ELSEVIER

Deep-Sea Research II 53 (2006) 793–802

DEEP-SEA RESEARCH
PART II

www.elsevier.com/locate/dsr2

The US JGOFS data management experience

David M. Glover[a,*], Cynthia L. Chandler[a], Scott C. Doney[a], Ken O. Buesseler[a], George Heimerdinger[a], J.K.B. Bishop[b], Glenn R. Flierl[c]

[a]*The US JGOFS Planning Office and Data Management Office, Woods Hole Oceanographic Institution, MS 43, Woods Hole, MA 02543, USA*
[b]*Earth Science Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, M/S 90-1116, Berkeley, CA 94720, USA*
[c]*MIT, Department of Earth Atmosphere and Planetary Science, Bldg. 54-1426, Cambridge, MA 02139, USA*

Accepted 28 January 2006
Available online 24 May 2006

Abstract

The US Joint Global Ocean Flux Study (JGOFS) database management system is an online, oceanographic database assembled from the research activities of the US JGOFS field and Synthesis and Modeling Program (SMP). It is based on modern, object-oriented programming, with a web-based user interface (http://usjgofs.whoi.edu/jg/dir/jgofs) that gives all users, regardless of the computer platform being used, equal access to the data and metadata. It is populated with an extensive set of biogeochemical data from the US JGOFS community along with the attendant metadata. This article summarizes the lessons learned that may serve as a primer for future oceanographic and earth science research programs. Good data management requires devoted resources, about 5–10% of the total cost of the program. A data management office should be established at the initiation of the program, conventions for standard methods, names, and units need to be established before the field program begins, and an agreed-to list of metadata must be collected systematically along with the data. Open and accessible data management depends upon investigators agreeing to share their data with each other, leading to more rapid scientific discovery. Data management should support data distribution and archival; interactions between the data managers and the principal investigators make the database a living database. Innovative use of commercial products in information technology can save time and money in scientific database management. Technology allows access to the database to be transparent in location and intuitive in use. Finally, the most important investments in data management are the people hired.

Keywords: Data collection; Oceanographic data; Data reports; Primer; Metadata; US JGOFS

1. Introduction

1.1. The JGOFS program

The Joint Global Ocean Flux Study (JGOFS), established under the auspices of the Scientific Committee for Ocean Research (SCOR) and the International Geosphere-Biosphere Programme

[*] Corresponding author. Tel.: +1 508 289 2656; fax: +1 508 457 2193.
E-mail address: dglover@whoi.edu (D.M. Glover).

0967-0645/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dsr2.2006.01.018

(IGBP), was a long-term (1989–2005), internationally coordinated program. The main goal of JGOFS was, "to determine and understand on a global scale the processes controlling the time-varying fluxes of carbon and associated biogenic elements in the ocean, and to evaluate the related exchanges with the atmosphere, sea floor and continental boundaries" (SCOR, 1987). This report goes on to note, "A long-term goal of JGOFS will be to establish strategies for observing, on long time scales, changes in ocean biogeochemical cycles in relation to climate change." Approximately, 250 principal investigators (PIs) participated from US institutions along with collaborators from 22 countries, and these JGOFS investigators generated unique and extensive data in amounts unprecedented in the marine biological and chemical communities. Rapid and effortless exchange of these data was important to the success of JGOFS.

The US JGOFS Data Management Office (DMO) employs a distributed, object-based data system (Appendix A), first created by Glenn Flierl, Jim Bishop, David Glover, and Satish Paranjpe under support from the National Science Foundation (NSF). The system permits PIs to swap data, both preliminary and final, and to analyze and synthesize their findings within a larger context by enabling them to retrieve desired information from an online data center. At first an interim data management team, consisting of the PIs of the development team and George Heimerdinger and Ray Slagel (NOAA NODC liaison officers), handled the US JGOFS data management needs with help from the US JGOFS Planning Office. However, by 1994 it was clear a more formal arrangement was needed, and the US JGOFS DMO was established within the Planning Office with Christine Hammond as manager. At the same time (1994) an ongoing partnership with the Global Ocean Ecosystems Dynamics (GLOBEC) project was established to combine programming expertise and share software.

