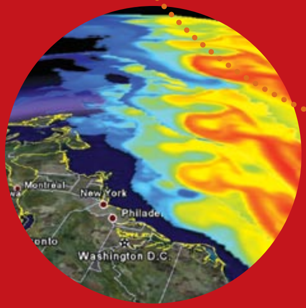
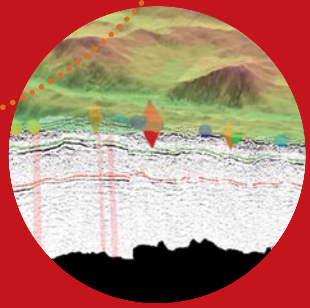
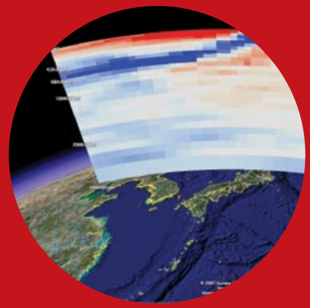
Building a Global Data Network for Studies of Earth Processes at the World's Plate Boundaries




Report of the International Data Exchange Workshop

Preferred Citation

Carbotte, S., K. Lehnert, S. Tsuboi, W. Weinrebe, and Workshop Participants. 2007. Building a Global Network for Studies of Earth Processes: Report of the International Data Exchange Workshop. May 9-11, 2007, Kiel, Germany, 44 pp. Online available at: http://www.nsf-margins.org/Datawkshp07.

Workshop Participants

An international group of marine scientists, data managers, and information technologists participated in this workshop. See Appendix 1 for a full list of the participants.

Building a Global Data Network for Studies of Earth Processes at the World's Plate Boundaries

Report of the International Data Exchange Workshop

Held May 9-11, 2007

in Kiel, Germany

Workshop Convenors

Suzanne Carbotte, LDEO Columbia University

Kerstin Lehnert, LDEO Columbia University

Seiji Tsuboi, JAMSTEC

Wilhelm Weinrebe, IFM-GEOMAR

Contents

Executive Summary

An international group of marine scientists, data managers, and information technologists convened a 2.5-day meeting in Kiel, Germany, to explore opportunities for international data exchange and to address the cultural and political challenges for building a freely accessible public data network for the global community that facilitates mid-ocean ridge and continental-margin-related research. Workshop participants discussed technical, procedural, and organizational issues of open global data sharing, and agreed on the following statements of principle and set of recommendations grouped broadly under the working group themes:

Science User Needs

- Open public access to data is fundamental to verifiable scientific progress. All data that are necessary to reproduce published scientific results, including field data, processed data, and laboratory (derived) data products, need to be published and stored in accepted archives. We need to advance a culture among scientists that is more open to public and transparent data sharing. (T1-R1[1]; T2-R5; T4-R4)

- Scientists studying earth processes require access to multidisciplinary data and data integrated from both the marine and terrestrial world. (T1-R2; T1-R3)

Data Documentation and Publication

- Uniform best practices and standards need to be developed, promoted, and used routinely within the international community for data acquisition, data submission to data centers, and data publication. Best practices should include formal submission agreements between individual institutions and respective national and international data centers and the use of globally unique identifiers for data and samples. Scientific societies should take an active role in formulating best practice guidelines for data publication. In addition, new mechanisms are needed to track the use of data sets to ensure academic recognition and to support scientific collaborations. (T1-R4; T2-R2; T2-R4; T2-R6; T2-R7; T2-R8; T4-R1)

- The ultimate responsibility for ensuring adequate documentation of a field program lies with scientists and it must be part of their obligation to funding agencies. Detailed, high-quality metadata creation and data submission should be made as easy as possible for ship operators and scientists, with development of new automated tools that support and further the implementation of best practices and standards. Funding agencies must be involved in enforcing standard practices for data documentation and submission to data centers. (T2-R1; T2-R3; T2-R4; T2-R6)

[1] Theme 1—Recommendation 1 (T1-R1), Theme 2—Recommendation 2 (T2-R2), etc.

Data and Metadata Interoperability

- The community must minimize the proliferation of metadata standards and work toward a uniform approach for scientific metadata. Processes need to be defined regarding how to develop high-quality, community-based standards, guidance, and profiles. New efforts to develop standards and protocols to support interoperability without loss of content and information should build upon and take advantage of existing community-based projects. (T3-R1; T3-R2; T3-R3; T3-R4)

- Development of a data discovery service across distributed marine geoscience data resources within the international community is an achievable initial goal. Data centers should work to expose their data resources via Web services using, for example, OGC or OAI protocols. (T3-R5; T3-R6)

2

Opportunities and Obstacles for International Data Sharing

- International programs and bodies, such as the Global Earth Observing System of Systems (GEOSS), the Electronic Geophysical Year (eGY), and the International Council for Science (ICSU), as well as ongoing International Polar Year (IPY) projects, that stimulate the development of global data sharing systems should be leveraged to promote an initiative for a global data network for marine and terrestrial geoscience data. (T4-R5)

- A dedicated task group should be established to advance implementation of a global data network. In addition, special interest groups that would share experience and solutions on issues concerning metadata and interfaces should be formed with tools to facilitate collaboration and science-based adaptive management. (T4-R6; T3-R7)

Based on these recommendations, the following next steps are identified: (1) develop test-bed sites for a data discovery service across globally distributed data resources; (2) establish forums for guidance and development of best practices in the areas of data acquisition, metadata, vocabularies, and interfaces; (3) formulate a dedicated task group to advance international alliances; and (4) establish opportunities for annual meetings of the international marine geoscience data management community.

Motivation for the Workshop

3

Rapid advances in database technology for scientific research, which have occurred over the past decade, are providing new access to data and new tools for data visualization and integration. Along with these advances in information technology has come the growth of digital collections of a broad suite of data across the sciences. Developments in database connectivity provide new opportunities for open data exchange across distributed data collections, greatly expanding the volume and diversity of data available to the scientist to address a particular scientific problem of interest. These advances hold great promise for the solid earth sciences, an inherently multinational and multidisciplinary field, which involves the collection of unique data sets during oceanic and terrestrial expeditions conducted by research institutions around the globe.

The international marine geoscience community is actively engaged in scientifically aligned goals through the InterRidge and InterMARGINS programs. These broad multidisciplinary initiatives focus on understanding fundamental processes of crustal formation, modification, and destruction at Earth's plate boundaries. InterRidge and InterMARGINS aim to coordinate efforts and priorities in mid-ocean ridge and continental margin research, respectively, across nations. Ridge2000 and MARGINS are US-funded programs that conduct focused investigations in a few geographic locations, most of which involve international partners. At present, there are no formal agreements for data sharing and data documentation within these international communities. Data exchange occurs primarily by informal agreements between scientists directly involved in specific projects. However, international marine-terrestrial geoscience research efforts would greatly benefit if data collections maintained as national efforts could be better linked and if broader access were initiated. New database technologies are available that enable independent, globally distributed sites to share, link, and integrate their data holdings and services while maintaining full ownership and credit for these holdings.

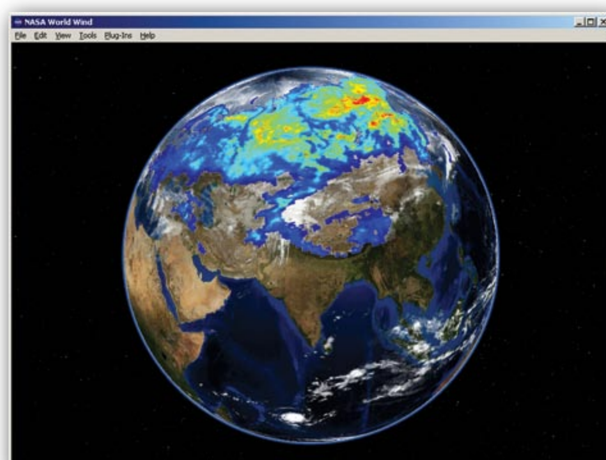


Figure 1. Northern Hemisphere distribution of snow water equivalent. Data from the National Snow and Ice Data Center (NSIDC), served through the Environmental Systems Science Centre (ESSC) WMS and displayed in NASA World Wind. Figure from Jon Blower, ESSC, University of Reading, UK.

Figure 2. The GeoMapApp map-based browser (www.geomapapp.org) provides live links to Alvin Frame Grabber and Jason Virtual Van Images that are publicly available through the National Deep Submergence Facility at WHOI. Black circles along dive tracks indicate locations of available online images that are selected to directly access the Virtual Van and/or Frame Grabber. This example is from the Mariner Vent Field, Lau Basin, Ridge2000 Integrated Study Site, and utilizes a high-resolution (10-cm grid) bathymetric map (from cruise TUIM05MV) as a backdrop. Figure provided by Vicki Ferrini, Marine Geoscience Data System.

To explore current opportunities and challenges for international data exchange in support of continental margin and mid-ocean ridge research, a workshop entitled "Building a Global Data Network for Studies of Earth Processes at the World's Plate Boundaries" was convened with two primary goals:

Goal 1. To explore current relevant data management efforts within partner countries.

Goal 2. To devise a strategy for building a global data network to support the sharing and exchange of data of greatest scientific interest for continental margin and mid-ocean ridge studies.

The primary desired outcome of this meeting was the development of new partnerships between marine geoscientists and data centers within the international community to establish enhanced access and exchange of data sets of broad interest for studies of Earth processes at the global plate boundaries.

Workshop Structure

Four scientists from Germany, Japan, and the United States convened the workshop, which was jointly supported and funded by InterMARGINS, MARGINS, InterRidge, and Ridge2000. The US National Science Foundation and the Cluster of Excellence "The Future Ocean" at the Christian-Albrechts-University in Kiel provided additional financial support. Seventy-one people from 14 countries attended the workshop, including scientists from the InterRidge and InterMARGINS communities, data managers representing data centers and data systems across a spectrum of primarily marine geoscience data, and information technologists involved in various aspects of interoperability development. Appendix 1 lists workshop participants. Prior to the meeting, participants were asked to provide a brief one-page summary describing their data system along with relevant URLs. Appendix 2 includes brief summaries of each data system or resource along with relevant URLs. Full summaries are available at the meeting Web site (http://www.nsf-margins.org/Datawkshp07/).

The workshop was held at the meeting facilities of the Hotel Birke in Kiel, Germany. The official program started on May 9 in the morning and lasted for 2.5 days. Interested participants were invited to continue discussions on May 11 in the afternoon. The workshop ended with an informal field trip to the historical town of Lübeck on May 12. The full agenda is included in Appendix 3.

The first 1.5 days of the workshop were devoted to presentations within three general areas:

a. Science Needs: Scientists outlined their needs for data access and defined data sets of broad interest for continental margin and ridge-related science.
b. Data Resources: Representatives of data centers presented existing data systems available for academic research. These presentations were complemented by poster presentations and live demonstrations of the systems.
c. Technologies: Information technologists reported about emerging technologies for interoperability and data sharing.

The afternoon of Day 2 and morning of Day 3 were devoted to working group sessions to discuss technological as well as organizational and cultural issues of global data exchange. The working group discussions were structured into four themes, each of which (except for the Science User Needs group) had two sessions:

1. Science User Needs and Concerns
2. Data Documentation and Publication
   a. Standards for Data Documentation
   b. Data Publication
3. Data and Metadata Interoperability
   a. Standards and Technologies for Metadata and Interfaces
   b. The "Low-Hanging Fruit" for Data Exchange
4. Opportunities and Obstacles for International Data Sharing
   a. Archives and Data Contributions
   b. Implementing an International Data Network

Each working group addressed a range of questions provided to the session leaders by the workshop conveners, and was charged to generate a set of recommendations that working group leaders presented in plenary sessions. Questions and recommendations are outlined in the following section.

Working Group Discussions

Theme 1: Science User Needs and Concerns

Scientists engaged in plate boundary research study the wide variety of active processes associated with the formation, modification, and destruction of Earth's crustal layer, which supports life on the planet. Plate boundaries transect the oceans, hug the continental margins, and penetrate into continental interiors. They are the locus of most earthquake and volcanic activity on Earth and of the pervasive fluid-chemical-thermal interactions associated with the development of unique ecosystems and the formation of economical metal deposits. Increasingly, these active environments are studied as integrated complex physical, chemical, and biological systems, subject to a variety of influences, rather than as primarily geological structures. To address these interdisciplinary goals, scientists increasingly require access to multidisciplinary data sets and from terrestrial and marine settings. These requirements make scientific data access and exchange challenging.

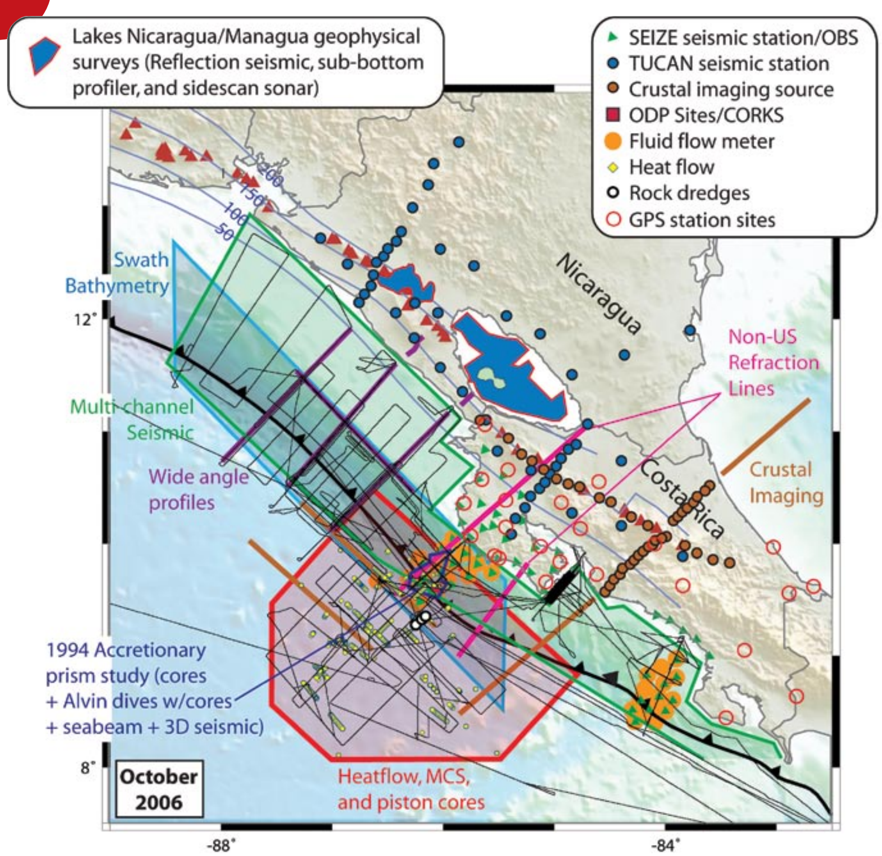


Figure 3. MARGINS and related data collected at the Central America SEIZE and SubFac site. Figure compiled by Paul Wyer, MARGINS Office, 2004, 2006.

The science user working group considered the following questions:

- What are science user needs and concerns with regard to data sharing?
- What are the key data sets needed for international exchange?
- What links exist and are desired between the marine and terrestrial world?
- What capabilities are desired that are currently lacking? What technologies are promising to scientists?

