Management of marine data with the PANGAEA-information system SEPAN

Michael Diepenbroek, Hannes Grobe, Manfred Reinke, Reiner Schlitzer, Rainer Sieger & Uwe Siems

Alfred Wegener Institute for Polar and Marine Research
27515 Bremerhaven, Germany
sepan@awi-bremerhaven.de

SUMMARY

To use the scientific resource of marine data effectively an information system was developed which guarantees longtime storage of the data in consistent formats and provides easy access for the scientific community via World Wide Web or a system specific client software with high functionality. The system is able to store data together with raw data, evaluated data and all related meta-information necessary for their understanding. The system provides standardized import and export routines, easy access with uniform retrieval functions, and tools for the visualization of data. The system is designed as a network with client/server technology providing access and data exchange through the Internet.

1. INTRODUCTION

The initiative for a marine data information system

The oceans play a key part in understanding climate as they are the major sink for carbon dioxide. The samples taken from the water column or sediment are analyzed to reconstruct oceanographic conditions. A number of analytical parameters can be determined from the samples or measured in situ. With the introduction of new and more efficient analytical methods the number of parameters and the amount of data obtained has increased by an order of magnitude during the last decade. Important groups of parameters are the percentages of species of marine fauna and flora, water and sediment chemistry or physical properties.

The emergence of an integrated earth systems science calls for a full knowledge of recent and past conditions, in both space and time, and for data sets that are drawn as composites from

different methods and techniques. The only way to obtain a useful system for this purpose is to collect as many data as possible, to store this data collection in a consistent format and to make it easily accessible to paleoceanographers. Tools for retrieving the data and for their visualization have to be closely linked to the collection. This data collection, implemented in a network between working groups, can then be used as the data source as well as a common interpretation tool. In the future it should also be used as a publishing and reference system for data related to new publications to ensure that all the relevant data are stored in the same system.

In 1993, scientists from various German research institutes working in the field of marine sciences initiated a project in response to the needs described above. The goal was the implementation of an information system which would allow an overview of the sampling material available with the related meta-information, and which would store oceanographic data of any kind in a consistent form and make these data easy accessible to the scientific community. Tools for import/export, graphical presentation and complex retrievals need to be closely related to data collection. Based on the discussion and recommendations of this group, the information system SEPAN (Sediment and Paleoclimate Data Network) was developed at the Alfred Wegener Institute for Polar and Marine Research (AWI), financed by the German Ministry of Education, Science, Research and Technology (BMBF) as part of a three year project (1994-1997).

2. SYSTEM DESCRIPTION

The most important generic aspects of SEPAN are the quality and availability of the data as well as the high adaptability and effective use of the system. Data quality can be described in terms of the validity of methods and the precision and objectivity of measurements. It is not essential to have only excellent quality data sets, however, it is important that the quality can be estimated. The completeness of the meta-information, including, in particular, the analytical method and the reference where the data have been published first is crucial in the understanding of the analytical data. The user of a specific data set must be able to verify the data by reading the reference and thus to make a decision about the quality and usefulness of the data.

The manual quality check is supplemented by an evolving system of generic and parameter-specific validation routines. These routines are based on the definition of analytical methods and parameters, which requires a given standard unit, possible minimum and maximum

values and the possible precision of the data. This information will be used by validation routines during the import to filter out suspect values, e.g. outliers.

To improve the data consistency, data sets can be stored at different levels of processing. The primary data, e.g. counts of a microfossil assemblage or weights of granulometric investigations, are the raw data for calculations and interpretations. Archiving the raw data allows future recalibration or new interpretations of the data sets. The secondary data are those values calculated from the raw data, and in many cases are percentages or other units of concentration. The secondary data will usually be the proxy data for the evaluation of parameters describing past environmental conditions. Parameters evaluated from the secondary data are defined as tertiary data.

When dealing with the publication and archiving of data, copyright has to be considered(Diepenbroek Reinke, 1995). If an information system also stores unpublished data, it is crucial for the acceptance and the trust of the database that the data are protected by a hierarchical system which can be organized and controlled by the user. The owner of a specific data set is the data producer (principal investigator), who must be able to either give copyrights to individual users/groups or to open data sets across the whole system. When using foreign data, a given reference has to be cited, thus giving a benefit to the data producers. For unpublished data the principal investigator has to be asked for permission to use the data.