1.2. Scope of data sets

The US JGOFS DMO was responsible for process study data (Ducklow and Harris, 1993; Murray et al., 1995; Smith et al., 1998; Smith et al., 2000) and Synthesis and Model Project (SMP) data (Doney et al., 2002). As an example of the scope of JGOFS data, the process studies produced heterogeneous data of: continuous CTD profiles; discrete bottle samples of nutrients, biological variables, inorganic compounds; moorings; sediment traps; satellite and airborne remote sensing data; underway data; and in situ and on-deck incubations. The sources of these data were from diverse sampling platforms with different time and space sampling regimens, had evolving and changing non-standard, investigator specific chemical and biological measurements, and were contributed by many participating PIs, post-doctoral investigators, graduate students, and technicians. The four major US JGOFS Process studies were:

- North Atlantic Bloom Experiment (NABE)—April/May 1989 (Ducklow and Harris, 1993).
- Equatorial Pacific Process Study (EqPac)—February/March and August/September 1992 (Murray et al., 1995).
- Arabian Sea Process Study—October 1994–January 1996 (Smith et al., 1998).
- Antarctic Environment and Southern Ocean Process Study (AESOPS)—August 1996–April 1998 (Smith et al., 2000).

In the end, the process studies generated about 600 MB of data, which was foreseen in 1988 and was considered daunting. Additionally, the SMP (Doney et al., 2002) produced models and unique synthesis data products. During this explosion of scientific creativity, the DMO was collecting the data, performing quality assurance in close coordination with the scientists, and making the data available, at first to collaborators and later to the global community, via the World Wide Web.

2. Problem definition

2.1. Why did we need it in 1988?

Early in the US JGOFS program it was recognized that desktop, computer workstations had dramatically altered the gathering and analysis of oceanic data. While the convenience and ease of use of these machines made them ideal for individuals working on their data, the process of exchanging data or collecting relevant information from archived data sets was difficult and time-consuming. There were relatively few and hard to find chemical and biological data sets available in the major archives. Different groups used different formats with different procedures for storing and manipulating data, multiple versions of key data sets existed within the community, no mechanism was in place for the original researcher to provide updates, and data often had to be ordered in batch and arrived on nine-track magnetic tape. Our system was dedicated to overcoming these difficulties and making it possible for the user of a small computer connected to the network to be able to locate and work with data stored anywhere in a distributed database without regard to its location or format on a time scale compatible with scientific discovery.

2.2. *Requirements*

Although commercial database vendors also were moving towards distributed systems (Codd, 1990), there are several ways in which a scientific database for a program such as US JGOFS differs from commercial products. First, as mentioned, the heterogeneous software and hardware environments of both data providers and users must be reconciled. Second, the update cycle is quite different; data should be altered only by the PI responsible, not by any of the other users. Third, a long-lived scientific data set requires additional documentation (metadata) that must remain linked with the data itself. These requirements were brought about by an iterative learning process between the DMO staff and the program PIs, facilitated through face-to-face meetings.

2.3. *Evolution with time*

Initially the US JGOFS DBMS was designed to exchange data via NASA's Space Physics Analysis Network (SPAN) between Digital Equipment Corp. (DEC) VAXs. The approach was object-oriented but pre-World Wide Web (WWW). As the system grew in complexity and function, the developers strove to expand the variety of platforms on which the system operated. Early on the UNIX operating system was adopted as an operating system of preference and development with the VAX operating system (VMS) was abandoned. But this still left a large number of JGOFS scientists using systems that were not supported, and the prospect of porting the code from UNIX to DOS and MacOS was intimidating. In 1989 and 1990 the WWW was quickly adopted by the computer networks. The universal acceptance of the WWW was quickly recognized as a solution to the multiple-platform-problem. The important lesson here is that hard creative work is always necessary, but often rapidly evolving developments in the marketplace can be made to serve scientific purposes.

3. Solutions

3.1. *PI participation*

In the end, only the principal investigator knows all of the details that pertain to the data he or she collected. Unrestricted access to this kind of information about the metadata including location, time, and methodology (e.g., filter pore size) is not always possible for the data specialist when compiling a large, heterogeneous database such as the US JGOFS Process Studies. Therefore clear and unfettered lines of communication from the data center to the individual PIs must be maintained for quite some time after the data are collected and submitted. This allows the data specialist in charge of data quality control to check with the PI about any abnormalities found in the data, collect any missing metadata, and allow PIs to update their data as needed in the course of data post-processing. An enormous amount of checking and rechecking, compiling and comparing, telephoning and e-mailing went on behind the scenes at the US JGOFS DMO between the staff and the PIs responsible for the data in order to produce the final product (data and metadata) and interface that the public sees. Therefore, it is one of our recommendations that similar, future programs empanel a collection of participating scientists whose task is to not only coordinate parameter names, units, and methodologies both before, during and after the field expeditions, but to also facilitate the exchange of these metadata between the scientists in the field and the data managers at the DMO.