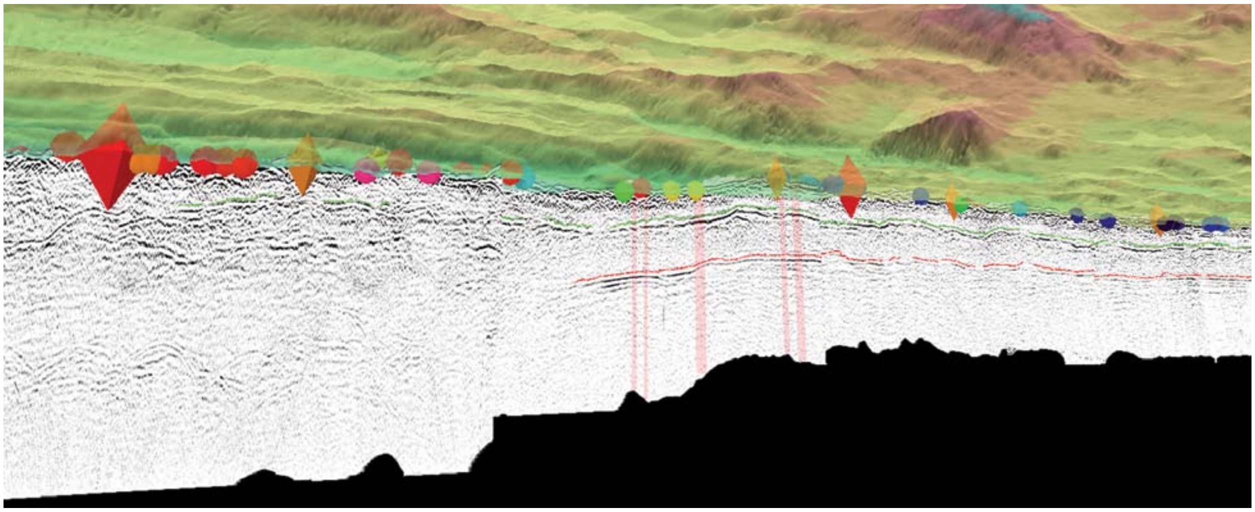


Figure 4. Screen shot of a 3-D visualization of the Eastern Lau Spreading Center highlights seismic, geochemical, and hydrothermal vent field data from around the Lau Integrated Study Site of the Ridge2000 program. Along-axis multichannel seismic data show a relatively continuous reflection from the interface between the pillow basalts and sheeted dikes (green line), while the axial magma chamber reflection does not appear until later (red lines). Red and orange diamonds denote known and potential hydrothermal vent fields (data from Baker et al., 2006). Spheres mark geochemical sample sites and are colored by MgO concentration (red being high, black being low; data from Bezos et al., 2005). Colors in upper portion of the image are bathymetric data. Three-dimensional scene available through RidgeView (http://ridgeview.ucsd.edu/). Figure provided by A. Jacobs, UCSD (Jacobs et al., in preparation, 2007).

The science user community strongly endorses the principle of fully open data access. They want access to all existing data relevant to their research problem. For programs conducted in the open ocean, scientists desire access to all data collected in a geographic area of study. Closer to shore, along the continental margins, there may be economic or national-security concerns that affect access to some kinds of data, but much data of value to basic science should be available. Easy access to a diverse suite of data is necessary for many studies; however, many data resources currently available represent disciplinary databases. More focus is needed on building data systems to support integrative science, providing access to multidisciplinary data. Although the fundamental science questions associated with continental margin studies transect the shoreline, the shoreline represents a major boundary in how data are collected, organized, and later archived. This disparity is a significant obstacle to scientific data access.

Recommendations

Workshop participants discussed technical, procedural, and organizational issues of open data sharing and agreed on the following statements of principle and general recommendations, grouped broadly under the working group themes.

T1-R1: Open public data access is fundamental to verifiable scientific progress. Full open public access to data is needed to support scientific progress and to enable the verification of research results. In general, geoscience relies on field observations. Thus, it differs from most experimental sciences in that measurements are difficult to repeat. With the typically unique data sets used to support plate-boundary studies, research results are often impossible to verify without open access to field observations and measurements.

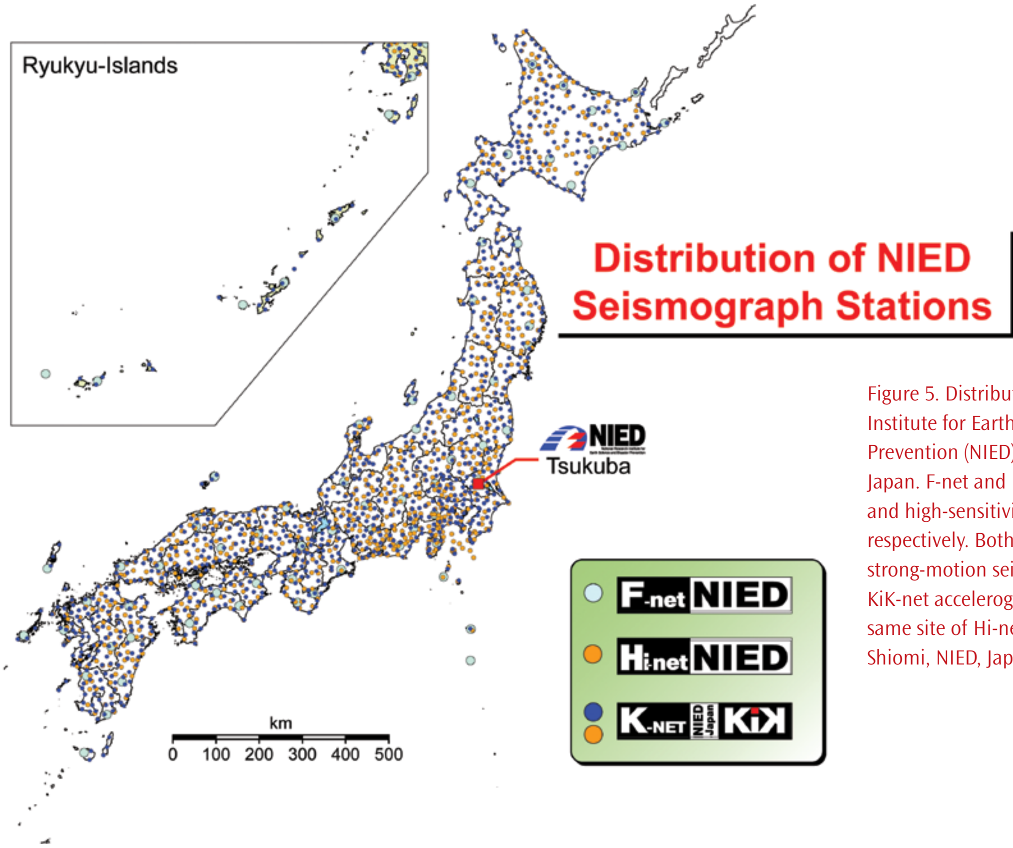


Figure 5. Distribution of National Research Institute for Earth Science and Disaster Prevention (NIED) seismograph stations in Japan. F-net and Hi-net are mean broadband and high-sensitivity seismograph networks, respectively. Both K-NET and KiK-net are strong-motion seismograph networks, although KiK-net accelerographs are installed at the same site of Hi-net. Figure from Katsuhiko Shiomi, NIED, Japan.

Scientists want unrestricted access to as much data as feasible within the framework of national requirements and proprietary periods of data collectors. National needs may require limitations for some data types and in some environments (e.g., ultra-high-resolution bathymetry in shallow coastal waters, on-land gravity, reflection seismics in petroleum-rich basins), but every reasonable effort should be made to release such data in a reasonable time frame. For research data subject to proprietary hold periods, scientists would like access to metadata describing the existence and location of the data at an early stage, with mechanisms that support interactions between data collectors and other scientists wishing to form collaborations.

T1-R2. Scientists require full and free access to multidisciplinary data. The integrative science programs that characterize modern studies at mid-ocean ridges and continental margins drive the need for integrated access to multidisciplinary data. More and more, scientists seek to work across traditional disciplinary boundaries either through developing collaborations or by acquiring interdisciplinary expertise. Data systems that support and facilitate collaborations and multidisciplinary access are required. Scientists need access to multidisciplinary databases of geographically referenced data and to physical property measurements, such as experimentally derived material properties. Derived data sets, including images and data-based models, have tremendous value for interdisciplinary studies, and these need to be preserved.

T1-R3. Complete and seamless integration of data resources from both the marine and terrestrial world is needed. Research along continental margins requires access to terrestrial and marine data. However, available data resources typically stop at the shoreline, with different agencies and organizations involved in terrestrial and marine studies. This ecologically and geologically artificial delineation has created major problems in the past and needs to be overcome. Significant obstacles to obtaining access to data across the shoreline relates to differences in how data are collected and organized. Whereas offshore work is usually defined and organized by cruise, onshore field studies are characterized in a variety of ways—by networks of instruments, by investigating group, by national or other geographic boundaries, or otherwise. Also, onshore and nearshore data sets tend to be spread through a wide array of national agencies with varying standards and missions. Data systems are needed that support the ability to search for and find related data objects in a variety of different frameworks that make sense for the problem at hand and which are not dependent on the platform or group collecting the data. While geographic data access makes sense for many problems, time-series data inherently require searches at a wide variety of time scales. The great variety of data set characteristics demonstrates the value of having several primary search categories.

T1-R4. Mechanisms are needed to track the use and publication of data sets to ensure academic recognition and to support scientific collaborations. While the existence of open data collections representing the accumulation of data from many individual studies provides important resources for scientists, an ongoing concern is how to ensure that credit to original data collectors is preserved. Within the current framework of citation supported by scientific journals, it is often not possible to cite the large number of original data sources used for a new analysis or other value-added product or syntheses (see also Theme 2: Data Documentation and Publication).

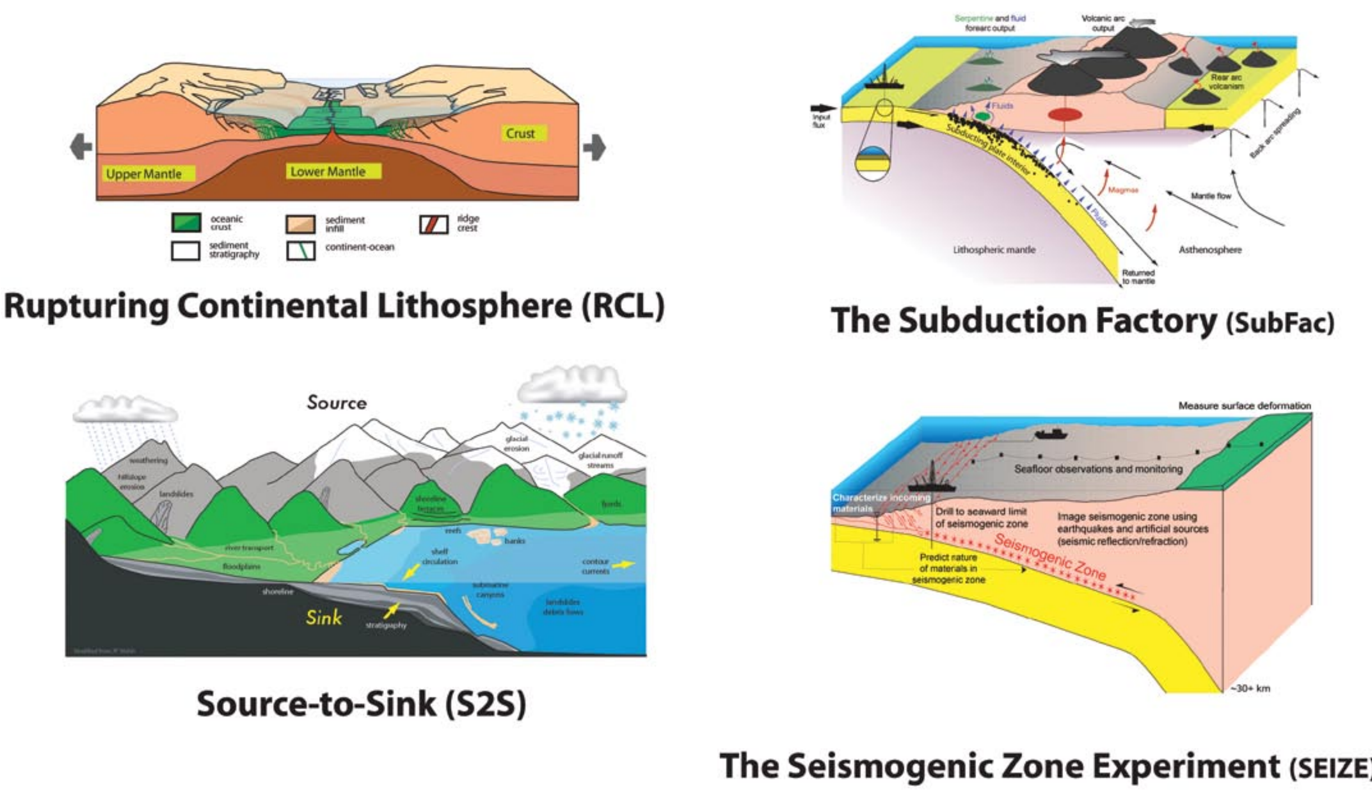


Figure 6. Schematic illustrations of the four science initiatives of the US MARGINS program. Figure from Paul Wyer, MARGINS.

Theme 2: Data Documentation and Publication

The development of digital data resources for scientific data, along with new technologies for data visualization and analysis, is changing the way marine geoscience research is conducted. These technological developments affect society, cultures, and businesses globally. An increasing number of scientists are making use of digital data collections as primary resources for studying an area of interest, to conduct global syntheses, and to facilitate new multidisciplinary studies. The utility of digital data resources fundamentally depends on the comprehensiveness and the quality of the data they provide. Therefore, data must be: (a) openly and fully accessible and (b) documented properly at all stages of the data life cycle, from initial acquisition, through processing, to primary and later secondary publication, to ensure evaluation of data quality. These requirements deeply impact the scientific data culture, imposing new obligations on scientists for comprehensive and transparent data description and analysis, and changing the way data is referenced and cited. This theme focused on issues of data documentation and publication.

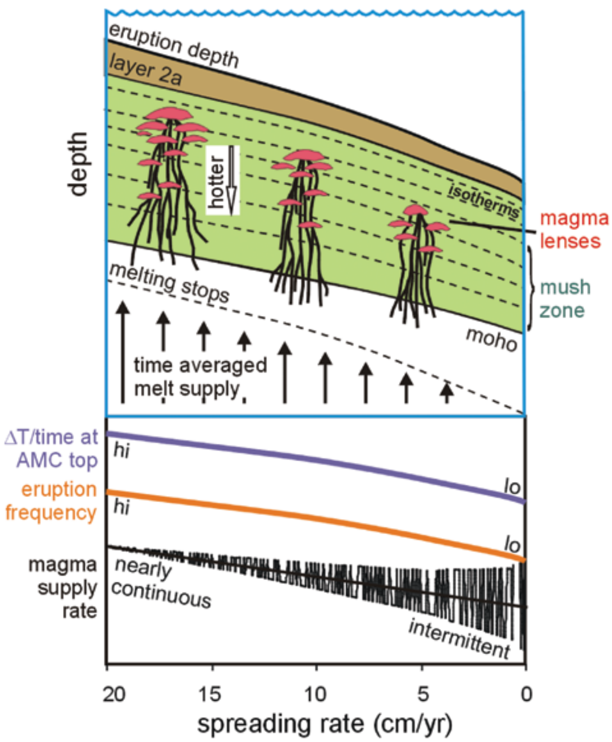


Figure 7. Using analytical data for > 11,000 samples from the PetDB database, Rubin and Sinton observe systematic, regional compositional variations in a global MORB data set. Based on this observation, they propose a new ocean ridge magma chamber model in which the number, size, and depth of shallowest melt segregations vary smoothly with spreading rate and magma supply. This new perspective on crustal magma bodies provides a framework for understanding structural, geophysical, hydrothermal, and volcanological attributes of ridges. According to the authors, the major element, trace element, and radiogenic isotope data within PetDB made this research possible. Figure from Rubin, K.H., and J.M. Sinton. 2007. Inferences on mid-ocean ridge thermal and magmatic structure from MORB compositions. *Earth and Planetary Science Letters* 260:257–276.

Session I. Data Documentation

The breakout group on Standards for Data Documentation addressed the following topics:

- Review current practices for different subdomains.
- How can we achieve standardized data documentation during acquisition in the field/at sea? For derived data?
- How do we ensure the highest level of data quality? What metadata requirements are necessary?
- What roles can and should agencies, ship operating institutions, and publishers play?

Working group discussions focused primarily on field data acquired during marine surveys. Current practices for data acquisition and documentation at sea are highly heterogeneous across the global marine geoscience community. In many cases, data documentation is the exclusive domain of the scientific party, but there is little support for ensuring that full documentation is achieved. While scientists must ensure adequate documentation of their data for their own use, this documentation is typically recorded in difficult-to-access workbooks or spreadsheets designed by scientists and is seldom captured for later incorporation into data systems. In addition, the documentation that a scientist may provide for their own data-reduction purposes is often insufficient to facilitate later use of the data by others. The Intergovernmental Oceanographic Commission (IOC) Cruise Summary Report (CSR) forms, formerly known as ROSCOP forms, widely used to report cruises within the European community, minimally documents cruise operations. Furthermore, on many modern expeditions, data other than those of primary interest to the scientific party may be routinely collected, but remain largely undocumented. The challenge is to more thoroughly and completely document data for all marine programs carried out within the international research community.

