Mediterranean Marine Science
Indexed in WoS (Web of Science, ISI Thomson) and SCOPUS
The journal is available on line at http://www.medit-mar-sc.net

Paper presented at the
"International Conference on Marine Data and Information Systems - IMDIS", Athens, 2008

Preservation and Access across Decades and Disciplines

S. P. MILLER, C. NEISWENDER and D. CLARK

Geological Data Center, Scripps Institution of Oceanography, UCSD, La Jolla, California 92093-0220 USA

Corresponding author: spmiller@ucsd.edu

Abstract

With the current interest in global processes and long-term changes, data in the marine sciences, and many other fields, are being used by a far broader community than ever before. Many of today's advances are driven by observations made across decades, platforms and disciplines. The era of traditional single-investigator research appears to be waning, replaced by multi-institution collaboration. While the Internet has dramatically improved the speed of collaboration, the re-use of data presents major new challenges.

We will review lessons learned from the building of the SIOExplorer Digital Library at the Scripps Institution of Oceanography (SIO), along with a number of related projects. Technical challenges have been largely overcome, thanks to a scalable, federated digital library architecture. Cultural challenges have been more formidable than expected. They became most apparent during attempts to categorize and stage digital data objects across multiple institutions, each with their own naming conventions and practices, generally undocumented, and evolving across decades. Whether the questions concerned data ownership, collection techniques, data diversity or institutional practices, the solution involved a joint discussion with scientists, data managers, technicians and archivists, all working together.

Keywords: Data preservation; Archiving; Marine sciences; Multi-disciplinary studies; Digital library; Metadata; Auto-harvesting; Controlled vocabularies; International data dissemination; Access policy.

Introduction

Long-term preservation of digital data is critical in the sciences and especially so in the ocean sciences where the cost of data acquisition is very high, and re-acquiring datasets is generally not feasible. Instrumentation, media and formats are rapidly evolving, more diverse types of sensors are being used, and the volume of data collected is increasing exponentially. While there have been dramatic advances in storage technology, raw storage capacity alone will not solve the long-term data retrieval and preservation challenges. Years after data were collected, researchers seeking to re-use the data struggle to discover information about the context of original observations. There is often a critical lack of metadata infrastructure, and substantial barriers often exist between individual projects, diverse computer systems, and different institutions.

With growing interest in global processes and long-term changes, data in the marine sciences are being used by a far broader commu-

nity than ever before. Furthermore, with the current rise in fuel costs, there is a greater urgency to make the broadest use of the data from shipboard expeditions, and to re-use existing data to the maximum extent. Many of today's advances need to be driven by observations made across decades, platforms and disciplines. The era of traditional single-investigator research is waning, replaced by multi-institution collaboration. While the Internet has dramatically improved the speed of collaboration, the re-use of data presents major new challenges in information technology and best practices, many of them revolving around the need to capture the complete original context of an expedition.

Metadata was a word rarely even mentioned in marine sciences in the past, but metadata are now critical for recording the provenance of marine data (who, what, where, when, which, why, how), including processing steps and quality control certification. The marine science community is only beginning to come to terms with the difficulties of auto-harvesting the information, selecting standards, exchanging metadata, and synchronizing metadata with evolving standards and needs. There is continuing debate on the relative metadata and data responsibilities of the chief scientist, the ship operating institution, and data repositories, especially in an era of limited budgets. Cyberinfrastructure tools are emerging to help solve the problems, and ideally insert instrumentation metadata at the moment of acquisition. However the technical aspects at times are dwarfed by the cultural obstacles faced by independent institutions, as they attempt to make a community whole that is greater than the sum of the parts.

In what follows in this brief paper, we will: 1) review some of the common issues facing our community, 2) present a case study of the SIOExplorer Digital Library project, 3) discuss examples of related projects, and 4) summarize some of the lessons learned.

Before launching into specifics, it may be helpful to set the stage by considering the four key words in the title "Preservation and Access across Decades and Disciplines," to see just how important they are to a broader audience.

Why Preservation?

Perhaps the most obvious reason preservation is so important is that digital media are so often at-risk. While anyone can pick up a 500-year-old book and read it, the nine-track magnetic tapes that were the backbone of data storage only two decades ago are now considered an "endangered species." Only part of the danger is from the deterioration of the physical media, and much of the challenge is from the rapid evolution of storage hardware and data formats.

Community practices also contribute to risky situations. When the funding for a research project expires, there is often no support left over for systematic archiving of data. Career steps, such as moving from one institution to another, or retirement, also introduce gaps in data stewardship. Perhaps worse, rarely is thought given to describing the data with understandable metadata, so that others can make sense of it in the future. In practice, much of preservation is left as an emergency procedure, rather than an essential and systematic component of the life cycle of data.

