Data and information management for the CMTT synthesis

Nicolas Dittert[†], Michael Diepenbroek[*], Hannes Grobe[‡]

[†] Laboratoire des sciences de l'environnement marin, IUEM, Université de Bretagne Occidentale, F-29280 Plouzané, France

[*] Centre for Marine Environmental Sciences, MARUM, P.O. Box 33 04 40, D-28359 Bremen, Germany

[‡] Foundation Alfred Wegener Institute for Polar and Marine Research, P.O. Box 12 0161, 27515 Bremerhaven, Germany

Manuscript to be published in

Carbon and nutrient fluxes in global continental margins by

L.Atkinson, K.K. Liu, R. Quinones, L. Talaue-McManus

Manuscript submitted/received: June 24, 2002 /_____

Revised version of ms submitted/received: _____/_____

Final version of ms submitted/received: _____/_____

For correspondence:

Dr. Nicolas Dittert

LEMAR – IUEM

Technopole Brest-Iroise, Place Nicolas Copernic

F – 29280 Plouzané, France

Tel: ++33 2.98.49.86.73 · Fax: ++33 2.98.49.86.45

E-mail: nicolas.dittert@univ-brest.fr

Abstract.
In 2001, the CMTT Global Synthesis group invited the World Data Centre for Marine Environmental Sciences (WDC-MARE) to take care for information and data management during and after the CMTT's synthesis work. Information should be circulated freely among all partners, which was accomplished through a private mailing-list and a set of web pages that both were updated at irregular intervals. Data were retrieved from participating scientists, publications, and online data collections, and then have been made available in a consistent format using the information system PANGAEA. Any work has been carried out respecting the guidelines for communication between JGOFS and LOICZ as expressed in the *Data and Information System Plan*.

Keywords.
Scientific data; data management; WDC; information system; PANGAEA

1. Introduction.

Gathering of scientific raw data has a long history. Benjamin Franklin in the 1770s and Matthew Maury a century later collected data directly from ship captains, which led to the first synoptic oceanographic study of the Gulf Stream. In the 18th and 19th centuries, scientific data were recorded in expedition reports. However, explicit data management plans were developed only in the late 1950s for the newly invented International Council of Scientific Unions' (ICSU) World Data Center (WDC) system, stating in detail type and format of data that were to be submitted to the WDCs (e.g., Ruttenberg et al. 1994). Since then, the gathering and exchange of data has been transformed by immense technological advances. The replacement of analogue devices with digital computers, the invention of relational data base management systems (RDBMS), digital communication networks etc. have made it possible to operate data at any place and any time.
This publication intends to demonstrate how WDC-MARE accompanied CMTT's synthesis work through actively supporting this international group of single scientists who are affiliated voluntarily to a common, yet inhomogeneous, organization (the CMTT) by information and data management during and after the final period.

2. Description of WDC-MARE and PANGAEA

In the broader sense, the WDC system takes care for the exchange of solar, geophysical and environmental data in order to describe the Earth's complex, non-linear and interactive

climate and biospheric systems with an ultimate goal of (1) understanding the workings of myriad individual physical and biospheric processes involved in the global system; (2) monitoring the progressive effects of those processes; and (3) predicting its evolution and future state.

Among some 50 WDCs worldwide, the World Data System for Marine Environmental Sciences (WDC-MARE; www.wdc-mare.org) is specialized on collecting, scrutinizing, and disseminating data related to global change science in the fields of environmental oceanography, marine geology, paleoceanography, and marine biology. WDC-MARE uses PANGAEA (Network for Geosciences and Environmental Data; www.pangaea.de) as operating platform. Besides an impressive list of WDC general duties, essential services supplied by WDC-MARE / PANGAEA comprise information infrastructure management, in particular project data management, data publication, and the distribution of visualization and analysis software (freeware products). Together with the WDC for Paleoclimatology, Boulder, USA, WDC-MARE forms the essential backbone within the IGBP/PAGES Data System (Eakin et al. 2002).

