the back catalogue of already existing works. A key technology to solve this task is the automatic creation of ontologies, which can be generated automatically by text mining applications. These ontologies can be combined with ontologies generated from reference lists and from metadata.

Conclusions

Data sharing is a key strategy in ICDP's data management and technical development over the years has aimed to make the data management processes as effective as possible. However, systematic data management is still additional work on the scientists for which they have so far not received the necessary recognition. Therefore, scientists need incentives to share data. A possible incentive may be to make data sharing a proper scientific publication. ICDP has implemented the tools necessary for persistent references to published data.

Data will only be re-used if their existence is known. Therefore data publications must be included into library catalogues and scientific portals. They must also be accessible to users that are not aware of the existence of the data or the database. This can be achieved by syndication of metadata among data portals and by semantic linking between literature, data and physical samples from which the data were derived.

The scientific workflow of creating knowledge is slowly adopting the new tools offered by the internet-based information revolution (Berman and Moore, 2006). However, data curation is not at the focus of scientific work and therefore scientists are not willing to invest much of their time in any work related to this task. Therefore it is essential to provide researchers with efficient, easy to use tools that are aligned with the scientific workflow. A high degree of automatisation will take away as much as possible of the workload of data curation and metadata editing from the scientists.

Acknowledgements

The development of the Scientific Drilling Database (SDDB) is funded by the International Scientific Continental Drilling Program (ICDP). The project "Publication and Citation of Scientific and Technical Data" (STD-DOI) is supported by the German Science Foundation, Libraries and Information Systems (DFG-LIS).

Literature

Alexander, W., Berlin, J., Cyr, P., Schofield, A. and Platt, L., 2004. Realities at the leading edge of research - Good practice and proper conduct in research pay off, scientifically and economically. EMBO Reports, 5(4): 324-329. doi:10.1038/sj.embor.7400137

Arzberger, P. et al., 2004. Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data Science Journal, 3: 135-152. doi:10.2481/dsj.3.135

Berlin Declaration, 2003. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities: Berlin, Germany. http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html

Berman, F. and Moore, R.W., 2006. Designing and Supporting Data Management and Preservation Infrastructure. CTWatch Quarterly, 2(2): 7. http://www.ctwatch.org/quarterly/articles/2006/05/designing-and-supporting-data-management-and-preservation-infrastructure/

Brase, J., 2004. Using Digital Library Techniques - Registration of Scientific Primary Data. In: M. Jones, E.A. Fox and R. Shen (Editors), Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science. Springer-Verlag, Heidelberg, Germany, pp. 488-494.

Butler, D., 2006. Virtual globes: The web-wide world. Nature, 439(7078): 776-778. doi:10.1038/439776a

Conze, R., Wallrabe-Adams, H.-J., Graham, C. and Krysiak, F., 2007. Joint ICDP and IODP Data Management for Scientifc Drilling Expeditions. Scientific Drilling, 4: 32-33. doi:10.2204/iodp.sd.4.07.2007

DFG, 1998. Regeln guter wissenschaftlicher Praxis, Deutsche Forschungsgemeinschaft. http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf

Dittert, N., Diepenbroek, M. and Grobe, H., 2001. Scientific data must be made available to all. Nature, 414(6862): 393. doi:10.1038/35106716

Heim, B., Klump, J., Fagel, N. and Oberhänsli, H., in press. Assembly and concept of a web-based GIS within the paleolimnological project CONTINENT (Lake Baikal, Siberia). Journal of Paleolimnology. doi:10.1007/s10933-007-9131-0

Heim, B., Oberhänsli, H., Fietz, S. and Kaufmann, H., 2006. The relationship between concentrations of chl-a calculated from SeaWiFS OC2 and chl-a calculated determined from ground truth measurements during field expeditions in Lake Baikal during 2001 and 2002. Potsdam, Germany. doi:10.1594/GFZ.SDDB.1043

Helly, J., Staudigel, H. and Koppers, A., 2003. Scalable models of data sharing in Earth sciences. Geochemistry, Geophysics, Geosystems - G (super 3), 4(1): 14. doi:10.1029/2002GC000318

Klump, J. and Conze, R., 2007. The Scientific Drilling Database (SDDB) - Data from Deep Earth Monitoring and Sounding. Scientific Drilling(4): 30-31. doi:10.2204/iodp.sd.4.06.2007

Lehnert, K., Vinayagamoorthy, S., Djapic, B. and Klump, J., 2006. The Digital Sample: Metadata, Unique Identification, and Links to Data and Publications. EOS, Transactions, American Geophysical Union, 87(52, Fall Meet. Suppl.): Abstract IN53C-07. http://www.agu.org/meetings/fm06/fm06-sessions/fm06_IN53C.html

Lyon, S.W., Lembo, A.J., Walter, M.T. and Steenhuis, T.S., 2006. Internet Mapping Tools Make Scientific Applications Easy. EOS, Transactions, American Geophysical Union, 87(38): 386. http://www.agu.org/journals/eo/eo0638/2006EO380003.pdf

NIH, 2003. Final NIH Statement on Data Sharing. NOT-OD-03-032, National Institute of Health, Bethesda, MD. http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

OECD, 2004. Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communiqué, Organisation for Economic Co-operation and Development, Paris, France. http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

OECD, 2006. Recommendation of the Council concerning Access to Research Data from Public Funding, Organisation for Economic Co-operation and Development, Paris,

France. http://webdomino1.oecd.org/horizontal/oecdacts.nsf/Display/3A5FB1397B5ADFB7C12572980053C9D3?OpenDocument

STD-DOI, 2003. Publication and Citation of Scientific Primary Data. 2005(2005-09-20): Hamburg, Germany. http://www.std-doi.de