The US JGOFS DMO exploited the trade-off between centralized data at a data center and distributed data under PI-control (or, equivalently, centralized data, but still with PI access to make modifications). In the former, the data are submitted to a centralized facility and all further changes in either content or format are made by data center personnel. In the latter, the PI always has control over the content and format of the data because they are stored on the PI's own computer. For the Process Studies the US JGOFS DMO adopted an intermediate case where the data are centralized, but the PI has access to the data, and the PI retains content and format control as though the data were stored on his or her own computer.

This is done as a convenience for PIs who may not have the hardware necessary to serve their data in a peer-to-peer sense between computers on the network. For SMP there is a critical mass in data volume, beyond which the PIs had the hardware, infrastructure, and desire to serve their data, in a distributed sense, on their own computers. In both cases, the data are apparent from the US JGOFS web site.

3.2. Merged products

The US JGOFS process study data, as initially collected and stored, are a large collection of data organized by cruise designation and individual PI. To make these data more useful to scientists seeking to synthesize these data, the individual cruises had to be put together in a common four-dimensional (4-D) space/time framework. Using the extensible nature of the underlying US JGOFS DBMS and the master parameter dictionary (see below) feature of the DBMS, we created "merged data products". Merged data products are hundreds of individual process study data objects (Appendix A) merged, with attending metadata data objects, into a small number (3 or 4) of common data objects. The merged data products consist of the alignment of samples along common axes (location, depth, time, etc.) only for those samples taken from a common sampling device (e.g., Niskin bottles of a CTD rosette). The result may be searched and sub-selected with the DBMS. Merged data objects can be recreated or updated by the DMO staff as new data become available, allowing the online versions to be kept up-to-date in a user transparent fashion.

Key to the production of merged data products is the master parameter dictionary. It acts like a thesaurus linking the many different names given to individual parameters to a single, consistent list of preferred parameter names (e.g. nitrate, no3, etc. are known consistently across the heterogeneous database as NO3). Additionally, the building of merged data objects is dependent upon the verification of the attendant metadata (parameter measurement techniques, standards, etc. must be consistent). Once the ingested data are quality-controlled and the metadata verified, the merged data objects are constructed from several related data objects. In this fashion, the master parameter dictionary acts as a filter to ensure overall consistency within merged data products. Future programs should consider the creation of a master parameter dictionary at the start of their program; it will help improve program oversight and facilitate the merging of data from separate program elements.

3.3. Web/LAS

During 1999, a Live Access Server (LAS) user interface was added to the US JGOFS data management system to access the gridded SMP products. The LAS interface (http://ferret.wrc.noaa.gov/Ferret/LAS) is developed by the University of Washington's Joint Institute for the Study of the Atmosphere and Ocean and NOAA's Pacific Marine Environmental Laboratory (UW/JISAO/PMEL). Initially, LAS was a graphical user interface for data stored in NetCDF format, a common format for large, gridded, model output (Rew and Davis, 1990). As the US JGOFS merged products grew in complexity and size, some additional way to examine the combined data sets in a more holistic fashion was needed. However, the merged data products are largely profile-oriented and modifications to LAS were necessary. In a joint effort between the US JGOFS DMO and UW/JISAO/PMEL, changes were made to LAS to enhance its compatibility with other data formats. The DMO is currently running a version of LAS serving both US JGOFS SMP data products and process study field data (http://usjgofs.whoi.edu/las/servlets/dataset). Since the metadata are treated like any other data in the US JGOFS DBMS, metadata can be used to search and sub-select the database with this new interface. These improvements to the LAS interface, along with the improvements made to the underlying US JGOFS DBMS (better space-time searches, additional property value searches, and expanded use of metadata for grouping and classifying data), have resulted in a tool that allows investigators to search, display, analyze field data and model results with a common interface.

3.4. CD-ROM as a data report

A key aspect of data management is to ensure the archival and availability of the field data beyond the life of the program. For US JGOFS, the data and metadata from the four process studies, along with the merged products created from the individual cruise data, have been placed on a CD-ROM and published as an electronic data report (United States JGOFS Process Study Data, 1989-1998, 2003). Synthesized data and model output from the SMP also have been placed on a CD-ROM (United States JGOFS Synthesis and Modeling Project Results, 2004) and a soon-to-be-released DVD-ROM (check http://usjgofs.whoi.edu/whatsnew.html for release date). The obvious advantage to this media is that the data are now searchable electronically.

4. Lessons

A number of lessons have emerged from our decade-long experience managing the US JGOFS data.