There are no particular system requirements since WDC-MARE is accessible online from any computer system (e.g., Linux, Macintosh, Unix and its derivatives, Windows), Internet connection (e.g., modem, LAN/WAN), and browser software (e.g., Internet Explorer, Netscape) at any place of the Earth.

WDC-MARE / PANGAEA is operated as a permanent facility by the Centre for Marine Environmental Sciences (MARUM) at Bremen University and the Foundation Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven, Germany.

3. Information management.

During the 2001 CMTT synthesis meeting in Taipei a need for three defined services clearly emerged: (1) Web pages that display topical information on Internet during and after synthesis work; (2) mailing list that fosters quick circulation of electronic discussion among all participants; (3) bibliographic service that provides exact citations and avoids distribution of erroneous information. All three requests are met by WDC-MARE / PANGAEA. Due to the Bremen High Speed Broadband Network (LBN, "*Bremen BriteLine*"), extremely rapid gateways and access to web pages and electronic mailing can be provided compared to other international educational facilities (Fig. 1).

All web pages (cf. www.pangaea.de/wdc-mare/Projects/CMTT/) are programmed using the current hypertext standard HTML 4.0, JavaScript, and CGI-Script (Perl). They are split up

into four categories (1) "*home page*", which corresponds to the outline of the synthesis book (i.e., link to publisher, design of content, links to authors, abstracts, data, figures) and an actual schedule for the production of the book); (2) *bibliography*, which comprises all essential references for easy exchange among the authors and for the attention of the scientific community interested on continental margins science. References may be visualized in HTML format, searched by browser internal functions, and downloaded according to their scientific field (e.g. Eastern Boundary Currents, "*EBC*") or as complete list by alphabetical order; (3) *forum*, which serves as pivotal information distribution and discussion platform. For rapid exchange of information among all participants a dynamic mailing list (cmtt@wdc-mare.org) is established at WDC-MARE using the Netscape® Messenger system; (4) a link to the Taiwanese CMTT home page.

4. Data management.

WDC-MARE / PANGAEA represent the European standard of Spatial Data Infrastructures (SDI) and operate under the umbrella of the ICSU WDC system. Due to a highly diversified affinity for data management within the CMTT scientific community, a dynamical approach is chosen that allows for a maximum of data with a minimum of scientist's exhaustion. Ergo, data can be readily shared between participants, which is a relatively straightforward consideration and meets the principles of the LOICZ Data & Information System Plan (Boudreau et al. 1996).

Data acquisition is carried out as retrieval form national data centers, submission as e-mail attachment, URL, hardcopy etc. Data archival is got to work by WDC-MARE staff at a central place. Depending on skill, data access is offered through (1) a simple data search engine *PangaVista*; (2) a sophisticated data mining tool *ART*; (3) a client/server data management tool *4th Dimension*®; (4) various PANGAEA related data analysis and visualization software. Since some data sets refer to existing publications, others are yet unpublished or non-public, all data sets were reconditioned under the same criteria to satisfy PANGAEA's formal RDBMS standard, the so-called data base model (Fig. 2), to ensure maximum quality, and to finally be disseminated at URL www.pangaea.de/PangaVista?query=CMTT (cf. Dittert et al. 2001)

4.1. Data preparation

Personal contact between data management staff and PI proves to be beneficial to both parties. Together with the PI, all important meta-information are collated: Project facts, cruise mnemonics, official station lists as well as notes on institutions and co-workers involved. The most critical performance is summing up and discussing the analytical data, which would serve as a first quality check: Are all values valid? Are all units consistent? Are all parameters adjusted? Are suspect numbers to be corrected or to be flagged or to be wiped out, respectively? Are methods checked for completeness? Are publications referenced? Finally, meta-information and analytical data both are converted into tab-delimited ASCII spread sheets following standardized input forms.