2.1 The SEPAN data model

The great variety of parameters, methods, calibrations and interpretations used in paleoenvironmental reconstruction, as well as the modification of established methods, are major obstacles to the integrative use of data sets in a common system. The challenge of managing these heterogenic and dynamic data was met in SEPAN through a highly flexible data model consisting of a relational data structure in combination with specialized server software (middleware) generating an object-oriented view of the data.

The simplified data structure of SEPAN is shown as a graphic on the opening screen of the client software (Fig. 1).

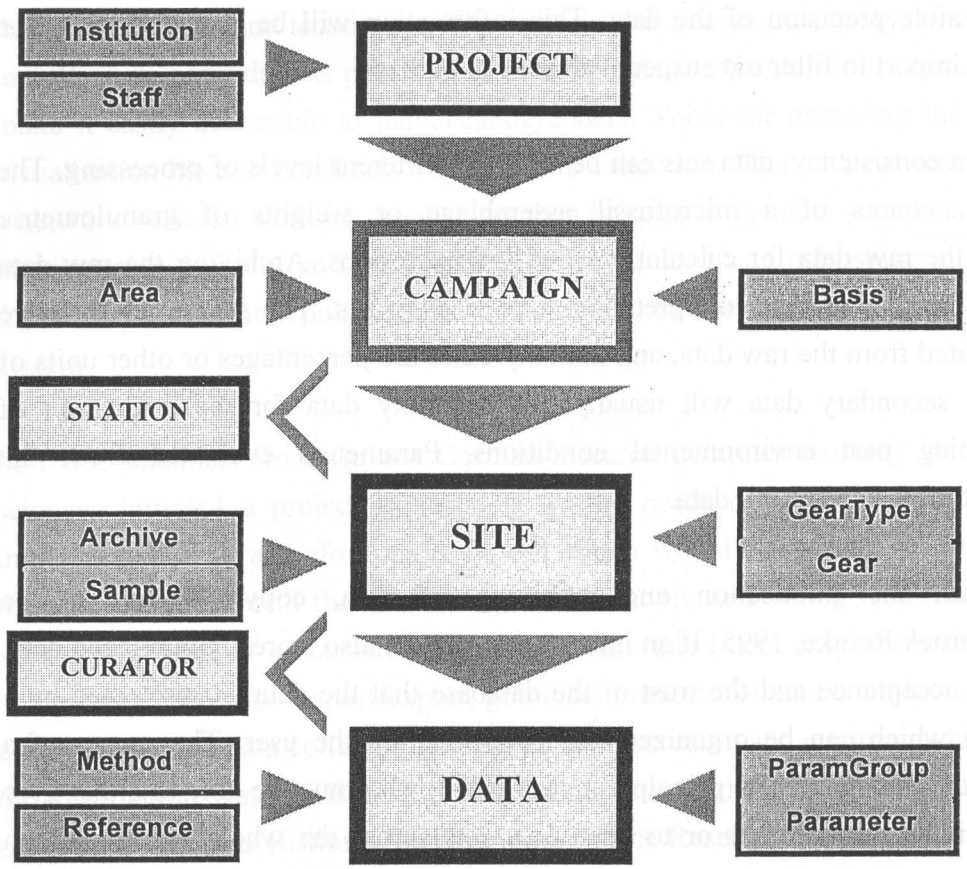

Fig. 1: The data model of SEPAN follows the path from collecting samples to the final analytical/environmental data. Within a PROJECT, different CAMPAIGNs are carried out to collect samples or measure environmental parameter at distinct SITEs to obtain DATA. The model is universal and can be used for any scientific data which are oriented to a geographical site.

The graphic allows users to enter all levels, tables and tools by selecting the required field. The structure reflects the standard processing steps for paleo-data. Lists including standardized meta-information are connected to the main data fields (e.g. gear, method). Different institutes/projects (PROJECT) working in the field of paleoceanography carry out expeditions (CAMPAIGN) for sampling. During a cruise at a number of locations (STATION) different samples may be taken or measurements made (SITE). At distinct points/intervals the medium to be investigated (e.g. sediment, water or ice) is subsampled or measured for different requirements. Information about the sampling procedure is stored at the CURATOR level. Down to this level, all data are considered to be meta-data. From each sample one or more analytical data points will be produced which can be found on the DATA level, with the related meta-information (Fig. 2).