4.1. A shift in culture

In any large research program, there is one thing that is often overlooked, the *culture* of the scientists participating in the program. *Culture* goes beyond science and data management; it also includes policies, in particular, those surrounding the issues of data submission and data sharing. Data sharing within US JGOFS begins with the PI's colleagues from the same project. As the database managed by the DMO expands, colleagues have access to the data to compare and contrast with their data. As the data age, access restrictions can be relaxed according to a data policy agreed to at the beginning of the program. In this fashion, timely access to the data (for both close and more distant collaborators) is guaranteed. Moreover, there is nothing like having someone else using your data to find errors overlooked at first check. The final product of this valuable interaction is a high-quality, complete data set with essential metadata ready for final submission to the national data archive.

Data submission is an issue of accountability, but also one of trust and new a culture of science. How does a program ensure the data collected by the individual PIs are actually submitted to the data system? At the start of the US JGOFS program there was a separation between PIs making measurements and data managers storing the results. By the end of the program scientists and data managers were working together in a mutually beneficial exchange of information and knowledge. Simply requiring, as a condition of receiving future funding, that the PI's data must be submitted (the "stick") to the congressionally mandated national data archive for that discipline is not enough to ensure this happens. A method must be found to guarantee both timely access to the data and long-term security of the data; this is where a DMO is invaluable. Through interactions with the DMO a type of trust develops; the DMO trusts the PI to submit their data and the PI trusts the DMO not to corrupt the data and to keep them private according to the guidelines established in the project data policy. Prompts from the DMO help busy scientists to contribute their data in a timely fashion to the growing collection of data. Additionally, PIs are able to receive the help they need to prepare their data for incorporation into a larger database without the financial burden of maintaining a data manager of their own. We have found that this "carrot" always works better than the "stick" referred to above.

4.2. Data management is more than a database

A Data Base Management System (DBMS) is a computer program designed for the projection and selection, sorting and extraction of data (Codd, 1990). Data management is done by a group of people whose primary task is to make available the highest quality data for users wishing access as soon as possible. The merger of the two goes beyond a top-down dictate that all scientists involved in the research program use a common data format to facilitate data exchange. In a recent National Research Council report (National Research Council (NRC), 1998), the NRC's Committee on Geophysical and Environmental Data found that data exchange is facilitated by having the data office staffed by people with a scientific understanding of the data they are managing. We have found the same thing to be true; having practicing scientists working with the data at the DMO has led to scientifically sound decisions of data treatment and has decreased the number of inquiries from the DMO to the PIs. Therefore, we strongly recommend the personnel composition of future program DMOs include working scientists in addition to data managers.

4.3. Public accessibility

The usefulness of a database is a direct function of data quality, access to, ease of use, and user confidence in the data. It has been our experience that one of the most effective ways to improve and guarantee the usefulness of a database is to have people using the data as soon as possible. Public accessibility is a cornerstone of good data management. We illustrate this issue through the US JGOFS example—we used staged, but short, periods for all aspects of data submission:

- data collection/submission;
- proprietary access for PI;
- password-protected access for program PIs;
- general public access.

The timing and description of the stages will depend on the nature of the particular program. What is important is the existence of a procedure. In our program, the data policy stated that data be submitted to the DMO within one year of collection, 6 months for "core" data as defined by the process study lead PI. These data were then made available to participating program PIs via a password-protected web interface. Scientists working with their own data gained invaluable insight by direct comparison to data collected by other scientists from the same program. After 2 years the data were made available to the public in general: other colleagues not part of the program, public officials seeking policy guidance, and school teachers whose lesson plans were greatly augmented with real scientific field data. All have a part to play in the final vetting of a database and cannot do so if the data are held proprietary with limited access. Flexible policies must account for cases where data measurement methodology does not lend itself to a rapid public accessibility mantra. The measurement of certain radioisotopes is a common example, and an exception to the submission schedule can always be made by a data center with sufficient scientific oversight. Therefore we strongly support staged, but short data submission periods for all large programs.

4.4. Synthesis and synergy

People working on data can become so focused on whether or not this one little bit of information is "correct" that they lose sight of how it fits into the over all picture of the study. Therefore we make the following recommendation: make the data publicly available as soon as possible and fund scientists to synthesize and model these data coincident with a program's field study phase. The coupling between data management and data synthesis is a broad issue; it is about more than improving the quality of the data by finding errors. Other issues include: availability of "undocumented" metadata (while they are still fresh in the minds of the PIs), modification of ongoing fieldwork, ownership of the data by the data originator (PI or technician), etc. As colleagues begin to synthesize disparate data sets into their framework, a unique kind of synergy begins to take place. By the time data become publicly available, this synergy has not only made the individual data sets better, it has strengthened the entire database, has ensured that information about the data (metadata) is accurate and complete, has provided important scientific insight into the processes being studied, and has potentially led to significant modification of data collection strategy.