Why Access?

Especially in the field sciences, progress is driven by data, with critical discoveries dependent on observations. Until recent years, in many cases oceanographic data was often considered the life-long property of the chief scientist, and very little effort was devoted to enabling the data for re-use by others. However, the field has undergone a paradigm shift. Rapid advances in the understanding of global and regional processes have been made by a broad range of researchers, not just the original data collectors, thanks to data access technology, and also to data sharing policies.

Data acquisition and scientific progress are, of course, fundamentally driven by funding, and access is critical for more than just technical research. Most disciplines need to be concerned with attracting the next generation of

bright young researchers, and with guaranteeing continued public support for their work. Those fields that effectively communicate their discoveries to the public, to funding agencies, and to legislative bodies may improve their chances. It is to be noted that US National Science Foundation (NSF) proposals are now required to address both "technical merit" and also "broader impact."

Why Decades?

There is a very fundamental reason why preservation and access need to function across decades. With time series observations, it is simply impossible to go back and fill in time gaps for data that are lost. In the marine sciences, data collected on cruises over the decades bear directly on our understanding of global climate change, tsunami and earthquake hazards, depletion of fisheries, destruction of coral reef habitats, and many other topics in the current news. Furthermore, even if a repeat expedition to replace lost data were possible, it would be very costly. Ship costs alone are generally greater than $US 25K per day, with most missions requiring 30 days to accomplish.

Why Disciplines?

In many fields, further advances require a systems approach, integrating the observations and expertise of a number of disciplines. The current trend in the Marine Sciences is toward multi-disciplinary investigations. It is not uncommon to find 30 or more berths on an expedition filled with an international team of geologists, biologists, chemists and physical oceanographers from many institutions, not just from the ship operating institution. It is to be noted that two new sections of the American Geophysical Union (AGU) are known for the most rapidly growing membership and number of contributed publications. One is "Biogeosciences." The other is "Earth and Space Science Informatics." The trend toward multidisciplinary investigations highlights the critical need for data and metadata practices that are understandable by a broader audience. Each discipline has its own jargon and set of acronyms, and readily accessible controlled vocabulary dictionaries will be required for the re-use of data.

Materials and Methods

We can illustrate some of the general problems of preservation and access across decades and disciplines with a case study of the SIOExplorer Digital Library Project that was launched in 2001 to make the data available online from more than 1000 SIO cruises, conducted worldwide since the 1950's (http://SIOExplorer.ucsd.edu) (MILLER*, et. al.*, 2001; 2003; 2004).

Origins of the SIOExplorer Project

SIO has played a pioneering role in taking computers to sea, beginning with IBM 1800 processing systems in 1967 (ABBOTT, *et. al.,* 1986; SHOR, 1978; SMITH, *et. al.,* 1988). More than one generation of seagoing computer technicians devoted themselves to the acquisition of data, carefully bringing it back to the home institution. Since 1970 the Geological Data Center (GDC) at SIO had taken on the role of performing quality control of underway-geophysical data and of mailing tapes of selected datasets to the US National Oceanic and Atmospheric Administration's National Geophysical Data Center (NGDC/NOAA) in Boulder Colorado. However, in the year 2000 the situation was at a crossroads, as external funding had lapsed and a generation of experts was heading into retirement. The primary archival storage was in file drawers of Exabyte magnetic tapes, along with boxes of CDs and paper records, augmented by limited storage on Sun workstations. A careful index of holdings was maintained on a local file system, but staff experts were required to perform searches. Homegrown software satisfied immediate needs, but there was no formal metadata, no database technology, and no web presence. It was also perceived that a broader solution was warranted, serving the data needs of a full range of disciplines, not just marine geology and geophysics.

What to do? Perhaps fortunately, this data

archiving situation was not that uncommon. Professional help was available. A team of experts was assembled from SIO, the San Diego Supercomputer Center (SDSC) and the UCSD Libraries to review the basic alternatives. A basic web site would provide access, but does not solve all the problems of metadata and general searching. A database would support searching, but could run into problems with massive amounts of binary data files. Larger disk systems were becoming available to provide storage for data files, but alone they do not address search issues.

Ultimately, the decision was made to build a digital library, as it combined integrated search, database, storage and persistence aspects. Drawing upon recent experience, John Helly of the SDSC designed the SIOExplorer Digital Library system architecture to provide a solution that was easily scalable in size, and extensible in terms of disciplines and metadata schemas (HELLY, 1998; HELLY *et. al.*, 1999; 2002; 2003). Initial funding for a first generation system was obtained in 2001 from the US National Science Foundation (NSF) Information Technology Research (ITR) and National Science Digital Library programs (NSDL) (MILLER, *et. al.*, 2001; 2003). A second-generation system was supported in 2004 by the Digital Archiving and Preservation (DIGARCH) program, jointly funded by NSF Computer Sciences and the Library of Congress (DETRICK, *et. al.*, 2005; MILLER, *et. al.*, 2004; 2006; 2007). The initial award has proven to be a catalyst and has led to a series of 11 subsequent related funded projects.