4.2. Data import/export

To upload data collections, any PANGAEA client offers import routines for CAMPAIGN, SITE/EVENT, SAMPLE, REFERENCE, PARAMETER, ARCHIVE, and analytical DATA (Fig. 2). These files are loaded online into a temporary table at the server side (Fig. 1) where a second quality check is initiated and meta-information are examined for coherence with entries already archived: Are station lists identical to the existing catalogue? Are geographical positions accurate? Then, analytical data are compared to pre-defined settings: Do there exist duplicates? Are the numbers situated between upper and lower range? Once a data series has cleared this hurdle, data are transferred to the import server and stored temporarily in order to update the data warehouse at regular intervals. Finally arrived at the export server, a uniform resource locator (URL) is generated and data are offered to the scientific community (Fig. 3).

4.3. Data retrieval

There exist three different ways to retrieve data (Fig. 3) from WDC-MARE, which may be distinguished by the skill that is required from the user. In any case, the general numeric output from the database are tab-delimited ASCII files that accommodate both analytical data and meta-information. All data can then be further processed in using commercial software. In any case, a help system and tutorial are provided for online support. (1) *PangaVista* (Fig. 3A), is a search engine similar to AltaVista. Any given keyword (i.e., author, journal, variable, project, campaign, etc.) results in a list of URLs to related data sets that can be displayed as HTML format (cf. Fig. 3A “HTML”) and downloaded as ASCII files (cf. Fig. 3A “TEXT”). Complementary to *PangaVista* is the download of lists that were compiled to portray whole projects, or institutes, or publishers, respectively (cf. ‘Projects’, DATA for the CMTT data collection). (2) Some more skill is required if the database is accessed by the Advanced

Retrieval Tool (*ART*, Fig. 3B) that gives full access to the system via the data model (Fig. 2) and allows the extraction of any individually configured subset of data from the inventory. To use *ART* efficiently, the user needs some experience with boolean terminology. All data retrieval tools introduced above exclusively read information. (3) The direct access to the server (Fig. 3C), however, is a sophisticated data processing tool that permits both reading and writing access. Since this application requires a particular measure of expertise, the gateway is limited to the data management staff.

4.4. Data presentation

PanMap, PanPlot, PaleoToolBox (Sieger et al. 1999), and Ocean Data View (ODV, Schlitzer 2000) are built to be used as standalone-applications on the scientists computer. These interactive software modules support direct access to any export format of PANGAEA. Maps, plots, and cross section profiles can be exported in platform specific interchange format and further processed in using commercial software. Since they belong to the information system PANGAEA rather than to the CMTT project they are explained elsewhere (Diepenbroek et al. 1999).

4.5. Data quality

Validation and verification of data are the two substantial aspects during data archiving: External discovery of errors will do more to destroy the credibility of the database and the data management group than anything else. However, the definition of what is correct is far from straightforward and can quite often be a matter of opinion and opinions are subject to change as scientific knowledge (Lowry et al. 1995). While it is not essential to have only excellent data sets archived it is, however, indispensable that exact information on the quality is provided. Complete meta-information, including, in particular, analytical method and reference where the data were published first is crucial to depict any data set. Manual quality checks are supplemented by an evolving system of generic and parameter-specific validation routines that base on the definition of parameters and analytical methods which require standard unit, upper and lower limit levels, precision and tolerance of the analytical value. Data collections from third parties must be treated even more carefully since data are trawled from many different sources, each with their own quality standards. In any case, data access is controlled by the data owner or Principal Investigator, respectively, since WDC-MARE archives published as well as non-public data.