4.5. People

Operating an effective and efficient DMO requires a careful investment in people. The correct mix of skill sets is required to achieve an operation that is both cost effective and satisfactory for the needs of the users, in this case, scientists. A properly functioning DMO should have an overall supervisor that sets the pace and direction of the work undertaken. For scientific DMOs, it is important that the supervisor have a well-grounded education in the sciences the DMO serves. And although the supervisor may not work full-time at the DMO, they should always be answerable to the users and have the user's best interests in mind as they make decisions regarding prioritization of DMO operations. Working closely with the supervisor is the chief data specialist, someone with a detailed knowledge of the computer software and hardware systems used in the DMO, an understanding of basic data management concepts, and at least cursory knowledge of the component science domains encompassed by the program. The chief data specialist also must have a fierce dedication to produce the most accurate, best possible database the DMO supervisor will allow within resource constraints. The implied tension between these two positions, when managed appropriately, will ultimately contribute to a more successful data management effort. Working along side these two are a group of data specialists who are, by their very nature, detail people. Issues of scale determine how these functions are distributed within a DMO, but it is important to ensure coverage of all necessary expertise and functions in a well-staffed DMO.

Table 1

DMO Timeline

🟦 Information systems technology and systems configurations
🟩 U.S. JGOFS Events
🟨 U.S. JGOFS DMO Activities

1981		"640K ought to be enough for anybody" (erroneously attributed to Bill Gates)
1982		Internet formed from existing TCP/IP networks
1984	September	U.S. NAS GOFS workshop
		DNS has 1000 registered hosts
1985		MS-DOS 3.0, 4.7 MHz Intel 386 processor Dual 5 1/4-inch floppy disk drives
1987		JGOFS launched
		First Sun SPARC system (10 MIPS)
1988		HOT and BATS initiated
		Data Manager: George Heimerdinger (NODC)
		Multitasking means 3 computers and a wheeled chair
		Intel 80286 with a 5 inch floppy disk drive and time on a Sigma-7 with 9-track tape, data transmitted via U.S. Mail
		U.S. JGOFS Planning Report 8 Data Management Working Group Report
1989	April	NABE Atlantis 119 cruise
	June	NABE Endeavor 198 cruise
		Data interface: seven function, X-based GUI for UNIX clients with TCP/IP
		World Wide Web invented at CERN, EPPL
		Intel 386 processor, 16 MHz Intel 80386, 1 MB RAM, 9-track tape drive
1990		CO$_2$ Survey initiated
		Windows 3.0 released
1991	February	NABE data report published
1992	February	Equatorial Pacific TT007 cruise
	November	Equatorial Pacific TT013 cruise
		16 MB Process Study data being served
1994	September	Arabian Sea TT039 cruise
		Netscape 1.0 released
		DMO created at WHOI (Manager Christine Hammond) 104 MB Process Study data being served Data server: Sun SPARC 5, 85 MHz 1 GB disk space, CD-ROM drive
1995	December	Arabian Sea TT054 cruise
		OS/2 system, 90 MHz Pentium 64 MB RAM, 880 KB floppy, 1MB HDD
1996	August	SMP workshop (Durham, NH)
	September	AESOPS NBP 96-4 cruise
		Windows 95 system, 133 MHz 64 MB RAM, 880 KB floppy, 170 MB HDD
		Netscape Navigator 2.0 released (JavaScript supported)
1997		CO$_2$ Survey complete
		176 MB Process Study data being served
1998	March	AESOPS NBP 98-2 cruise
	May	Data server: SGI Origin 200, dual CPU, 128 MB RAM, three 9 GB disk drives
1999		JGOFS Arabian Sea CTD CD-ROM published
		PowerMac G4, 400 MHz processor, 64 MB RAM, CD-ROM, 10 GB HDD
2000	January	U.S. Timekeeper (USNO) reports year as 1910
	September	DMO Director: David M. Glover DMO Manager: Cyndy Chandler
2001	December	514 MB Process Study data being served 80% NABE, 90% EqPac, 95% Arabian Sea, and 80% AESOPS data acquired
2002	July	16. million DNS registered hosts
	December	602 MB Process Study data being served (final data set submitted to DMO) 100% NABE, 99% EqPac, 95% Arabian Sea. And 96% AESOPS data acquired 3.8 GB Process Study and SMP data being served
2003	April	JGOFS DMTT publishes final Data Collection
		U.S. JGOFS CD-ROM Data Report published
	May	JGOFS Open Science Conference (Washington, DC)
		Windows XP, 2.4 GHz Intel Xeon processor, 1 GB RAM, DVD-RW, 40 GB HDD
	July	Final SMP workshop (Woods Hole, MA)

Computer timeline info taken from: White, S. (2001) A Brief History of Computing (http://www.ox.compsoc.net/~swhite/history/timeline.html)

4.6. "This is hard!"

