Alfred Wegener Institute

"There is no other scientific data model capable of handling such volumes and this high level of complexity. Since many diverse aspects can be combined in a single query, the system opens entirely new avenues of scientific examination. This allows us contribute significantly to the further exploration of our earth."

—Dr. Hannes Grobe, Scientist, AWI

INDUSTRY

- Life Sciences

KEY BENEFITS

- Facilitates fast, flexible evaluation based on many different criteria
- Delivers high compression rates and automated indexing
- Enables researchers to find specific subsets of information from a wide range of search criteria
- Enables new areas of scientific focus

SYBASE TECHNOLOGY

- Sybase IQ

The German Alfred Wegener Institute for Polar and Maritime Research (AWI) and several research centers at Bremen University have jointly established an information system for geo-referenced measuring and monitoring data that will make climate and environmental research information available to the general public, as well as to the scientific community for academic study. Named PANGAEA, the long-term archiving system will store billions of data sets from earth-system research. This information is made available to scientists via web portals for further evaluation and analysis. As a data warehouse, this system has become a major knowledge base for many geoscientific disciplines.

THE INSTITUTIONS

Founded in Bremerhaven, Germany in 1980, the Alfred Wegener Institute for Polar and Maritime Research (AWI) conducts research in the Arctic and Antarctic regions, as well as marine areas in temperate regions. The public law foundation, a member of the Helmholtz Association of German Research Centers (HGF), coordinates German polar research projects and provides the logistics support needed for polar expeditions. The institute has approximately 780 employees.

The Alfred Wegener Institute's Marine Environmental Science Center (MARUM) and the German Research Foundation's Ocean Margins Research Center (RCOM) centers strive to understand the role of oceans in the "earth system," applying advanced geoscientific methods. For over 10 years, these centers have collaborated with scientists from the University of Bremen, Max Planck Institute for Marine Microbiology, Bremen Center for Tropical Marine Ecology, Senckenberg Research Institute, and Wilhelmshaven.

FLEXIBLE TIME SERIES ANALYSES ON BILLIONS OF DATA POINTS

Nutrients in the oceans, such as nitrate and silicate, influence plankton growth, which then directly impacts our climate. Consequently the distribution of plankton in the oceans of the world is a major parameter in climate research. In order to find out how the nitrate and silicate concentrations changed during the 1980's, Dr Rainer Sieger, an AWI scientist, selects these two criteria from a list of 46,000 parameters offered by the PANGAEA data system. He set a time period from January 1st, 1980 to December 31, 1989, and limited the data to be evaluated to readings taken at a depth between 45 and 55 meters below sea level. Within seconds the requested "information slice" is extracted from roughly 1.6 billion datasets in the information system. The resulting data table is formatted by a special graphics application, allowing Sieger and other scientists to study local concentration readings on a multi-colored world map.

In an attempt to better understand environmental changes, climate and environmental researchers are constantly developing new ways to query the data. Yet such queries are still far from being standard procedure in the scientific world. While data assets keep growing, they are often scattered across many different archives around the world. The AWI's scientific tools, such as the research vessel "Polarstern" or the atmospheric chemistry observatory at Neumayer research station in Antarctica, supply millions of datasets during each expedition. And these datasets constitute only a small portion of the overall scope of international earth system research efforts.

It is exactly this wealth of data that has become an obstacle rather than an asset for scientific learning. Dr. Hannes Grobe, an AWI scientist, describes the problem, "Data is gathered with enormous effort and cost but frequently remains inaccessible to most of the scientific community for a long time." While some of the information may appear in a tabulated format in appendices of scientific publications, only a small portion is stored in systems that are generally accessible—usually in heterogeneous structures and proprietary formats unfit for systematic evaluation. "The storage structure chosen for specific data is usually derived from the original scientific interest that led to its capture," explains Grobe.

