GeoResJ 6 (2015) 98–105

Contents lists available at ScienceDirect

GeoResJ

journal homepage: www.elsevier.com/locate/GRJ

Curating the web's deep past – Migration strategies for the German Continental Deep Drilling Program web content

Jens Klump [a,*], Damian Ulbricht [b], Ronald Conze [b]

[a] *Commonwealth Scientific and Industrial Research Organisation, 26 Dick Perry Avenue, Kensington, WA 6151, Australia*
[b] *GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany*

ARTICLE INFO

Article history:
Received 22 October 2014
Revised 20 February 2015
Accepted 21 February 2015
Available online 21 March 2015

Keywords:
Data rescue
Long-term preservation
Data curation
Metadata
Data publication
Continental Deep Drilling

ABSTRACT

On timescales beyond the life of a research project, a core task in the curation of digital research data is the migration of data and metadata to new storage media, new hardware, and software systems. These migrations are necessitated by ageing software systems, ageing hardware systems, and the rise of new technologies in data management. Using the example of the German Continental Deep Drilling Program (KTB) we outline steps taken to keep the acquired data accessible to researchers and trace the history of data management in KTB from a project platform in the early 1990ies through three migrations up to the current data management platform. The migration steps taken not only preserved the data, but also made data from KTB accessible via internet and citable through Digital Object Identifier (DOI). We also describe measures taken to manage hardware and software obsolescence and minimise the amount of maintenance necessary to keep data accessible beyond the active project phase. At present, data from KTB are stored in an Open Archival Information System (OAIS) compliant repository based on the eSciDoc repository framework. Information packages consist of self-contained packages of binary data files and discovery metadata in Extensible Mark-up Language (XML) format. The binary data files were created from a relational database used for data management in the previous version of the system, and from websites generated from a content management system. Metadata are provided in DataCite, GCMD-DIF, and ISO19139/INSPIRE schema definitions. Access to the KTB data is provided through download pages which are produced by XML transformation from the stored metadata.

1. Introduction

Speaking to people about long-term curation, one thing that is mentioned in almost all conversations is the floppy disk. To many, the floppy disk epitomises what they see as the core challenges in long-term data curation: bit stream preservation and media obsolescence. It is not only this particular medium that is problematic; the general still rapid development of information technology requires regular migrations of content, media, hardware and software. These challenges have been recognised early on and a very well written overview can be found in Rothenberg [21].

Many analytical data in the geosciences can be represented as tables and encoded as character separated value (CSV) files. Accompanied by descriptive metadata, these files pose a relatively minor challenge to format migration and their often small size does not demand large computing resources for migration processes. The challenge lies in the system migrations and describing the contents in metadata for discovery and reuse.

Initially, the web based components of projects in the 1990s were solitary systems, today often termed "silos", run in the context of large projects or as efforts by individual researchers. This is also true for most projects of the International Continental Scientific Drilling Program (ICDP). Notable exceptions are systems like PANGAEA [6] which, from going online in 1995, curates and disseminates data from many different projects in marine environmental research. Project based systems all face the challenge of curating the data long past the end of the project when resources, such as contextual knowledge of the project and funding, may no longer be available. In this sense, this paper does not describe the rescue of data that might have been lost to media obsolescence or had to be digitized from analogue media, but rather the challenges posed by technical obsolescence. In the course of this paper we will discuss the strategies employed in successive projects over 25 years to migrate the data dissemination platform of the German

[*] Corresponding author. Tel.: +61 8 64368828.
E-mail addresses: jens.klump@csiro.au (J. Klump), damian.ulbricht@gfz-potsdam.de (D. Ulbricht), ronald.conze@gfz-potsdam.de (R. Conze).

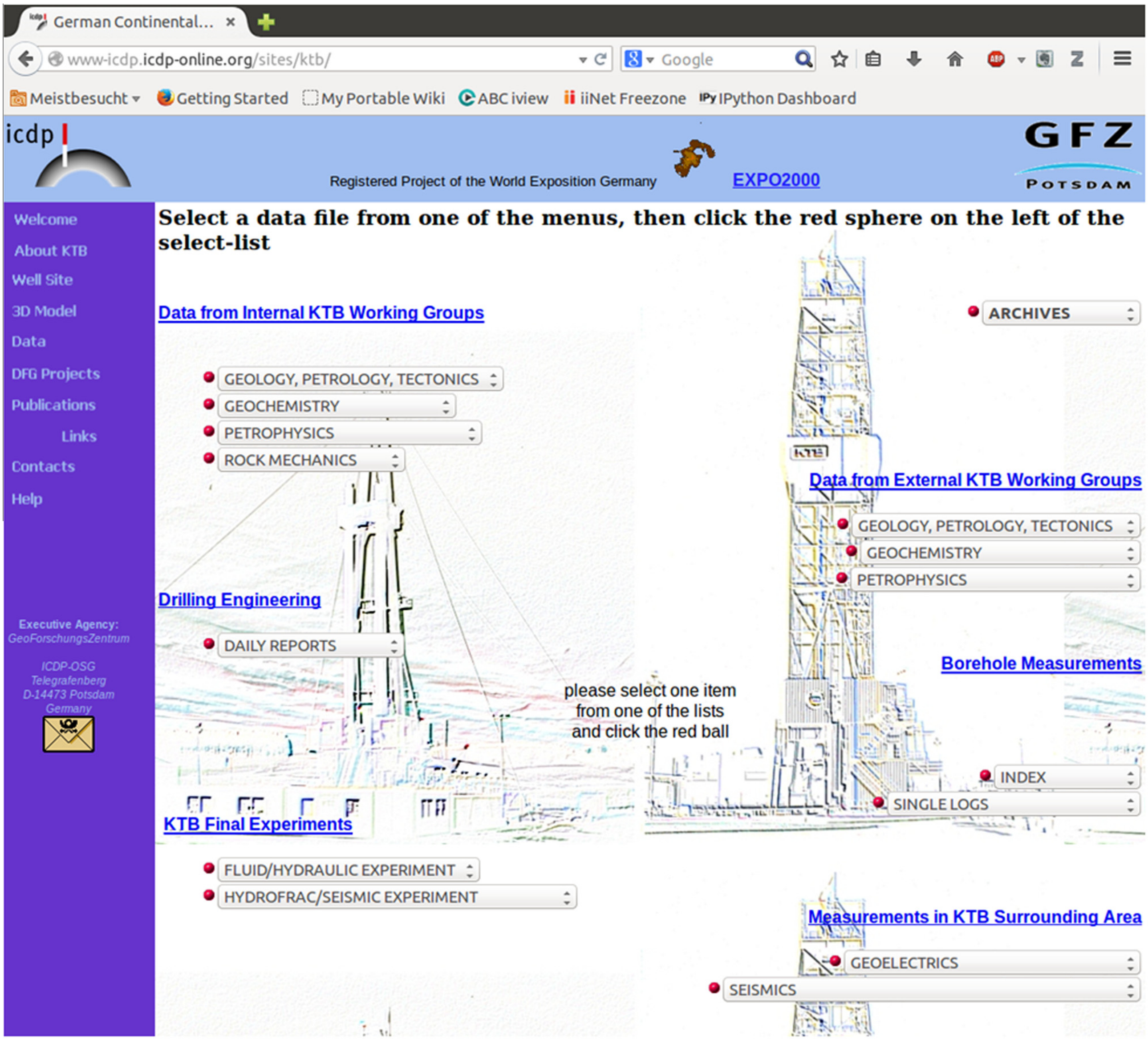
http://dx.doi.org/10.1016/j.grj.2015.02.011
2214-2428/© 2015 Elsevier Ltd.