The working group's consensus is that while the collection of cruise metadata is often incomplete and that this is a global issue, improving data documentation at sea can be readily addressed with broader adoption of standardized forms and procedures. The needed information is collected in some form during a field program. The challenge is to find relatively easy ways to get this information out of the notebook or personalized electronic file of the scientist or technician, and into a standardized format, and to formalize the transfer of this record-keeping to the relevant database system.

Procedures for capturing this information need to be of obvious benefit to the scientists themselves and must minimally impact their existing responsibilities. The current bureaucratic overhead of research for scientists is high and it is important to design documentation procedures that add minimum extra burden to their responsibilities.

Figure 8. Launching the Japanese submersible KAIKO 7000 operated by JAMSTEC. Photograph provided by JAMSTEC.

To facilitate more complete documentation of data acquisition at sea, standardized metadata forms and acquisition procedures have been developed within some communities. For example, metadata forms have been developed by the Marine Geoscience Data System (MGDS) for the US MARGINS and Ridge2000 programs (http://www.marine-geo.org/metadata_forms.html) to ensure adequate documentation of data collected during these programs. The French Research Institute for Exploitation of the Sea (Ifremer) has established a data-quality plan that outlines procedures for standard data acquisition aboard their ships. The System for Earth SAmple Registration (SESAR) provides unique identifiers (the International Geo Sample Number, or IGSN) for samples to ensure that all sample analyses can be ultimately tied to a unique sample. The existing standardized MGDS forms were examined during breakout group discussions as possible working models for basic data documentation at sea. Working group participants agreed that the information requested is generic and a basic minimum for scientists to provide.

Marine expeditions involve a wide array of data-collection activities in addition to the standard underway geophysical data streams, such as multibeam, gravity, and magnetics, and all of them must be documented (e.g., cores and dredges, biology samples from dives, ocean-bottom seismometer deployments). Ideally, standard digital forms should be used and, if lacking, designed so that they can replace scientists' personal records.

Recommendations

T2-R1: The ultimate responsibility for ensuring adequate documentation of a field program lies with scientists and must be part of their obligation to funding agencies. Standard practice should include the identification of a "data liaison" from within the science party, who works with the ship's support staff to ensure capture of all needed information. On many ships and for many data types, the shipboard science support staff will produce the needed data documentation as part of their routine operations. But, the shipboard support staff is unlikely to have access to all information on the full suite of data acquired during a program. Scientists bring sensors on board, and they are typically in charge of station operations associated with sampling or instrument deployment. As the primary interest and responsibility for the scientific data acquired during an expedition reside with the scientific party, the ultimate responsibility for ensuring comprehensive documentation for all data should also lie with the scientists. For some ships, (e.g., UK Natural Environment Research Council [NERC] cruises and National Science Foundation Office of Polar Programs [NSF-OPP]-funded ships within the United States) a data/metadata specialist who is responsible for generating complete documentation of survey operations often sails on each cruise, particularly cruises with participants from multiple laboratories.

Figure 9. Left. Water sampling on Lake Baikal. Photo provided by Jens Klump. Bottom. Photograph of the Japanese riser drilling platform *Chikyu,* which is capable of drilling to 7000 m beneath the seafloor. Photo provided by JAMSTEC.

T2-R2: Routine use of standardized data documentation procedures should be adopted by ship operators and scientists. Comprehensive and standardized data documentation at sea is a tractable goal. The standardized electronic metadata forms provided by the MGDS, the data-quality plan of Ifremer, and assignment of IGSNs to samples are steps in the right direction and provide models for wider adoption. While ships are operated by different agencies in different countries, each with its own procedures and requirements for survey operations, the concept of standard metadata forms should be generally applicable. Metadata forms need to be developed in close collaboration with users. Easy mechanisms for users should be provided to customize forms for specialized use. Data documentation procedures need to be designed to fulfill requirements of existing metadata standards (e.g., Federal Geographic Data Committee [FGDC] and International Organization of Standards [ISO]). Adequate documentation is needed of the field program (e.g., participants, roles and affiliations, projects conducted), of all digital data acquisition and sampling events (e.g., data type, device used, position and temporal information, OSDS: Data Inventory Map

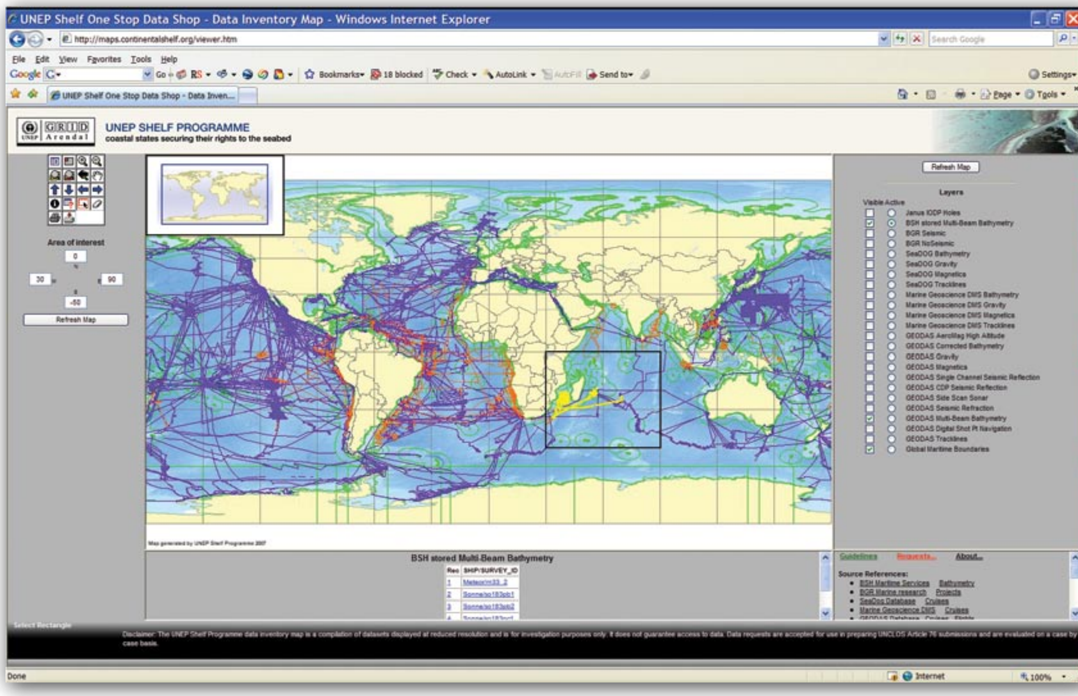


Figure 10. United Nations Environmental Programme (UNEP) Shelf Programme online metadata portal. The One Stop Data Shop's Web portal allows users to examine the public marine geoscientific research data available in their areas of interest. Survey lines and associated metadata are displayed to allow users to evaluate data availability. Access to actual data, including seismic and bathymetry, is restricted to Article 76 requests and is facilitated in conjunction with the collaborating institutes. Figure from UNEP Shelf Programme/UNEP-GRID-Arendal.

unique identifiers for samples), and of all instrumentation used (e.g., manufacturer, make, model, serial number, sensor history, and calibration information).

T2-R3. Automated tools for metadata creation at sea are needed. Metadata creation suitable to support long-term data preservation is time consuming for scientists to produce and they lack sufficient incentive. Assessment and ingestion of metadata are also time-consuming procedures and are error prone. New automated methods to tag data with required metadata at the time of data acquisition are needed[2]. The long-term vision to support marine geoscience data acquisition is a Web-based shipboard event-logging system that pulls in the required information, such as navigation, person, sampling event or operation, and sample or data type confirmed by the science party. The shipboard event-logging system should include pull-down menus of controlled vocabularies to describe operations. A comprehensive shipboard data acquisition system is in use for Integrated Ocean Drilling Program (IODP) expeditions and is a model for wider application.

T2-R4. Funding agencies must be involved in enforcing standard practices for data documentation and submission to data centers. Requirements for the standard documentation and submission of data acquired during all field programs will need to be enforceable through funding agency actions.

[2] See for example Morpho, a framework for storing and serving ecological data and metadata using the Ecological Metadata Language (EML).

Session II: Data Publication

Discussions in the Data Publication Working Group were concerned with issues relating to policies and procedures for data publication:

- What data need to be accessible (raw vs. derived, published vs. unpublished)?
- How should data be identified (use and granularity of unique identifiers for data)?
- How can new requirements for data publication be implemented? What are the special disciplinary issues?

Issues concerning data publication are a key concern to both individual scientists and to data system providers. Scientists publish the data they acquire through analytical, experimental, or computational procedures as a major product of their research, "marketing" them to gain credit and reputation that ultimately form the currency of their careers (Edwards et al., 2007)[3]. In many scientific cultures, data have traditionally been treated as private intellectual property and have typically been shielded carefully, often even after publication. Journal articles frequently contain only fragments of a "published" data set (tables with "representative analyses"). Publication of raw data has been a rare exception and data documentation in general is poor and quite heterogeneous. Edwards et al. (2007) state that the "private-ownership practice has led to a plethora of data collection practices and data formats, many of them idiosyncratic, as well as an absence of the metadata needed by other scientists to understand how the data was originally produced."

While many scientists now recognize the benefits of digital data collections and support their existence, they are rightfully concerned that access via digital data resources to data generated through their research will circumvent the original journal publication of the data and leave them without proper citation and credit for their contributions. Policies and procedures for data publication as well as the design of a global data network need to address these concerns. The appropriate use of globally unique identifiers for data that allow a data set to be identified and cited independent of a journal publication, but also allow data in digital collections to be linked to the original publication in the scientific literature can contribute to a satisfactory solution.[4] Mechanisms are also needed to link original data sets to higher-level data products or syntheses such as gridded bathymetric compilations or Geographic Information System (GIS) layers.

Scientific data come in many different types. The main differences relate to their origin (e.g., sensors, observation, experiment, modeling), their nature (digital data, physical specimens, numerical models, images, video, sound), and the level of processing (raw data, corrected, reduced, or "derived" value-added data). Data related to oceanic expeditions can range from geophysical, to geochemical, to biological. Data acquired shipboard range from raw to processed, for example, underway geophysical data streams (e.g., multibeam, gravity, magnetics); CTD casts; and rock, fluid, or biological samples. "Derived" data are mostly generated on shore in laboratories with a wide range of processing procedures applied to raw geophysical data or analyses

[3]Edwards, P.N., S.J. Jackson, G.C. Bowker, and C.P. Knobel. 2007. *Understanding Infrastructure: Dynamics, Tensions, and Design. Report of a Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures*. 50 pp. Online available at: http://www.si.umich.edu/InfrastructureWorkshop/documents/UnderstandingInfrastructure2007.pdf

[4]For example, the German project "Publication and Citation of Scientific Primary Data" (http://www.std-doi.de) has prototypically implemented a system for the publication of scientific data, which is open to the scientific community in any scientific field. This project uses persistent identifiers (DOI, handle.net, and URN) to identify data sets available in a digital format.

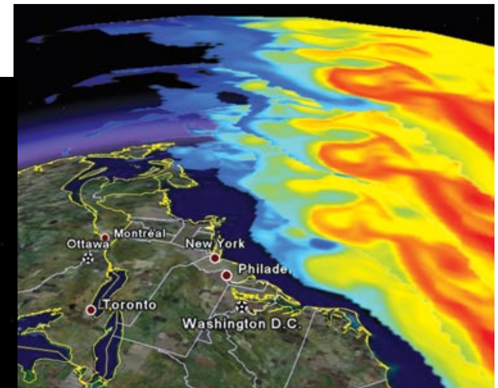


Figure 12. Three-dimensional temperature structure of Gulf Stream shown in Google™ Earth (imagery displayed apparently above sea level, owing to limitations of Google™ Earth). Figure from Jon Blower, ESSC, University of Reading, UK.

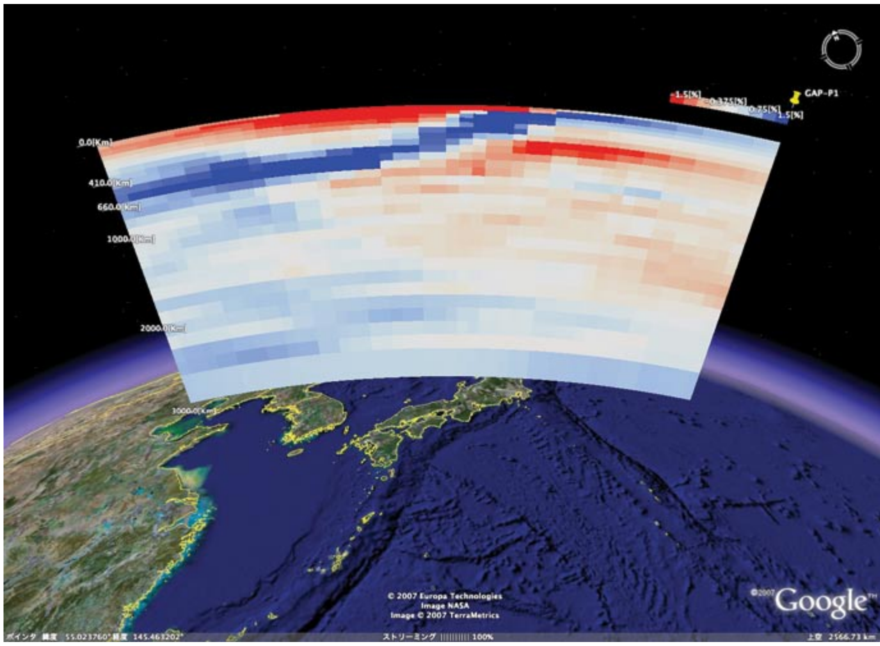


Figure 11. Vertical cross section of seismic tomography model, GAP-P1 (Obayashi et al., 2006, EPSL. 243, 149-158), beneath Japanese Islands displayed in Google™ Earth. The KML file used here is produced using conversion tool developed by Jamstec (see http://www.jamstec.go.jp/pacific21/TMGonGE/top.html). Figure prepared by Yasuko Yamagishi, JAMSTEC.

of samples collected during a cruise. Guidelines are necessary to define criteria for identifying data that should be preserved, data that should be published, and whether data should be "discarded" after use. An example of such guidelines are the "Rules of Good Scientific Practice" adopted by the Max Planck Society that take a general perspective on the data-preservation issue:

Scientific examinations, experiments and numerical calculations can only be reproduced or reconstructed if all the important steps are comprehensible. For this reason, full and adequate reports are necessary, and these reports must be kept for a minimum period of ten years, not least as a source of reference, should the published results be called into question by others.

A large part of the working group discussion was related to who should submit the data to the archive (database), revealing cultural differences among countries on how the ships are operated. Discussion also emphasized the principle that data submission requires standardized data documentation and input.

Recommendations

T2-R5: All data necessary to reproduce published scientific results need to be published and archived in an accepted data archive. Raw data from sensors and research activities should be archived along with all needed high-quality metadata to allow future processing and appropriate interpretation of the data. In addition, standard (routine) corrections should be applied to the "raw" data to make the data more easily usable by a larger community. These corrected data should be archived as well. Physical samples are considered "raw" data, for example, for geochemical measurements, and should be archived to ensure that analytical data are reproducible and can be complemented by new measurements. So far, repositories

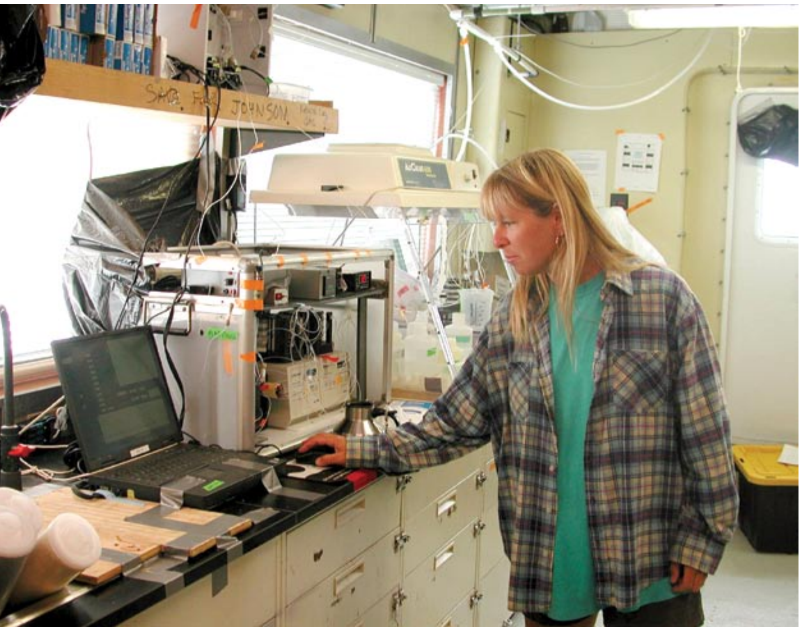


Figure 13. Ginger Elrod works with an iron measurement system on board MBARI's *Western Flyer.*

barely exist for samples from ocean-going expeditions, and are virtually absent for land-based expeditions. It is critical that samples carry globally unique identifiers to ensure unambiguous identification and allow tracking their analytical history.

