Published in: Fütterer, D K, Damaske, D, Kleinschmidt, G, Miller, H & Tessensohn, F (eds.)
Antarctica: contributions to global earth sciences, Springer, Berlin, Heidelberg, New York, 403-406
doi:10.1007/3-540-32934-X

The original publication is available at http://www.springerlink.com see
doi:10.1007/3-540-32934-X_51

Archiving and Distributing Earth-Science Data with the PANGAEA Information System

Hannes Grobe[1] · Michael Diepenbroek[2] · Nicolas Dittert[2] · Manfred Reinke[1] · Rainer Sieger[1]

[1] Alfred Wegener Institute for Polar and Marine Research, Am Alten Hafen 26, 27568 Bremerhaven, Germany
<hannes.grobe@awi.de>

[2] Center for Marine Environmental Sciences, Leobener Str. 26, 28359 Bremen, Germany

Abstract. PANGAEA - Publishing Network for Geoscientific and Environmental Data (*http://www.pangaea.de*) is an information system aimed at archiving, publishing, and distributing data related to climate variability, the marine environment, and the solid earth.The system is a public “data library” distributing any kind of data to the scientific community through the Internet. Data are stored in a relational database in a consistent format with related meta-information following international standards.Data are georeferenced in space and/ or time, individually configured subsets may be extracted. Any type of information, data and documents may be served (profiles, maps, photos, graphics, text and numbers). Operation by Alfred Wegener Institute for Polar and Marine Research (AWI) and Center for Marine Environmental Sciences (MARUM) is assured in the long-term. Both institutions provide the technical infrastructure, system management and support for data management of projects as well as for individual scientists. Most important collections from Antarctic research archived in PANGAEA so far are the data of the Cape Roberts Project, geological maps and age determinations of rock outcrops, a complete set of JGOFS,WOCE, DSDP and ODP data including those from the Southern Ocean, any marine sediment cores, documentation and analytical data from German expeditions and an increasing inventory of data published by the running EPICA project.

Introduction

In 1990 SEPAN, the predecessor of the PANGAEA information system, was established at AWI for the administration of the geological samples archived in the “Polarstern” Core Repository. The system was reorganized to archive data from paleoclimate research in 1994. After the inclusion of a full spatial and temporal georeference for each single value, the system was able to handle any kind of geodata. Since 1998 the system has been accessible on the Internet via the domain *www.pangaea.de*. During the last six years, PANGAEA has been used by 23 major projects and many individual scientists for the archiving of scientific primary data (Diepenbroek et al. 1998, 2002). By the end of 2005, the system had an inventory of about 250 000 data sets. Major international projects served with PANGAEA are IMAGES (International Marine Global Change Study) and several EU projects (see *http://www.pangaea.de/Projects/*). Of major importance for the Antarctic are CRP (Cape Roberts Project) and EPICA (European Project for Ice Coring in Antarctica). PANGAEA has also archived all marine geological data from samples and sediment cores taken by the research vessel “Polarstern” in the Southern Ocean as well as in the Arctic Ocean. The World Data Center for Marine Environmental Sciences (*http://www.wdc-mare.org*), which was founded in 2001 in Germany as a member of the ICSU World Data Center System (WDC 1996), is using PANGAEA as its central archiving system. PANGAEA operates the European web and publication mirror for the Ocean Drilling Program (ODP). The institutional framework for PANGAEA, including the World Data Center,is supplied by the Alfred Wegener Institute for Polar and Marine Research (AWI) in Bremerhaven and the Center for Marine Environmental Sciences (MARUM) at the University of Bremen. PANGAEA is structure and function unique to any other geoscientific database available to date on the Internet.