Continental Deep Drilling Program onto new technical platforms. Unlike the ocean drilling programmes, ICDP drilling projects do not have a common structure and differ widely in form and extent of involvement of the ICDP Operational Support Group. This heterogeneity makes it difficult to apply the same approach to all ICDP legacy projects. The scientific review of ICDP in 2014 recommended following the data curation procedures developed for KTB and CONTINENT in future ICDP projects.

2. Pre-web and web KTB

The German Continental Deep Drilling Program (Kontinentale Tiefbohrung, KTB) was a large scale geoscience project conducted from 1987 to 1995 in Windischeschenbach, Germany. Its two super-deep boreholes (4000 m and 9101 m) are worldwide unique masterpieces of drilling engineering. The programme yielded essential insights in the structure and processes of the upper crust of the Earth. For this reason it is one of the most important geoscientific and geotechnical research projects. The great success in geosciences and drilling engineering induced the scientists to establish the International Continental Scientific Drilling Program (ICDP). An overview of the scientific achievements from KTB can be found in Emmermann and Lauterjung [8].

The KTB Information System was set up to perform two major functions in the information management in the context of KTB: (1) document and store project data, and (2) support interdisciplinary dissemination of the data. Data were originally stored on tape and at the end of the KTB project migrated from tape storage onto optical storage media. Already in times before the worldwide web data were accessible over the German Research Network (Deutsches Wissenschaftsnetz, WiN), a precursor system to the worldwide web based on the X.25 network protocol. Wächter [26] gives a comprehensive overview of the system at the KTB site during the peak of its operation. After the end of the project, and with the emergence of the modern internet, significant proportions of the data were ported to a web application with a browser based user interface (Fig. 1). The focus of this first migration was on tabulated data. Raster data, such as images of drill cores, and seismic exploration data were deemed as being too large in volume at the time and were stored offline. This study will only discuss the original



Select a data file from one of the menus, then click the red sphere on the left of the select-list

please select one item from one of the lists and click the red ball

Fig. 1. Screenshot of the first KTB web presence, data access page.

KTB web portal and its migration to new technical platforms. Photographs and optical scans of drill cores were not included.