First Generation SIOExplorer Solution

SIOExplorer was designed to archive "arbitrary digital objects" (ADO's), so that any type of data, image or document file could easily be stored and recovered. Furthermore, a basic archival ASCII "metadata interchange file" (MIF) was created for each ADO. The information architecture of the digital library is specified in a "metadata template file" (MTF) that defines a set of modular metadata blocks used to organize the collection, contain Dublin Core metadata, and to record discipline-specific information as needed. The SDSC Storage Resource Broker middleware (http://www.sdsc.edu/srb/index.php) was selected to manage ADO data files as well as MIF metadata files. The information from each metadata file was loaded into an Oracle database. A graphical Java "CruiseViewer" application was written to provide interactive geospatial and keyword search and download capabilities, using a detailed Global Topography (SMITH & SANDWELL, 1997) map underlay, or alternatively a global map of crustal age (MÜLLER, *et al.,* 1997).

While individual researchers may maintain their own archives, and national and project repositories may store special data sets, what is unique about SIOExplorer is that it preserves the complete context of an expedition, with all the data, images and documents from a wide range of disciplines and sensor systems all gathered in one persistent location, tagged with comprehensive metadata to enable searches and uses beyond the scope of the original scientific party activities (Fig. 1).

The auto-harvesting of data and metadata was a critical requirement, since each cruise produced up to 10,000 data files, and there would never be enough time or funding for a manual approach. Furthermore, over 50 years of expeditions there were major variations in data types, formats and naming conventions. It soon became apparent that a "canonical cruise data structure" (CCDS) approach would be needed, as it would provide about twenty reliable categories into which the various files over 50 years could be sorted (CLARK, *et. al*., 2003; 2003).

Data from 753 cruises were made available online in the first-generation SIOExplorer Digital Library. As the project matured it grew to include five federated collections: Cruises, Photo Archives, EarthRef Seamounts, Marine Geological Samples, and the Educator's Collection. Public access to SIOExplorer is considerable, with 795,351 files (206 GB) downloaded over the last year (Fig. 2).

Findings from the First Generation

A number of data problems were encountered during the construction of SIOExplorer, and several of them are common to many other data projects. Data diversity turned out to be much more of a challenge than data volume. Media failure can be catastrophic, unless there is a backup. Although in the end only very little data was ultimately unrecoverable, problems were encountered with 9-track, Exabyte and DAT tapes, with floppy disks and hard disk drives, and also even with CD's and DVD's. More troublesome than media failure was the problem of sorting out multiple versions of input data stored on different media, at different times, by different people, almost all retired. Intermediate processing results were often left on file systems, mixed with data products, with only semi-consistent naming. Processing documentation was exceedingly rare. Significant amounts of data needed to be re-processed, especially multibeam swath bathymetry sonar data that made up the bulk volume of the collection, requiring sound velocity correction and pitch-, roll- and yaw-bias correction.

As with many other projects that attempt to re-use legacy data, we found that the effort spent on metadata was much greater than the effort spent on the actual data files. For example, out of 100,000 candidate data files for SIOExplorer, there were absolutely no original metadata files. All metadata had to be extracted from data files, or from other sources of information scattered across diverse paper and magnetic media, as well as from anecdotal memory.

Metadata accuracy is another cause for concern. Over the years, and across projects and disciplines, there is an unfortunate tendency for descriptive terminology to wander. Some of the variation is due to evolution in sensor technology, but some may be due to odd abbreviations, typographical errors on rolling decks, institutional practices, or a momentary inspiration to use a new term. As a consequence, we now face challenges in searching digital collections, and in designing re-usable tools that can be applied to multiple institutions. In practice, we found

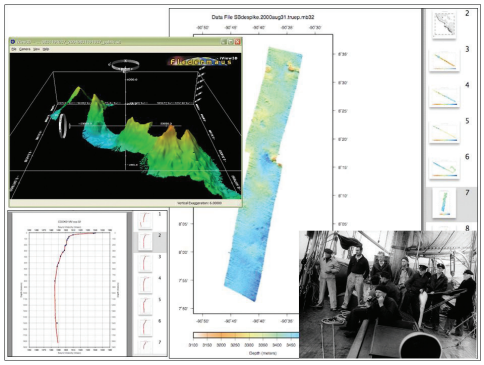

Fig. 1: The mission of SIOExplorer was to capture the complete context of each expedition, preserving an authoritative version of all observations to meet the needs of future researchers. Examples are illustrated above, including seafloor maps and visualization scenes for every multibeam bathymetry sonar file, sound velocity profiles, and historic photographs. Other data include cruise reports, navigation, underway gravity and magnetics, subbottom profiler, current profiler, high resolution meteorology, as well as information on biological, chemical, dredged rock and sediment core samples.