The Data Model

The challenge of managing any kind of georeferenced data was met in PANGAEA through a flexible data model. This was implemented by a combination of a simple fully normalized relational database frontended by middleware components and various clients for upload and download. The model reflects the standard activities for data collection in the geosciences (Fig. 7.9-1). Collaborative activities in a PROJECT carry out expeditions (CAMPAIGN) for sampling. During an expedition samples may be taken or measurements are made (EVENT) at a number of locations (SITE). The medium to be investigated (e.g.,sediment,rock, water or ice) is subsampled or measured for different analytical procedures (SAMPLE). Finally from each sample or measurement analytical DATA result, organized in DATA SETs. These main levels are supplemented by related tables comprising information about items as personal, references, parameters or methods.

The essential part of the model is the combination of the "Data", "Parameter" and "Method" tables, which allows the definition and storage of new, unique parameters at any time. Up to a maximum of four different geocodes can be used simultaneously for the description of data points in space and time. These are selected from latitude, longitude, elevation, altitude, date/time, geological age, and depth in different media like water, sediment, rock or ice. Due to the complete georeferencing of each single value, the system allows the combination of any data types and the extraction of individually defined quantities of data. It is therefore a useful prerequisite for the analysis of complex data inventories or data mining (Han and Kamber 2001).

How to Find Data in PANGAEA

A number of clients allow the user to access metainformation and analytical data from the system on different levels of information and technical complexity. The data import client is the central management interface for the data curators and written in a proprietary software; all other clients for export are web-based.

PangaVista

This is a simple web-based search engine which allows the retrieval of predefined datasets in the relational system, referenced by web links. PangaVista makes use of a thesaurus, comprising all the metainformation related to the data, thus allowing a retrieval for any given keyword like an author's name, a parameter, a project or a sample label. Keywords may be combined to create bolean expressions, the syntax of which is similar to the one used by other search engines on the Internet. A map server allows the user to set geographical constraints and shows the locations of those datasets found by a retrieval. The

result of a query is a list of short descriptions of the datasets found, with a link to the complete set at the end of each header. Data can be downloaded to a dynamically produced web page in html-format or as a tab-delimited text file. A login is required if data are unpublished. Data sets are typically composed of a number of data series, accompanied by a meta-data description conforming to the ISO19115 standard. The ability to download all data sets found by a retrieval in one step is given through an external program, a web based version is in preparation.

Advanced Retrieval Tool (ART)

This tool provides full access to all tables of the relational system and enables the user to retrieve individually configured subsets of data from the inventory. This provides functionality such as the ability to compare several paleoclimatic records from different archives versus time. The simplified data model is used as the graphical user interface (Fig. 7.9-1). ART is designed as a "data mining tool" to support the production, use and interpretation of comprehensive data collections. As a Java application it runs on any platform and with the most common web browsers. Users are advised to study the "Help" provided or to contact <*info@pangaea.de*> in case specific data mining requests are needed.

Direct Download Interface (DDI)

This tool provides the functionality to easily distribute and publish data. Each data set in the system has a unique identifier that may be obtained as a URL such as *http://doi.pangaea.de/10.1594/PANGAEA.132796*. This link can be used in publications as a precise reference to a data set. This technology was used for the first time in a publication edited by Fischer and Wefer (1999) where each publication refers to its primary data through a given link. Links can be defined for any query or retrieval on data as well as to metadata and may also be distributed via email or placed on web pages.

PanCore

This is a web-based interface to search for locations and metadata of sampling sites. Geographical constraints can be set within the included map, which is also used for the display of the resulting list of sites; any result set can be downloaded as a text file. For geological samples, the curator, responsible for the repository where the samples are archived, is given. Thus PanCore enables the user easily to search for samples in a certain area and submit a sample request to the appropriate curator.

4th-Dimension-Client (4D-Client)

This is the administrative tool for the processing and maintenance of any information stored in PANGAEA. It supplies routines through a graphical user interface for the import and editing of analytical data and the definition of all types of meta-information with its relations to the data. The 4Dclient is mostly used by the data curators and librarians of PANGAEA and other projects and institutes.