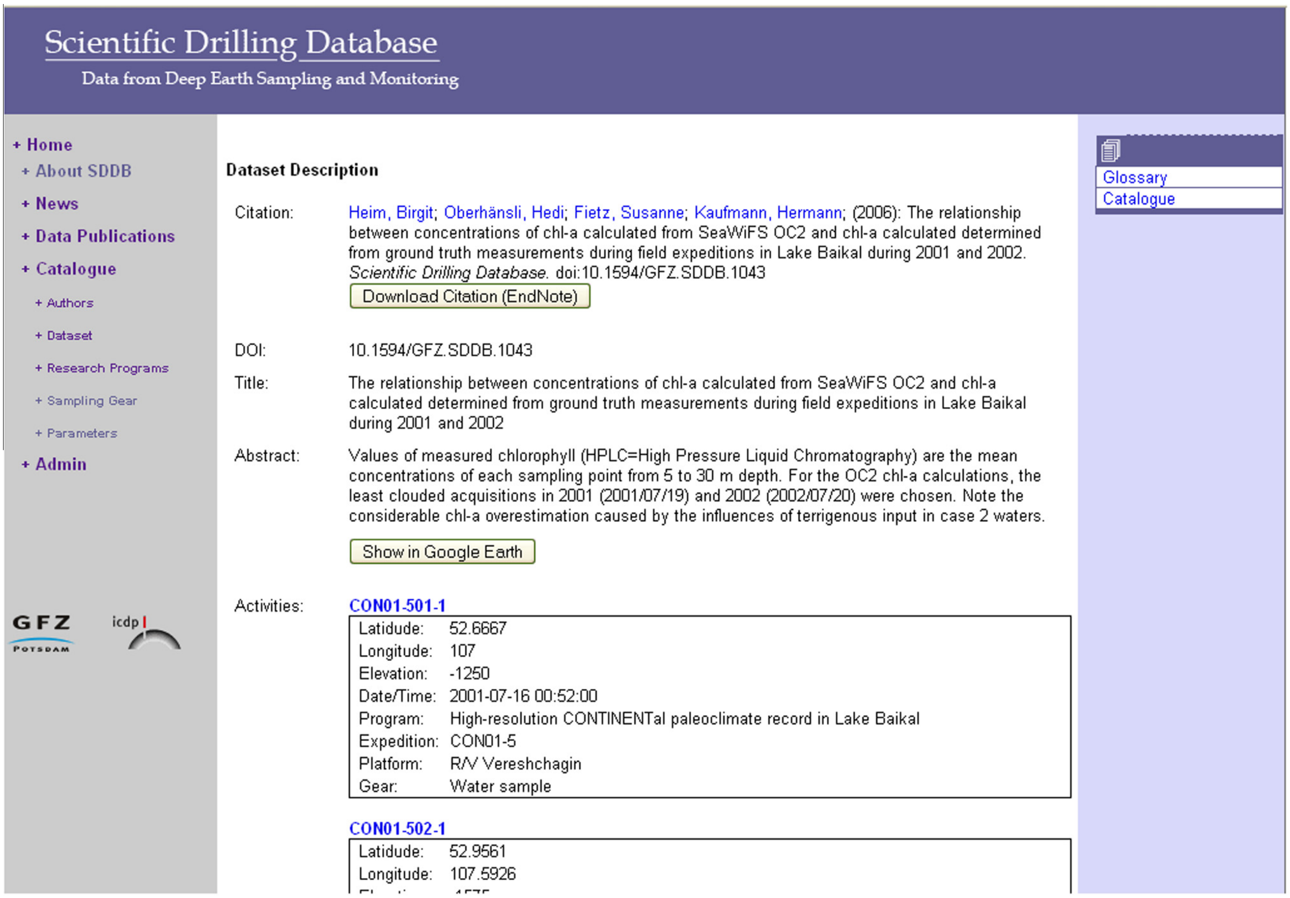
3. Scientificdrilling.org

The KTB web presence, as outlined above, became the template for the web presence of other ICDP projects. Even though based on templates, the web pages had to be maintained manually. With an increasing number of ongoing and completed projects the web pages became increasingly difficult to maintain. Also, ICDP projects up to 2001 were centred on single drill holes with limited activity outside the immediate vicinity of the drill site. These requirements changed with the start of a number of lake drilling projects supported by ICDP at the German Research Centre for Geosciences (GFZ), the first being the EU project "High-resolution CONTINENTal paleoclimate record in Lake Baikal" (CONTINENT). An overview of this project can be found in Oberhänsli and Mackay [18].

The project CONTINENT had many drill sites and a host of supporting investigations in a large field area across central Siberia and north-western Mongolia. Its multi-disciplinary and distributed nature required a departure from the data model used in previous projects. The data from field and laboratory measurements were stored in a single table following the data warehouse model [2]. This new data model followed the example of PANGAEA as a fully normalised relational model [6]. Copies of the incoming data tables were stored as files on the GFZ data backup system, the data were uploaded through a staging system and manually annotated with metadata using web forms.

As ICDP evolved, more and more projects needed their own web presence. To reduce the work required to maintain many concurrent project web sites we introduced a content management system (CMS). A CMS separates data storage, editorial texts, web based user interface and business logic. For the ICDP projects the separation of data entry and web editorial work from software and hardware maintenance allowed a clear separation of tasks between scientific and technical staff. The CMS was required to offer parallel client spaces to allow independent management of the web content and access rights for each project. Our choice for a CMS fell on the open source product Contenido, a project initiated by *four for business AG*. The project is still active and can be found at http://www.contenido.org. The CMS software is based on PHP with a mySQL database for the business logic. Analytical data from the project and metadata were stored in a Sybase database. The Sybase database was chosen because it was the database management system offered by the GFZ Computing Centre at the time and it had to be shared with other projects. The rendering of the project web pages took place on a dedicated web server. Fig. 2 shows an example of a landing page for a dataset rendered by the Contenido CMS in the first version of SDDB.

Concurrent with the project CONTINENT, procedures and technologies for publication and citation of research data using Digital Object Identifiers (DOI) were developed in the context of the project "Publication and Citation of Scientific Data" (STD-DOI) [3]. GFZ became one of the first publication agents for data publications using DOI through the German National Library for Science and Technology (TIB Hannover). This ICDP data portal, which included the data from KTB, became known as the Scientific Drilling Database (SDDB). A more detailed description of the SDDB can be found in Klump and Conze [13].



Dataset Description

Citation: Heim, Birgit; Oberhänsli, Hedi; Fietz, Susanne; Kaufmann, Hermann; (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. *Scientific Drilling Database*. doi:10.1594/GFZ.SDDB.1043

DOI: 10.1594/GFZ.SDDB.1043

Title: 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

Abstract: Values of measured chlorophyll (HPLC=High Pressure Liquid Chromatography) are the mean concentrations of each sampling point from 5 to 30 m depth. For the OC2 chl-a calculations, the least clouded acquisitions in 2001 (2001/07/19) and 2002 (2002/07/20) were chosen. Note the considerable chl-a overestimation caused by the influences of terrigenous input in case 2 waters.

Activities: CON01-501-1

Latidude:	52.6667
Longitude:	107
Elevation:	-1250
Date/Time:	2001-07-16 00:52:00
Program:	High-resolution CONTINENTal paleoclimate record in Lake Baikal
Expedition:	CON01-5
Platform:	R/V Vereshchagin
Gear:	Water sample

CON01-502-1

Latidude:	52.9561
Longitude:	107.5926

Fig. 2. Screenshot of the Scientific Drilling Database (SDDB) showing supplementary data to a publication from the project CONTINENT in the first version of SDDB. The dataset in this example is identified by doi:http://dx.doi.org/10.1594/GFZ.SDDB.1043.

By this time, KTB had been completed many years earlier and it became uncertain how much longer the original KTB web presence could be maintained. We therefore decided to migrate all data available from the original website to the SDDB and publish the KTB datasets. The migration was an entirely manual process of downloading data from the original website and uploading them to the SDDB, adding metadata in the process.

The introduction of a CMS did reduce system maintenance for serving many projects, but not all anticipated benefits materialised. In practice it turned out that scientists still found the web editing and data upload services offered by the system too complicated. Data upload and metadata editing was primarily done by student assistants under supervision from scientific staff involved in ICDP drilling projects.

One of the reasons to depart from using the Sybase database was that a fully normalised database has a rigid data model that does not easily adapt to changes in the research processes and thus may slowly degrade by having more and more features added to it. Our experience also showed that researchers were interested mainly in downloading datasets but were not using any of the features a relational database management system (RDBMS) can offer. Therefore we decided to part from a RDBMS and move to a file based data storage and management.

To make online data mining for researchers useful, the data need to have something in common. While in ocean scientific drilling derives a lot of information from summarising data across many sites, the continental drill sites were too diverse to offer any meaningful application for data mining across drill sites. At the same time, the types of data collected in projects became more and more diverse and many data types were no longer adequately represented in the SDDB data model.