During a cruise, some data types may be processed. Files with processed data should be submitted to the relevant databases, accompanied by adequate metadata about the processing method. For post-cruise processed data, the situation can be very different. While it is unclear how to proceed, there was consensus that principal investigators should notify collecting institution database groups when they submit processed data to relevant data banks.

T2-R6: Data submission should be streamlined and standardized. Procedures are needed to seamlessly integrate data into databases, and make the process of data submission as easy as possible for scientists, while ensuring comprehensive and consistent data documentation. Data submission requires standard data input, like cruise name, dates, location, and participants, which is already available in some form to the ship operator. This standard data should be easily available so that researchers submitting their data do not have to re-enter this information.

Data types such as geochemical measurements need a standard set of parameters (sample and analytical metadata) at the time of publication to accompany the sample information before a paper is accepted. Editors need to link acceptance of a manuscript to the submission of the data and accompanying metadata to a public "accepted" archive. Whenever possible, published derived data should be in a reusable format (e.g., electronic data table).

T2-R7: Unique identifiers for data should be used at the level of a study or publication. The working group reached consensus that unique identifiers for data should be applied at the level of a "study" or "publication," and not at finer granularity, such as a single analysis. This recommendation pertains to raw data as well as peer-reviewed published data, which is often derived data. Modern publications already have unique identifiers (DOI). Older publications might not, and incorporation of those data in databases might require "new" unique identifiers.

T2-R8: Scientific societies should take an active role in formulating best practice guidelines for the publication of data. There is general recognition that the existence of databases has improved the quality and documentation of the published data. Societies should take on the role of formulating best practice guidelines for data publication. These best practice guidelines need to be enforced by funding agencies and journal editors through policies established based on the guidelines.

Theme 3: Data and Metadata Interoperability

In a world of heterogeneous data formats, different metadata formats and standards, diverse terms or vocabularies, and varying interfaces or protocols for metadata and data transport, true interoperability of data exchange requires global standards, and tools and services to support them. To the extent each community or subcommunity has its own data management needs and practices, those groups must collaborate to establish agreements on the common approaches they will follow to enable global interoperability.

Session I: Standards and Technologies for Metadata and Interfaces

The Session I working group was asked to review:

- existing standards for metadata and interfaces, their current use, success, advantages, and disadvantages
- existing registries for data resources, their use, success, and other attributes
- whether new technologies or standards are needed

Metadata

Group discussions began with metadata and metadata requirements. The need for metadata, and the range of metadata required, vary depending on the intended application. It is important to define the use applications before considering what metadata standards should be adopted. Capturing metadata is motivated by the desire to describe data (who, what when, where, how, data quality); facilitate data discovery and new scientific collaborations; reprocess and synthesize data; exchange data, including harvesting it at one location for specialized use; and generate user interfaces.

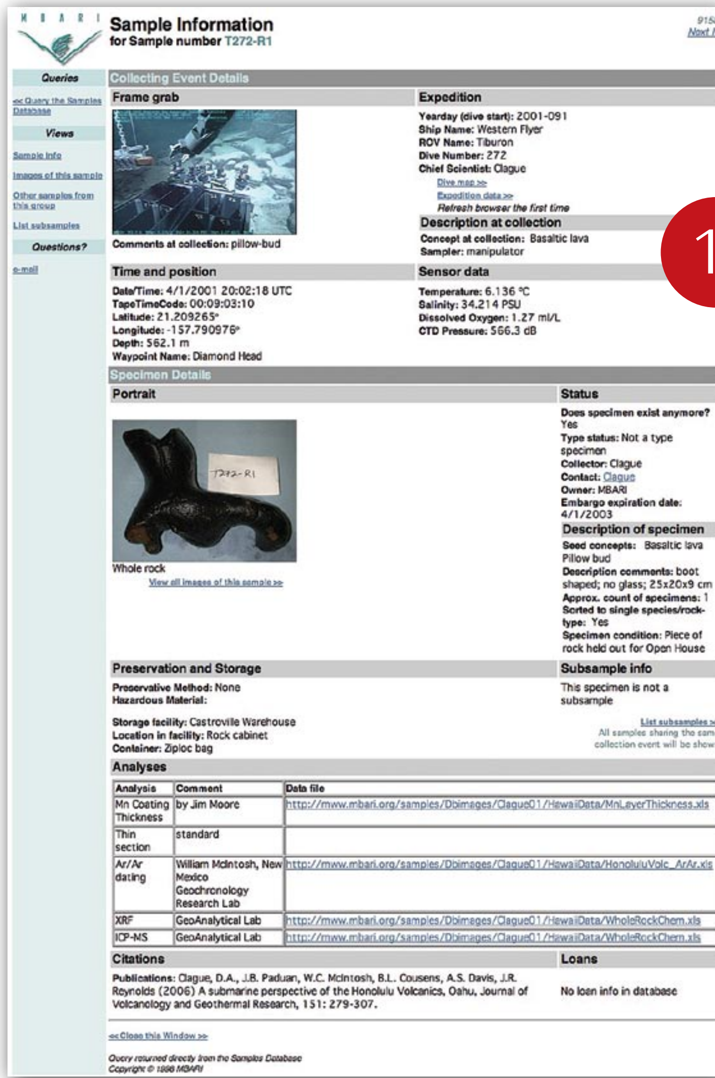


Figure 14. Screenshot of Monterey Bay Aquarium Research Institute (MBARI) Samples Database, a data resource for samples collected during MBARI cruises. Figure from John Graybeal, MBARI; see http://www.mbari.org/samples/docs.

Extensions, Profiles, and Vocabularies

Extensions are additions to a metadata standard that allow users to provide information in additional fields that were not mentioned in the original standard. In standards such as ISO 19115, extensions include:

- addition of a new metadata section
- alteration of the domain of a metadata element (e.g., assigning a code list to specify allowable responses for that metadata element)
- addition of terms in a code list
- addition of a new metadata element to an existing metadata element
- addition of a new metadata entity
- changing the obligation of a metadata element from optional to mandatory (but not the reverse, which would break the core standard)

Constraints are considered a specialized subset of extensions, in which additional restrictions are placed on the standard. (In the above list, items 2 and 6 are constraints.) In this case, the term "extension" describes the addition of information to the standard, even though the metadata instances that follow the standard are restricted.

Profiles are the community-specific application of the metadata standard. In a sense, profile = metadata content standard + extensions. Profiles must meet the core requirements of the metadata content standard (that is, provide the mandatory elements that the standard requires) but can include extensions (described above). Since we also know a metadata content standard is composed of the core metadata set, a profile also can be thought of as profile = core metadata set + optional elements + extensions.

The developers of most content standards expect and encourage the development of extensions and profiles, and may direct how they are to be specified and/or registered. A community that adopts a profile increases the interoperability of its metadata internally. It even increases its interoperability with communities that use other profiles, because the use of the core metadata elements is shared.

An important way that content standards may be constrained is through the use of vocabularies. Vocabularies can be used to fill out particular fields within the standard. The vocabulary used may be specified within the standard itself (e.g., some fields in ISO 19115 define possible entries); or the standard may describe how to specify the vocabulary or vocabularies used (netCDF COARDS/CF allows users to specify the "standard vocabulary"); or the standard may be silent about vocabularies (the CSDGM is fairly open about how many fields are filled out). As noted above, extensions are a common way to narrow the options for filling out fields requiring textual responses. From the Sensor Metadata Interoperability Workshop Report, 2007 (http://marinemetadata.org/smireport).

The development of a consistent community practice with respect to metadata is hindered by a wide range of problems:

- Benefits of metadata may not be adequately understood by those who originally document a data set, leading to inadequate metadata for most reuse.
- Interpretation of standards differ, and for some standards there is little guidance on how to fill them out.
- To make metadata fully discoverable and usable by scientists from other fields, it may be necessary to satisfy a number of sophisticated standards and vocabularies, even for a single data set. This level of refinement is not fully supported by current tools and data models (with some exceptions), and is not expected by users.
- For legacy data, it may be very difficult to recover all needed metadata after observations have been made.
- The desire to control what information is exposed sometimes constrains the metadata that is provided (e.g., the location of a ship working in an ecologically or financially sensitive area).
- Initial creation of metadata by users can be time-consuming, confusing, and unrewarding (due to the amount of metadata requested, poor tools and user interfaces, and limited infrastructure supporting metadata creation).

Common practice for how metadata are provided also varies greatly among disciplines and data types. For some data types, metadata may be embedded in formatted data (e.g., GeoTIFF, HDF, NetCDF, NITFS, SEGY, MGD77, GRIB). For embedded metadata, additional challenges include inconsistent metadata formats in file headers and the often inadequate models and structure for information (metadata/data) adopted in the file format. For other data types, metadata are provided external to data. Currently used standards include FGDC, DIF, Dublin Core, and ISO 19115 (following the implementation approach of ISO 19139 in XML).

Most data and metadata centers are moving to work with ISO 19115, but it is a somewhat general-purpose standard. To become more useful for a particular community, a profile or extension (see sidebar on p. 18) must be developed that meets community needs. Of course, such tailored enhancements of the standard will not work with the ones developed for other communities unless specific measures are taken to ensure interoperability. In addition, ISO standards are not freely available (and in fact are somewhat costly). Workshop participants expressed concern that these issues might inhibit widespread adoption of ISO 19115.

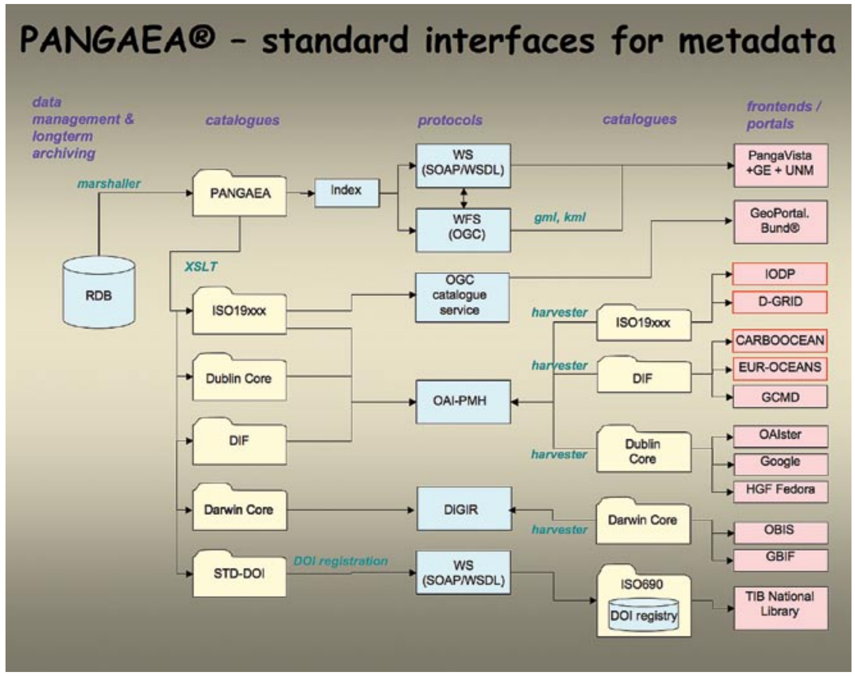


Figure 15. Illustration of the PANGAEA data system components, including metadata standards and interfaces supported. Figure from Hans-Joachim Wallrabe-Adams, World Data Center for Marine Environmental Sciences.

Interfaces

To develop an interoperable system requires more than standardization of data and metadata formats. It requires consideration of the interfaces to data catalogs or data servers that facilitate data transport between distributed repositories, and of the interfaces to services, such as vocabulary list servers, unique reference systems (that generate unique identification numbers or strings for objects and data sets), and universal resource name resolvers (that can translate a URN to a Web site, or to other information as appropriate). The specification for these interfaces includes *transport protocols,* which describe how the connection is made between systems, and is likely to include a specification of the content that is transferred using the protocol. That content specification is analogous, and in some cases the same as, the content specifications described above.

Just as there are a wide variety of data and metadata formats currently in use, there are also a wide range of protocols in common use for interfaces (e.g., SOAP, REST, OAI-PMH, UDDI, WSDL, OPeNDAP, THREDDS). General needs with respect to interfaces are for a well-defined, overarching architecture that is open for neighboring communities to access; consistent ways to discover data; coherent, consistent, and complete standards with respect to a science domain; better tools to work with standards; and better collaborative tools that gracefully integrate appropriate interfaces, or can be used to develop new ones. Interfaces must be chosen and implemented appropriate to use requirements and current state-of-the-art practice.

Registries

Registries provide searchable lists of "objects," which are typically computation resources but may range from Web sites, to metadata, to data sets, to data systems. An overview of some existing registries relevant for marine, and more broadly geoscience, data are listed in Table 1. Registries for a variety of other kinds of "objects" are currently lacking. For example, registries of Web Map Services, online KML resources, or of sensor information are all needed.

Principles for Selection

When selecting the protocol, content, and vocabulary specifications and tools for a community, consideration should be given to the needs of the community and characteristics of the available resources (specifications and tools). Factors to consider include the degree of adoption of each resource (within the community, and as a whole); the degree to which the resource describes or satisfies the characteristics of interest to the community, or can be extended to do so; and the degree to which the resource will be used in automated systems. Another important consideration is whether the agreement is intended to come up with a working solution as quickly as possible, or is able to develop a solution that can support future growth of both the community and the larger environmental cyberinfrastructure. More capability is possible, and required, for systems to support anticipated advances in cyberinfrastructure.

There are several existing community-based efforts relevant to the selection and development of standards and protocols to support data exchange within the marine science community. Examples include the SeaVox project (www.bodc.ac.uk/data/codes_and_formats/seavox/), the Ocean Biogeographic Information System (OBIS, www.iobis.org), and the Marine Metadata Interoperability Project (MMI, www.marinemetadata.org). SeaVox is a joint SeaDataNet/IOC Vocabulary Content Governance Group, moderated by the chair of the IOC MarineXML Steering Group (currently Roy Lowry from BODC), for the development of controlled vocabularies in the marine data domain. These vocabularies comprise topics such as parameters, platforms, instrumentation and spatio-temporal coverage. The MMI hosts a wide range of information on specifications and tools and encourages contribution of information developed by the community for others (in that and other communities) to use. They also encourage community

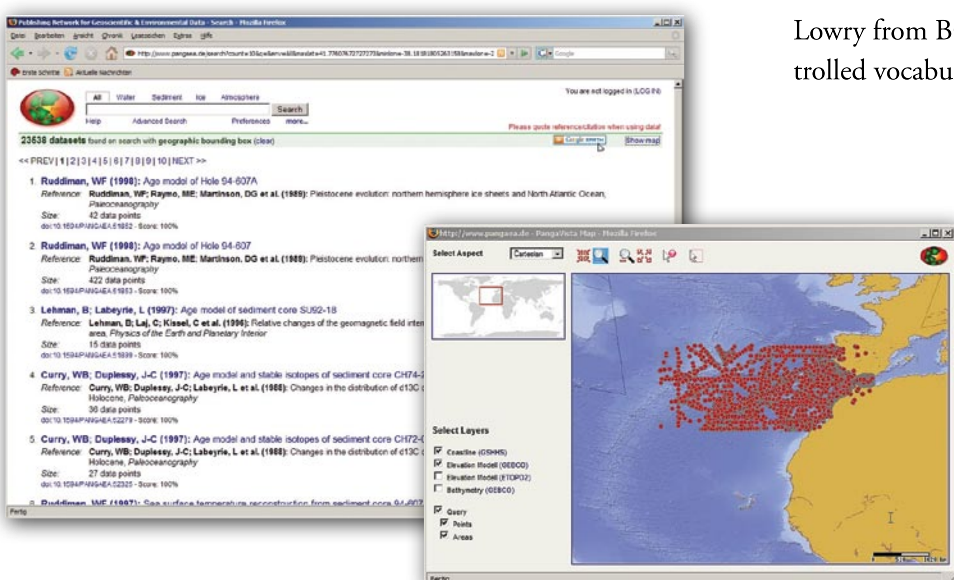


Figure 16. Example of results using PANGAEA's "PangaVista" search interface. The search is for all data within a specified geographic box. Figure from Hans-Joachim Wallrabe-Adams, World Data Center for Marine Environmental Sciences.