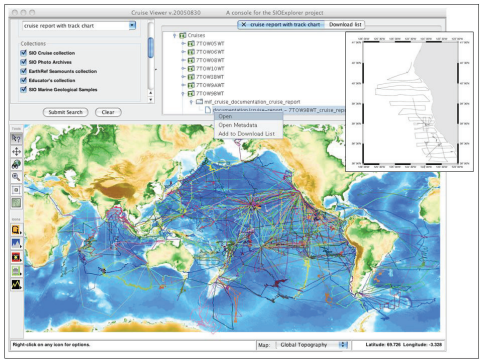

Fig. 2: The SIOExplorer Digital Library provides discovery and download interfaces for a suite of federated collections, illustrated above in a search for cruise reports. Google Earth and MapServer interfaces are becoming available, in addition to Java and webform searches. Data from 1079 cruises are now available.

that up to about 10 percent of the metadata harvested from original human-generated sources was incorrect.

Second Generation SIOExplorer Solution

The second generation SIOExplorer Digital Library has been migrated to higher performance servers, with the open source PostgreSQL database system as part of a "Multi-Institute Testbed for Scalable Digital Archiving" in collaboration with the SDSC and the Woods Hole Oceanographic Institution (WHOI) (DETRICK, *et. al.*, 2005). The project has implemented a prototype digital library of data from both institutions, including cruises, Alvin submersible dives, and ROVs. The SIO cruise collection has now grown to include data from 1079 cruises. In addition, significant progress has been made toward preserving a laboratory full of historic records from 38 years of operation of the DeepTow submergence, including 1463 physical objects from 101 cruises, as well as the recovery of critical data from 9-track magnetic tapes. A new federated "Instrument Collection" is being added to the digital library, including the processing procedures, troubleshooting guides, software, data format descriptions, instrument manuals for all the instruments used on SIO vessels over 50 years.

A more sophisticated staging system was implemented for harvesting metadata and data in two separate institutions, each with their own terminology and best practices. The new system is database-driven, and managed by controlled vocabularies, thus allowing much more flexibility to function in an imperfect world. For example, it is much easier to add new instrumentation. There are new fields for calibration, processing steps, and methods of deployment. It is now easy to create code for a series of search interfaces, each customized for the needs of an individual community, since the php search interface code is generated by an XML template that manages screen real estate, tool tips and query prompts, as well as dropdowns for metadata and controlled vocabularies.

During metadata harvest, the use of controlled vocabularies greatly tightens up the terminology, inhibiting the traditional sprawl of descriptions. Controlled vocabularies are accessed by user interfaces to tighten up queries and guarantee the accuracy and completeness of searches. In other projects, some controlled vocabularies have syntax with two fields ("formal metadata parameter name", "allowed metadata value"). For this multi-institution project a third "brief description of allowed value" field was added, significantly reducing the number of emails and phone calls between institutions. Controlled vocabulary dictionaries for critical metadata parameters are available for fields such as port names, formats, and instrument types, approved for use by both SIO and WHOI (MILLER*, et. al.*, 2006). Whenever possible, these vocabularies are harmonized with existing resources from other authorities and repositories, such as WHOI, the University-National Oceanographic Laboratory System (UNOLS) the Lamont-Doherty Earth Observatory (LDEO), SeaDataNet (http://www.seadatanet.org/) and of course the Marine Metadata Interoperability Project (MMI; http://marinemetadata.org/).

With the new staging database we are able to scan across all the initial entries for a metadata parameter, perhaps across 100,000 metadata records. An analysis of all the unique results from the search easily detects outliers due to typographic errors and blunders. Values were found to be incorrect, missing, misspelled, an unapproved abbreviation or synonym, or misplaced from another parameter. Common examples include the names of chief scientists, port names, operational areas, science themes, image types, sample types, data types, format designations and processing techniques. In addition, the analysis may reveal important values that need to be added to the original controlled vocabulary for that parameter. Finally, the database can be used to correct the detected errors during staging, thus allowing a clean set of metadata to be published in the digital library. Correcting the metadata errors in the many diverse source documents would be cost-prohibitive.