Eventually, the computer hardware, on which the SDDB was running, became rare and it was impossible to obtain spare parts, the hardware architecture became obsolete and its operating system was no longer maintained. To avoid data loss we had the option to migrate our customised Contenido installation to recent hardware and software. Another option was to rethink the data management, since most of the SDDB functionality was not used. Despite the initial gains from the introduction of a CMS and the data warehouse model it became apparent that the next system migration could not be avoided.

4. Migration to eSciDoc

In the management of institutional data centres the age of dedicated hardware servers for individual projects came to an end and was being replaced by hardware virtualization. By the year 2010 the number of projects managed at GFZ reached approximately 100 concurrent projects, 25 new projects starting every year and an equal number of projects ending every year. The challenge was how to provide data management systems for as many projects as required, and what to do with the systems of projects that had ended. Maintaining individual data silos after the end of a project was beyond the means of the GFZ infrastructures and a solution had to be found. An additional challenge arose from the question how to ensure the long-term availability of these data, which had all been gathered at great cost and were not recoverable by repetition of the measurements.

In order to achieve better conditions for maintaining the system components and for a greater persistence we decided to modularise the data management infrastructure. The conceptual model behind this modularisation is to divide the data management space into “domains of responsibility”. [23] found that in large, heterogeneous organisations, such as universities or research centres, the requirements towards research data infrastructures are broad and could be described as a “Data Curation Continuum”. Since a continuum cannot be managed in an organisation, Treloar et al. suggested dividing this continuum into “Domains of Responsibility”. These “Domains of Responsibility” in research data management not only help to delineate the responsibilities of the actors involved but also outline the context of shared knowledge about data. The domains are characterised by different degrees of shared contextual knowledge. The more of the context we share, the less metadata we need to understand the data. In this way, the “domains” help to determine how detailed descriptive metadata need to be in order to allow the reuse of research data and at which point implicit contextual metadata need to be encoded into the stored metadata. In addition, the structure of this model is conformable with the reference model for Open Archival Information Systems (OAIS) [4], while other commonly used models of the data curation lifecycle (e.g. DCC) [5] make it more difficult to delineate areas of responsibility.

In the case of GFZ we divided the data management space into four “domains” (Fig. 3) [12]. The “Private Domain” is the domain of the individual researcher. Little metadata are needed because the researcher has all contextual information. As the researcher shares data with collaborators, not all contextual knowledge is shared by all individuals in the group. On transfer into the “Group Domain” some metadata need to be added to make this context more explicit. This process can be formalised in research data management systems such as PMD [14]. The transition from the “Group Domain” to the “Persistent Domain” is the most critical. Here, almost all contextual knowledge is lost over time, or not accessible for reuse of data. Technical solutions can be used to add standardised metadata to the research data to facilitate discovery and reuse. The “Access Domain” provides discovery mechanisms, regulates access and may provide additional services in the data access process.

Crucial in this model is the transfer of data from one domain into another. In some cases it is possible to transform contextual information and existing metadata into standardised metadata

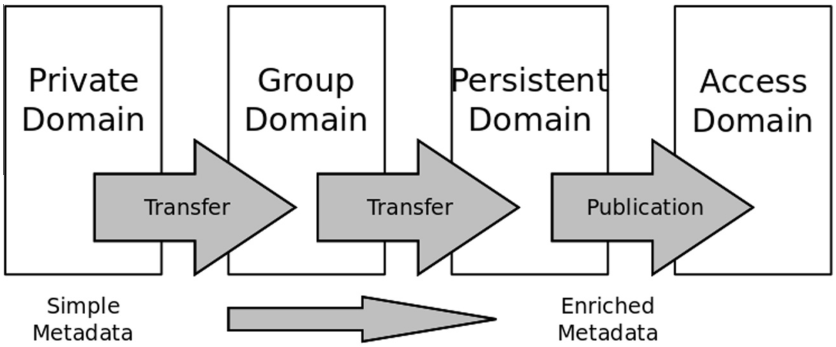


Fig. 3. “Domains of Responsibility” in research data management. The domains not only help to delineate the responsibilities of the actors involved but also outline the context of shared knowledge about data and how detailed descriptive metadata need to be in order to allow the reuse of research data.

for discovery and reuse. In many cases this transfer does require effort from scientists involved in the project and information specialist to enrich the metadata.

4.1. Designing a persistent system

As mentioned earlier, for an institution managing many projects it becomes impossible to preserve the hardware and software stacks of projects long after the projects have ended. To find a solution to this challenge we analysed the lifecycle of research data infrastructures and saw that the most complex and specific software was used during data entry. After the end of the project, this component becomes obsolete. On the other hand, the software requirements for data dissemination after the end of the project are fairly generic.

Extending on the concept of the multi-client capability of the Contenido CMS we decided to introduce a service oriented architecture that would help us to further modularise the research data infrastructure and allow us to separate data ingest, data storage and data dissemination and link the different components through common organisation wide data infrastructure. Our choice fell on the software eSciDoc, a joint project by the Max Planck Digital Library and FIZ Karlsruhe [19].

At GFZ eSciDoc is used as institutional repository for publications and data. The software offers a representational state transfer (REST) interface [9] to store binary file-based data and characterise the data with Extensible Markup Language (XML) based metadata in so-called "eSciDoc items". Each item can be composed of an arbitrary number of data files as well as an arbitrary number of metadata records, allowing, for instance, to store different representation forms of the same item, but associated with specific metadata. Since metadata schemas tend to have a complementing set of information and an intersecting set of information we are able to describe datasets in more detail than by using only one schema, but at the cost of storing redundant metadata for the intersecting part. The service oriented modularisation on top of an enterprise service bus enabled us to provide an infrastructure that allows us to map the domains of responsibility for research data. In this way data and metadata can be added when the dataset is transferred between domains of responsibility.

eSciDoc offers detailed management of access rights to its objects which make use of internal and external user authentication methods. It supports the basic methods of creating, reading, updating and deleting objects (CRUD services) and offers version control mechanism. It also offers a service for controlled names and entities (CONE services). The purpose of this service is to provide methods to deal with controlled lists of named entities to assure data quality and facilitate data access and data entry. All items in eSciDoc are identified by a unique identifier. At GFZ, all items created and edited through eSciDoc can be entered into a publishing workflow and issued with a DOI [24]. This workflow is designed to interface with the workflow for publication of research manuscripts.