Table 1. Compilation of some existing registries for marine data sets and other resources

Registry	Objects	Services	Interface Protocol	Metadata
GCMD	Data sets	WxS		DIF
STD-DOI	Data sets		SOAP	
OceanPortal	Web sites			
SESAR	Samples		WSDL/SOAP	
Pangaea	Data sets		OAI-PMH	DIF, DC, ISO
WDC	Data sets			
GeoNetwork	Data sets		Z39.50	ISO, FGDC, DC
GeoConnections	Data sets			FGDC, ISO
SEDIS	Data sets	WMS	OAI-PMH	ISO
NDG	Data sets		OAI-PMH, SOAP, REST	MOLES, FGDC, ISO, DIF, DC, CSML
OAIster	DOIs		OAI-PMH	
GEON	All	WxS	WSDL/SOAP	

projects, which are developing their own standards to consider using the MMI site to host their materials and publish their deliberations.

Recommendations

T3-R1. The community must minimize the proliferation of metadata standards and work toward a uniform approach for high-quality scientific metadata. There are two basic approaches to the problem of proliferating metadata standards: (1) develop a single uniform specification for scientific metadata and (2) facilitate mediation or crosswalks among a limited number of different metadata standards. A single universal specification may be unattainable, but a coherent, consistent, science-focused approach, ideally centered on building a minimum subset of profiles around a single standard, will limit the proliferation of profiles and ensure that the concept of developing crosswalks is viable.

T3-R2. The community must create agreed-upon processes for community development of standards, guidance, and profiles. Governing structures are needed to enable the development of a community consensus about overall standard(s) and approaches, and to establish processes for developing "official" extensions as needed for different specialized fields.

T3-R3. Community-based best practices for adoption of the ISO 19115 standard are required. As many groups within the global geoscience community are moving to adopt the ISO 19115 standard, there is a strong desire to avoid fragmentation and adopt a common solution to the problems of interpretation associated with this standard. To address these issues, a subcommittee of geoscience data-metadata users needs to be established to come up with a best practice document with clear examples for application of the ISO 19115 standard (and ISO 19139). These guidelines would provide recommendations developed by the scientific community to resolve the interpretation ambiguities of the ISO standard, provide examples, and make the current standard more portable between data and metadata centers.

T3-R4. New efforts within the marine geoscience community to develop standards and protocols to support interoperability should build upon and take advantage of existing efforts. Community-based efforts such as OBIS, the SeaVox project, and MMI offer relevant services, as well as forums for participation and contribution.

Session II: The "Low-Hanging Fruit" for Data Exchange

The Session II working group focused their discussions on identifying opportunities for interoperability in the near future given the existing data resources within the global marine geoscience community. This group was asked to:

- explore realistic opportunities for the implementation of international data exchange
- define a plan for easy start

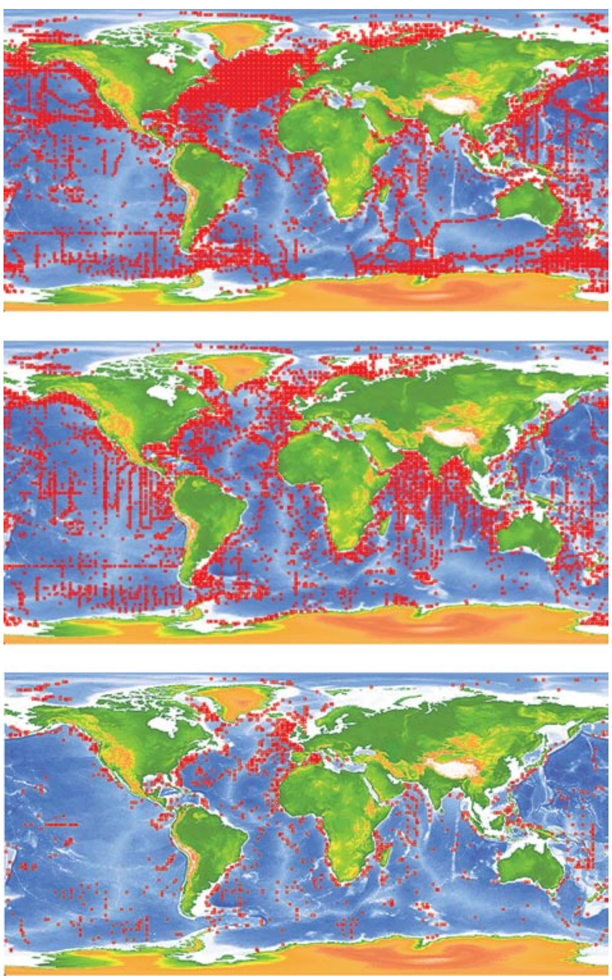


Figure 17. Figure illustrates results obtained using the OBIS data portal to search for all data points available through OBIS at different depths: above 100 m, between 100 m and 1000 m, and below 1000 m. The results show the decrease in the number of available data with increasing depth. Figure provided by Eva Ramirez Llodra from Ocean Biogeographic Information System. April 2007. http://www.iobis.org

A growing variety of data resources relevant for marine geoscience research now exist within the international community. Each provides varying levels of data discovery and data delivery through their own custom search interfaces. At present, to find data of interest across these distributed data centers, a user must first be aware of all relevant data resources, visit each site, and learn how to use the particular search interfaces provided (often in a language other than their own) just to determine whether data of interest exist at that data center. In contrast to the current scenario, users desire the ability to seamlessly discover (and then access) data of interest across distributed centers without the need for pre-existing knowledge of each resource and how to use their search tools.

The general consensus was that an achievable initial goal is to develop a data discovery resource across a subset of the distributed and heterogeneous data resources now available within the international community. Discussions regarding how to implement a resource discovery interface focused on its scope, as well as organizational and technical issues.

Scope

One approach for building a resource-discovery-only interface would be to harvest online metadata from distributed resources across the marine geoscience community into a central repository (e.g., through the World Data Center system), which would build the discovery interface. Metadata could be gathered by harvesting from distributed data centers or through centers contributing to the central repository. The European Union has adopted the model of a central metadata resource through the SeaDataNet project (see http://www.seadatanet.org/). However, a centralized metadata repository for the broader global community is unlikely to be an optimal solution in the short term. Working group participants agreed that a more practical approach would be to identify

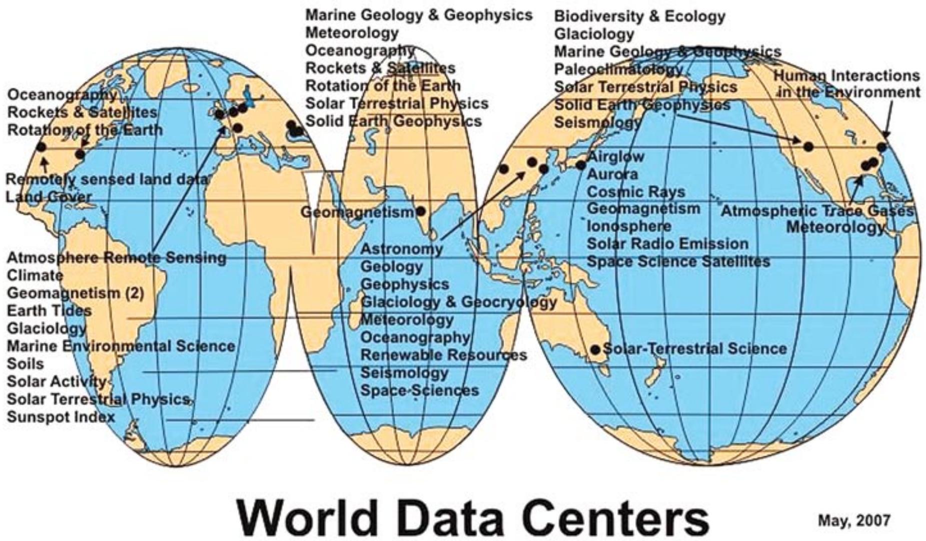


Figure 18. The World Data Center (WDC) system encompasses 51 centers in 12 countries. Its holdings include a wide range of solar, geophysical, environmental, and human-dimensions data. Figure from Ferris Webster, International Council of Science WDC panel.

a few select focus sites for building a discovery-only interface as a proof of concept (e.g., the MoMAR site on the Mid-Atlantic Ridge and/or the Nankai subduction zone). Existing international programs such as InterRidge, InterMARGINS, or IODP could be used to host the data discovery service.

Organizational Considerations

An organization structure for the discovery of marine data across the European community already exists with SeaDataNet. There are currently several marine geoscience data providers within North America, Asia, and the UK with significant data holdings that could participate to bring in a larger suite of resources across the global community. There was a general consensus that an online forum or process to support group collaboration would be valuable.

Technical Issues

Technical issues include how to obtain the needed metadata from distributed resources. Metadata could be harvested by a central portal in an agreed upon standard format on a regular basis (e.g., like the standardized collection level metadata provided via the Cruise Summary Report within SeaDataNet). Harvesting is preferred over the submission of data by providers as it encourages them to invest in themselves and develop Web services for their data resources. Some data centers have deployed OGC Web services for serving some elements of their data holdings (e.g., the Publishing Network for Geoscientific and Environmental Data [PANGAEA], the National Oceanic and Atmospheric Administration National Geophysical Data Center [NGDC], the Incorporated Research Institutions for Seismology [IRIS], Marine Geosciences Data System [MGDS], Petrological Database of the Ocean Floor [PetDB]). An alternative approach would be to serve metadata through the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH). SeaDataNet will be using ISO 19139 metadata standards. The broader marine geoscience community could move to adopt this standard to facilitate interchange with the EU community. As part of

Figure 19. SeaDataNet partnerships include major oceanographic institutes (including National Oceanographic Data Centres [NODCs]) of 35 countries within the European Union and northern Africa. Figure from Dick M. A. Schaap, SeaDataNet.

developing a common metadata standard, there is a clear need within the community to harmonize and map vocabularies for key parameters, including platforms, devices/sensors, and data types. In this context, the MMI initiative or SeaDataNet itself could play a role. International Oceanographic Data and Information Exchange (IODE) members usually offer services as well.

The advent of Google™ Earth/Google™ Maps as a tool for locating data is an attractive option for a community of distributed data providers to enable quick visualizations of location of their data resources. Data providers could provide a KML service with their collection metadata to show locations of their data, purely for discovery. Serving a visualization of the actual data through KML is also low cost ("this is what the data look like") as existing images can be readily wrapped in KML (e.g., using PHP). However, the value of this service depends on data type and quality.

Recommendations

T3-R5. Development of a data discovery service across globally distributed marine geoscience data resources is an achievable initial goal. First steps should focus on collecting metadata, starting with cruise-level information (e.g., geographic extent, expedition information, list of parameters/data types, instrument, temporal extent). Sample, station, and track locations should be provided to enable data resources to be discovered in map-based searches. A few selected mid-ocean ridge and continental margin test-bed sites could be adopted for building a proof-of-concept, discovery-only interface. For example, the MoMAR site on the Mid-Atlantic Ridge, Nankai or Costa Rica subduction zones, or New Zealand margin would all be excellent candidate sites given the current interest in these areas within different groups of the international community. The existing relevant data resources for these sites can be readily identified. A simple search interface could be built that might be hosted for the international community at InterRidge, InterMARGINS, or IODP. An initial low-cost–of-entry data discovery service could take the form of a repository of existing KML resources with encouragement for others to offer their resources in this format.

T3-R6. Data centers should work to expose their data resources via Web services (e.g., OGC or OAI protocols). Workshop discussions pointed in the direction of the creation of Virtual Organizations (VO) in which their members provide independent ways to expose their resources to generic portals. Web services enable data resources to be readily harvested by other services and provide scientists with the flexibility they desire to discover and access data in the front-end analysis and visualization tool of their choosing. A large and increasing number of GISs can interface with OGC-compliant Web services so that data from many different sources can be discovered, visualized, inter-compared, analyzed, and shared with other applications. Developing a "critical mass" of OGC-compliant services is an important strategic goal for achieving the vision of truly interoperable federated systems.

T3-R7. Development of special interest groups with tools to facilitate collaboration is needed. The range of experience and level of expertise/resources available to different segments of the global marine geoscience data management community varies widely. At the same time, technologies for information management are rapidly evolving. Development of special interest groups to share experiences and solutions, and to provide guidance, would be valuable for this community. An online forum or process to support group collaboration is needed (e.g., Google™, Elgg).

T3-R8. A dedicated task group is needed to harmonize and map vocabularies for key parameters, including platforms, devices/sensors, and data types. There are existing processes that could be used, but harmonization is not a trivial task. In addition to facilitating interoperability among existing data centers, harmonization of vocabularies and development of a publicly accessible vocabulary service would be very valuable as new data resources are being built. Where possible, existing community efforts should be leveraged to advance this goal (MMI, SeaVox, SeaDataNet, OBIS).

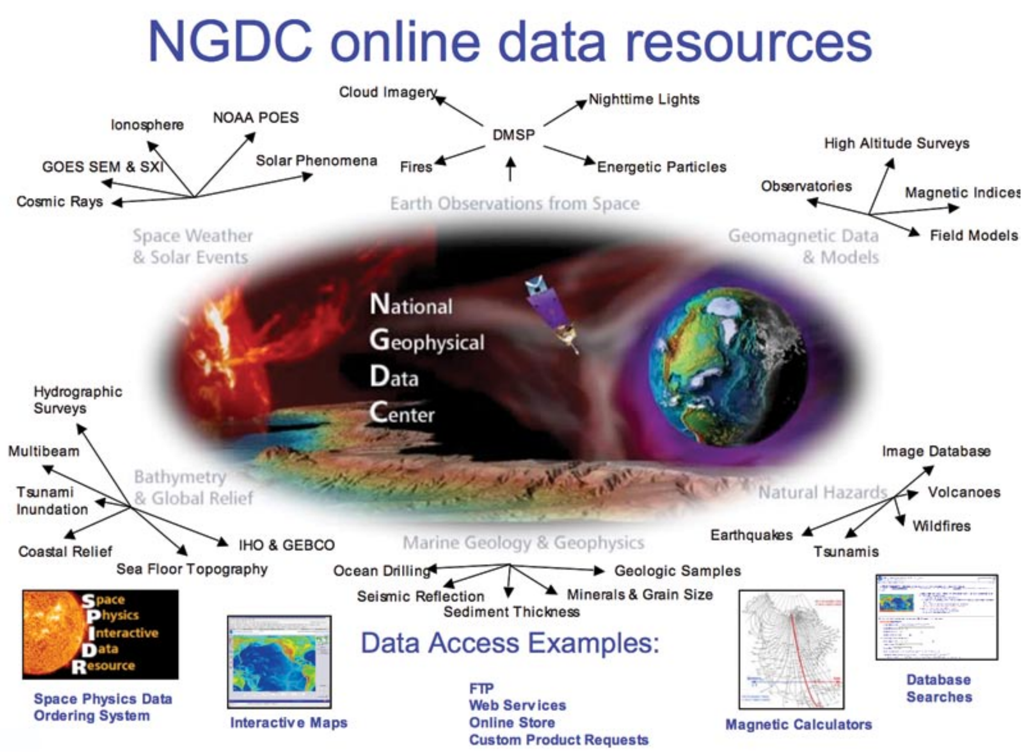


Figure 20. The US National Geophysical Data Center (NGDC) provides stewardship, products, and services for over 400 geophysical data sets describing the solid Earth, marine, and solar-terrestrial environment, as well as Earth observations from space. A variety of mechanisms exist to access data, ranging from file systems, to database query methods, and include use of Web service protocols for machine-to machine-transactions. Figure from Chris Fox, NGDC.

