Information System for the ESF/QUEEN Programme

QUEEN/PANGAEA

MAS3-CT98-0185

FINAL REPORT

Summary

The QUEEN/PANGAEA project compiled recent and quaternary environmental data from a range of international and national projects in the Eurasian Arctic. More than 1000 data sets including metainformation and references were imported to the information system PANGAEA. The first data library related to the quaternary environment of the Arctic is now available to the public through the Internet (http://www.pangaea.de/Projects/QUEEN). The data sets are from all earth systems (atmosphere, geosphere, hydrosphere, biosphere, kryosphere) and allow new insights into the causes, impacts, and feedback mechanisms which determined the Arctic climate system during the younger geological history, and still do today.

Technically the system is operated by the AWI and MARUM institutes; both are committed to run the system on a long-term basis. The system is still in use by running and future projects, new data are imported every day. Different clients have been developed to allow user friendly and flexible access to the data. The system is a further step towards a standard in archiving geological and environmental data and thus will help to investigate scientific data more efficiently. The organization of information in a relational system with different access clients on the Internet and the opportunity to investigate a comprehensive data inventory in detail, provide a new and unique opportunity for state of the art data management in scientific projects.

Introduction to the QUEEN programme

The Arctic is one of the most important regions concerning processes related to Global Change and thus is also of eminent importance for understanding European climate. The high northern latitudes comprise some of the most sensitive elements in the global systems, and are considered to respond rapidly to climate change, as well known from e.g. the decreasing sea ice coverage of the Arctic Ocean. However, our knowledge of the processes driving the Arctic system in the past and today is still very limited, thus making it difficult to predict the behavior of these processes and related future climate scenarios.

To safeguard the future for the European population, it is highly important to understand the climatic history of the Arctic. For this reason the European Science Foundation (ESF) decided to support cooperation and initiate synergisms in the complex network of

environmental and climate research projects in the Arctic and, in early 1996, established QUEEN (Quaternary Environment of the Eurasian North) as one of its programmes.

The aim of QUEEN is to facilitate the exchange of new discoveries, data and knowledge of the Late Quaternary environmental history in the Arctic. QUEEN attempts to synthesize the new results gained from the many former and current projects and to stimulate new investigations in areas which may hold clues for understanding environmental processes better. The focus for QUEEN was set on the Arctic regions of Europe and Russia, an area which in former times was accessible to environmental research only under circumstances of great difficulties because of political reasons. Major progress in studying the QUEEN regions can now be seen, mainly because of the establishment of a number of bilateral projects between research institutions in Russia and in western countries.

QUEEN attempts to correlate established and dated records of the Quaternary history of the Eurasian Arctic and thus to contribute to the reconstruction of the former climate. From these records it is well known that peak interglacial times - as we have them today - were rare exceptions from the more normal glacial conditions. Different atmospheric and oceanic circulations resulted in a substantially colder climate over Europe than we have today, mainly because the extension of the Gulf Stream current did not enter the Norwegian-Greenland Sea. The instability and variability of the regional extension of the temperate climatic zones into high northern latitudes affects mostly northern Europe. Due to this scenario northern Europe experienced the fastest most dramatic climatic changes over the youngest geological past than any other environment on Earth. The most recent studies and climate models indicate that the high latitudes may also be subject to substantial changes in the future if mankind-induced climate change continues as predicted.

The most prominent expression of the properties of the Late Quaternary climate over the northern hemisphere are the waxing and waning of large ice sheets on the circum-Arctic continents. As part of QUEEN, the EU-project *'Ice sheets and climate in the Eurasian Arctic during the last Glacial maximum'* is based on a multi-disciplinary approach where Russian, German and Scandinavian scientists concentrate on the reconstruction of the Late Weichselian ice sheet in Eurasia. It combines new field research in the Russian Arctic with glaciological modeling to reconstruct the growth and decay of the last ice sheets in the Kara and Barents Sea region and the contemporaneous climatic conditions. A second EU-project, also funded within ENRICH, was established to improve the

visualization of glacier and ice sheet-related observations, in using a geographical information system (GISICE).

Through this project (QUEEN/PANGAEA), the EU has provided financial support for the establishment of a technical infrastructure to archive and exchange data from the Arctic. Due to the complexity of environmental research and the variety and amount of different parameters, the database established through this relatively small project can not be more than a show case. Anyhow this project is able to demonstrate the power of a common information system for the long term archiving of data, for an easy exchange of information through the Internet and the value of a consistent data inventory for the interpretation and visualization of comprehensive data collections. This might also help increase the motivation of scientists to establish common data management in future projects.