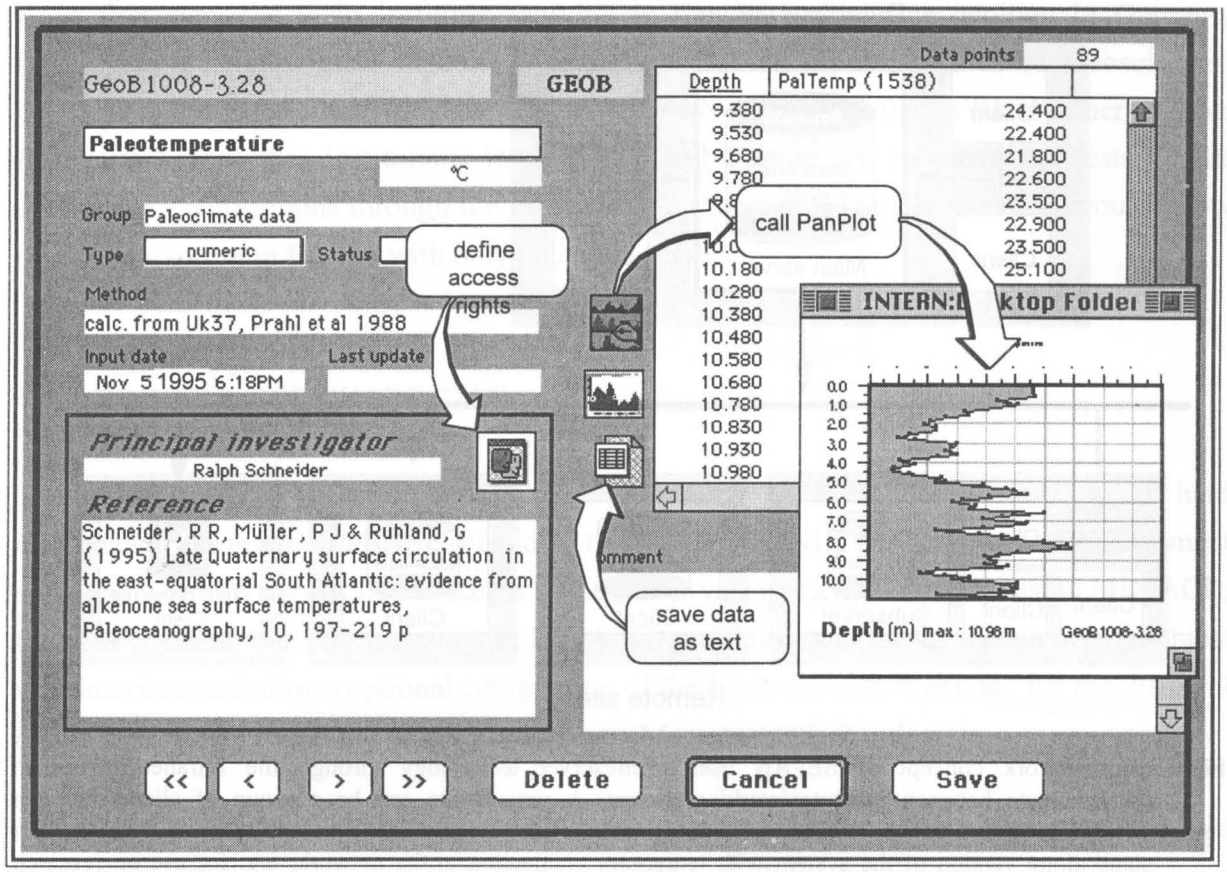

Fig. 2: The DATA level of SEPAN provides the user with the analytical data of a selected data set in combination with all related meta-information necessary for understanding the data. The window also allows the definition of access rights, editing of the data and the export of data as text files or graphics.

From this field, data can be exported as a table or plotted graphically. The parameters are gathered into parameter groups for a better overview and data are grouped as primary, secondary and tertiary data as described above - data types can be numerical, textual or pictures. The combination of the DATA, 'Parameter' and 'Method' fields is the essential part of the model, which allows the definition and storage of new, unique parameters by the user at any time. The middleware allows the user to retrieve complex data matrices, e.g. time slices.

3. NETWORK

SEPAN uses client/server technology through various local networks and the Internet to communicate between working groups (Fig. 3).