5. Conclusion.

The amount of publications in natural sciences doubles every about ten years, the estimated growth rate in related scientific data is even higher. This evolution is accelerated by technological progress as well as by growing public interest. Since printing of scientific data is economically no more acceptable, the context (scientific *text-data* unit) gets more and more lost. Moreover, binding data standards are (if at all) poorly developed and rather confusingly established. Global Change science, however, requires a good availability of enormous amounts of analytical data (e.g., Alverson et al. 2001). Respecting the international WDC standard, WDC-MARE / PANGAEA accompanied CMTT during the final synthesis phase to ensure:

(a) Philosophy of consistent data sets
(b) Geo-coding of analytical data in a RDBMS environment
(c) Quality check of data according to existing meta-standards
(d) Mutual effect of the unit CMTT scientific community – WDC – Publisher
(e) Maximum of information exchange among participants

6. Acknowledgements.

The authors wish to express their sincere appreciation for successful cooperation with all Principal Investigators during the CMTT synthesis work. L. Corrin (UBO, Brest), R. Sieger and S. Makedanz (AWI, Bremerhaven) provided invaluable assistance. We acknowledge comments from <space for reviewer names>. This research was funded by the European Commission through grant EVK2-CT-2001-00100 “ORFOIS”.

7. References.

Alverson K, Eakin CM (2001) Making sure that the world's palaeodata do not get buried. Nature 412:269

Boudreau PR, Geerders PJF, Pernetta JC (1996). LOICZ Data and Information System Plan (LOICZ Reports & Studies No. 6 (II), 62 pp).

Diepenbroek M, Grobe H, Reinke M, Schlitzer R, Sieger R (1999). Data management of proxy parameters with PANGAEA. In: Fischer G, Wefer G (eds.): Use of proxies in paleoceanography: Examples from the South Atlantic: 715-727, Berlin Heidelberg (Springer-Verlag)

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

Eakin C, Diepenbroek M, Hoepffner M (2002). The PAGES data system. In: Alverson K, Bradley R, Pedersen T (eds.): Paleoclimate, global change and the future: in press, New York (Springer-Verlag)

Lowry RK, Loch SG (1995). Transfer and SERPLO: powerful data quality control tools developed by the British Oceanographic Data Centre. In: Giles JRA (ed.):, Geological data management: 109-115, (Geological Society Special Publication)

Ruttenberg S, Risbeth H (1994) World Data Centres - past, present and future. Journal of Atmospheric and Terrestrial Physics 56(7):865-870

Schlitzer R (2000) Electronic Atlas of WOCE Hydrographic and tracer data now available. EOS Trans AGU 81(5):45

Sieger R, Gersonde R, Zielinski U (1999) A new extended software package for quantitative paleoenvironmental reconstructions. EOS Trans AGU Electronic Supplement, http://www.agu.org/eos_elec/98131e.html

Figure captions.

Figure 1. The World Data Center for Marine Environmental Sciences (WDC-MARE) and its information system PANGAEA are strictly organized in terms of technical and scientific design. The network concept uses client/server technology through Intranet/Internet communication. Remote sites are (a) groups of clients using a subserver (mirror site); (b) single clients inter-connected with the main server; (c) stand-alone devices for temporary connection to the network (e.g., on research vessels). Any client has full access to the information system.

Figure 2. PANGAEA's (simplified) data model is converted into a straightforward, flexible RDBMS scheme. The hierarchy of the data model is classified into four level that redraw the evolution of analytical values. The *PROJECT* level contains all meta-information on the project, its scientists and affiliated institutions; the *CAMPAIGN* level consists of the regional objective and its task basis; the sub-level *SITE* serves as the superimposed layer of sampling areas; the level *EVENT* accommodates equipment and includes the sub-level *SAMPLE* that reflects the custody; the level *DATA* comprises analytical data sets and series, analytical methods, variables and units, and their publication. The data model is universal and can be employed for any scientific, geo-coded data. The arrows show principle relations between tables.

Figure 3. Access to data is - depending on skill - presented through (A) the simple data search engine *PangaVista*; (B) the sophisticated data mining tool *ART*; (C) the client/server data management tool *4th Dimension*®; plus various PANGAEA related data analysis and visualization software. The general numeric output from the database are tab-delimited ASCII files that accommodate both analytical data and meta-information. All data can then be further processed in using commercial software.