In Table 1 we present a time line of developments in the IT industry and in the US JGOFS DMO (United States JGOFS Planning Report 8, 1988). While this time line is not meant to be definitive for either, it does set the context of the rapidly changing computer world in which the DMO operated. This context makes a very important point: change, often rapid and dramatic, is part and parcel of the infrastructure upon which data management activities depend. In this complex web of changing standards, hardware, software, and infrastructure, corporate memory becomes incredibly important. The investment in people can be as imperative as the investment in creating a well-functioning combination of hardware and software to support the storage, accessibility, and distribution of data. Whether or not the challenge is retaining qualified people on the staff or dealing with Moore's Law (Moore, 1965), the truth of the matter is doing scientific data management is hard work. This underlines the difference between data storage and data management. It may not be flashy; it may not grab headlines; and it certainly will not guarantee tenure (and is, perhaps, antithetical to obtaining tenure). But the rewards of proper data management come from the awareness that the promise of tomorrow's knowledge is safely secured in the data of today. While data management has certainly benefited from recent technological advances, proper stewardship of data by dedicated people is what ensures the sanctity of that promise.

4.7. Costs money

We have observed that the combined cost of a good data management and planning office is approximately 5–10% of the total cost of the program. The co-location of both offices in Woods Hole, MA, led to vastly improved communication and a higher-order synergy than is normally experienced in this sort of endeavor. How much did the US JGOFS Planning Office and DMO cost the US JGOFS program? The US JGOFS process studies and SMP cost approximately 200 million dollars (US). Of that \$200 M, approximately \$12 M was spent on data management and planning office operations, or 6% of the total cost of the US JGOFS program.

5. The future

5.1. Take advantage of lessons learned in US JGOFS

As we look to the future of data collection and management, we see a future where the users will expect more of everything: greater bandwidth, larger data collections, and better access to these larger data collections both in terms of speed and granularity. These are just a few of the trends that are already apparent. For example, in the Ocean Drilling Program all of the "at sea" data go directly into a relational DBMS *while at sea.* These data are available to all the PIs on the ship and are being managed from shore (via a high-bandwidth satellite link). Eventually, all of these data will be available to shore-based scientists in real- or near-real time. Recent plans for developing the Ocean Observatories Initiative (ORION Executive Steering Committee, 2005) embrace the concepts of remote control of, access to, and interoperability of data collection within a network of globally distributed sensors. As these trends continue we see a future data stream of ever growing size and complexity.

One aspect of the US JGOFS DMO that has had little notice is the interrelationship between the DMO and the US JGOFS Planning Office. Coordination between the Planning Office and DMO is essential for the successful execution of their respective responsibilities. A vibrant, well-informed DMO working in close coordination with a planning office achieves more than a static web site or a data dump archive. Collaboration between these two offices is far more cost effective than disparate centers of operation separated by more than an afternoon's drive.

As the US funding agencies look to the future of data collection and management, it would seem prudent to take advantage of the lessons learned from the US JGOFS DMO experience. We have tried to summarize the key findings (lessons) the DMO has made during the last 10-plus years and hope to pass this knowledge on to promote better data management of large, interdisciplinary, field-oriented research programs. Any program that spends large amounts of money to collect information about our planet cannot deem itself successful without proper stewardship of the resultant data.

Acknowledgements

All CD and DVD data-ROMs are available from the US JGOFS Planning Office in Woods Hole, MA and can be ordered from either http://usjgofs.whoi.edu/contacts/RequestForm.html or by writing to: US JGOFS Project Office, Woods Hole Oceanographic Institution, Mail Stop 43, Woods Hole, MA 02543. We thank Mary Zawoyski, David Schneider, Jeff Dusenberry, Steve Hankin, Jon Callahan, and Joe Sirott whose hard work behind the scenes made the US JGOFS DMO a great success. Satish Paranjpe programmed the original version of the US JGOFS DBMS and contributed greatly to the final version. We also thank Neal Anderson, Phil Taylor and Don Rice for having the foresight to fund our activity. This work was carried out under NSF Grants OCE-8814310 and following grants for the Data Management Office and OCE-8820638 and following grants for the Planning Office.