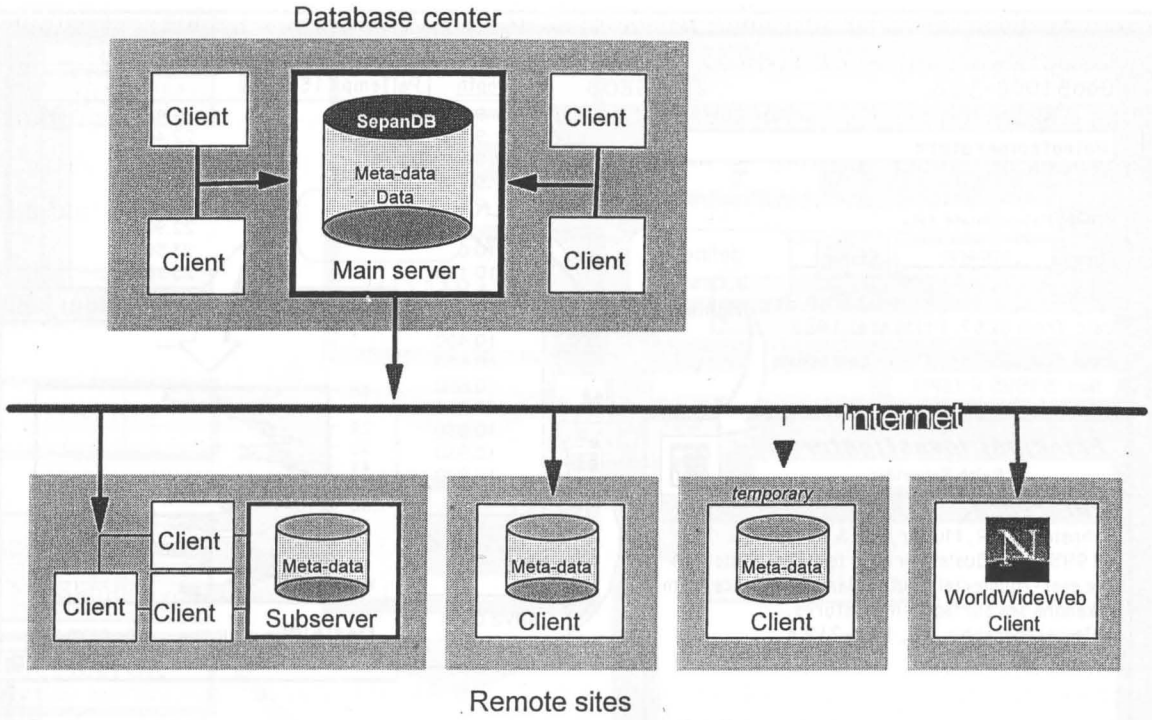

Fig. 3: The network concept of SEPAN uses client/server technology through the Intranet/Internet to communicate between institutes/working groups. A remote site can be a group of clients using a subserver where all meta-data are mirrored, a single client directly connected with the main server, or a stand-alone version of the system with temporary connection to the network, e.g. for use on research vessels. Any of these clients will have full access to the functionality of the system. For retrieving published data from SEPAN, a WWW interface will be provided.

The main server, located in the computer center of the AWI, is connected through the Internet with subservers at various external institutes (remote sites). The clients, which are the personal computers of the scientists, are connected through the local network to the subserver of their related institute. To increase the access speed all meta-information is mirrored on each local subserver. The mutual update of newly imported meta-information is made in the background through the network. The update is based on optimistic strategies, thus avoiding the problems connected with the handling of complex data dictionaries in distributed databases. With the development of reliable high speed electronic networks this feature may become obsolete. Clients may also connect directly to the main server. In this case the meta-information is mirrored on the client (Fig. 3). In addition, the system will provide read only access for metadata and published analytical data via the World Wide Web (WWW).

As a first step, the main institutes working in the field of marine geology in Germany (Geological-Paleontological Institute, Kiel; GEOMAR, Kiel; Department of Geosciences, Bremen) were connected with the central SEPAN server at the AWI. Further remote installations for other institutions in Germany and Europe are in progress. Tests of the client/server connections through the Internet have shown that the transfer rates would allow this system to run in Europe with only one main server.

4. HARDWARE, SOFTWARE, DATA

The main server is a DEC Alpha 8200 (four processors, 2 GB internal memory, 50 GB hard disk capacity) running SYBASE Version 11 under DEC/UNIX as the database management software. The client software for access to the server was written in 4th Dimension (4D, ACI); the WWW-client software is written in JAVA. 4D provides tools for the design of a graphical user interface and allows optional compilation of the front end software code for the different operating systems found in personal computers (MacintoshOS, Windows).