Figures.

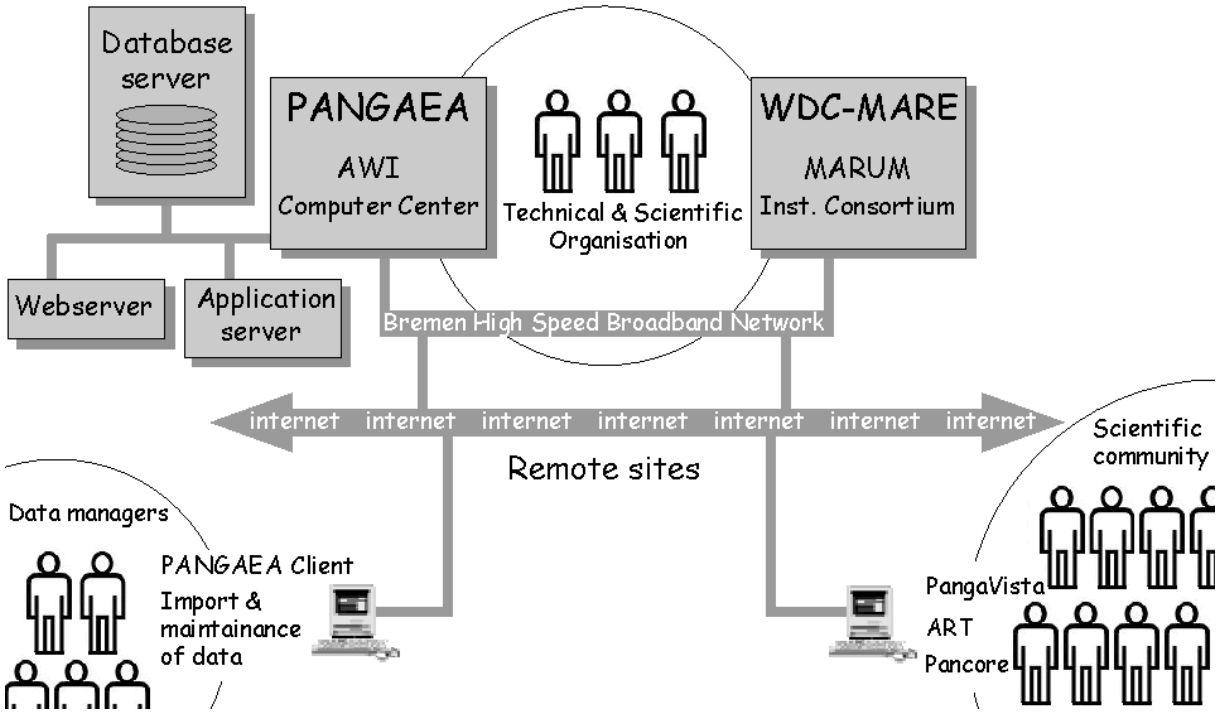

Figure 1.