Theme 4: Opportunities and Obstacles for International Data Sharing

Session I: Archives and Contributions

As a first step toward identifying opportunities and obstacles for international data sharing, the working group assessed the range of policies and practices for data archiving and data access on an international scale, addressing the following questions:

- What are the national, institutional, agency, and society policies with regard to data contributions and enforcement?
- Are available archives adequate? Are there orphan data types?
- What is the status of contributions to archives in practice? Do they need to be improved? If yes, how can that be achieved?

Countries represented in the working group included Spain, France, Norway, United Kingdom, Japan, New Zealand, Taiwan, United States, Oman, Canada, and Germany. During a round-table discussion, working group members described—to their best knowledge—data policies of their country for the data types and data centers relevant to the workshop topic. Several issues of note are summarized here:

- Data policies, where they exist, vary widely among and within countries and on all levels.
- Many countries still do not enforce data contributions from individual investigators at private/academic institutions, even if official policies require it. One notable exception offered was the UK, where a NERC-funded investigator was penalized for noncompliance.
- Government agencies typically have stronger policies and better enforcement than private/academic institutions, even in cases where investigators at such institutions receive government funding.
- Several countries have comparatively stricter policies for data within their Exclusive Economic Zone (EEZ), particularly while the United Nations Convention on the Law of the Sea (UNCLOS) mapping and claims are underway. In some cases, a country may require any research vessel traversing its EEZ, foreign or domestic, to submit a copy of all data collected.
- Overall, the situation has improved from that of five to ten years ago. Improvements in technology, such as faster network connections and larger storage systems, have made it easier for investigators to post their data online and/or contribute them to data centers.
- In many countries, it is still often necessary to "know the right person" in order to find and obtain data sets. In addition, use of the data may be restricted in some manner, making it very difficult to obtain data of interest.

Encouragingly, the overall trend in the last decade is toward greater openness in data sharing. Some countries still "guard" valuable data sets by imposing processing fees and intellectual property claims, but there is growing consensus to build interoperable systems and to adopt data standards. An example is the recent series of EU initiatives, including SeaSearch (2002–2006) and SeaDataNet (2006–2010). Also, recent natural disasters have caused some countries to more fully acknowledge the need for broad and open access to data. In the academic community, data management systems that are developed and operated through science initiatives such as the NSF-supported Ridge2000 and MARGINS programs are recognized by their target community as providing a highly useful service. Appreciation of such systems substantially contributes to a culture change in the science community toward more open data sharing.

While the number and variety of data centers all over the globe are growing, anecdotal evidence suggests that many countries lack data centers for particular data types. Examples offered include paleoclimate data in the UK; ocean bottom seismometer data in France; undersea acoustics, hydrology, and volcanology in the United States; and wildlife observations in New Zealand. Many countries also lack facilities for curation of physical specimens. Further, some data centers (or networks of centers) exist but are incomplete, such as sea-level (tide-gauge) data in the United States.

With the growing number of data centers, it becomes increasingly harder for scientists to easily find all the data in their area of interest. Perhaps the most significant and universal problem with existing data centers (and networks) is a lack of standard registration and discovery. No mechanism is known to exist for truly comprehensive, interoperable, international search across global data holdings.

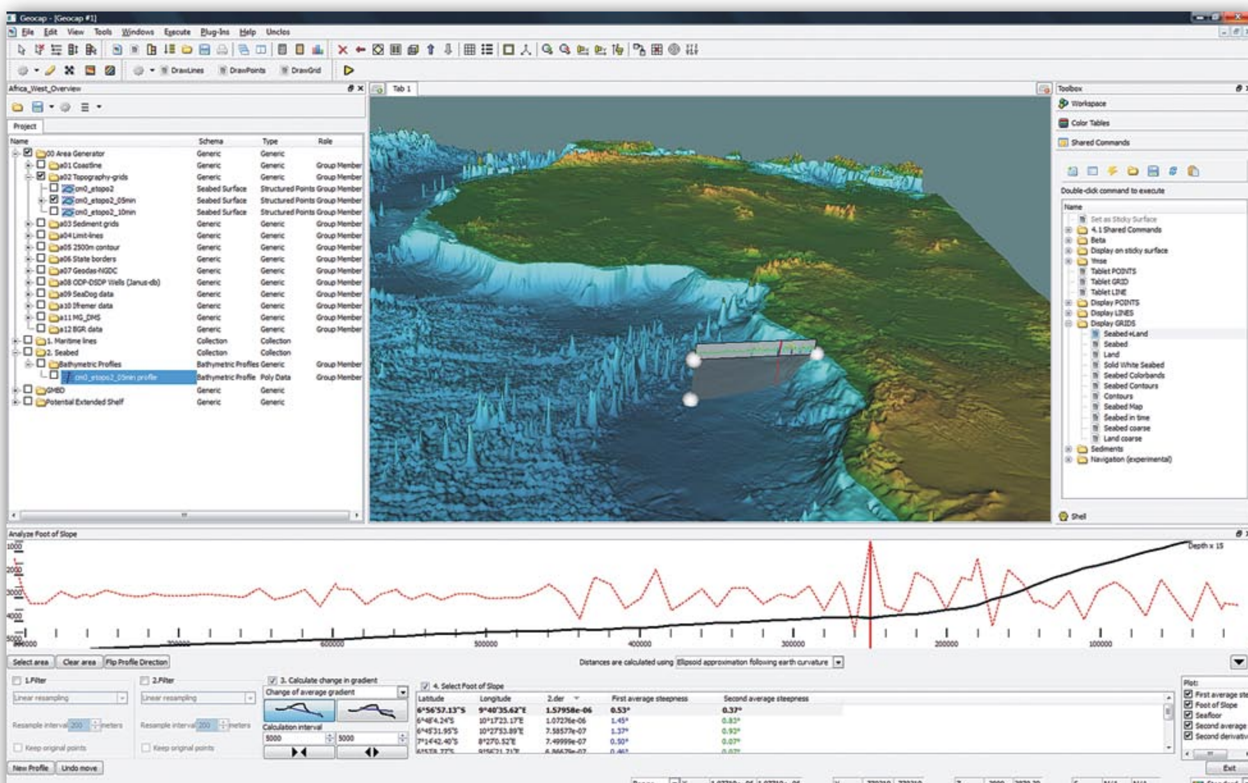


Figure 21. Article 76 software, Geocap as used by the United Nations Environment Programme (UNEP) Shelf Programme to provide complete project data packages to states delineating the outer limits of their continental shelf. Data types that can be imported and manipulated include various forms of analogue and digital seismic and single and multibeam bathymetry. Figure from UNEP Shelf Programme/UNEP-GRID-Arendal.

Recommendations

T4-R1: Uniform best practices and standards for data acquisition and data submission should be adopted on a global scale. To achieve a higher level of data contributions to data archives and to facilitate the enforcement of data policies, ship operators and scientists worldwide are encouraged to adopt consistent best practices for data acquisition and submission. Metadata should be collected in a standardized way during data acquisition, and automated wherever possible. For example, where possible, metadata should be encoded directly into data streams from sensors and other data-acquiring devices. As part of best practices, formal submission agreements should be established between individual institutions and respective national and international data centers in order to effectively aid regular, timely, and standardized contributions to data centers.

T4-R2: Real-time (field) data, processed data, and laboratory data products should all be archived. Experience has shown that raw data need to be archived because they become useful for applications that were not anticipated during the original acquisition. At the same time, it is important to also archive processed data, such as edited multibeam sonar data, because in this form they are most useful to the broadest range of users beyond the specialists who are experienced with handling the raw field data types.

Archives for derived data (value-added data products) are glaringly missing, especially for products that are never formally published or are only published in print journals (unavailable online). Solutions should be explored to parallel efforts in other science fields (e.g., astronomy) to archive derived data products in collaboration with university libraries or journals.

T4-R3: The strong and continued interest in data from coastal waters should be leveraged to attract funding for data systems and standards. Strong interest in shelf mapping and benthic habitats within EEZs in particular can help increase resources to operate data centers and advance interoperability technologies and standards.

Session II: Implementing an International Data Network

A common vision, broad community support, an organizational framework, and resources are required to implement the envisioned international data network. To address these issues, the working group discussed the following questions:

- What levels of data access are desired? How can we achieve this? What technologies and agreements are needed?
- What are the appropriate organizations to advance the goal of data sharing (e.g., eGY, GEOSS)? Do we need separate ones?
- What funding is needed? What are the funding opportunities?

Users ultimately want to have open access to all data for their field of research, easily discovered and accessed on a global scale via a central portal and downloadable for free in a common format that is supported by many standard applications. Modern Internet and database technologies now allow construction of an overarching infrastructure/architecture that can act as an umbrella network over all relevant data centers in the geological, geophysical, and geochemical community and that supports interoperability or uniform communication by applying technical and content standards to give users insight into its inventories and access to its data sets. At the most basic level, metadata should provide

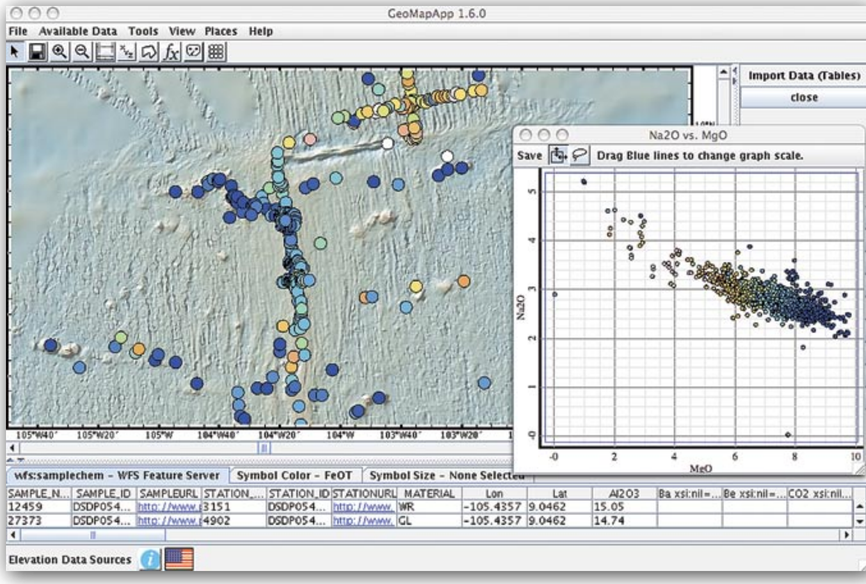


Figure 22. GeoMapApp data visualization interface illustrating the capability to access data that reside externally through OGC-compliant Web Feature Services (WFS). With the GeoMapApp, data served from PetDB through a WFS are directly accessed, displayed in map view, and analyzed. This example shows the East Pacific Rise near the Clipperton Transform Fault. Rock sample locations are color coded by FeOT (orange-reds are high values, and blues are low). Chemical parameters can be graphed; here MgO is plotted verses Na_2O, with dots colored by FeOT content.

information on the originators, location of the data sets, and data quality. Metadata are not only needed to ensure that data originators and original data centers receive credit for their work, and to make future users and funding agencies aware of their value, but also to ensure the quality level of available data sets for global use. Web services are now being used to set up overarching systems. This technology supports development of standard application programming interfaces (APIs), so that other users can build their own applications on top of the databases that are managed by the Web service. This development facilitates sharing and interoperability. However, references to data originators and original data centers need to be safeguarded.

The challenge is to identify sensible ways to organize the development of a global data network and establish the right institutional culture that allows such an endeavor to move forward with support from both scientists and decision-makers. The scientific culture needs to become more open toward data sharing and servicing the public (e.g., with raw and interpreted data and information). Global and regional issues, developments in technology, more multidisciplinary research, and understanding of the benefits will be drivers for this culture change. National and international governments are stimulating more open data sharing and exchange by directives and guidelines. Institutes, organizations, and agencies are more open to data exchange than ever before. For example, the ocean-observing programs of Canada and the United States are adopting policies of immediate open public access to all data, with few exceptions. In Europe, the Infrastructure for Spatial Information in the European Community (INSPIRE) directive calls for easy access to public domain data (http://www.ec-gis.org/inspire/). At a global level, the Group on Earth Observations (GEO) initiative by the G8 countries promotes the development of GEOSS (http://www.earthobservations.org/index.html). Backed by policy-makers and decision-makers, these initiatives also help significantly in obtaining appropriate funding for data facilities. Various reports on access to research data have been published recently, for example, a report by the Organization for Economic Co-operation and Development (OECD) called "OECD Principles and Guidelines for Access to Research Data from Public Funding" (OECD, April 2007[5]).

[5] http://www.oecd.org/document/55/0,3343,en_2649_37417_38500791_1_1_1_37417,00.html

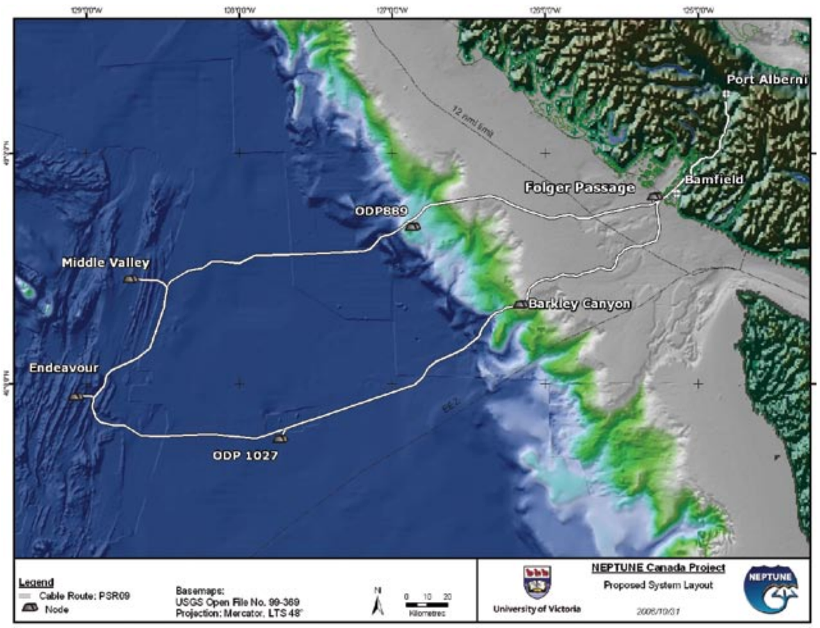


Figure 23. Proposed backbone route and sites of scientific interest for the NEPTUNE Canada regional cabled observatory. The cable is to be laid between the end of August and the beginning of October 2007. Instruments, nodes, and junction boxes will be deployed in late summer of 2008. Figure from NEPTUNE Canada.

Data policies still exist and will remain in place at institutional or national levels that restrict access to specific data types and specific data domains (e.g., data sets from hydrographic surveys in the EU, Asia, or in military circles). At the international level and in specific domains, data policies have been formulated that aim for openness, at the same time respecting local policies. Another way to overcome restrictions would be to create data products with specific resolution that make use of the original data sets for their production. These data products then can be made free and open, and in most cases already satisfy the needs of the users. For scientific purposes, a moratorium period of two years is normal practice to ensure that the originating scientists can make exclusive use of the data they acquire for their scientific work. However, the existence of these data sets should be discoverable by entries in metadata systems.

Large international bodies, such as GEOSS, ICSU, and the Committee on Data for Science and Technology (CODATA), can help the community initiate a global data network for geoscience data; these groups can inform decision-makers and policy-makers about the motivation for such a network and the community's position. But, programs such as GEOSS cannot steer development of the network itself. This has to be done by the players (data centers, scientific communities), who need to prepare a plan for an overarching network that will connect existing data centers and include a critical mass of users. This network will not only provide access to metadata and data, but also establish uniform vocabularies and standard protocols.