4.2. Data and metadata

The eSciDoc content model requires self-contained packages made of file-based binary data and XML metadata. Data and metadata had to be extracted and transformed into elements of eSciDoc items in the process of migration from the SDDB database structure to the eSciDoc content model. Fig. 4 illustrates the flow of data and metadata in the migration of SDDB from the Contenido CMS to the eSciDoc data management system.

SDDB landing pages for data files offered data in several different download formats that were generated on the fly by a PHP-script from the values stored in the database. For maximum interoperability with software of researchers that download the data, we decided to export CSV files from SDDB. The CSV format has a simple structure and is on the list of safe preservation formats published by the Library of Congress [15].

From the wide range of metadata schemas available we decided to use only discovery metadata schemas, since our exported CSV files contain a header with additional reuse information. Our discovery metadata schemas are GCMD-DIF [10], the INSPIRE profile of ISO19139 [7] and DataCite [22]. GCMD-DIF is popular for its well curated vocabulary – in particular the science parameters – and for

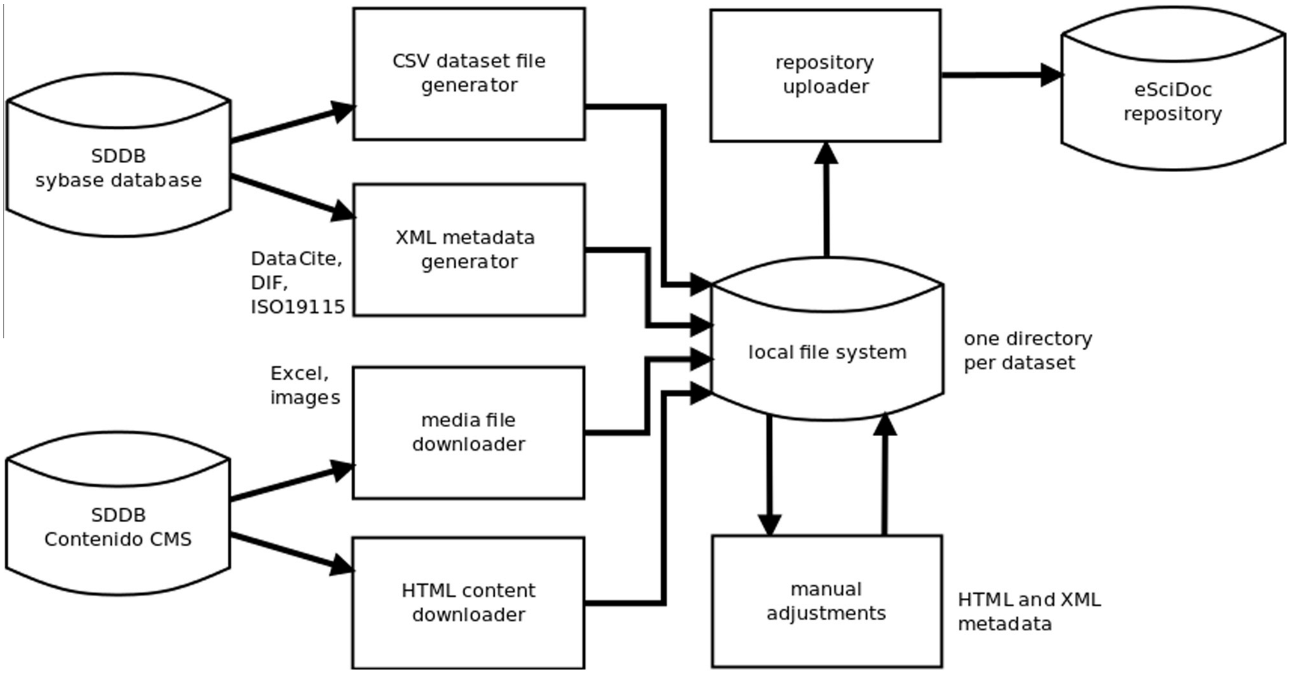
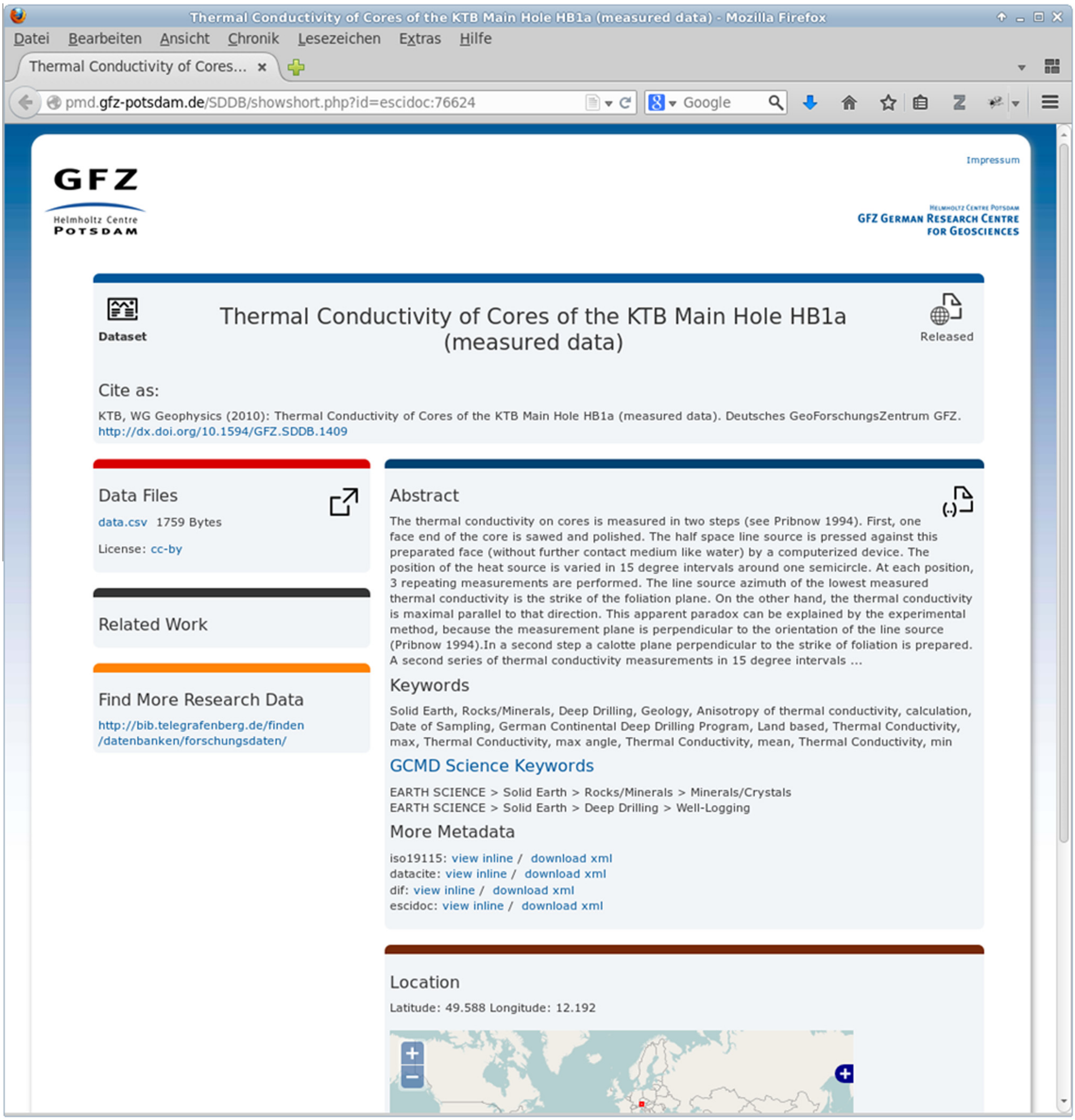