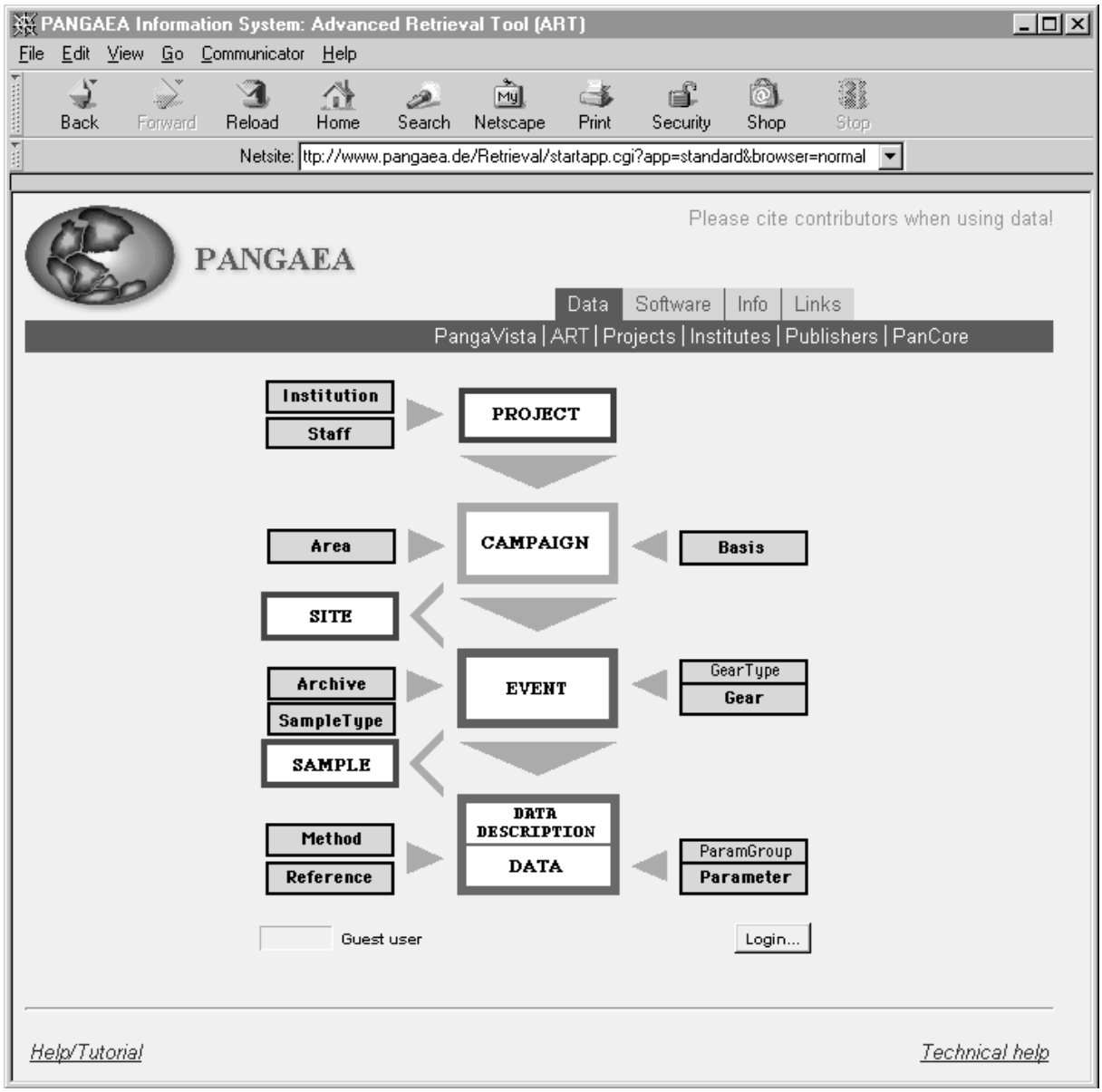
PANGAEA Information System: Advanced Retrieval Tool (ART)
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Netsite: ttp://www.pangaea.de/Retrieval/startapp.cgi?app=standard&browser=normal
Please cite contributors when using data!
PANGAEA
Data Software Info Links
PangaVista | ART | Projects | Institutes | Publishers | PanCore
Institution
Staff
PROJECT
Area
CAMPAIGN
Basis
SITE
Archive
SampleType
EVENT
GearType
Gear
SAMPLE
Method
Reference
DATA DESCRIPTION
DATA
ParamGroup
Parameter
Guest user
Login...
Help/Tutorial
Technical help

Figure 2.