Appendix A. The JGOFS Data System

The US JGOFS data system is a distributed, object-based data system for multidisciplinary, multi-institutional programs. It provides the capability for all the scientists to work with the data without regard for the storage format or for the actual location where the data resides. The approach used yields a powerful and extensible system, in the sense that data manipulation operations are not predefined. The system has proven successful in unifying and providing simple access to distributed, heterogeneous data sets.

A.1. System description

The basic elements of the JGOFS system (Fig. 1) are:

- PIs can keep the data sets on their own machines in their own formats—from ASCII tables, to multiple file databases, to binary floating point, to full database systems—and manipulate the data with their own programs. Quality control and decisions about updating the data remain the responsibility of the PI, working with the data managers.
- The JGOFS data system creates a viewport by which other investigators can access the same data set. This viewport is provided by an executable program (called a "method" in object-based terminology—it also can be thought of as a "translator"), which converts (on-the-fly) the requested data into a uniform structure and format. These programs are

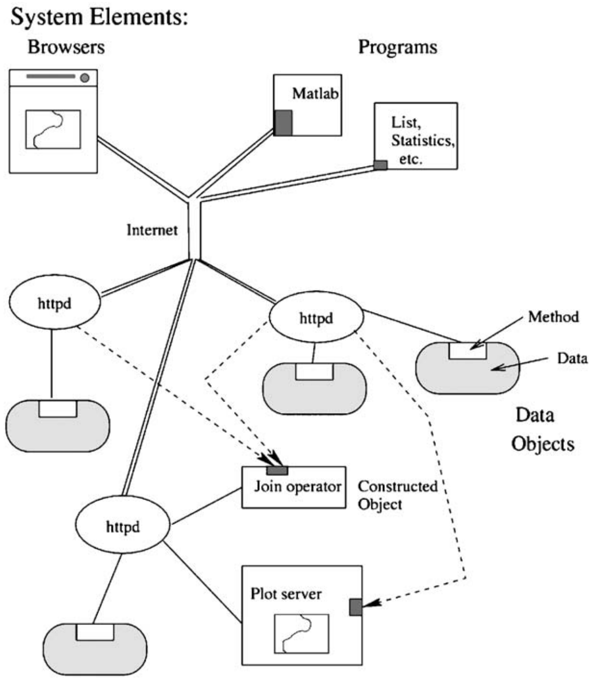

Fig. 1. Sketch of system showing clients at the top, which send requests via HTTP to servers on various machines. The requests are passed to the methods/translators, which gather the information from the data and return it to the client. Constructed objects (e.g., join) have only the method and take their data by calls to one or more other data objects (dashed lines). The Data Management Offices also maintain services such as plotting, which again request their information from the user-selected (and subsetted) object(s).

responsible for:

1. receiving requests for subsetting of the data, both in terms of the desired fields and in terms of ranges or values for particular fields. Thus the request could ask for chlorophyll and depth in the range depth $\geqslant 20$ & depth $\leqslant 50$ & month = jul;
2. gathering the requested information from the data set;
3. translating the information into the appropriate common form used for transferring data (HTML for browsers or flat files or a structured stream for other client programs);
4. sending the information through **httpd** to the process which made the request.

One translator may serve several different data sets—the translators depend on the format chosen by the PI, but generally not on the information itself.

- Data are transmitted with a common protocol, which implies that programs can work with any data in the system. The protocol includes:
  - Comments (text)
  - Variable descriptions
    - Name
    - Attributes (e.g., units)
    - Hierarchical structure
  - Data
    - Strings or numbers

The hierarchical structure allows the data to be organized (for example, by cruise/station number/depth) with each level containing the relevant information. This structure is reflected in the HTML pages, so that the user can readily work down to the particular data desired. It also makes transmission of the data more efficient since common information is not sent repeatedly.

- Servers work with dictionaries that can translate brief names into object references either on the server or on a different machine. Access generally starts at the main data management portal, which maintains a dictionary/directory for all of a program's data.
- "Constructed objects" behave as filters, taking the output from one or more other objects, transforming it, and passing it to the caller. Examples would be data transformations (e.g., rescaling or math operations on multiple columns) and joining data from multiple objects into merged products. Since the functions are defined by independent programs, they can be added to at any time without affecting existing functions. This extensibility allows the data system to do specifically oceanographic functions (e.g., AOU, dynamic calculations).