Several international standardization projects are underway, but they will take time to complete. "Because of the increasing flood of data since the 1980's, a lot of the data is never published, let alone printed. This means it is lost for scientific research because it is not centrally available," explains Michael Diepenbroek, Managing Director of the World Data Center for Marine Environmental Sciences (WDC-MARE). WDC-MARE likewise uses PANGAEA as a centralized data archive. Diepenbroek concludes, "In view of the rapidly growing number of scientific publications and data, information must be presented in a more focused manner. The most focused manner is a structured data model that enables free access to its contents."

A LONG-TERM ARCHIVE FOR EARTH-SYSTEM RESEARCH

The concept devised to improve the availability of information focuses on publishing data and making it available in a quotable form. It was developed by the WDC-MARE center which, like PANGAEA, is operated jointly by AWI, MARUM and RCOM. As a library open to the public, PANGAEA also serves as a publication system.

From the very beginning, the development team intended PANGAEA to serve as a publication system in a bibliographic sense. All data is saved together with its descriptive metadata (such as the identity of the person who captured the data, the type of data measured, etc.). The data model concepts were kept generic and open to allow additional parameters to be included at any time to adapt to new scientific developments. All data is geo-referenced in terms of time and space. Each measuring point is defined by three spatial coordinates (latitude, longitude and elevation/depth). The metadata also reveals which reading was taken when, by whom, how and in what medium (water, air, ice, sediment etc.).

The data model is strictly regulated so new variables can be easily integrated. For example, if a new plankton type needs to be included, it is not necessary to add a new column to the table (changing the table structure); rather, a new line is added, defining how the new type is to be interpreted. An internal descriptor establishes a reference to the large table of measured values containing the actual reading. This means that the table will always retain an identical structure. This is an important prerequisite for flexible queries in a scientific data warehouse that integrates many different types of measured values.

Besides datasets and tables, PANGAEA also stores binary large objects (BLOBs) such as images or models. These are also geo-referenced and available for database searches. Each object is referenced by a URL. "Thanks to this concept, PANGAEA can capture nearly any type of data encountered in fundamental scientific research, as long as it can be assigned a time and geography," Grobe emphasizes.

SYBASE IQ—THE BASIS FOR THE DATA STORE

Sybase IQ stores the validated, published (therefore unalterable) data. The ability to provide a multitude of attributes and measured values from a single, large table is due to Sybase IQ's high compression ratios. This allows raw data to be compressed to between one-sixth and one-ninth of its original volume.

The PANGAEA system also leverages Sybase IQ's inherent indexing architecture. A traditional relational database requires multiple indices to be created before qualified searches can be performed. This inflates the data volume; limiting the indices, and compromising query flexibility. In contrast, Sybase IQ organizes data in columns rather than rows, allowing any field to serve as an index. This enables very flexible searches: whatever search criteria the user selects, the database always offers features that support access. Since queries do not have to read the entire row but only the selected value in the respective column, queries are processed very rapidly even in very large datasets.

USING THE LIBRARY CONCEPT TO PUBLISH DATA

To encourage scientists to reference databases in their publications, the data must be quotable. PANGAEA supports quotable data by emulating the established workflow from author to publisher to library. To ensure reliable data availability, a code for labeling scientific publications (Digital Object Identifier – DOI) was adopted for the data world. Issued by the TIB library in Hannover, the world's largest scientific and technical library, these reference numbers have been assigned to all datasets published since 2004. This allows each dataset to be located reliably so it can be quoted in compliance with bibliographic standards. In cooperation with other partners, PANGAEA has helped establish a global DOI system for data publishing.

Authors of data are now referenced properly when their data is quoted in scientific publications. In turn, the user of the quote can be certain that the data in question has been validated and published by an author and is therefore quality-assured. Using the DOI, the respective dataset can be retrieved through the Internet directly at any time. Since the DOI is a permanent link, the physical address of the data will be found even if the location has changed in the meantime.

EFFICIENT TIME SERIES ANALYSIS

Sybase IQ also handles data administration within PANGAEA. Interested parties can access the server using a client or Web portal (and the AWI itself is currently designing several portals for specific topics). All data is available to the scientific community through a general-purpose search engine using a Google-like syntax under the URL http://www.pangaea.de.