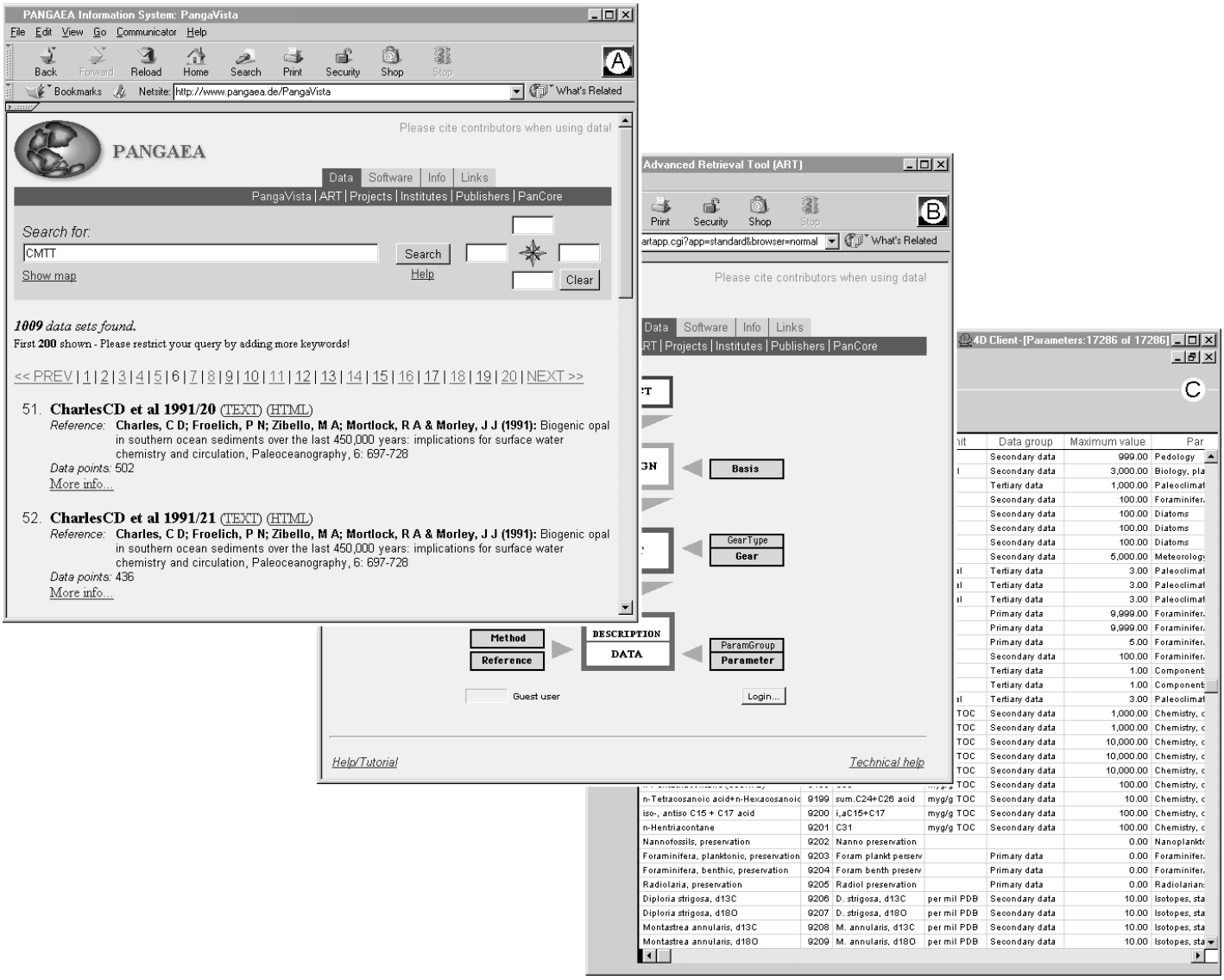
PANGAEA Information System: PangaVista
File Edit View Go Communicator Help
Back Forward Reload Home Search Print Security Shop Stop
Bookmarks Netsite: http://www.pangaea.de/PangaVista What's Related
Please cite contributors when using data!
PANGAEA
Data Software Info Links
PangaVista | ART | Projects | Institutes | Publishers | PanCore
Search for:
CMTT
Search
Help
Show map
Clear
1009 data sets found.
First 200 shown - Please restrict your query by adding more keywords!
<< PREV | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | NEXT >>
51. CharlesCD et al 1991/20 (TEXT) (HTML)
Reference: Charles, C D; Froelich, P N; Zibello, M A; Mortlock, R A & Morley, J J (1991): Biogenic opal in southern ocean sediments over the last 450,000 years: implications for surface water chemistry and circulation, Paleoceanography, 6: 697-728
Data points: 502
More info...
52. CharlesCD et al 1991/21 (TEXT) (HTML)
Reference: Charles, C D; Froelich, P N; Zibello, M A; Mortlock, R A & Morley, J J (1991): Biogenic opal in southern ocean sediments over the last 450,000 years: implications for surface water chemistry and circulation, Paleoceanography, 6: 697-728
Data points: 436
More info...
Advanced Retrieval Tool (ART)
Print Security Shop Stop
artapp.cgi?app=standard&browser=normal What's Related
Please cite contributors when using data!
Data Software Info Links
RT | Projects | Institutes | Publishers | PanCore
Basis
GearType
Gear
Method
Reference
DESCRIPTION
DATA
ParamGroup
Parameter
Guest user
Login...
Help/Tutorial
Technical help
4D Client - [Parameters: 17286 of 17286]
Data group
Maximum value
Secondary data 999.00 Pedology
Secondary data 3,000.00 Biology, pla
Tertiary data 1,000.00 Paleoclimat