Examples of Related Projects

As with many other digital library projects, SIOExplorer has produced benefits beyond its original scope, acting as a catalyst to a number of additional efforts. The technology has been extended to the Site Survey Data Bank (SSDB; http://ssdb.iodp.org) creating a digital library for data sets that support the global community of the Integrated Ocean Drilling Program (IODP) proposals throughout their lifecycles, from initial ideas through proposal review and on to operations, as well as for educational use (MILLER*, et. al.*, 2006b). In the past, international review panels would meet twice a year and spend much of their time shuffling through analog documents, trying to find the data associated with each proposed site. The SSDB now supports hundreds of proponents worldwide with user interfaces to upload maps, seismic sections and other proposal supporting documents (EAKINS, *et. al*., 2006). With the new technology, literally tons of analog material are now available online digitally, identified by metadata that allow rapid discovery and viewing of essential objects (Fig. 3).

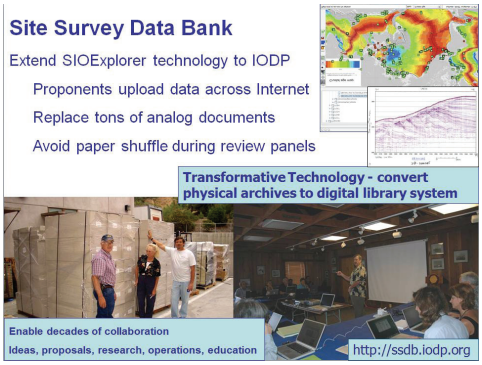

Fig. 3: The Site Survey Data Bank is an extension of the SIOExplorer Digital Library technology, converting rooms full of physical archives to an online scientific decision support system. It enables worldwide proponents to upload data across the Internet, in support of seafloor drilling proposals, and avoids immense paper shuffles during review panel meetings.

Education and outreach projects have made use of SIOExplorer, which provides expert level metadata for every digital object. The Enduring Resources for Earth Science Education (ERESE; http://earthref.org/ERESE/) (SYMONS, *et. al.,* 2005) conducted two workshops for teachers from across the USA, introducing them to the world of research and a suite of resource materials in plate tectonics. The project has been extended to provide three national live web seminars under the auspices of the National Science Teachers Association (NSTA) (SYMONS, *et. al.,* 2007).

The creation of a number of widely used data products has been enabled by SIOExplorer contents, by external projects, with no additional effort from the SIOExplorer project. For example, the second-generation 1-minute global topography model by Sandwell and Smith was driven gridded data from 300 multibeam cruises, automatically downloaded from SIOExplorer. The 1-minute model is available from http://topex.ucsd.edu/marine_topo/. The first generation 2-minute resolution model is one of the most frequently cited works in the Earth Sciences and used in the public media, commonly used for base maps and expedition planning (SMITH & SANDWELL, 1997). Likewise, the underway-geophysical data from SIOExplorer have been automatically extracted by the United Nations Environment Programme (UNEP) Shelf Programme (www.continentalshelf.org), to assist developing states in completing the activities required to establish the outer limits of their continental shelves according to Article 76 of the United Nations Convention on the Law of the Sea (UNCLOS).

One recent example of a basic international goal is the definition of a standard set of metadata to define an oceanographic cruise, and a mechanism for exchanging the information. The US University-National Oceanographic Laboratory System (UNOLS) oversees the vessel and submergence activities of 18 operating institutions, and has recently convened a Data Management Best Practices Committee (http://data.unols.org). One of their first goals is cruise level metadata, and they are working in collaboration with SeaDataNet to establish an inter-

national approach. The efforts are currently be extended to develop a prototype metadata and data dissemination system for all UNOLS vessels, which will provide information to end users, as well as project and national repositories, with a combination of webform data gathering interfaces, interactive search interfaces, and automatic computer-to-computer web services.

The Marine Metadata Interoperability project (MMI; http://marinemetadata.org) provides an opportunity for community building by the creation of a "common body of knowledge" across multiple disciplines and institutions. The project has more than 350 members, worldwide. SIOExplorer metadata profiles, controlled vocabularies and case studies are posted on the MMI site, and considerable effort has been devoted to creating introductory and advanced guides to metadata usage. In addition, the SIOExplorer project has been used as an example in three national workshops on interoperability (http://www.sdsc.edu/~hellyj/papers/Interop_III.pdf).

Results

The Scripps Institution of Oceanography (SIO) which has been operating research vessels, and collecting data for 104 years, supporting a wide range of disciplines: marine geology and geophysics, physical oceanography, geochemistry, biology, seismology, ecology, fisheries, and acoustics (MILLER, *et. al*., 2007). In recent years the efforts have grown to include observatories and regional networks, with more collaborative and multidisciplinary projects, involving multiple institutions and broad international community initiatives. From this broad perspective, it appears that current research faces three types of barriers: technical, social and financial.