Examples from Geoscientific Investigations in Antarctica

The following examples are given as a short tutorial to show how geoscientific data can be retrieved and how the data are downloaded to the users computer in a consistent format with metaheader. The PANGAEA search engine can be accessed at *http://www.pangaea.de*. Any expression included in the description of data sets can be used for a search, e.g., names, parameters or labels. The search is not case sensitive; results of a search example are shown in Fig. 7.9-2.

Example 1

The aim of CRP was to drill three stratigraphically overlapping cores in the Ross Sea to provide full coverage of Antarctic glaciation history. The cores were labeled CRP-1, CRP-2/CRP-2A and CRP-3. To retrieve all data from the project simply type in “crp”. To obtain data from one hole type in its identifier e.g., “crp-1”. If you are looking for a certain parameter in one of the cores, a combination of two expressions with a blank in between can be used (The syntax is the same as used in common search engines). For example, if looking for pollen of the plant family Caryopyllaceae a retrieval in PangaVista can be made with “crp-1 caryopyllaceae”. If looking for data from an author, the name and the label can be combined, e.g., “crp-2 kettler”. Download in text or html-format starts by clicking on the links given at the end of each metaheader.

Example 2

The EPICA project was running two drill sites on the East Antarctic ice sheet to recover the Pleistocene climate history of Antarctica. Any data related to EPICA can easily be found by typing in “epica”. Looking for the age models of the EPICA cores would produce a query like “epica age model”. The main cores are accompanied by several short firn cores. To see its distribution ask for “epica firn core” and click on“Show map”. Switch to stereographic (S) projection and zoom into the map by using the magnifying glass. If looking for data of a specific site, use the “?” button, click on the dot of interest and find related data sets listed in a new window.

Example 3

To search for geological ages measured by the fission-track method in Antarctica, a retrieval may be started using the map provided with PangaVista. Switch to stereographic south projection, choose the button with the arrow/rectangle and drag a rectangle above the Antarctic to set the geographical constraints. The geographical limits will be included in the four fields besides the PangaVista search line. Type in “age fission-track” and press “Search”. The retrieval should list the data set of Meier (1999).

Example 4

PANGAEA has archived most of the marine geological data of AWI resulting from the analysis of sediment cores taken by its research vessel “Polarstern” in the Atlantic and Pacific parts of the Southern Ocean. A search for “polarstern sediment” will list more than 2 500 data sets with only the first 200 shown. In such a case, the search has to be defined more restrictive.A search on“polarstern sediment pachyderma” (Fig. 7.9-2) will prompt the user with a list of all published data sets containing data about e.g., the planktonic foraminifera *Neogloboquadrina pachyderma*. The user may scan through the list of sets by using the, «prev and next» buttons.

Example 5

PANGAEA is also used to archive georeferenced graphics or images.A retrieval on“ant-VIII documentation” will show a list of data sets, each containing the metadata of a sediment core and including just a “georeferenced” link to a directory. This directory contains all photos and descriptions of the sediment core in standard file formats (txt, pdf, jpg), presented in an overview with thumbnails. The grafic/photo is available for download in full resolution by a click on a thumbnail.

Conclusion

PANGAEA is an information system for the long-term archiving and distribution of georeferenced data. Due to the flexibility of the data model analytical data from many fields of basic research in natural science can be stored consistently together with the related metainformation necessary for their understanding and usage. With its comprehensive graphical user interfaces and the built-in functionality for upload and download, PANGAEA is an efficient system for scientific data management and data publication. Web-based interfaces to retrieve information from the system range from a simple search engine to a sophisticated data mining tool, the latter allowing the retrieval and combination of any subquantity of analytical data from the full inventory. For the visualization of data, software tools are distributed as freeware from the PANGAEA web site (Schlitzer 1997; Grobe et al. 2000; Sieger et al. 2001; see http://www.pangaea.de/Software). The internal consistency in combination with the use of clients and tools optimized for the users’ needs, give data an added value if they are archived in PANGAEA. The major advantage of PANGAEA is its easy accessibility on the Internet providing the scientific community with a library of thousands of valuable data sets even from remote areas like the Antarctic. Any scientist, project or institute is encouraged to contribute to and make use of the PANGAEA library to establish a long-term archive and publication system for earth science data.