A.2. Putting data on the system

Data can be added to the system by serving it directly; this requires a WWW server and a set of CGI scripts. The scripts call the methods that take in and subset the data and produce the output for **httpd**. In many cases, the existing methods for common tabular formats or for MATLAB files will be satisfactory. If the data cannot be handled by an existing method, a new one must be constructed, a process involving fairly straightforward programming. Users who do not maintain a server can send data directly to the data management office, which will then serve it directly.

A.3. Getting data off system

There are also many ways to get data out of the system:

- Browsers that allow selection of particular information and return it as **HTML** pages or flat files suitable for import into spreadsheets or analysis programs.
- Plots generated by programs using a simple API to obtain the data over the network (this API is documented and called within other programs as well).
- Production of **Matlab**-readable files.

A.4. Summary

Our data base system thus is distinguished from conventional and available systems by five important features:

1. the ability to handle data in arbitrary formats;
2. data transfer from remote, networked data sets;
3. extendible—data manipulation routines or relational functions can be added at any time;
4. new data can be added to the system in a simple way without a lengthy conversion;
5. this system can be used either interactively or with user-written programs.

We have constructed servers, a number of different translators, and constructed objects. The US JGOFS and GLOBEC experiences confirm the system's ability to handle a wide range of data types and internal formats. For on-going projects, online access to current data sets is essential, and the providers must have considerable freedom in how they build their files and systems. Likewise, the capability of building "extensible" data systems, analysis packages, and graphics packages offers significant improvements in our abilities to share software. The US JGOFS system provides a solid basis for information management in multi-institutional, multi-disciplinary programs. For more information about the US JGOFS DBMS see the online system documentation at http://usjgofs.whoi.edu/datasys/jgdb_docs/jgsys.html.

References

Codd, E.F., 1990. The Relational Model for Database Management, Version 2. Addison Wesley Publishing Co., Reading, MA (538p).

Doney, S.C., Kleypas, J.A., Sarmiento, J.L., Falkowski, P.G., 2002. The US JGOFS Synthesis and Modeling Project—an introduction. Deep-Sea Research II 49 (1–3), 1–20.

Ducklow, H.W., Harris, R.P., 1993. Introduction to the JGOFS North Atlantic bloom experiment. Deep-Sea Research II 40 (1–2), 1–8.

Moore, G.E., 1965. Cramming more components onto integrated circuits. Electronics 38 (8), 114–117.

Murray, J.W., Johnson, E., Garside, C., 1995. A U.S. JGOFS process study in the equatorial Pacific (EqPac): introduction. Deep-Sea Research II 42 (2–3), 275–293.

National Research Council, 1998. Review of NASA's Distributed Active Archive Centers. Committee on Geophysical and Environmental Data. National Academy Press, Washington, DC, 233pp.

ORION Executive Steering Committee, 2005. Ocean Observatories Initiative Science Plan. Washington, DC, 102pp.

Rew, R., Davis, G., 1990. NetCDF: an interface for scientific data access. IEEE Computer Graphics and Applications 10 (4), 76–82.

SCOR, 1987. The Joint Global Ocean Flux Study: Background, Goals, Organization and Next Steps. Report of the International Scientific Planning and Coordination Meeting for Global Ocean Flux Studies, Paris, 2/17—19/87, Available from SCOR Secretariat, Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J1, 42pp. See also The Joint Global Ocean Flux Study: North Atlantic Planning Workshop, Paris, 9/7–11/87.

Smith, S.L., Codispoti, L.A., Morrison, J.M., Barber, R.T., 1998. The 1994–1996 Arabian Sea expedition: an integrated, interdisciplinary investigation of the response of the northwestern Indian Ocean to monsoonal forcing. Deep-Sea Research II 45 (10–11), 1905–1915.

Smith Jr., W.O., Anderson, R.F., Moore, J.K., Codispoti, L.A., Morrison, J.M., 2000. The US Southern Ocean Joint Global Ocean Flux Study: an introduction to AESOPS. Deep-Sea Research II 47 (15–16), 3073–3093.

United States JGOFS Planning Report 8, 1988. Data Management, Report of the U.S. GOFS Working Group on Data Management, 52pp.

United States JGOFS Process Study Data 1989–1998, 2003. CD-ROM vol. 1, version 2. U.S. JGOFS Data Management Office, Woods Hole Oceanographic Institution, USA, April 2003.

United States JGOFS Synthesis and Modeling Project Results, 2004. Part 1: CD-ROM vol. 2, version 1. U.S. JGOFS Data Management Office, Woods Hole Oceanographic Institution, USA, June 2004.