Introduction to the Information System PANGAEA

In 1987, a group of marine geoscientists at AWI started to develop a database on a mainframe to archive marine geological data, including any metainformation needed for the administration of a repository for sediment cores. The system was redesigned to a client/server system in 1992. Based on the resulting data model, requests from the community and the requirements for research, the system was extended between 1993 and 1997 to a universal information system for geocoded data: PANGAEA - Network for Geological and Environmental Data; additional funding was supplied through the German Ministry for Research and Technology. Since 1996, the system has been used for project data management, allowing storage, sharing and publishing of data on the Internet. In particular it allows the evaluation of data in a new scientific way, once a large inventory has been built. Technically the system is operated by AWI (Alfred Wegener Institute for Polar and Marine Research) and MARUM (Center for Marine Enviornmental Sciences); both are committed to run the system on a long-term basis.

In April 1998, the domain *www.pangaea.de* was established with a very modern and sophisticated technology (Java applet) to access the tables of the relational database. The development of the different access clients on the Internet ended with this project in 2000, with the invention of the first universal search engine on a relational database for scientific data (PangaVista). During the last years several international and national projects have used PANGAEA to archive their data and to make results available to the

public (http://www.pangaea.de/Projects/). The number of hits from different hosts increased since the opening of the pangaea-domain in 1998 to about 7000 per month.

The most recent operational version of PANGAEA is designed as a network which uses client/server technology through the Internet. Clients for import are connected to the central database server via local sub-servers with mirrored metainformation for faster import, retrieval and navigation. The challenge of managing the heterogenic and dynamic data of environmental and geological research was met through a flexible data model which reflects a strictly generalized 'world' of scientific data.

The logical hierarchy of the data model follows the way environmental research is mostly organized. Within a PROJECT different CAMPAIGNs are carried out for investigations. At a number of SITEs one to several EVENTs are carried out to take samples or measure something. At distinct points/intervals, the medium to be investigated (e.g. sediment, water, ice) is sub-sampled or measured for different requirements (SAMPLE). From each sample analytical DATA may result. To ensure consistency, additional tables are used to fill fields with predefined metadata (e.g. method, gear, PI). The data model allows the definition of new parameters at any time and thus can be extended ad libitum.

Each single data point is geocoded in time and space. Spatial codes are *latitude*, *longitude* and any vertical position or lateral extension in the medium to be investigated. The system provides different choices for this third geocode which can be *depth in sediment*, in *water* or in *ice*, *elevation*, *altitude* or just relative *distance* on a measured profile e.g. through a tree or a shell. If no third spatial dimension is given, just an *ordinal number* can be used to keep data in its original sequence. Orientation in time can be given by *date/time* with a precision of milliseconds for time series of measurements (e.g. daily temperature in permafrost profiles). Also *age* in years before present can be used to plot data on a common geological time axis. The geocoding of all data allows the extraction of any individually configured subset from the inventory by defining spatial and temporal limits with the advanced retrieval tool.

Data gathering and work flow

The collection and publication of data as the major work do be done in this project was organized and mostly executed by the data manager at MARUM. In total more than

1000 data sets, comprising 36 000 data series with nearly 3 million data points in total from members of the QUEEN community and from related projects were made available. Besides the data stored in the relational system, a considerable amount of data sets, necessary as background information for the interpretation of results were made available as flat files on the PANGAEA web server.

(In the definition of the PANGAEA data model, a data set is a matrix which consists of several lines (sample points) and columns (parameters), e.g. the results of a chemical analysis of a sediment section or a water profile. A data series is one column within a set, e.g. just the iron content in the data set as described above. A data series might have one to n single data points.)

Different sources of data had to be included in the gathering process:

(1) Due to the limited access on already published data in electronic form from the principal investigators, a significant amount of data was typed in by hand from printed tables in publications. In a few cases also the geocodes for the position had to be digitized from the printed map. Data have mostly been made available in an output format as they have been printed, even though they were fully integrated in the internal consistent standard format of PANGAEA.

(2) The data manager has tried to keep track of ongoing work within the various QUEEN groups to get data sets of recently published or submitted new papers. PANGAEA allows the definition of access rights to PIs or groups and thus data have been password-protected on request of the authors until the final paper was printed.

(3) A major contribution to QUEEN and the *Eurasian Ice Sheet* project for the reconstruction of former ice sheets are the facies logs and age determinations from outcrops and profile sections in Siberia. The logs as the primary and most important descriptive documentation from the field work existed in hand-written form only. Prior to the digitization, the format of the tables and the abbreviations for the lithologic descriptions have been standardized by one of the partners (Per Möller, Lund). The standard files have been submitted to the data manager as tab-delimited text files, formatted for import.

(4) Important published data collection of general value to the community have also been imported; e.g. data from the GRIP ice core are of importance to those groups working on sediment profiles for the interpretative comparison of climatic records from different archives or for stratigraphic correlation.