Fig. 4. Flowchart of data and metadata in the migration process of SDDB from Contenido CMS to eSciDoc.

its simple XML structure, which makes it easy to classify and navigate datasets and for software developers to add dissemination and harvesting functions to software systems. The XML structure of ISO19139 is much more complex and the US National Oceanic and Atmospheric Administration and the Australian Antarctic Data Centre have published Extensible Stylesheet Language Transformation (XSLT) files [1,17] to automatically transform GCMD-DIF to ISO19139 metadata. In our case, this simple transformation could not be used to generate ISO metadata since the European INSPIRE directive requires metadata elements that are not present in GCMD-DIF and had to be added in an editorial process. In addition to the aforementioned formats, metadata are also stored in the DataCite schema for registration of DOIs through DataCite.

While operating the SDDB at scientificdrilling.org we investigated ways of exporting metadata from the relational database underlying the SDDB to populate the GCMD-DIF schema. This was done primarily to be able to disseminate the metadata to data portals and thus address a broader audience through metadata syndication. The result was a direct mapping from distinct database table rows to GCMD-DIF entities which was possible because GCMD-DIF already informed the design of the original database schema. The extraction of GCMD science keywords could be automated to a large degree by extracting entries from the database columns defining analytical methods and parameters measured. Some editorial work was required because parameters and methods did not always map unambiguously to the science keywords, e.g. the SDDB did not encode explicitly whether the investigated


Thermal Conductivity of Cores of the KTB Main Hole HB1a (measured data) - Mozilla Firefox
Datei
Bearbeiten
Ansicht
Chronik
Lesezeichen
Extras
Hilfe
Thermal Conductivity of Cores... ×
pmd.gfz-potsdam.de/SDDB/showshort.php?id=escidoc:76624
Google
Impressum
GFZ
Helmholtz Centre
POTSDAM
Helmholtz Centre Potsdam
GFZ German Research Centre for Geosciences
Dataset
Released

Thermal Conductivity of Cores of the KTB Main Hole HB1a (measured data)

Cite as:

KTB, WG Geophysics (2010): Thermal Conductivity of Cores of the KTB Main Hole HB1a (measured data). Deutsches GeoForschungsZentrum GFZ.
http://dx.doi.org/10.1594/GFZ.SDDB.1409

Data Files

data.csv 1759 Bytes

License: cc-by

Related Work

Find More Research Data

http://bib.telegrafenberg.de/finden/datenbanken/forschungsdaten/

Abstract

The thermal conductivity on cores is measured in two steps (see Pribnow 1994). First, one face end of the core is sawed and polished. The half space line source is pressed against this preparated face (without further contact medium like water) by a computerized device. The position of the heat source is varied in 15 degree intervals around one semicircle. At each position, 3 repeating measurements are performed. The line source azimuth of the lowest measured thermal conductivity is the strike of the foliation plane. On the other hand, the thermal conductivity is maximal parallel to that direction. This apparent paradox can be explained by the experimental method, because the measurement plane is perpendicular to the orientation of the line source (Pribnow 1994).In a second step a calotte plane perpendicular to the strike of foliation is prepared. A second series of thermal conductivity measurements in 15 degree intervals ...

Keywords

Solid Earth, Rocks/Minerals, Deep Drilling, Geology, Anisotropy of thermal conductivity, calculation, Date of Sampling, German Continental Deep Drilling Program, Land based, Thermal Conductivity, max, Thermal Conductivity, max angle, Thermal Conductivity, mean, Thermal Conductivity, min

GCMD Science Keywords

EARTH SCIENCE > Solid Earth > Rocks/Minerals > Minerals/Crystals
EARTH SCIENCE > Solid Earth > Deep Drilling > Well-Logging

More Metadata

iso19115: view inline / download xml
datacite: view inline / download xml
dif: view inline / download xml
escidoc: view inline / download xml

Location

Latitude: 49.588 Longitude: 12.192

Fig. 5. Screenshot of KTB data served by the current version of the SDDB. The example in this figure is identified by doi:http://dx.doi.org/10.1594/GFZ.SDDB.1409.

material was derived from land or from a lake. This information was only implied by the context of the data, e.g. the location of where the sample was taken. Another problem was the missing convention for "not applicable" in the database. Thus, fields describing methods or platforms sometimes contained empty strings, "other", or "not specified". The alternatives for "not applicable" used in the database were limited to these three cases that could easily be covered by the conversion script.