The client software was modularized into a database front end together with tools developed individually for processing specific data sets. The modularization and open environment facilitates the future adaptation of the system. The entry requirements for handling the software are low because the functionality is uniform for all tables and tools. Updates of the 4D front end software only have to be made on the subservers, so that the PCs are not affected by the update procedure.

The system requirements for running SEPAN at an institute are an Internet connection and a fast Macintosh or Windows computer with at least 80 MB of RAM, which has to be used for storing the mirrored meta-information. For subserver systems, additional licenses for the 4D server software are needed, depending on the number of clients.

The available meta-data comprises related information about expeditions, sampling sites/sets and storage facilities. In addition to the sampling/investigation site label, the most important meta-data are the location (latitude/longitude) and the elevation. For archiving profiles the definition of two locations/elevations, including the date/time, is possible. All scientists and institutes related to the data in the system are stored with their full addresses. Other related items, such as the names of ships, gears or sample types, are defined in lists which are regularly updated. The parameter list is organized in groups and consists of about 2000

parameter definitions of different types used in marine research. Accessible from the SITE field, a site description can be stored as a graphic. Analytical methods can be defined with all the necessary information. References for cruise reports or published data can be typed in or imported from professional bibliography software.

5. TOOLS

The import of meta-data is organized through predefined form tables which are available for the import of references, cruises, stations/sites and curatorial information. Analytical data are imported via simple tabulator delimited text tables with the name or the SEPAN-ID of a specific parameter in the header of the input matrix. Meta-information related to the data (method, owner, comments) have to be defined before the import and are also updated during the import procedure.

The retrieval tool for finding and extracting data from the system is uniformly designed for all levels and allows the use of complex combinable search criteria relevant to the desired data. Data can be exported as tables or plotted with one of the graphic tools. Tables can be sorted and configured individually. Multiple data sets can either be displayed with identical parameters and locations in one column, or the data can be split by data sets and location into separate columns, thus allowing the comparison of data sets from different investigators or multiple versions of a single data set.

For the geographical presentation of data the SEPAN tool PanMap was developed, which is either directly connected to the database front end or can be used as a stand-alone application (Fig. 4).