(5) PANGAEA also allows the definition of data set descriptions with links to external data sets, sources or centers. If the mirroring of data on a second site seemed not to be useful (e.g. because the data content is dynamic), only a static link to the data source was stored. It is known, that a link (URL) is seldom static on a longer perspective and thus this technique was only used for well-established web sites.

(6) The definition of a data set description with a link pointing to a certain data archive was also used for the storage of flat files in the directory structure of the PANGAEA web server where appropriate (core photos, graphics etc.).

The technical work flow of data acquisition, collection and import can be subdivided into different steps from the source to the final data link:

1a) Data as defined in the 'letters of intent' had to be prepared at the partner institutes.

1b) Data from various other institutes, cooperating in QUEEN, were acquired.

1c) Important data, only available in printed form, had to be typed in spread sheets.

2) All electronic data files were collected at MARUM.

3) The data files had to be quality controlled, harmonized and converted to the import format.

4) The metadata were collected and imported to PANGAEA.

5) The import of data included the definition of the relations between metadata and data and the configuration of the output formats.

6) The accessibility and output format of all data sets were verified on the web client.

Data management problems in general include missing data set documentation, errors in data, inconsistent formats, delayed or no submission of data, complicated definition of access rights, unknown quality or validation status and communication. Most of these problems have been discussed in detail by Lowry (1998) who is thought to be one of the most experienced data managers in Europe. All problems as discussed in his publication as an example of an international EU project are common in data management and have also been encountered during this project. They will not be discussed again in this report.

Technical improvements of PANGAEA

The use of information systems is still not very popular in most scientific disciplines. Scientists prefer to deal with their personal data organization which, in the best case, might be a local, proprietary data base, in the worst -and more common- case is just a collection of spread sheet files in different formats at different locations and different storage media. The heterogeneity of the data and the increasing complexity of science hampers the invention of common data systems and standards in data formats. There are few positive examples from oceanography, were data are well organized and easily accessible on the web (e.g. from the WOCE project at *http://www.cms.udel.edu/woce* or *http://www.awi-bremerhaven.de/GEO/eWOCE/*); mostly in this case the

oceanographic data consist of just a few well defined parameters. In other disciplines with more complex results, e.g. in bio- or geosciences, very few web sites are known where a considerable amount of data is available to the public. One of the first successful steps towards publicly available environmental data archives was done in a marine research project by the BODC (British Oceanographic Data Center) with the data of the OMEX project (Ocean Margin Exchange). The amount of work and problems to solve were remarkable (Lowry, 1998, see above).

It was also experienced in several projects that even after demonstrations of the information system on how to define a query and download data, the participants still refused to work with it in the sense of the projects objectives. Even though the retrieval is easy to use, some training is still needed due to the complexity of the data model necessary for environmental data. Almost none of the participants had gained any deeper insights into the system, even though the projects aim was to evaluate a comprehensive data compilation with a tool like PANGAEA.

To improve the acceptance and the common usage of the data base by the scientific community, new access clients have been developed, some of which are unique so far. The clients are provided in different levels of complexity and can be used depending on the needs of the user and the definition of the query (*http://www.pangaea.de/Retrieval/*).

In a modern research project any information to the public, its exchange between the participants and the publication of results is organized through web pages. The duration of a project can be between 1 and more than 10 years but is generally around 3 years. This also implies that the life time of the web pages is limited and thus not too much effort should be put into their design and functionality. If a significant part of the data management resources were used to develop web pages even with some functionality instead of gathering and archiving data, this task of the project is obsolete after its termination.

Also providing the data with proprietary access tools and in special formats, not following any standards, might be useful for the lifetime of the project or for individual scientists but can very seldom be handled after the project ends and the developers, web masters and data curators have moved to another project or institute. In consequence, after a short time, data are mostly lost and web pages are out of order. This process is intensified by the increasing speed of progress in computer technology and software.

During the last 2 years, the PANGAEA group has tried to find solutions, to minimize the effort to be put into the design of project web pages for data access and also to establish standards in data archiving. The following paragraphs will give a short review of the different versions of clients which can be used, depending on the requirements of projects and users.

4th Dimension client for import/export

Data are imported to PANGAEA by the data curator with a proprietary graphical user interface. This client is written in 4th Dimension, a software which is available for the most common operation systems on PCs (MacOS and Windows). The relations between data and metainformation are made during the import procedure. Shortly after the import, data can be retrieved worldwide on the Internet with one of the access clients listed below. The 4D-client also allows retrieval and export of data; due to the client/server concept, users are able to import and investigate data very quickly and efficiently.