To build this network, two approaches are possible: (1) start building a network among a small number of ICSU World Data Centers to demonstrate its feasibility and (2) start with a wider group of data centers, including the World Data Centers, to get a broader foundation and more involvement at the institutional level (e.g., major marine geoscience surveys and relevant institutes in the EU, Asia, North America, and elsewhere).

Recommendations

T4-R4: Advance a culture among scientists that is more open to data sharing. Scientists need to agree that data sharing is beneficial to research and knowledge circulation so that they actively support and contribute to a data infrastructure that is based on open data sharing. This culture change can be advanced through practical examples and cases. For example, participation in international and multidisciplinary projects such as the EU Research Framework Programme or the IODP requires researchers to be more open towards data sharing. Participation in these projects is attractive to researchers because they provide funding opportunities at a time when many have to "fight" for research funding. A critical aspect of advancing an open data exchange culture is that data infrastructure guarantees appropriate credit to data authors and data providers.

T4-R5: International programs and bodies, such as GEOSS, eGY, IPY, GBI, and ICSU, that stimulate the development of data-sharing systems should be leveraged to promote our initiative for a global data network for marine and terrestrial geoscience. There are ongoing international research programs, such as CODATA, the eGY, and the EU Frameworks Programmes, as well as a number of international policies, adopted and driven by governments, that are intended to encourage and support international cooperation toward a global data infrastructure. International bodies such as IOC, ICSU, the World Meteorological Organization (WMO), and the United Nations Environmental Programme (UNEP), with membership at the country level, adopt and support these programs and plans. Reference to these programs and bodies is important to help decision-makers and policy-makers understand the motivation and position of an initiative for a global data network that is emerging from this workshop.

T4-R6: A dedicated task group should be established to advance implementation of a global data network for marine geoscience data. In many regions of the world, the time is ripe for starting to construct a network of data resources on an international level. A task group composed of the organizers of the workshop and some of the workshop attendees should prepare a plan for broad access to metadata and data by means of an overarching network that will connect existing data centers. In particular, the task group should:

- formulate a precise definition of the aims and scope of the overarching system
- prepare a matrix of relevant organizations from all over the globe that should be invited and engaged in its further planning and proposal development
- explore funding options with various agencies (e.g., NSF, Japanese agencies, and EU Framework); seek to formulate an overarching proposal to one or more of these agencies that demonstrates clear community interaction and complementarity of data

It will be helpful to know how ongoing international programs, such as IODP and eGY, achieved international recognition and funding from various complementary resources.

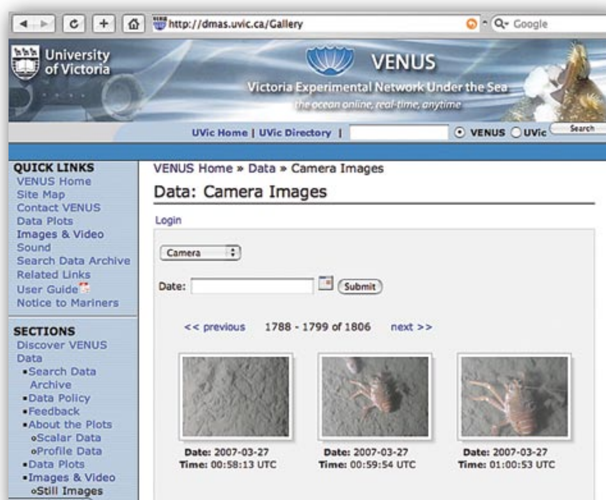


Figure 24. Example of VENUS still camera images provided through the Data Management and Archiving System (DMAS) browsing interface. The interactive camera is located in the Saanich inlet off Vancouver Island at a depth of 95 m and is cabled to shore. It provides real-time access to seafloor images. Figure is screenshot of DMAS interface, courtesy of VENUS/University of Victoria.

Next Steps

From the working group recommendations, the following immediate next steps are identified:

1 Adopt test-bed sites for development of a data-discovery service across distributed marine geoscience data resources within the international community (T3-R5).

Form alliances focused on a few select mid-ocean ridge and continental margin test-bed sites (e.g., the MoMAR site on the Mid-Atlantic Ridge, Nankai or Costa Rica subduction zones, or New Zealand margin) where relevant data centers work to expose available data within these sites. A KML repository of data resources accessible in Google™ Earth for these sites could be developed, perhaps hosted through InterRidge, InterMARGINS, or IODP. A centralize resource or registry of relevant Web services is needed.

2 Establish forums for guidance and development of best practices.

Special interest groups should be established to share expertise and solutions for development of interfaces and for metadata standards (T3-R7). As many groups within the global geoscience community are moving to adopt the ISO 19115 standard, there is a strong desire to avoid fragmentation and adopt a common solution to the problems of interpretation associated with this standard (T3-R3). A task group to establish best practices for implementation of this standard is needed. A dedicated working group is also needed to develop best practices for data documentation at sea (T2-R2; T4-R1). Existing procedures need to be assessed in light of data documentation requirements to establish guidance for routine shipboard operations across the global marine geoscience community.

There is an immediate need and opportunity to harmonize and map vocabularies for key parameters, including platforms, sensors/devices, and data types (T3-R8). Interested data centers should form alliances, building upon existing efforts and initiatives to move forward on development of publicly accessible vocabulary services to facilitate interoperability.

3 Create higher level task force (T4-R6).

Assembly of a high-level task force focused on forming international alliances among data centers within the marine geoscience world is needed. National and institutional marine data centers should work to align with current efforts within the terrestrial world involving national geologic surveys and efforts such as the FDSN, and join at a high level.

4 Establish opportunities for follow-up meetings of the international marine geoscience data management community.

Building a global data network requires regular forums for the international marine and terrestrial geoscience community to meet, assess progress, evaluate new opportunities, and define future directions. A follow up workshop should be held in one year with possible focus on specific task areas identified in the Kiel meeting.

Appendix 1. Workshop Participants

Abers	Geoff	MARGINS Office, Boston University (USA)
Ahern	Tim	IRIS (USA)
Al-Habsi	Harib	Sultan Qaboos University (Oman)
Arko	Bob	LDEO, Columbia University (USA)
Bach	Wolfgang	University of Bremen (Germany)
Baker	Maria	National Oceanography Centre, Southhampton (UK)
Barckhausen	Udo	BGR-German Geological Survey
Baumann	Peter	Jacobs University Bremen (Germany)
Beaudoin	Yannick	UNEP-Continental Shelf Programme (Norway)
Blackman	Donna	Ridge 2000 Office, Scripps institution of Oceanography (USA)
Blower	Jon	University of Reading (UK)
Briones	Katiusca	Oceanographic Institute of Ecuador
Cannat	Mathilde	Institut de Physique du Globe de Paris (France)
Carbotte	Suzanne	LDEO, Columbia University (USA)
Chen	Bob	CIESIN, Columbia University (USA)
Chiao	Ling-Yun	National Taiwan University
Clark	Dru	University of California, San Diego (USA)
Condit	Christopher	University of California, San Diego (USA)
Cogan	Christopher	Alfred Wegener Institute for Polar and Marine Research (Germany)
Damm	Timo	University of Kiel (Germany)
Devey	Colin	IFM-GEOMAR (Germany)
Diviacco	Paolo	Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (Italy)
Dransch	Doris	GeoForschungsZentrum Potsdam (Germany)
Fox	Christopher	National Oceanic and Atmospheric Administration (USA)
Galkin	Anastasia	GeoForschungsZentrum Potsdam (Germany)
Goldstein	Steven	LDEO, Columbia University (USA)
Goodwillie	Andrew	LDEO, Columbia University (USA)
Graybeal	John	Monterey Bay Aquarium Research Institute (USA)
Haak	Katherine	Ridge 2000 Office, Scripps Institute of Oceanography (USA)
Halvorsen	Oystein	UNEP-Continental Shelf Programme (Norway)
Hanafusa	Yasunori	JAMSTEC (Japan)
Haq	Bilal	National Science Foundation (USA)
Hosseini	Keivan	Ferdowsi University of Mashhad (Iran)
Huettmann	Falk	University of Alaska, Fairbanks (USA)
Javidpour	Jamileh	IFM-GEOMAR (Germany)

Jones	Craig	GNS Science (New Zealand)
Kandel	Cary	MARGINS Office, Boston University (USA)
Khan	Shuhab	University of Houston (USA)
Klump	Jens	GeoForschungsZentrum Potsdam (Germany)
Le Bas	Tim	National Oceanography Centre (UK)
Lehnert	Kerstin	LDEO, Columbia University (USA)
Lezaeta	Pamela	MARGINS Office, Boston University (USA)
Lowry	Roy	British Oceanographic Data Centre (UK)
Matsuda	Shigemi	The Center for Deep Earth Exploration (Japan)
Meier	Thomas	Bochum University (Germany)
Miller	Stephen	University of California, San Diego (USA)
Miville	Bernard	International Ocean Drilling Program (Japan)
Moussat	Eric	IFREMER (France)
Neben	Soenke	BGR-German Geological Survey
Nygard	Atle	University of Bergen (Norway)
Pirenne	Benoát	University of Victoria (Canada)
Ramirez-Llodra	Eva	CMIMA-CSIC (Spain)
Ranero	Cesar	CMIMA-CSIC (Spain)
Ryan	William	LDEO, Columbia University (USA)
Sadeghi	Hossein	Ferdowsi University of Mashhad (Iran)
Salters	Vincent	Florida State University (USA)
Sarbas	Baerbel	Max-Planck-Institute for Chemistry (Germany)
Schaap	Dick	SeaDataNet (Netherlands)
Schaefer	Angela	Jacobs University Bremen (Germany)
Schirnick	Carsten	IFM-GEOMAR (Germany)
Schoolmeester	Tina	UNEP-Shelf Programme (Norway)
Shiomi	Katsuhiko	National Research Institute for Earth Science and Disaster Prevention (Japan)
Shipley	Thomas	University of Texas (USA)
Stransky	Julia	IFM-GEOMAR (Germany)
Trueger	Mickael	IFREMER (France)
Tsuboi	Seiji	JAMSTEC-IFREE (Japan)
Unnithan	Vikram	Jacobs University Bremen (Germany)
Venuti	Fabio	National Oceanography Centre (UK)
Wallrabe-Adams	Hans-Joachim	University of Bremen (Germany)
Weatherall	Pauline	British Oceanographic Data Centre (UK)
Weinrebe	Wilhelm	IFM-GEOMAR (Germany)

Appendix 2. Database Systems

> One-page descriptions for each data resource are available at the meeting Web site.
>
> www.nsf-margins.org/Datawkshp07

Arctic Ocean Diversity Project (ArcOD) Data

An international collaborative effort to inventory biodiversity in the Arctic sea ice, water column and sea floor from the shallow shelves to the deep basins using a three-step approach: compilation of existing data (sea ice algae, phytoplankton, zooplankton, zoobenthos, fish, and birds), taxonomic identification of existing samples, and new collections focusing on taxonomic and regional gaps.
www.coml.org/descrip/aobio

British Oceanographic Data Centre (BODC)

The UK National Oceanographic Data Centre in the IODE network and the Natural Environment Research Council's designated data centre for marine data.
www.bodc.ac.uk

ChEssBase

A web-based relational database integrated with OBIS (listed below). The aim is to provide taxonomical, biological, ecological and distributional data of all species described from deep-water chemosynthetic ecosystems, as well as bibliography and information on the habitats. These habitats include hydrothermal vents, cold seeps, whale falls, sunken wood and areas of minimum oxygen that intersect with the continental margin or seamounts. Sample data are also about to be added to the database.
www.noc.soton.ac.uk/chess

CODATA

Committee on Data for Science and Technology of the International Council for Science (office based in Paris, France).
www.codata.org

EarthChem

Data infrastructure for geochemistry, NSF-funded; data archives hosted by the Geoinformatics for Geochemistry program; portal services to international federation of geochemical databases.
www.earthchem.org

Earthquake Research Center (EQRC), Iran

Management of seismological data generated by the Khorasan Seismic Network (eight stations in the Khorasan provinces, Iran), operated by Ferdowsi University.
seismo.um.ac.ir/

General Bathymetric Chart of the Oceans (GEBCO)

Global ocean floor bathymetric data sets, operated under the auspices of the International Hydrographic Organization (IHO) and the Intergovernmental Oceanographic Commission (IOC).
www.gebco.net

GeoForschungsZentrum Potsdam, Germany

Provides data management for several international projects in earth science, such as the International Scientific Drilling Program, data from CHAMP and GRACE satellite missions, the GEOFON seismological network and other sensor networks and geophysical observations (e.g., GPS, geomagnetic and tide gauges and the German Indonesian Tsunami Early Warning System).
www.gfz-potsdam.de

Geoinformatics for Geochemistry

NSF-funded program at the Lamont-Doherty Earth Observatory and the Center for International Earth Science Information Network CIESIN to operate geochemical and sample data collections PetDB (Petrological Database of the Ocean Floor – www.petdb.org), SedDB (Information System for Marine Sediment Geochemistry – www.seddb.org), EarthChem (see above), and SESAR (see below).

GEO (Group on Earth Observations)

An intergovernmental group leading a worldwide effort to build a Global Earth Observation System of Systems (GEOSS) over the next 10 years. GEO is established on a voluntary and legally non-binding basis, with voluntary contributions to support activities, currently with 69 member countries, the European Commission and 46 participating organizations.
www.earthobservations.org

GEOROC (Geochemistry of Rocks of the Oceans and Continents) Database

An online resource hosted by the Max-Planck-Institut fuer Chemie in Mainz, providing published geochemical data from volcanic rock samples, glasses, minerals and inclusions from ocean islands and several other active and passive tectonic regions. Since 2003, GEOROC joined with PetDB and NAVDAT databases to form the EarthChem consortium (described in this list).
georoc.mpch-mainz.gwdg.de

GEOREM (Geological and Environmental Reference Materials) Database

An online resource hosted by the Max-Planck-Institut fuer Chemie in Mainz, providing reference materials and isotopic standards from samples of rock powder, glasses, minerals, isotopes, river water and seawater.
georem.mpch-mainz.gwdg.de

GNS, New Zealand

GNS Science collects geological, geochemical & geophysical information from around the New Zealand region.
www.gns.cri.nz

International Federation of Digital Seismographic Networks (FDSN)

www.fdsn.org

IRIS Data Management Center (DMC)

Global (and beyond) time series data, an NSF-funded program to manage time series generated by the IRIS Global Seismic Network, the Program for Array Seismic Studies of the Continental Lithosphere (PASSCAL) and other seismological data. In addition, approximately 24 other kinds of time series data are available from globally distributed sensors, such as weather, hydrologic, gravimetric and magnetotelluric data.
www.iris.edu

JAMSTEC (Japan Agency for Marine-Earth Science and Technology) Database

Access to 19 data portals from JAMSTEC database page, such as observation data (e.g., ROVs, submersibles, mooring sites, cable stations and buoys), numerical model output, samples, etc.
www.jamstec.go.jp/e/database

MARGINS Data System

A multiple-data file repository, access tools and web services for data collected in studies funded under the US MARGINS programs.
www.marine-geo.org/margins

Marine Geoscience Data System

NSF-funded program at the Lamont-Doherty Earth Observatory of Columbia University composed of the Ridge 2000 and MARGINS Data Systems, the Antarctic and Ridge Multibeam Bathymetry Synthesis projects, and the Seismic Reflection Field Data Center.
www.marine-geo.org

Marine Metadata Interoperability Project (MMI)

Promoting the exchange, integration and use of marine data through enhanced data publishing, discovery, documentation and accessibility.
marinemetadata.org

Marine Seismic Data Center (MSDC)

Free access to seismic images and data for education and research, hosted by Jackson School of Geosciences, University of Texas at Austin in collaboration with the Field data center from MGDS (LDEO).
www.ig.utexas.edu/sdc and
www.marine-geo.org/seismic

National Geophysical Data Center (NGDC), U.S.