Acknowledgments We are grateful to Roy Lowry and Hartmut Schulz for review and comments. Technical infrastructure and operation of the system is ensured by the computer centre of the Alfred Wegener Institute for Polar and Marine Research (AWI). AWI is a Helmholtz Research Centre, funded by the German Ministry for Education and Research and the State of Bremen.

References

Diepenbroek M, Fütterer DK, Grobe H, Miller H, Reinke M, Sieger R (1998) PANGAEA information system for glaciological data management. Annals Glaciol 27:655–660

Diepenbroek M, Grobe H, Reinke M, Schindler U, Schlitzer R, Sieger R, Wefer G (2002) PANGAEA an information system for environmental sciences. Computer Geosci 28:1201–1210

Fischer G,Wefer G (eds) (1999) Use of proxies in paleoceanography: examples from the South Atlantic. Springer, Berlin Heidelberg New York, http://www.pangaea.de/Projects/SFB261/Use_of _Proxies

Grobe H, Sieger R, Diepenbroek M (2000) PanMap. http:// www.pangaea.de/Software/PanMap

Han J, Kamber M (2001) Data mining, concepts and techniques. Morgan Kaufmann Publishers

Schlitzer R (1997) Ocean-Data-View. http://odv.awi-bremerhaven.de

Sieger R, Grobe H, Diepenbroek M (2001) PanPlot.http://www.pangaea.de/ Software/PanPlot

WDC (1996) Guide to the World Data Center System. Secretariat of the ICSU Panel on World Data Centers,http:// www.ngdc.noaa.gov/ wdc/guide/wdcguide.html

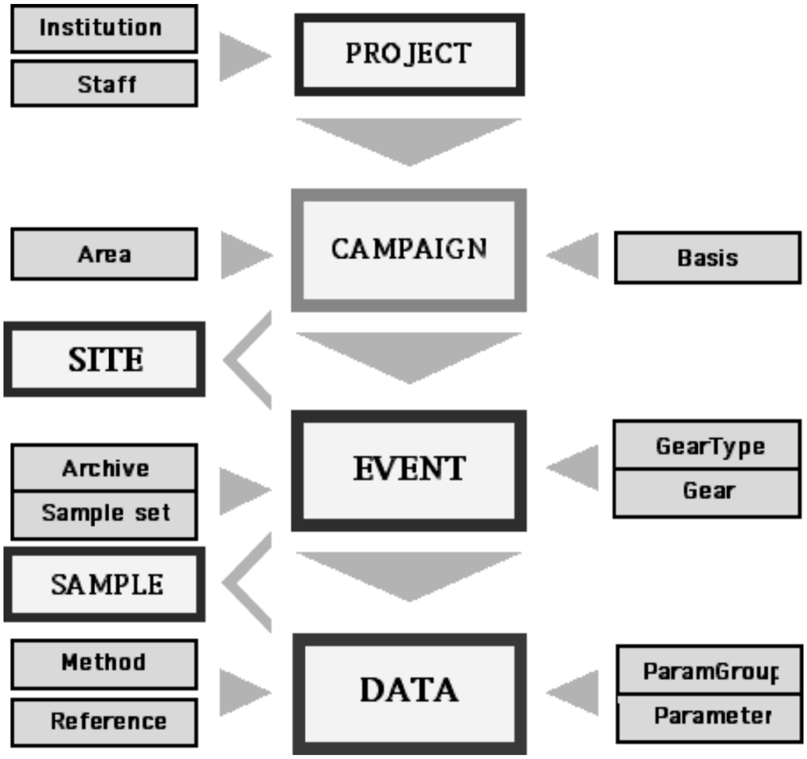
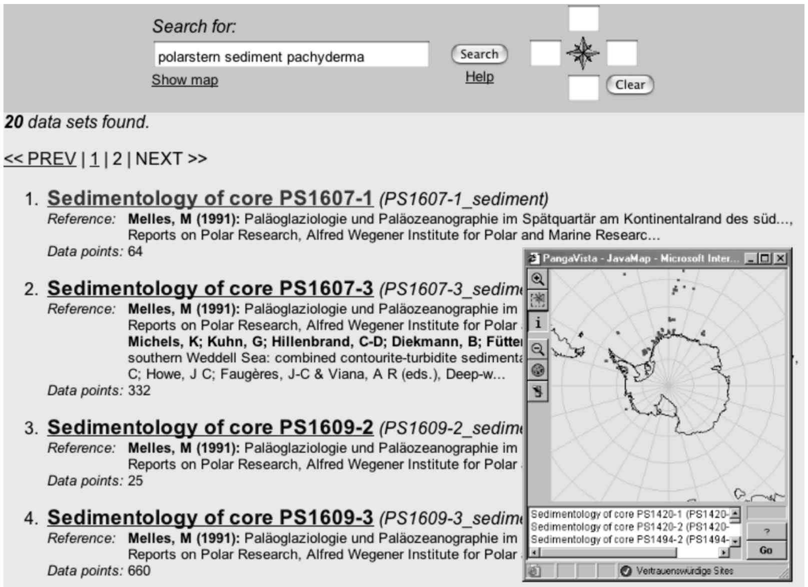