Advanced Retrieval Tool (ART)

ART is the 'high end' web client, based on Java technology. It enables the user to access any table of the model, to individually configure output tables and to search through the complete inventory of available data. ART was designed as a tool to investigate comprehensive data collections for scientific purpose which presumes that the user has a minimum of knowledge about the data model, how to formulate a retrieval and about the definition of the project related parameters. Even though this is one of the best database clients on the Internet, it has the problem that most of the users refuse to invest half an hour of time to learn how to use it.

Metadata query definition

ART also enables the user or curator to save any specified output format of a data matrix and to define and save the related query to extract a required subset of data from the system. Format and query file can be stored on a web server and related to a link on a web page. This procedure requires that the curator designs some web pages, including at least the most important metainformation for the data set, but ultimately users have access to any individually configured data set just by a mouse click. Access in this way is used for dynamic and 'on demand' produced lists of metadata, e.g. sampling sites, references or project members.

Direct Download Interface (DDI)

In the next step, a data set was defined by the relational system through a unique ID which could later be used in a web link to access and download it to a users PC. This technology already enabled the user to access the data from the relational data base just by a mouse click on a web page but still the metainformation around the link had to be written in html-code to produce the web page. The DDI is used for the design of individual but well established data pages which will be used for a longer time (e.g. personal publication home pages, which are offered to the community as a service of the data center).

PangaVista

The 'easy to use' solution for individual access is the PANGAEA search engine called PangaVista, the name in imitation of AltaVista. From one given expression or a combination of expressions the system produces a list of related data sets including documentation and links to download the data in text- or html-format. The query can be defined as a URL (web address) and thus can be used to build up web pages dynamically. This functionality is now extensively used in project data management with PANGAEA, the QUEEN data access page being the most prominent example. There is no easier or more efficient way of searching for data or producing web pages. Any administration of metadata, data, output configuration of tables, access rights and formats is done solely within the information system. Even though it looks as if the data are downloaded as flat files from a directory structure, using the links will always provide the user with data from the relational structure and, as the new advantage, the resulting lists will show the current status of the inventory at any time.

Performance and safety

Important for the acceptance of a system delivering data to a user are also a short response time and high transfer rates. In PANGAEA this is ensured by fully indexed tables of the relational data model, resulting in an index tree with four times the storage capacity compared to the real data. But due to the constantly decreasing prices for storage media, modern information systems will always have access to sufficient storage capacity.

As described above, data are organized in series and sets; several sets might belong to a reference and are listed on a dynamic web page as a result of a request. The sets including its metaheader are stored in a memory cache of the database management system once they have been retrieved for the first time. In consequence any following request is ten times faster in delivering the data directly from the cache.

Concerning its integration in the infrastructure of the operating computer center at AWI, the data base mainframe is connected with a band width of 34 Mbit to the Internet and thus transfer rates between host and users might only be limited by bottle necks at certain parts of the Internet or in the Intranet of other institutions.

To gard against loss of data, a backup procedure is installed with an incremental backup every day and a weekly full backup on tape, stored in a fire proof safe.

Common software for mapping

The comparison and common interpretation of results from different groups on maps or plots are often hampered by using different scales, geographical data and software. This problem becomes evident in a project like QUEEN, dealing with large scale reconstructions on maps. To improve the exchange of information and the process of discussion of results on maps, during the QUEEN workshop in Potsdam we discussed how common map software and geographical resources could be found to harmonize the presentation of geocoded data and thus improve discussions of results between different groups. It was decided that the PANGAEA group will look for software and provide geographical resources for the area of interest.

This Mini-GIS (Geographical Information System) PanMap was recommended which can plot vector and point data in maps of different projections; any ASCII-file, where scientific data are accompanied with the latitude/longitude information can be imported and converted to the internal layer format. PanMap was developed as a visualization tool for data from PANGAEA, but it can also be used as a stand-alone running application with any geocoded data on both PC operation systems (MacOS, Windows).

PanMap is freeware and can be downloaded at *http://www.pangaea.de/Software/PanMap*. A manual in German and English was written and additional geographical data, in particular from the Arctic, have been converted for use with PanMap. We provide GTOPO30 for land elevations and GEBCO for ocean bathymetry. The recent compilation of both GEBCO and GTOPO30, including new Russian bathymetry for the Arctic Ocean was published by Martin Jakobsson et al. in the International Bathymetric Chart of the Arctic Ocean (IBCAO). This gridded data set has been vectored in 100 m contour lines, using a professional GIS system (ArcInfo). The resulting vectors in GF3-format were converted to the layer format of PanMap and listed as compressed file archives per depth contour on the PanMap web page for download.

PanMap was developed as part of the PANGAEA project some years ago; it was proven to be useful prior to this project (it is listed in the top 10 of an international GIS Directory at *http://www.tenlinks.com/mapGIS/products/mac.HTM*). The geographical resources for the Eurasian Arctic and the manual have been prepared as part of the

QUEEN projects; the QUEEN community was informed about the availability of the new PanMap web page in June 2000.

QUEEN has three different adresses on the Internet:

1) Under the umbrella of the ESF, QUEEN is listed as one of its programmes on the web site of the ESF at *http://www.esf.org/lp/lp_016a.htm*

2) The home page of the QUEEN programme for the exchange of information is provided by the QUEEN secretary at GEOMAR
http://www.geomar.de/~hbauch/king/html/queen.html
This page includes a link to the data management page on the PANGAEA web server.

(3) ***http://www.pangaea.de/Projects/QUEEN/***
is the home page for this project (Annex). The page provides an example list of QUEEN related references and/or data set descriptions with direct links to download the data sets. The links are mostly build dynamically by using PangaVista, which can be used by those who do not need to work with ART. A password protection had to be established for some unpublished data due to ongoing publication or thesis work.

World Data Center for Marine Environmental Sciences

MARUM and AWI have successfully applied to get the status of a World Data Center for Marine Environmental Sciences (WDC-MARE). The WDC was established on the 2nd February 2001 and the web pages will be accessible in the near future. The WDC-MARE will cooperate closely with the WDC for Paleoclimatology in Boulder concerning any paleoenvironmental data from QUEEN. The cooperation is already defined through an agreement between Boulder, the PAGES office in Bern and the Bremen institutes, to share the PAGES part of the data inventory. Any marine data of QUEEN will also be available via the web server of the new WDC.

Workshops & presentations

1) 4/5.11.1997 During the first QUEEN data management workshop at AWI, Bremerhaven, the PANGAEA system was demonstrated to potential users and data contributors. Data flow and publication were discussed, parameters and formats were defined.

2) 13./14.12.1998 A demonstration of the system was given to the partners during the Eurasian Ice sheet/QUEEN workshop at the AWI in Potsdam. In the resulting discussion the standardization of facies logs and the establishment of a common map software were discussed and the responsibilities defined.

3) 14.-18.04.1999
During the QUEEN meeting in Bergen, a comprehensive overview about the status of the project was given, including a demonstration on how to use the Internet clients.

4) 01.-09.04.2000 The status of the project was reported during the QUEEN workshop in Lund, including a final call for data, given by the chairman of the QUEEN steering committee, Prof. Dr. Jörn Thiede.

Ongoing work and future projects

This project can be seen as the nucleus to a comprehensive library of environmental data from Arctic research, operated on a long-term basis and open to further projects. Additional data will be available through the web clients in the future and the inventory will grow through the following projects.

A new German successor project related to environmental research in the Arctic is Siberian River Runoff (SIRRO) which has started in 2000 and also uses PANGAEA as its data archive and scientific tool. The major aim of the project is the reconstruction of the rivers drainage following the last glacial maximum. Parts of the data inventory from QUEEN will also be helpful to SIRRO.

The new international project Arctic Coastal Dynamics (ACD) refers to the process-response system which operates on permafrost-affected coasts. The steering committee of ACD has decided that PANGAEA will be used to archive and distribute

observational time series from coastal key sites and geomorphological profiles, also providing the data in consistent formats for further GIS mapping and interpretation.

PANGAEA will also be used for the EPILOG initiative. EPILOG was formed by members of the paleoclimate community in order to establish a new and comprehensive look at the paleoenvironment of the last Ice Age, in particular to foster a revision of the CLIMAP reconstructions of the Ice Age earth. An information system, providing the available data in a consistentently formatted and geocoded inventory, is essential to extract geologic time slices of paleoclimate proxy parameters.

Within the PEP III community (Pole-Equator-Pole Transect) there are efforts to establish an information system for paleodata from different archives ('multi-proxy-parameter-network') including a common geological time scale to map all records. This system is also expected to become the new electronic home for the European Pollen Database and should function as an archive for all PEP III data. PANGAEA was offered to the PEP III community several times, but was rejected for political reasons. Concerning the highly specialized and relatively small amount of paleo-data and in order to attain a standard in data archiving, the establishment of additional systems for paleo-data must be avoided.

Recommendations for future data management projects

The data collection is by no means complete, a data collection from the Arctic can never be complete. During the course of the project new sources of data were identified which should be included in future projects. This includes not only data from American and European countries, but in particular from Russia. It is well known that comprehensive environmental data inventories from the Arctic, collected during several decades, exist in some Russian institutes. Those should be made available to the community, even though they are hard to get due to required permissions. It should also be mentioned that important parts of known Arctic data could not be included because some of the data holders did not even answer the requests which were made during this project via e-mail, fax or at workshops.