Supported by the U.S. Department of Commerce, National Oceanic & Atmospheric Administration. Provides scientific stewardship for data from the ocean floor (coastal and open ocean areas), and for data related to natural hazards including tsunamis.
www.ngdc.noaa.gov

National Oceanographic Data Center, Oceanographic Institute of Ecuador

Maintains a 50-year database of continuous oceanic, coastal and atmospheric data collections from cruises, permanent meteorological stations and other observatories along the Ecuadorian coast and Galapagos Island, supporting also the Central-South American ODINCARSA project.
www.odincarsa.net and www.inocar.mil.ec

National Research Institute for Earth Science and Disaster Prevention (NIED, Japan)

Seismograph Networks in Japan.
www.hinet.bosai.go.jp/register/ENGLISH
(registration required)

Nautilus, the IFREMER Web portal

Gives access to multidisciplinary in situ data sets of: chemistry, physical oceanography and geophysics, and is being widened to new types of data such as from geology.
www.ifremer.fr/nautilus

NEPTUNE, Canada (The North-East Pacific Time-Series Undersea Networked Experiments)

An 800 km ring powered fiber optic cabled observatory on the seabed over the northern part of the Juan de Fuca tectonic plate, a 200,000 sq km region in the northeast Pacific off the coasts of British Columbia, Washington and Oregon, operated by Univ. of Victoria, nodes and instruments ready by 2008.
www.neptunecanada.ca

Network of Expertise in Long-Term Storage of Digital Resources (NESTOR)

The project's objective is to create a network of expertise in long-term storage of digital resources for Germany. As the perspective of current and future archive users is central to the project, the emphasis is put on long-term accessibility.
www.langzeitarchivierung.de/index.php?newlang=eng

OBIS (Ocean Biogeographic Information System)

A web-based provider of global geo-referenced information on marine species. Contains expert species level and habitat level databases and provides a variety of spatial query tools for visualizing relationships among species and their environment. OBIS strives to assess and integrate biological, physical, and chemical oceanographic data from multiple sources.
www.iobis.org

Ocean Data Bank, National Center for Ocean Research (ODB, NCOR), Taiwan

Limited access for the purpose of academic research to a variety of geophysical and hydrographic data, some of them in development.
www.ncor.ntu.edu/ODBS

OceanLife Project

A testbed collaboration underway between Rutgers University Marine Lab and San Diego Supercomputer Center designed to assist marine scientists to integrate biological and physical oceanographic data, dealing with data sets such as the World Ocean Atlas.
scirad.sdsc.edu/datatech/aqp_ocean.html

PANGAEA (Publishing Network for Geoscientific & Environmental Data)

Public library for dissemination of data and metadata via portal networks using georeferenced data related to basic research on earth & environment. Operated by AWI and MARUM, Germany.
www.pangaea.de

Publication and Citation of Scientific Primary Data (STD-DOI)

A project funded by the German Science Foundation. Its aim is to make primary scientific data citable as publications. In this system, a data set would be attributed to its investigators as authors like it would be done for a work in the conventional scientific literature.
www.std-doi.de

Ridge2000 Data System

A multiple-data file repository, access tools and web services for data collected in studies funded under the US Ridge2000 program.
www.marine-geo.org/ridge2000

SDDB (Scientific Drilling DataBase) of the International Scientific Continental Drilling Program (ICDP)

Holds Deep Earth Sampling and Monitoring data from ICDP operations (www.icdp-online.org) and associated projects, and publishes data sets through the STD-DIO system. It is operated by GeoForschungsZentrum Data Centre and by the ICDP's Operational Support Group (OSG) (www.scientificdrilling.org).

SeaDataNet

A Pan-European Infrastructure for Ocean and Marine Data Management, undertaken by the National Oceanographic Data Centres (NODCs). Data types comprise in-situ and remote sensing, metadata and data products from the disciplines of physical oceanography, marine chemistry and geology, bathymetry and marine biology.
www.seadatanet.org

SeaDog (Stewardship of Deep Ocean Geophysical)

Database available per request (www.noc.soton.ac.uk/cgi-bin/seadog/seadog.pl) and RODIN (Repository of Oceanographic Data and Information at NOCS – National Oceanographic Center Southampton; www.noc.soton.ac.uk/meta) as its portal to aid the integration of the UK Natural Environment Research Council (NERC) data grid (ndg.badc.rl.ac.uk).

SEDIS (Scientific Earth Drilling Information Service)

A web based information system in development by the Integrated Ocean Drilling Program (IODP), to facilitate access to all data and information related to scientific ocean drilling, regardless of origin or location of data. SEDIS will be designed to integrate distributed scientific drilling data via metadata.
sedis.iodp.org

SESAR (System for Earth Sample Registration)

Registry for Earth samples, generates and administers the International Geo Sample Number IGSN as a global unique identifier for GeoObjects (holes, cores, dredges, rock samples, etc.).
www.geosamples.org

SIOExplorer Digital Library Project

Web-accessible data, documents and images from 822 expeditions by the Scripps Institution of Oceanography (SIO) since 1903, and will become part of the overall NSF-funded National Science Digital Library (NSDL). The effort is a collaboration between Scripps, the San Diego Supercomputer Center (SDSC) and the UCSD Library.
SIOExplorer.ucsd.edu

Site Survey Data Bank (SSDB)

To support the review of IODP ocean drilling proposals, worldwide (e.g., backscatter, documents, electromagnetics, fluid flux, navigation, seafloor imagery, seismic, etc.), at the Geological Data Center, Scripps Inst. Of Oceanography, La Jolla, CA.
ssdb.iodp.org

SNAP (Seismic database Network Access Point), Istituto nazionale di Oceanografia e di Geofisica Sperimentale

A web content manager devoted to seismic data handling and is currently used within several International Initiatives such as the SCAR-SDLS (Seismic Data Library System), the ECORD (European Consortium for Ocean Research Drilling), the EU project CO2GeoNet and the internal OGS seismic data access facility.
snap.ogs.trieste.it

STD-DOI

Homepage of the project "Publication and Citation of Scientific Primary Data."
www.std-doi.de

Stratigraphy Net's (Snet) Connotea library

A core discipline of the geological sciences. Since stratigraphy is an interdisciplinary approach to geology, Stratigraphy.net strives to provide a forum to discuss and share latest developments and ideas in various branches of stratigraphy
www.connotea.org/wiki/Group:Stratigraphy%20Net

TaxonConcept Biostratigraphy Database

As part of Snet, an open web based system to store and retrieve taxonomic data with special emphasis on the creation of online taxonomic dictionaries and fossilum catalogues style.
taxonconcept.stratigraphy.net

UNEP Shelf Programme, a One Stop Data Shop (OSDS)

For use by coastal states preparing submissions for an extended continental shelf beyond 200 nautical miles. The OSDS was developed in response to a United Nations resolution that: "Calls upon the United Nations Environment Programme, working within the Global Resource Information Database (GRID) system for data and information management.
www.continentalshelf.org

World Data Center for Marine Environmental Sciences (WDC-MARE)

Aimed at collecting, scrutinizing, and disseminating data related to Global Change in the fields of environmental oceanography, marine geology, paleoceanography, and marine biology, using the PANGAEA information system. Maintained by AWI and MARUM, Germany.
www.wdc-mare.org

World Data Center (WDC)

Operated under the guidance of the International Council for Science (ICSU), originally planned to manage geophysical and solar data on a global scale, now data sets are broadly environmental (e.g., soils, paleoclimate, land processes) supported by host institutions (51 centers in 12 countries).
www.ngdc.noaa.gov/wdc

Appendix 3. Workshop Agenda

Building a Global Data Network for Studies of Earth Processes at the World's Plate Boundaries

May 9-11, 2007, Kiel Germany

Powerpoint presentations from the meeting are available at the meeting Web site.

www.nsf-margins.org/Datawkshp07

May 9

Science presentations on needs of InterRidge/InterMARGINS scientists for data and visualization tools.

09:00-09:05	Introduction
09:05-09:30	InterRidge – Colin Devey IFM-GEOMAR, Germany
09:30-09:55	Ridge2000 – Donna Blackman SIO, USA
09:55-10:20	MARGINS – Geoff Abers – BU, USA

10:20-10:50 Coffee

Science Use Case Scenarios.

10:50-11:10	Science Use Case Scenarios 1 – Mathilde Cannat IPGP, France
11:10-11:30	Science Use Case Scenarios 2 – Wolfgang Bach, Universität Bremen, Germany

Overview of Existing Data Systems (Project, National, and International) Relevant for InterMARGINS/InterRidge research.

11:30-12:30 Presentations Session I

- NEPTUNECanada – Benoît Pirenne, Data Management and Archiving System (DMAS), Victoria, BC, Canada
- IODP Data Systems – Bernard Miville, IODP-MI Sapporo, Japan
- SeaDataNet – Pan-European marine data system – Dick Schaap MARIS, Netherlands
- FDSN – Tim Ahern, IRIS, USA
- NIED Data Systems – Katsuhiko Shiomi, National Research Institute for Earth Science and Disaster Prevention, Japan

12:30-13:30 Lunch

Overview of Existing Data Systems (Project, National and International) Relevant for InterMARGINS/InterRidge research (continued).

13:30- 14:30 Presentations Session I, continued

- IFREE/JAMSTEC Data Systems – Seiji Tsuboi, JAMSTEC, Japan
- ORFEUS – Thomas Meier, Bochum University
- Marine Seismic Data Center – Tom Shipley, UTIG, USA
- Seismic Database Network Access Point – Paolo Diviacco, Istituto Nazionale di Oceanografia e di Geofisica Sperimentale, Italy
- IGNS Data Systems – Craig Jones, IGNS, New Zealand
- ChEssBase & OBIS – Eva Ramirez, Llodra &Maria Baker, Southhampton, UK

14:30-16:30 Coffee, Demonstrations Session I

16:30-17:30 Presentations Session II

- SeaDOG – Tim LeBas, Southhampton, UK
- Ridge2000/MARGINS DMS – William Ryan, Columbia, USA
- EarthChem & SESAR – Kerstin Lehnert, Columbia, USA
- IFREMER Data Systems – Mickael Trueger & Eric Moussat, IFREMER, France
- SIO Data Systems – Steve Miller, SIO, USA

17:30–19:00 Reception & Posters

May 10

Overview of Existing Data Systems (Project, National, and International) Relevant for InterMARGINS/InterRidge research (continued).

08:30-09:00 Presentations Session II continued

- WDC-MARE, Hans-Jürgen Wallrabe-Adams, MARUM, Germany
- World Data Center – Marine Geophysics – Christopher Fox WDC-NGDC, USA
- UNEP One Stop Data Shop, Yannick Beaudoin, Øystein Halvorsen, & Tina Schoolmeester UNEP-GRID Arendal, Norway

09:00-10:20 Demonstrations Session II & Coffee at 10

Technologies and Approaches for Interoperability

10:20-10:50 Every bit counts - Data management and data publication in the earth sciences, Jens Klump Postdam, Germany

10:50-11:10 Metadata 5: Content Standards for Marine Geoscience (Basics +) – John Graybeal (MMI), MBARI, USA

11:10-11:40	Plaintext to governed vocabularies: restoring order to anarchic metadata – Roy Lowry, BODC, UK
11:40-12:00	Metadata 105: Ontologies for Marine Geoscience (What You Really Want) – John Graybeal (MMI), MBARI, USA
12:00-12:30	Sharing and visualizing earth science data with Web Services and Virtual Globes – Jon Blower NERC, UK
12:30-14:00	Lunch
14:00-16:00	Working Group Session 1
16:00-16:30	Coffee

Interagency/intergovernmental efforts within the geosciences regarding data access and current implementation.

16:30-16:50	An Overview of the Global Earth Observing System of Systems (GEOSS) – Bob Chen, CIESIN
16:50-17:10	CODATA Strategic Plan and Activities, 2007-12 – Bob Chen, CIESIN
17:10-17:30	WDC Overview – Chris Fox, NGDC

May 11

08:30-10:00	Working Group Session 2
10:00-10:30	Coffee
10:30-12:00	Plenary Session Summary of working group discussions and formulation of key recommendations to science programs, to scientists, to data centers.
12:00	Meeting Adjourns
Afternoon	Tentative depending on interest Meeting of Data Managers to develop plan for next steps towards developing data system interoperability.

Appendix 4. Acronyms

API..........................Application Programming Interface
COARDS/CF ....Cooperative Ocean/Atmosphere Research Data Service/Climate & Forest
CSDGM ....Content Standard for Digital Geospatial Metadata
CSML.....................Climate Science Modelling Language
CTD ....Conductivity-Temperature-Depth sensor
DC..........................Dublin Core
DIF ....Directory Interchange Format
DOI ....Digital Object Identifier
EEZ ....Exclusive Economic Zone
eGY.........................Electronic Geophysical Year
EML ....Ecological Metadata Language
FDSN ....International Federation of Digital Seismograph Networks
FGDC ....Federal Geographic Data Committee
GEO ....Group on Earth Observations
GEOSS ....Global Earth Observations System of Systems
GeoTIFF.................Metadata standard which allows georeferencing information to be embedded within a TIFF Tagged Image File Format file
GIS ....Geographic Information System
GRIB ....GRIdded Binary data format commonly used in meteorology to store historical and forecasted weather data
HDF ....Hierarchical Data Format
ICSU ....International Council of Science
Ifremer ....Institut Français de Recherche pour l'Exploitation de la Mer (French Research Institute for Exploitation of the Sea)
IGSN ....International Geo Samples Number
IOC ....Intergovernmental Oceanographic Commission
IODE......................International Oceanographic Data and Information Exchange
IODP......................International Ocean Drilling Program
IPY..........................International Polar Year
IRIS ....Incorporated Research Institutions for Seismology
ISO ....International Organization for Standardization
KML.......................Keyhole Markup Language
MGD77..................Marine Geophysical Data Exchange Format (Bathymetry, Magnetics, and Gravity)
MGDS....................Marine Geoscience Data System
MMI.......................Marine Metadata Interoperability Project
MoMAR ....Monitoring the Mid-Atlantic Ridge; project by the InterRidge programme
NDG ....NERC Data Grid

NERC......................Natural Environment Research Council
NetCDF..................Network Common Data Form; machine-independent, self-describing, binary data format standard for exchanging scientific data
NITFS ....National Imagery Transmission Format Standard
NSF ....US National Science Foundation
OAI..........................Open Archives Initiative
OAI-PMH ....Open Archives Initiative – Protocol for Metadata Harvesting
OBIS........................Ocean Biogeographic Information System
OBS ....Ocean Bottom Seismometer
OECD ....Organization for Economic Co-operation and Development
OGC........................Open Geospatial Consortium
OPeNDAP ....Open-source Project for a Network Data Access Protocol
OPP ....Office of Polar Programs (Directorate of the US National Science Foundation)
PHP ....Hypertext Preprocessor; server side programming language originally designed for producing dynamic Web pages
REST ....Representational State Transfer; style of software architecture for distributed hypermedia systems such as the World Wide Web

SEDIS......................Scientific Earth Drilling Information Service; web-based information service of the International Ocean Drilling program
SEG Y ....Society of Exploration Geophysicists' exchange format for demultiplexed seismic data on 9-track tape
SESAR ....System for Earth Sample Registration
SOAP.......................Simple Object Access Protocol; protocol for exchanging XML-based messages over computer networks, normally using HTTP/HTTPS
STD-DOI ....Digital Object Identifier for Primary Scientific Data
THREDDS.............Thematic Realtime Environmental Distributed Data Services
UDDI ....Universal Description, Discovery and Integration; standard interoperable platform that enables systems and applications to find and use Web services over the Internet
UNCLOS ....United Nations Convention on the Law of the Sea
UNEP......................United Nations Environmental Programme
URL.........................Uniform Resource Locator
URN ....Uniform Resource Name
WDC.......................World Data Center
WMO......................World Meteorological Organization
WMS ....Web Map Service
WSDL ....Web Service Definition Language
WxS ....Web Exchange Service; XML schema language by the World Wide Web Consortium (W3C)
XML ....Extensible Markup Language

Editing and design by
Geosciences Professional Services, Inc.

Funding

The workshop was jointly supported and funded by InterMARGINS, MARGINS, InterRidge, and Ridge2000. The US National Science Foundation and the Cluster of Excellence "The Future Ocean" at the Christian-Albrechts-University in Kiel provided additional financial support. The views herein are those of the contributing authors and do not necessarily reflect the views of the NSF.

For more information, including Data Center descriptions and PowerPoint presentations from the meeting, visit

www.nsf-margins.org/Datawkshp07.

This report is available for download at
www.nsf-margins.org/Datawkshp07.
Printed copies may be requested from
S.M. Carbotte or K. Lehnert,
LDEO at Columbia University,
61 Rte 9W, New York , NY 10964

August 2007