Technical Obstacles

For many projects, the preservation of information on obsolete media is a critical but ongoing problem, due to a lack of equipment or funding. While there is widespread awareness of the need to be routinely copying data to newer media, and to create redundant backups, in practice we find that there are generations of media types that have become isolated, with little likelihood of recovery. Since funding is limited, it is recommended that institutions draft a risk vs. value matrix for their digital media inventory to prioritize those objects that are extremely valuable, and also likely to become unreadable in the near future. The common practice of preservation as an emergency activity will lead to data catastrophes.

For a number of projects and institutions, initial concerns regarding large data volume has become less critical, due to advances in storage technology. At the same time there is a growing awareness of the problems associated with data complexity, especially as institutions struggle to integrate decades of legacy holdings with the flood of data from new multi-sensor instruments and multidisciplinary expeditions.

Metadata currently appears to be much more of an obstacle to many projects, compared to data itself. Given the current lack of experience with metadata among practicing scientists, the lack of supporting tools, and the evolving nature of metadata standards and profiles, in the face of newly emerging mandates for metadata, many institutions find themselves spending far more time and money on metadata generation than they had anticipated.

Data ingest appears to be more problematic than final archival storage. While considerable expertise is required to implement a digital library, much greater effort is usually required to stage and harvest metadata and data in preparation for loading into the system. Traditionally, a human expert would make decisions and catalog individual objects, but this approach is not cost effective or scalable with today's large, diverse, distributed and multidisciplinary digital data contributions.

Quality continues to be a troublesome issue, especially in an era of greater re-use of data by persons not involved in the original data collection. Very little progress has been made on systematic approaches to identifying the qual-

ity status of digital objects in large and diverse collections, and even less on auto-detection or auto-repair of problems in metadata or data. One promising solution may be to create an institutional quality XML certificate that travels with a data object and is updated as appropriate throughout its lifecycle.

Social Obstacles

As if the technical obstacles were not challenging enough, the social obstacles are daunting for a community that is only beginning to transition from single-investigator research to collaborative efforts. It remains true that only very few people "love metadata" for itself. There is a general lack of communication between scientists and information practitioners. Although data have been generated for decades, there is often a marked lack of metadata sophistication among both data providers and data users. As they consider adopting new approaches, many institutions and research groups are victims of the "not invented here" syndrome.

Financial Obstacles

Two different business models commonly exist for data access. As mandated by funding authorities, one is driven by an open access policy, and the other follows a pay-as-you-go approach. Although both are well intentioned, in practice each can introduce pitfalls for data preservation and access.

For example in the USA, currently almost all research funded at public expense mandates the release of the data to the public at no cost, or at most the minimal cost of delivery. A proprietary hold period of restricted access is often allowed for the original collector of the data, generally two years, to publish the results and for students to work on a thesis project, as specified for example in the NSF Division of Ocean Sciences Data and Sample Policy (http://www.nsf.gov/pubs/2004/nsf04004/nsf04004.pdf). This open data access policy has led to rapid and broad use of immense amounts of information, and to a number of collaborative and remarkable discoveries. However, the success of this approach depends on adequate and reliable funding being available to support individual or systematic dissemination. Unfortunately, in an era of limited resources, the best of intentions may lead to critical gaps. With rapidly rising fuel costs, there is a conflict of interest between proponents of new field programs and advocates of community archiving of existing data. In the domain sciences, NSF proposal success ratios are now in the 15-30 percent level, and peer review panels often tend to value new programs more highly than archiving. Furthermore, even if funds are awarded for an individual research project, the period of performance is rarely more than 2 years, hardly in keeping with long-term preservation needs. Specific community projects such as the Marine Geoscience Data System (http://www.marine-geo.org/) have achieved success in sustainable archives for selected types of data and geographic areas, but at the moment there is no comprehensive systematic long-term support for shipboard data archiving.

The second approach to sustainability involves charges to end users, which are often mandated by the governmental agencies that funded the original acquisition, especially in Europe. While this approach in principle can offset preservation and access costs, in practice at times it may be a challenge to recover costs with only a modest number of transactions.

Analysis of Strengths and Weaknesses

In the USA, the Council on Library and Information Resources (www.clir.org) is sponsored by 223 major universities with a mission to "expand access to information, however recorded and preserved, as a public good." To further that goal, CLIR convened a national conference on Digital Scholarship and Digital Libraries at Emory University in Atlanta, GA in November 2007 (www.metascholar.org/events/2007/dsdl/) with a combined audience representing three types of communities: faculty scholars, librarians, and information technologists. Over the wide range of disciplines that were represented, from oceanography to astronomy, history and archaeology, it became apparent that each

individual community had it's own strengths, and weaknesses, and that by combining forces there is considerable potential for advancement. In general terms, faculty and technologists are effective at making rapid responses in frontiers of research, and at approaching diverse funding agencies for support, while library institutions tend to lag in those areas. Regarding sustainability, traditions of collaborative effort, and a heritage of preservation, it is the libraries that generally excel, compared to faculty and technologists. Faculty, of course, generally lag the other communities in metadata expertise.