4.3. Data in web Pages

In addition to numerical data in the database, the web CMS Contenido was used to present individual datasets with research data stored in Microsoft Excel sheets, ESRI shapefiles, and images from drilling projects other than KTB. The metadata for these datasets were not stored in the database and GCMD-DIF and ISO19139 metadata had to be written from scratch. While the Excel sheets could be attached to the metadata inside the eSciDoc repository similar to the CSV files before, the situation for the shapefiles and images was different. The shapefiles and images were used in the past as elements of a Web-GIS [11] of the Lake Baikal region based on ESRI ArcIMS. Snapshots of the rendered maps were used as illustrative previews of the shape file contents and incorporated into web pages describing the data.

Since this visualization supports the reuse of these data, it was decided to also migrate these Hypertext Markup Language (HTML) formatted contents into the eSciDoc repository. This required additional manual work to create self-contained packages, because HTML is designed to link resources and linked resources are not necessarily self contained. Thus, all presentation pages had to be reviewed for linked entities and had to be customized. Referenced images were integrated into the package containing the HTML page and external links were modified to be persistent identifiers. An excerpt of the HTML page, with all Contenido CMS headers removed, is stored and is now used as default display page for these objects. This procedure follows practices developed for long-term archiving of web pages [16] but is not commonly encountered in the long-term preservation of research data.

4.4. Access and presentation

Currently we store ISO19139, GCMD-DIF and DataCite metadata together with the dataset files inside the eSciDoc items. This way the files are characterized as citable spatial datasets. Using additional metadata schemas the data could be described in more detail in other situations. For presentation to researchers we use XSLT to assemble information from the XML metadata, such as keywords, spatial information and citation details, and convert it to HTML. Fig. 5 shows a screenshot of the data presentation in its current form. Links are added to download files. Contents of migrated web pages are displayed inside an HTML "iframe". The stylesheet used in these transformations is available for download [25].

4.5. Context and digitisation of KTB reports

A drawback of the CSV format is that information on the data types of table columns is not recorded in the data themselves but has to be explicitly recorded in the header of the file. To make this header information readable in an automated way, it has to be systematically structured. The Full Metadata Format proposal [20] outlines such an easy to read and easy to parse format.

In the current format the data tables stand by themselves with little contextual information and no explanation, unless the data are supplementary material to publications. Where data are supplementary materials to publications, as is the case for most data from the CONTINENT project, this context is recorded in the literature. In the case of the KTB data most context is recorded in the project reports of KTB.

Reports produced during the original deep drilling project were produced as printed reports. GFZ is currently retro digitizing the KTB Report series to make them available to a broader audience via the internet. It is planned to link KTB Reports and KTB data publications through the cross-linking capabilities in the DataCite metadata schema. The DataCite metadata schema features an element <relatedIdentifiers> that facilitates the cross-linking of globally uniquely identifiable resources [22], which will allow cross-linking of data sets with corresponding reports.

5. Conclusions

The long history of data originating from the KTB programme makes this an interesting case study for the long-term curation of research data well beyond the end of the original project. The successive migration processes provided us with valuable insights into the practical aspects of long-term data curation because it spanned a period of intense technological development.

Over the years the storage of data and metadata changed from a project database to databases for internet access and publishing. Currently data, reports, and papers are published and the data are unlikely to be modified – making the maintenance of a database system and the associated software an additional effort. Creating distinct information packages from the relational database and storing the data and metadata inside an eSciDoc repository simplifies the maintenance of software systems for dataset access and reduces archiving costs by making maintenance of legacy data entry software obsolete.

The process of generating the self-contained packages from the web pages generated by the CMS showed that a missing separation between a dataset and its online presentation makes the creation of information packages difficult. Furthermore, hyperlinks from CMS web pages had to be substituted by persistent identifiers (PID) to enable stable references between self-contained packages and make it easier to delineate the content of an information package.

Our approach to the migration of the SDDB data from a relational, fully normalised data model to a file based model does introduce limitations when compared to the original data warehouse model. Following the curation domain model, these limitations are acceptable because in the "persistent domain" of data curation the focus is on data preservation, not on data analysis and processing. On the other hand, migration to a file based data model significantly simplified the repository structure and this supports preservation and future reuse.

In this sense, this paper did not describe the rescue of data that might have been lost to media obsolescence or had to be digitized from analogue media, but rather the challenges posed by technical obsolescence and the strategies employed in successive projects over 25 years to migrate the data dissemination platform of the German Continental Deep Drilling Program onto new technical platforms.

Currently only KTB data and measurements from Lake Baikal are stored in formats that are identified by the Library of Congress as safe for preservation [15]. Future work includes converting proprietary or more complex data formats, such as Microsoft Excel Sheets, ESRI shapefiles, and images into formats that lend themselves to long-term preservation.

Acknowledgements

Initial work on the Scientific Drilling Database was supported by EU project CONTINENT (EVK2-2000-00057). Further development was supported by the International Scientific Continental Drilling Program (ICDP) and by the German Research Centre for Geosciences (GFZ). Implementation of DOI for data publication was supported by the project "Publication and Citation of Scientific Primary Data" (STD-DOI), work on long-term preservation of research data was supported by the project "Development of Workflow Components for Long-term Archiving in the Geosciences" (EWIG), both funded by the German Research Foundation (DFG). Significant insights into the long-term preservation of digital content were gained through our participation in the German competence network for digital preservation (nestor). The authors would also like to thank Jens Ludwig, Matthias Razum, Frank Schwichtenberg and Andrew Treloar for their contributions to our discussions on data management infrastructures. Lastly, we would like to thank the reviewers for their constructive comments that helped us improve our manuscript.

References

[1] AADC (2014), Metadata guide, Australian Antarctic Data Centre – Data management and spatial data services. [online] Available from: <https://data.aad.gov.au/aadc/metadata/#writing> (Accessed 13 October 2014).

[2] Behrends K, Conze R. Das ICDP Data Warehouse. In: Bauer A, Günzel H, editors. Data Warehouse Systeme: Architektur, Entwicklung, Anwendung. Heidelberg, Germany: dpunkt Verlag; 2001. p. 496–502.

[3] Brase J. Using digital library techniques – Registration of scientific primary data. In: Jones M et al., editors. Research and Advanced Technology for Digital Libraries, vol. 3232. Heidelberg, Germany: Springer-Verlag; 2004. p. 488–94. http://dx.doi.org/10.1007/978-3-540-30230-8_44.