In addition, Bremen University's Michael Diepenbroek and Uwe Schindler are developing a universal access tool that will allow user-defined subsets to be extracted from the pool of data. Sybase IQ supports this through an efficient feature for indexing time information. Diepenbroek explains the feature, "Users who need more than 40,000 measured values, whether individually or in any combination, will be able to use the filtering features to retrieve the required data in any space-and-time-related quantity—with response times of just a few seconds," So, for example, the tool will allow scientists to build and view custom subsets of time series.

Christian Schäfer-Neth, a scientist at the AWI, demonstrates another example: analyzing data from Neumayer station. In the exceptionally clean air of Antarctica, the atmospheric chemistry observatory can track long-term developments without being compromised by local, human influences. Measurements of ozone, particles, greenhouse gases, trace gases, aerosols, turbidity and other parameters are taken at regular intervals (minutes and hours). Present-time atmospheric measurements are complemented by ice core drilling samples to trace back the evolution of the earth's atmosphere during the last one hundred thousand years. All of the measurements gathered are archived in PANGAEA as a basis for a long-term series analysis comparing historical and modern-day values.

For example, to determine ozone and particle concentrations during the 1990s, Schäfer-Neth enters these two parameters, along with the time frame. Within seconds, the system generates a diagram of the requested mean values, based on roughly 11 million atmospheric chemistry datasets.

A CENTRALIZED ARCHIVE FOR SCIENTIFIC DATA

Time series, oceanographic and seismic profiles, sediment profiles, geological maps – are examples of the data from PANGAEA, it offers nearly unlimited possibilities for evaluation. The volume and diversity of the data assets are increasing steadily. Initially a repository of AWI data, PANGAEA soon evolved into a central archive for a wide range of scientific data. Based on a model of universal query options, the sheer bulk of the data stored to date, and the appeal of the DOI concept are points of attraction for external archives, some of which are actually receiving public funding to help them archive their data in PANGAEA. One of the world's largest tree-ring archives has been integrated, an important contribution to paleoclimate research, the study of the world's climate in prehistoric times. Numerous national, European or international projects have entrusted PANGAEA with their data management tasks.

The data entry procedures and technical routines support data imports. To guarantee data consistency, a curator is appointed for each project, assuming responsibility for the technical quality of the project data. Before a dataset can be entered into the database, it must first be approved by the curator and tagged with a DOI number. In the future, a peer review will be added to this process to help ensure the scientific quality of data. Concentrating vast amounts of information in one data warehouse has also optimized quality assurance. Diepenbroek says, "If we enter a dataset of, say, 5,000 values that contains one anomaly, the cause of that may be either a faulty measurement or an extreme event. We can only determine that if we have a sufficiently-sized data pool available that is capable of reflecting comparable phenomena."

Over the last few years, the contents of PANGAEA have increased exponentially. Since 1996, the data volume has doubled each year. By November, 2006, 136 research projects with a total of 470,000 datasets and 1.7 billion data points had been entered, amounting to a volume of 1.2 terabytes. And the amount of data keeps growing; several observatories are feeding in new values every minute. The stream of data will be augmented by one or two orders of magnitude once the sensor networks currently under construction will begin delivering additional oceanographic, meteorological and geophysical data.

"By storing such a large amount of data together with the metadata in one centralized system, we have increased our efficiency enormously," says Grobe. "There is no other scientific data model capable of handling such volumes and this high level of complexity. Since many diverse aspects can be combined in a single query, the system opens entirely new avenues of scientific examination. This allows us contribute significantly to the further exploration of our earth."

SYBASE, INC. WORLDWIDE HEADQUARTERS, ONE SYBASE DRIVE, DUBLIN, CA 94568 USA 1 800 8 SYBASE Copyright © 2007 Sybase, Inc. All rights reserved. Unpublished rights reserved under U.S. copyright laws. Sybase and the Sybase logo are trademarks of Sybase, Inc. or its subsidiaries. All other trademarks are the property of their respective owners. ® indicates registration in the United States. Specifications are subject to change without notice. 06-07

www.sybase.com

SYBASE®