Fig. 7.9-1. The simplified data model of PANGAEA is used as the graphical user interface (GUI) for data import and mining on the Internet. Each *box* represents a table in the relational database. The user can use a uniform retrieval tool to find information in each of the tables.


Search for:
polarstern sediment pachyderma
Search
Show map
Help
Clear
PangaVista - JavaMap - Microsoft Inter...
?
Go
Vertrauenswürdige Sites

20 data sets found.

<< PREV | 1 | 2 | NEXT >>

1. Sedimentology of core PS1607-1 *(PS1607-1_sediment)*
   Reference: **Melles, M (1991):** Paläoglaziologie und Paläozeanographie im Spätquartär am Kontinentalrand des süd..., Reports on Polar Research, Alfred Wegener Institute for Polar and Marine Researc...
   Data points: 64
2. Sedimentology of core PS1607-3 *(PS1607-3_sedim*
   Reference: **Melles, M (1991):** Paläoglaziologie und Paläozeanographie im Reports on Polar Research, Alfred Wegener Institute for Polar **Michels, K; Kuhn, G; Hillenbrand, C-D; Diekmann, B; Fütte** southern Weddell Sea: combined contourite-turbidite sedimenta C; Howe, J C; Faugères, J-C & Viana, A R (eds.), Deep-w...
   Data points: 332
3. Sedimentology of core PS1609-2 *(PS1609-2_sedim*
   Reference: **Melles, M (1991):** Paläoglaziologie und Paläozeanographie im Reports on Polar Research, Alfred Wegener Institute for Polar
   Data points: 25
4. Sedimentology of core PS1609-3 *(PS1609-3_sedim*
   Reference: **Melles, M (1991):** Paläoglaziologie und Paläozeanographie im Reports on Polar Research, Alfred Wegener Institute for Polar
   Data points: 660

Sedimentology of core PS1420-1 (PS1420-
Sedimentology of core PS1420-2 (PS1420-
Sedimentology of core PS1494-2 (PS1494-

Fig. 7.9-2. When using the search engine PangaVista any combination of keywords is possible. The results are listed and the location of data sets found are plotted on a map. Using the "?" button by pointing on one of the sites, a list with related data sets will be given for download.