[4] CCSDS (2012), *Reference Model for an Open Archival Information System (OAIS). Magenta Book*, Recommendation for Space Data System Practices, Recommended Practice, Consultative Committee for Space Data Systems, Greenbelt, MD. [online] Available from: <http://public.ccsds.org/publications/archive/650x0m2.pdf>.

[5] DCC (2010), *DCC Curation Lifecycle Model*, Resources for digital curators, Digital Curation Centre, Edinburgh, UK. [online] Available from: <http://www.dcc.ac.uk/resources/curation-lifecycle-model>.

[6] Diepenbroek M, Grobe H, Reinke M, Schindler U, Schlitzer R, Sieger R, Wefer G. PANGAEA – an information system for environmental sciences. Comp Geosci 2002;28(10):1201–10. http://dx.doi.org/10.1016/S0098-3004(02)00039-0.

[7] Drafting Team Metadata, and European Commission Joint Research Centre (2010), *INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119*, European Commission Joint Research Centre, Ispra, Italy. [online] Available from: <http://inspire.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf>.

[8] Emmermann R, Lauterjung J. The German Continental Deep Drilling Program KTB: overview and major results. J Geophys Res 1997;102(B8):18179–201. http://dx.doi.org/10.1029/96JB03945.

[9] Fielding, R. T. (2000), Architectural Styles and the Design of Network-based Software Architectures, Ph.D., University of California Irvine. [online] Available from: <http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf>.

[10] Global Change Master Directory (2008), *Directory Interchange Format (DIF) Writer's Guide*, National Aeronautics and Space Administration. [online] Available from: <http://gcmd.nasa.gov/User/difguide/>.

[11] Heim B, Klump J, Fagel N, Oberhänsli H. Assembly and concept of a web-based GIS within the paleolimnological project CONTINENT (Lake Baikal, Siberia). J Paleolimnol 2008;39(4):567–84. http://dx.doi.org/10.1007/s10933-007-9131-0.

[12] Klump, J. (2011), Langzeiterhaltung digitaler Forschungsdaten, in *Handbuch Forschungsdatenmanagement*, edited by S. Büttner, H.-C. Hobohm, and L. Müller, pp. 115–122, Bock + Herrchen, Bad Honnef, Germany. [online] Available from: <http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:kobv:525-opus-2339>.

[13] Klump J, Conze R. The Scientific Drilling Database (SDDB) – Data from Deep Earth Monitoring and Sounding. Sci Drill 2007;4:30–1. http://dx.doi.org/10.2204/iodp.sd.4.06.2007.

[14] Klump, J., and D. Ulbricht (2011), PanMetaDocs - A tool for collecting and managing the long tail of 'small science data', EOS, Transactions, American Geophysical Union, 92(53, Fall Meet. Suppl.), IN23C–1461.

[15] Library of Congress (2014), *Recommended Format Specifications*, Request for Comments, Library of Congress, Washington, D.C. [online] Available from: <http://www.loc.gov/preservation/resources/rfs/index.html>.

[16] Neuroth, H., A. Oßwald, R. Scheffel, S. Strathmann, and K. Huth (Eds.) (2010), *nestor-Handbuch - Eine kleine Enzyklopedie der digitalen Langzeitarchivierung*, nestor Materialien, Version 2., Niedersächsische Staats- und Universitätsbibliothek, Göttingen, Germany. [online] Available from: <http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf>.

[17] NOAA (2009), Extensible Stylesheet Language Transformation (XSLT), *NOAA EDM*. [online] Available from: <https://geo-ide.noaa.gov/wiki/index.php?title=Extensible_Stylesheet_Language_Transformation_(XSLT)&oldid=10712> (Accessed 13 October 2014).

[18] Oberhänsli H, Mackay AW. Introduction to Progress towards reconstructing past climate in Central Eurasia, with special emphasis on Lake Baikal. Global Planet Change 2005;46(1–4):1–7. http://dx.doi.org/10.1016/j.gloplacha.2004.11.003.

[19] Razum M, Schwichtenberg F, Wagner S, Hoppe M. ESciDoc Infrastructure: A Fedora-Based e-Research Framework. Research and Advanced Technology for Digital Libraries, vol. 5714. Heidelberg, Germany: Springer Verlag; 2009. p. 227–38. http://dx.doi.org/10.1007/978-3-642-04346-8_23.

[20] Riede M, Schueppel R, Sylvester-Hvid KO, Kühne M, Röttger MC, Zimmermann K, Liehr AW. On the communication of scientific data: The Full-Metadata Format. Comp Phys Commun 2010;181(3):651–62. http://dx.doi.org/10.1016/j.cpc.2009.11.014.

[21] Rothenberg J. Ensuring the longevity of digital information. Sci Am 1995;272(1):42–7. http://dx.doi.org/10.1038/scientificamerican0195-42.

[22] Starr J, Gastl A. IsCitedBy: a metadata scheme for DataCite. D-Lib Mag 2011;17(1/2). http://dx.doi.org/10.1045/january2011-starr.

[23] Treloar A, Groenewegen D, Harboe-Ree C. The data curation continuum – Managing data objects in institutional repositories. D-Lib Mag 2007;13(9/10):13. http://dx.doi.org/10.1045/september2007-treloar.

[24] Ulbricht, D., J. Klump, and R. Bertelmann (2012), Publishing datasets with eSciDoc and panMetaDocs, in *Geophysical Research Abstracts*, vol. 14, pp. EGU2012–7058–2, Copernicus Society, Vienna, Austria. [online] Available from: <http://meetingorganizer.copernicus.org/EGU2012/EGU2012-7058-2.pdf>.

[25] Ulbricht D, Klump J, Conze R. Supplement to: Curating the web's deep past - Migration strategies for the German Continental Deep Drilling Program web content., Supplementary Material. Potsdam, Germany: German Research Centre for Geosciences; 2014. http://dx.doi.org/10.5880/GFZ.CEG.2014.001.

[26] Wächter, J. (1990), KTBase (KTB database) – The core of a scientific/technical database, KTB-Report, Projektgruppe Kontinentales Tiefbohrprogramm der Bundesrepublik Deutschland, Hannover, Germany.