A great deal of environmental and geological data is still in a preliminary state. These data need to be validated before they can be imported into an information system. Since the costs for investigations in the Arctic as well as the costs for analysis are high, this hidden treasure should be made available. Data archaeology is more and more part of projects

with large scale investigations. A comprehensive overview and access to all data available from a specific scientific field or a geographical area of interest is essential for the success of such projects. One problem is that the necessary resources for final archiving in a database in most cases are not implemented or at least underestimated.

The major problem of any project dealing with data management is the aquisition of the data from the individual scientist, independent of whether the data management forms just a small part of the project or the major goal is to establish a data inventory. This has been experienced in any national, European and international project. The motivation of scientists to make data available to the community is still very low, the added value of data archives is not at all popular. This problem can only be solved on a long-term basis with different mechanisms to be discussed and finally to be implemented in the evaluation processes of projects and in the scientific work.

As experienced during this project as well as during former projects, the major work for the data manager was the collection of data from the various partners (30 %) and the harmonization of data, quality control and completion of metadata (40 %). Only the remaining 30 % of time is needed for import, publication and design of web pages if a well organized information system is used. If scientists would be more supportive in providing data, or even better, providing data in the requested formats, costs for data management could be reduced significantly. From the experience of data managers, there is no way to improve the data flow in the near future. This also implies that the amount of work needed for data management is still underestimated within the scientific community and in the funding agencies.

Also in QUEEN the amount of work for the collection and harmonization of data was underestimated. Nevertheless, an impressive amount of data was archived and published on the Internet. The recent inventory of data from the Arctic in PANGAEA is probably the most comprehensive one available in a relational database with retrieval and public accessibility through the Internet. Depending on the amount of data to be archived from future projects in consistent formats, PANGAEA is envisaged to establish a data standard for environmental data in general, with the paleodata being an integral part of it.

The partnership

Partner 1: Alfred Wegener Institute for Polar and Marine Research (AWI)

Columbusstrasse, 27568 Bremerhaven, Germany
http://www.awi-bremerhaven.de

Dr. Hannes Grobe (coordinator),
Prof. Dr. Hans Hubberten (AWI, Potsdam)
Dr. Rainer Sieger (data curator)
Prof. Dr. Jörn Thiede (Chairman of the QUEEN Steering Committee)

Partner 2: Department of Quaternary Geology, Lund University

Sölvegatan 13, 223 62 Lund, Sweden

Dr. Christian Hjort
Dr. Per Möller

Partner 3: Geological Institute, University of Bergen

Allégt. 41, 5007 Bergen, Norway

Dr. Mona Henriksen
Prof. Dr. Jan Mangerud

Partner 4: Institute of Geography, Russian Academy of Sciences

Staromonetny pereulok, 29, Moscow 109017, Russia

Dr. Serguei Arkhipov
Prof. Dr. Vladimir M. Kotlyakov

Partner 5: Arctic and Antarctic Research Institute

38 Bering Street, St. Petersburg 199397, Russia

Dr. D. Bolshiyanov
Dr. S. Pryamikov

Partner 6: Center for Marine Environmental Sciences (MARUM)
World Data Center for Marine Environmental Sciences (WDC-MARE)

University of Bremen, 28334 Bremen, Germany
http://www.marum.de

Dr. M. Diepenbroek (system manager and principal data scientist)
Prof. Dr. G. Wefer (Director of WDC-MARE)

The ESF was involved in beeing the umbrella organisation of the QUEEN programme.

European Science Foundation (ESF)

1 quai Lezay-Marnésia, 67080 Srasbourg Cedex, France

Dr. Michele Fratta (ESF Scientific Secretary)

Annual Report of the ENRICH project

Information System for the ESF/QUEEN Programme

(QUEEN/PANGAEA) MAS3-CT98-0185

submitted by coordinator Hannes Grobe

Workshops

3.-4. Nov 1997: workshop I at AWI Bremerhaven, discussion on the principles of
data management for QUEEN with representatives of all potential partners,
preparation of proposal.
13.-15. Dec 1998: workshop II at AWI Potsdam, demonstration of the functionality of
the system to members of 'Eurasian Ice Sheet' and other QUEEN groups.
Discussion of common geographic data resources to interpret and present all
data of the project in standardized, uniform maps.
15.-18. Apr 1999: QUEEN workshop in Bergen, introduction of the system and web
site to the full QUEEN community.
22. Jun 1999: visit to partner 2 (Hjort/Möller, Sweden) to define standards of profile
descriptions and forms for the submission of data, introduction to web site,
geographical resources and software.
7-10. Apr 2000: workshop for QUEEN and 'Eurasian Ice Sheet' in Lund, Sweden.