Secondary data 100.00 Foraminifer.
Secondary data 100.00 Diatoms
Secondary data 100.00 Diatoms
Secondary data 100.00 Diatoms
Secondary data 5,000.00 Meteorology
Tertiary data 3.00 Paleoclimat
Tertiary data 3.00 Paleoclimat
Tertiary data 3.00 Paleoclimat
Primary data 9,999.00 Foraminifer.
Primary data 9,999.00 Foraminifer.
Primary data 5.00 Foraminifer.
Secondary data 100.00 Foraminifer.
Tertiary data 1.00 Component
Tertiary data 1.00 Component
Tertiary data 3.00 Paleoclimat
TOC Secondary data 1,000.00 Chemistry, c
TOC Secondary data 1,000.00 Chemistry, c
TOC Secondary data 10,000.00 Chemistry, c
TOC Secondary data 10,000.00 Chemistry, c
TOC Secondary data 10,000.00 Chemistry, c
TOC Secondary data 100.00 Chemistry, c
n-Tetracosanoic acid+n-Hexacosanoic 9199 sum.C24+C26 acid myg/g TOC Secondary data 10.00 Chemistry, c
iso-, antiso C15 + C17 acid 9200 i,aC15+C17 myg/g TOC Secondary data 100.00 Chemistry, c
n-Hentriacontane 9201 C31 myg/g TOC Secondary data 100.00 Chemistry, c
Nannofossils, preservation 9202 Nanno preservation 0.00 Nanoplankt
Foraminifera, planktonic, preservation 9203 Foram plankt preserv Primary data 0.00 Foraminifer.
Foraminifera, benthic, preservation 9204 Foram benth preserv Primary data 0.00 Foraminifer.
Radiolaria, preservation 9205 Radiol preservation Primary data 0.00 Radiolarian
Diploria strigosa, d13C 9206 D. strigosa, d13C per mil PDB Secondary data 10.00 Isotopes, sta
Diploria strigosa, d18O 9207 D. strigosa, d18O per mil PDB Secondary data 10.00 Isotopes, sta
Montastrea annularis, d13C 9208 M. annularis, d13C per mil PDB Secondary data 10.00 Isotopes, sta
Montastrea annularis, d18O 9209 M. annularis, d18O per mil PDB Secondary data 10.00 Isotopes, sta

Figure 3.