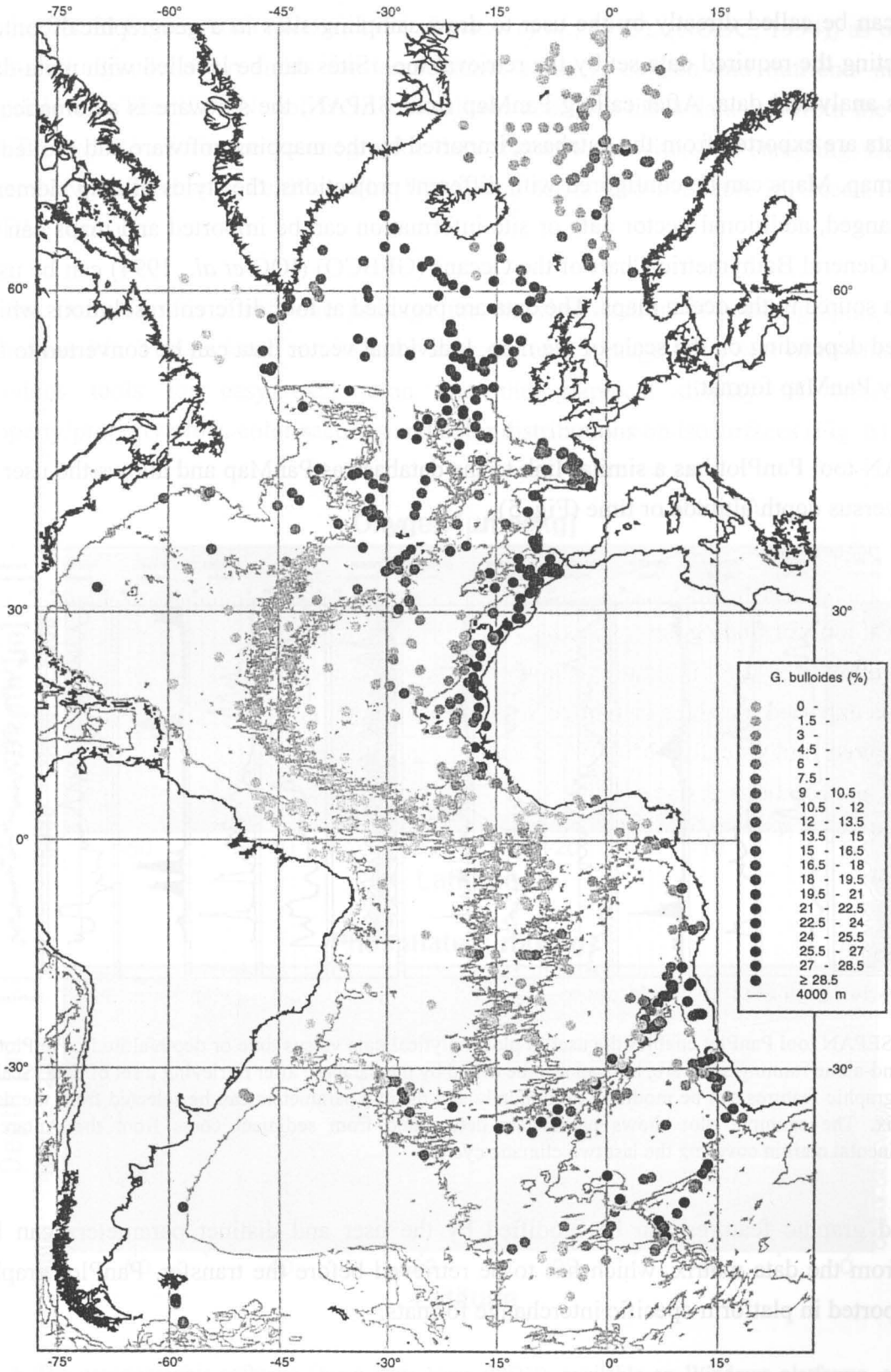

Fig. 4: For the presentation of data in geographical maps the SEPAN tool PanMap was developed, which is either directly connected to the database front-end or used as a stand-alone application. PanMap can be called by SEPAN to draw sampling sites labeled with meta-data or analytical data. The map shows sampling sites in the Atlantic where planktonic microfossils were investigated. The grey shading of dots is related to the percentage of a specific species. The bathymetric data source is the General Bathymetric Chart of the Oceans (GEBCO) (IOC *et al.*, 1994) .

PanMap can be called directly by the user to draw sampling sites in a geographical context after selecting the required data set by the retrieval tool. Sites can be labelled with meta-data as well as analytical data. After calling PanMap from SEPAN, the software is automatically started, data are exported from the database, imported by the mapping software and plotted in a default map. Maps can be configured with different projections, the styles of map elements can be changed, additional vector data or site information can be imported and maps can be exported. General Bathymetric Chart of the Oceans (GEBCO) (IOC *et al.*, 1994) can be used as the data source in the ocean maps. The data are provided at four different resolutions which can be used depending on the scale of the map. Individual vector data can be converted to the proprietary PanMap format.

The SEPAN tool PanPlot has a similar link to the database as PanMap and allows the user to plot data versus depth/altitude or time (Fig. 5).