Scientific Report

During the first workshop the principles of data management for QUEEN with
representatives of all potential partners were discussed. The partners identified the
data sources and parameters. It was agreed, that the data to be collected should
comprise all archives of paleoclimatic information: marine sediments, terrestrial
records and continental ice. This would allow on a later step to correlate information
from different archives, implying, that geological datings are available to each
record to map data on a common time axis (calendar years before present).

Also a detailed introduction to the system PANGAEA was given and the buildup of a
data exchange and information network was discussed. Discussions and the flow of
information is coordinated by the secretary of the QUEEN project at GEOMAR; data
have to be delivered to the PANGAEA-group at AWI.

During the first year, communication lines with partners were established and a call
for the submission of data including definition of file formats for data and metadata
with example files (see attachment) was made. First test data sets were compiled by

partners and imported to PANGAEA. Tables of the data model with metadata needed to be extended for the purpose of terrestrial research; several new methods and parameters had to be defined. The system was primarily designed to store data from single points on earth, given with geographical coordinates, e.g. a profile in an outcrop, a CTD in the ocean or an ice core from a glacier. The system was extend to store also coordinates of vector information as a scientific parameter. This extension was e.g. important for the inclusion of the different versions of reconstructed ice limits from the last glacial in Europe and Asia. Mostly the ice limits are reconstructed from the interpretation of terrestrial datings, profiles and landscape observations; an ice limit is finally drawn by hand in maps which are published. The lines were digitized to coordinates and imported to PANGAEA.

Access on the World Wide Web

QUEEN has three different addresses on the Internet:
(1) http://www.esf.org/life/lp/Queen/lp_016a.htm
is an information page of the European Science Foundation (ESF), because QUEEN is a programme under the umbrella of the ESF.
(2) http://www.geomar.de/~hbauch/king/html/queen.html
which is the page of the QUEEN programme to distribute any programme related actual information. The page is managed by the QUEEN secretary at GEOMAR.
(3) http://www.pangaea.de/Projects/QUEEN/
is the home page for this EU-project. The pages will provide lists of QUEEN related references and data set descriptions with direct links to download the data. These pages are designed for users, who do not want to use the Advanced Retrieval Tool (ART). Password protection for some unpublished data are included. Operation of this web server and the URL for the QUEEN-data is long-term ensured by AWI and MARUM.
For individual queries on all data the Advance Retrieval Tool (ART) has to be used. Storing all data in a relational database also enables the user to extract any required individually configured subset of data with the ART through retrievals on all tables of the data model. A tutorial is available on the retrieval pages to learn how to use the ART.

Development and progress of the information system PANGAEA

The use of information systems is still not very popular in most scientific disciplines. Scientists prefer to deal with their personal data organization which, in the best case, might be a local, proprietary working data base, in the worst case is just a collection of spread sheet files in different formats. Mostly the complexity of the data of a specific scientific field hampers the invention of common data systems. There are few examples in oceanography, were data are well organized and accessible on the Web (e.g. WOCE, http://www.cms.udel.edu/woce); the reason in this case is, that oceanographic data consist of just a few well defined parameters and thus can be organized in an easy way. WOCE has 44 defined standard parameters, marine and terrestrial paleoclimate research may have several thousands. In other disciplines with more complex data 'worlds', e.g. in bio- or geosciences, very few sites are known, where 'real' data are available to the public. A first step towards data archiving is publishing a project related CD-ROM as done by the British

Oceanographic Data Center (BODC) or the Irish Marine Data Center. More complex marine data can be found on the web pages of US JGOFS (Joint Global Ocean Flux Studies, http://usjgofs.whoi.edu/jgdms_info.html); data of paleoclimate research are collected by the World Data Center for Paleoclimatology (http://www.ngdc.noaa.gov/paleo/data.html).

The purpose of these systems is to permit scientists to use data without concern for storage technique, location, or format. The problem is, that these sites are providing flat files on storage media, just archiving the existing diversity in data handling. It is not possible to send a query to extract a well defined subset from all existing data of such an inventory. If dealing with real information systems instead of data archives the query functionality seemed to be the most important part of the system if a 'data base' should be used also as a scientific tool to 'work' with the data through the Internet.

Data are imported to PANGAEA by the projects data managers with the graphical user interface which is provided through the 4th Dimension proprietary front end software. Relations between data and meta-information are also made during this procedure. A few seconds after the import, data can be retrieved on the Internet with the ART of PANGAEA (http://www.pangaea.de/Retrieval/) which is the scientific tool to handle, retrieve and extract any data in the system. The use of the ART implies, that the user has a minimum of knowledge about the data model, on how to formulate a retrieval and about the definition of the project related parameters. In case he has this knowledge, he can do nearly everything with the data in the inventory.

The problem is, that already a minimum of knowledge about the archives functionality hampers its acceptance. In many cases, just a central archive without any functionality is needed. It was experienced in the data management part of projects, that most of the participants refused to work with it in the sense of the projects objectives. Even though the retrieval is easy to use, it still needs some training which is mostly due to the complexity of the data. Nearly none of the participants had gained some deeper insights into the system, even though if the projects aim was to evaluate a comprehensive data compilation with a tool like this. In most cases, scientists refuse to use data bases if it takes more than a minute to find out how it works.

To improve the acceptance and the usage of the common data base within the projects using PANGAEA, a new tool has been developed which is unique so far. The Direct Download Interface (DDI) enables the user to access the data from the relational data base just by a mouse click on a web page. To establish the links, the curator has to use the ART, which enables him to define and save specific queries and output formats to extract a required subset of data from the system. Format and query can be stored on a web server and related to a link on a web page. This procedure requires that the data manager designs some web pages, including at least the most important meta-information describing each single data set (e.g. the full reference). Finally any user has access to the projects data just by clicking on links.

On the QUEEN data pages (http://www.pangaea.de/Projects/QUEEN), the users will have the choice to access data with the DDI. Even though it looks as if the data are downloaded from a flat file directory of the web server, using the links will always

provide the user with the actual data status in PANGAEA. The great advantage of the DDI is, that a click on a link e.g. related to a certain parameter will include any data fitting the query behind the link, also including those, which were imported after the establishment of the link.

The import of data sets and related meta-information and the establishment of links with citations on web pages can easily be done and is not very time consuming if a well organized information system like PANGAEA is used. As already experienced from other projects, also in QUEEN the major work for the data manager is
(1) the collection of data from the various partners and
(2) the harmonization of data and completion of metadata.
The amount of work of the data manager has been estimated as 30 % collection of data, 40 % harmonization, 15 % final quality control and import, 10 % publication of data with DDI, 5 % support/workshops/reports.
If partners would be more supportive in providing data, or even better, providing data in the requested formats, costs for data management could be reduced significantly.

Common geographical resources

Most scientists of the QUEEN community need to visualize data in a geographical context for further interpretation. To compare results between groups and to bring final results together in one map, it was agreed during the workshop in Potsdam, that it would be helpful to have a mapping tool for all participants which is easy to use. It should have detailed geographical resources for those areas of interest.

During the second phase of this project, we will try to invent the software PanMap as the common mapping software for the QUEEN community. PanMap was primarily designed as an easy to use Mini-GIS to visualize georeferenced data, coming out from PANGAEA. The software is available for the most common operating systems on PCs, MacOS and Windows (http://www.pangaea.de/Software/) and is freeware. To keep PanMap simple, it can only process point and vector data.

The major work to be done is the conversion of public available geographical information as elevations, rivers, lakes and cities from its proprietary formats to the PanMap-format. This 'quatrie'-format was created to organize large datasets in a way, that any subset of the data from a selected geographic area can be loaded in a short time. The software allows to build up a map in layers of information as usually available in GIS systems.

For elevation lines, the GTOPO30 data set was chosen to be the most suitable. GTOPO30 is a global digital elevation model (DEM) resulting from a collaborative effort, led by the U.S. Geological Survey. It was developed to meet the needs of the geospatial data user for regional and continental scale topographic data. GTOPO30 data are freely available on the Internet (http://edcwww.cr.usgs.gov/landdaac/gtopo30/gtopo30.html). Elevations are regularly spaced on a 30-arc seconds grid and thus had to be converted to 100 m elevation vectors. The conversion from grid to vector was done in the GIS ArcInfo® on a workstation. Data are available in 20x20° tiles through a clickable map.

The best available source for global bathymetry data is the General Bathymetric Chart of the Oceans (GEBCO). The fifth edition of GEBCO was compiled since 1972, initiated by the International Hydrographic Organization (IHO) and the Intergovernmental Oceanographic Commission (IOC) under the auspices of the British Oceanographic Data Center (BODC). The World Vector Shoreline (WVS), released by the US Defense Mapping Agency (DMA) was adopted as the standard coastline for use in GEBCO. The WVS is available in a high resolution and a draft resolution PanMap layer-format, GEBCO data are available in 4 different resolutions which can be used depending on the required scale; (reference: IOC, IHO, and BODC, 1994: GEBCO Digital Atlas, 5th edition, British Oceanographic Data Center, Birkenhead; http://www.nbi.ac.uk/bodc/gebco.html).

The attached map was produced using PanMap. It shows more than 500 sites of different kinds of investigations related to QUEEN from the area of interest, the 'Eurasian North'. PANGAEA has already stored more than thousand data series related to the sites shown. Sets include data from marine sediments, the marine water column, lakes, glaciers and terrestrial records comprising several hundred different parameters.