Discussion

Despite the predominance of discussions of problems in this review, on balance it is actually quite encouraging to compare the State-of-the-Art when the SIOExplorer project started with the present condition. In mid-2001 most of the SIO cruise data resided on tapes in boxes, scattered small disk drives on various workstations, or on paper. There were no formal databases in use, virtually no metadata anywhere, and the experts were rapidly retiring. We now have an organized collection of more than 1000 cruises, a second-generation streamlined archiving process, and multiple high performance servers and RAID storage systems, including one in the high-bandwidth SDSC machine room. It has taken a group effort by a dedicated team of technicians, programmers, computer scientists, archivists, librarians, researchers and students, not only at SIO, but also at WHOI and LDEO.

The technical challenges have been largely overcome, thanks to a scalable, federated digital library architecture from the San Diego Supercomputer Center, implemented at SIO, WHOI and other sites. The metadata design is flexible, supporting modular blocks of metadata tailored to the needs of instruments, samples, documents, derived products, cruises or dives, as appropriate. Domain- and institution-specific issues are addressed during initial staging. Data files are categorized and metadata harvested with automated procedures. In the second-generation version of the project, much greater use is made of controlled metadata vocabularies. Database and XML-based procedures deal with the diversity of raw metadata values, detect and repair errors, and map the information to agreed-upon standard values, in collaboration with the MMI community. Metadata may be mapped to required external standards and formats, as needed. All objects are tagged with an expert level, thus serving an educational audience, as well as research users. After staging, publication into the digital library is completely automated.

The cultural challenges have been more formidable than expected. They became most apparent during attempts to categorize and stage digital data objects across multiple institutions, each with their own naming conventions and practices, generally undocumented, and evolving across decades. Whether the questions concerned data ownership, collection techniques, data diversity or institutional practices, the solution involved a joint discussion with scientists, data managers, technicians and archivists, working together. Because metadata discussions go on endlessly, significant benefit comes from dictionaries with definitions of all community-authorized metadata values.

Acknowledgements

The authors wish to gratefully acknowledge the efforts of many devoted ship captains, crews, scientists and technicians and the time spent on rolling decks to help make decades of discoveries possible. We are fortunate to benefit from the pioneering and conscientious work of Stu Smith, Uta Peckman and Ginny Wells in more than 30 prior years at the Geological Data Center, and to computer scientist John Helly of the San Diego Supercomputer Center for the visionary design of the information architecture to take the center into the next era. Primary support has been provided by the NSF National Science Digital Library program (DUE 0121684) and the Digital Archiving and Preservation program of the Library of Congress and NSF (IIS 0455998), as well as the Scripps Institution of Oceanography.

References

ABBOTT, J. L., SMITH, S. M., CHARTERS, J. S., DOWNES, P. G., HYLAS, T., MOE, R. L., MOORE, J. M. & STUBER, D. V., 1986. Scripps Seagoing Computer Centers: Real-time Data Acquisition and Processing, IEEE Proceedings 4th Working Symposium on Oceanographic Data Systems, pp 123-129, San Diego, CA, Feb. 1986.

ARKO, R., MELKONIAN, A., CARBOTTE, S., LEHNERT, K. & VINAYAGAMOORTHY, S. 2007. Web Services for Geoscience Data: Experiences and Lessons, Proc. of the Geoinformatics 2007 Conference, San Diego, California, May 17-18, 2007, Abstract 122326.

CLARK, D., MILLER, S.P., PECKMAN, U., CHASE, A. & HELLY, J., 2002. SIOExplorer: Managing data flow into a digital library, *Eos Trans*. AGU, 83 (47), Fall Meet. Suppl., Abstract OS62B-0255.

CLARK, D., MILLER, S. P., PECKMAN, U., SMITH, J., AERNI, S., HELLY, J., SUTTON, D. & CHASE, A., 2003. Streamlining Metadata and Data Management for Evolving Digital Libraries, *Eos Trans*. AGU, 84 (46), Fall Meet. Suppl., Abstract U41B-0018.

DETRICK, R.S., CLARK, D., GAYLORD, A., GOLDSMITH, R., HELLY, J., LEMMOND, P., LERNER, S., MAFFEI, A., MILLER, S. P., NORTON, C. & WALDEN, B.,2005. WHOI and SIO (I): Next Steps toward Multi-Institution Archiving of Shipboard and Deep Submergence Vehicle Data, EOS Trans. AGU, 86(52), Fall Meeting Suppl., Abstract IN44A-07.

EAKINS, B., MILLER, S. P., HELLY, J. & ZELT, B., 2006. The Fully Electronic IODP Site Survey Data Bank, Scientific Drilling, vol 1, pp 40-43.

HELLY, J., 1998. "New concepts of publication". Nature 393: 107.

HELLY, J., ELVINS, T. T., SUTTON, D. & MARTINEZ, D., 1999. A Method for Interoperable Digital Libraries and Data Repositories, Future Generation Computer Systems, Elsevier, 16, pp. 21-28.

HELLY, J., ELVINS, T. T., SUTTON, D., MARTINEZ, D., MILLER, S., PICKETT, S. & ELLISON, A. M., 2002. "Controlled Publication of Digital Scientific Data". CACM May (5): 97-101.

HELLY, J., STAUDIGEL, H. & KOPPERS, A., 2003. "Scalable Models of Data Sharing in the Earth Sciences". Geochemistry, Geophysics, Geoscience. DOI number 10.1029/2002GC000314.

MILLER, S.P., STAUDIGEL, H., KOPPERS, A.A.P., JOHNSON, C., CANDE, S., SANDWELL, D., PECKMAN, U., BECKER, J.J., HELLY, J., ZASLAVSKY, I., SCHOTTLAENDER, B.E., STARR, S. & MONTOYA, G., 2001. Building a digital library for multibeam data, images and documents. EOS 82: F591.

MILLER, S. P., STAUDIGEL, H., JOHNSON, C., MCSHERRY, K., CLARK, D., PECKMAN, U., HELLY, J., SUTTON, D., CHASE, A., SCHOTTLAENDER, B., DAY, D. & HELLY, M., 2003. Launching Discovery through a Digital Library Portal: SIOExplorer, *Eos Trans*. AGU, 84 (46), Fall Meet. Suppl., Abstract ED51B-1195.

MILLER, S. P., CLARK, D., HELLY, J., SUTTON, D. & HOUGHTON, T., 2004. SIO-Explorer: Advances Across Disciplinary and Institutional Boundaries, *Eos Trans. AGU, 85*(47), Fall Meet. Suppl., Abstract SF42A-08.

MILLER, S. P., CLARK, D. & NEISWENDER, C., 2006. The Role of Controlled Vocabularies in Digital Archiving, EOS Trans. AGU, 87(52), Fall Meeting Suppl., Abstract IN51A-0807.

MILLER, S. P., SYMONS, C. M. & HELLY, J. 2006. The IODP Site Survey Data Bank: Lifecycle Cyberinfrastructure for drilling projects, in "Recent Advances and the Road Ahead," SEG Annual Conference Technical Program Expanded Abstracts vol. 25, p 3528-3530.

MILLER, S.P., CLARK, D., NEISWENDER, C., RAYMOND, L., RIOUX, M., NORTON,

C., DETRICK, R., HELLY, J., SUTTON, D. & WEATHERFORD, J., 2007. Lessons Learned from 104 Years of Mobile Observatories, EOS Trans. ATU, 88 (52), Fall Meet. Suppl., Abstract IN13B-1211.

MÜLLER, R. D., ROEST, W. R., ROYER, J.-Y., GAHAGAN, L. M. & SCLATER, J. G.,1997. Digital isochrons of the world's ocean floor, *J. Geophys. Res.*, *102*(B2), 3211–3214.

SHOR, E. N., 1978. *Scripps Institution of Oceanography: Probing the Oceans 1936 to 1976.* San Diego, Calif: Tofua Press, 1978. http://ark.cdlib.org/ark:/13030/kt109nc2cj/

SMITH, S.M., CHARTERS, J. S. & MOORE, J. M., 1988. "Processing and management of underway marine geophysical data at Scripps," *OCEANS '88. "A Partnership of Marine Interest". Proceedings*, vol., no., pp.385-390 vol.2, 31 Oct - 2 Nov 1988.

SMITH, W. H. F & SANDWELL, D. T., 1997. Global Sea Floor Topography from Satellite Altimetry and Ship Depth Soundings, Science, 277, 1956-62.

SYMONS, C.M., HELLY, M., STAUDIGEL, H., KOPPERS, A., REINING, J., HELLY, J. & MILLER, S.P., 2005. The ERESE Workshop: a Unique Opportunity for Collaboration Between Classroom Teacher and Research Scientist, Eos Trans. AGU, 86(52), Fall Meet. Suppl., Abstract ED23A-1251.

SYMONS, C. M., KOPPERS, A., HELLY, M., STAUDIGEL, H. & MILLER, S. P., 2007. ERESE: An online forum for research-based Earth Science inquiry, Eos Trans. AGU, 88(52), Fall Meet. Suppl., Abstract ED42A-04.