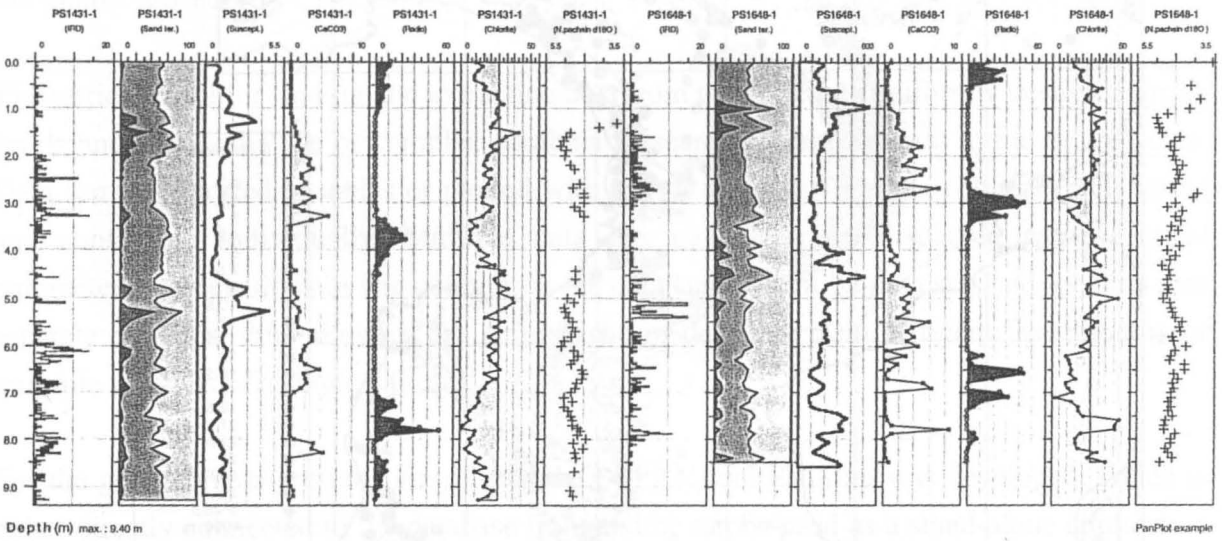

Fig. 5: The SEPAN tool PanPlot enables the user to plot analytical data versus time or depth/altitude. PanPlot is a stand-alone running software, but can also be called by the database after retrieving a set of data. Scales and graphic features can be modified individually and distinct parameters can be selected from the data matrix. The example plot shows parameters determined from sediment cores from the Antarctic continental margin covering the last two climatic cycles.

Scales and graphic features can be modified by the user and distinct parameters can be selected from the data matrix, which has to be retrieved before the transfer. PanPlot graphs can be exported in platform-specific interchange formats.

Paleoceanographic research also requires access to the actual oceanographic data of the oceans. The National Oceanographic Data Center (NODC, Washington), provides these data

in the World Ocean Atlas (WOA) 1994 (Levitus *et al.*, 1995), (NODC, 1994)) as one degree latitude-longitude mean fields for temperature, salinity, oxygen and nutrients at standard depth levels. The SEPAN tool WOAccess allows access to these data sets from the SITE and DATA level to obtain the oceanographic data closest to the site(s) of sampling. Data can be directly visualized with PanPlot. The WOA data set is also available in Ocean Data View (ODV) format.

Ocean Data View is a software package for the visualization of oceanographic data on a Windows PC. The package can be used to create and manage large sets of marine data and provides tools for easy exploration and the graphical display of these data as property/property plots, color sections and color distributions on isosurfaces (Fig. 6).

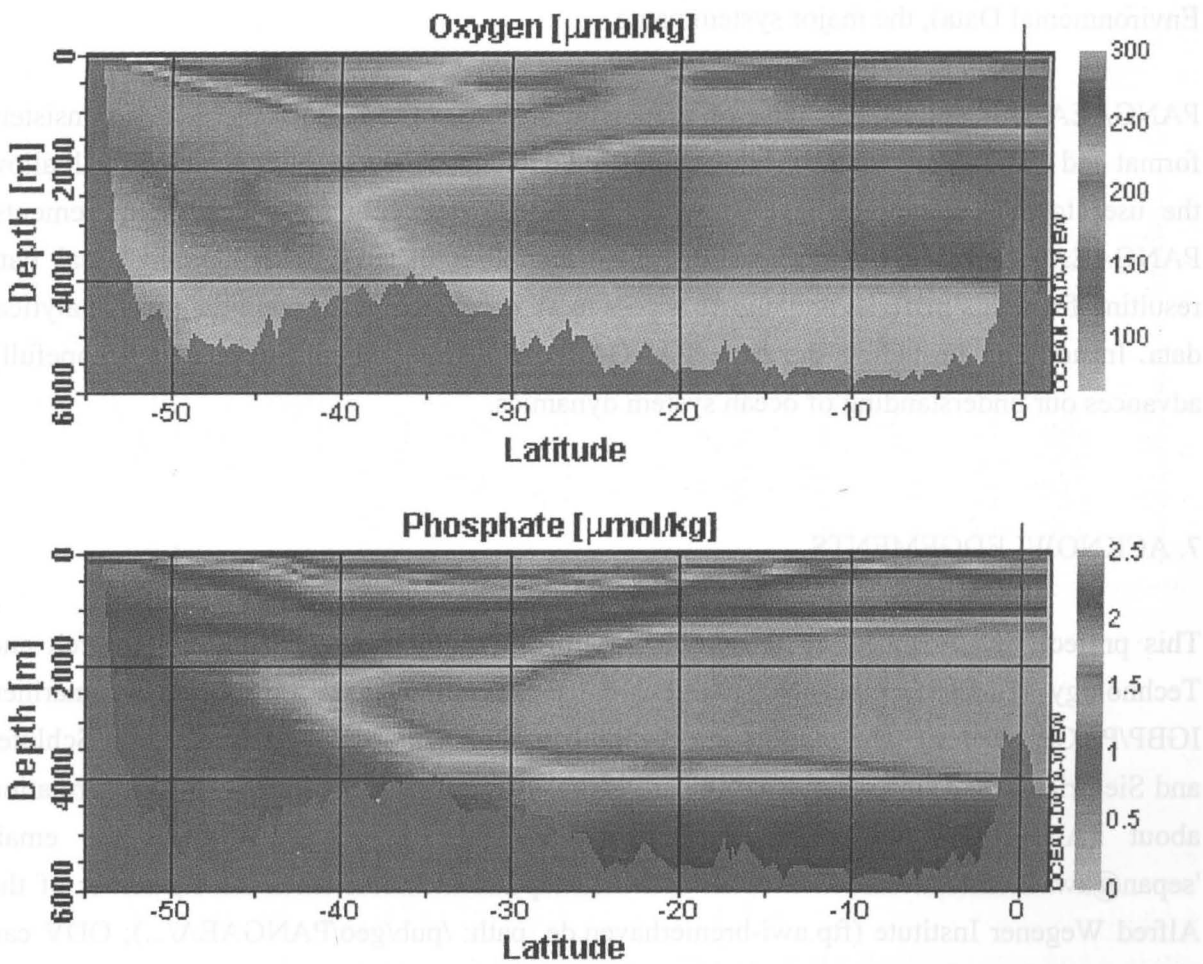

Fig. 6: The oceanographic software Ocean Data View (ODV), available on Windows platforms, can also be linked to SEPAN. ODV draws data in depth/time slices or in sections. Simple interpolation routines are included in the software. In the sections given the oxygen and phosphate concentrations are plotted along a N-S transect through the western South Atlantic.

The data collection format of ODV is optimized for dense storage and direct data access allowing the storage of large station collections on desktop computers. ODV can be called from SEPAN and data will be provided in a format which can be imported into ODV.

6. CONCLUSIONS

The data model of SEPAN is universal and allows the storage of any site-oriented marine data. Because it is not useful to save data from any site or data source in one system, SEPAN can be cloned and the clone used to store, for example data from ice, lakes or terrestrial sources (Lake and Terrestrial Data Information Network, LATIN). All clones will be subsystems of the PANGAEA information systems (PaleoNetwork for Geological and Environmental Data), the major system name.

PANGAEA will collect any data of a specific scientific field, store them in a consistent format and give overall access. The structure and data model of a PANGAEA clone will allow the user to start comprehensive retrievals to extract data sets for specific requirements. PANGAEA combines the information about sampling material with the analytical data resulting from this material and allows access to all combinations of meta-data and analytical data. In addition providing the data, PANGAEA is a new scientific tool which hopefully advances our understanding of ocean system dynamics.

7. ACKNOWLEDGEMENTS

This project was financed by the German Ministry of Education, Science, Research and Technology (BMBF), Fund No: 03F0131B (Paläoklima-Datenzentrum für die marinen IGBP/PAGES-Daten). The authors are thankful to Hans Krause, Beate Marx, Jens Schlüter and Siegfried Makedanz for managing hardware, software and network. Detailed information about PANAGAEA information systems is available on request through the email 'sepan@awi-bremerhaven.de'. PanPlot and PanMap are available from the ftp-server of the Alfred Wegener Institute (ftp.awi-bremerhaven.de, path: /pub/geo/PANGAEA/...); ODV can be accessed through http://www.awi-bremerhaven.de/GPH/ODV/.

8. REFERENCES

Diepenbroek M, Reinke M (1995) Publishing scientific data - a strategy for the integration of heterogenous and dynamic data environments. IGBP Informationsbrief 19: 7-9

IOC, IHO, BODC (1994) Supporting volume to the 'GEBCO Digital Atlas' published on behalf of the Intergovernmental Oceanographic Commission and the International Hydrographic Organization as part of the General Bathymetric Chart of the Oceans (GEBCO). British Oceanographic Data Centre, Birkenhead with CD-ROM:

Levitus S, Conkright ME, Gelfeld RD, Boyer T (1995) World Ocean Atlas 1994. IGBP Newsletter 20: 4-6

NODC (1994) World Ocean Atlas 1994. National Oceanographic Data Center set of 9 CD: