U.S. JGOFS Data Management Lessons Learned
"Data Management in the Wild"

Authors: Cyndy Chandler and David Glover (Woods Hole Oceanographic Institution)

The U.S. JGOFS Data Management Office at Woods Hole Oceanographic Institution was funded by the U.S. National Science Foundation.

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 (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). A long-term goal of JGOFS was 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 Data Management Office (DMO) was formed to serve this need. The metadata (supporting descriptive information about the data) was critical because JGOFS was a trans-disciplinary program and investigators would require access to data and results from other participating PIs. Sampling and analytical methodologies were key components of the metadata as these methods were themselves in flux over the duration of the program.

U.S. JGOFS Research Components

Four Process Studies: (1989 -1998)
- North Atlantic Bloom Experiment (NABE)
- Equatorial Pacific Process Study (EQPAC)
- Arabian Sea Process Study
- Antarctic Environment and Southern Ocean Process Study (AESOPS)

Time-series: (ongoing)
- Hawaii Ocean Time-series (HOT)
- Bermuda Atlantic Time Series (BATS)

Global CO_2 Survey

Satellite Observations of Ocean Color

Synthesis and Modeling (SMP)

Data Management

The U.S. JGOFS Data Management Experience

Lessons Learned

1. scientists will generate data in a format useful for their needs
2. archived oceanographic data sets are best organized in terms of metadata (temporal and geographical)
3. users should be able to obtain all the data they require from one source and in a consistent format
4. data interchange formats should be designed for the convenience of scientific users

The first four 'lessons learned' are confirmations of four of the basic principles of data management listed in a report from the 1988 JGOFS Working Group on Data Management.

5. metadata is critical
6. a published data policy developed through collaboration with the principal investigators helps set expectations
7. researchers will want to download the data to create a local copy
8. data synthesis is also a valuable part of the ongoing data quality assurance process
9. it is important to involve participating investigators in data management efforts
10. there are pros and cons associated with distributed and single-server data systems
11. technology and tools will continue to change
12. data management practices have an enormous effect on the ultimate utility of the data collection

Best Practices Guidelines

1. the process by which results are ingested into a database must be flexible and accommodating
2. collect and report accurate and complete spatial and temporal metadata with the data (discovery metadata)
3. provide access to distributed data resources from a single interface that aggregates distributed resources
4. data delivery system should provide user control of data delivery format (variety of export/download formats) on the fly
5. make metadata mandatory and include methodology and quality assurance protocols
6. develop, adopt and publicize a data policy
7. whatever other functionality a data access system provides, it must also include export/download capability, including user-controllable sub-selection of data
8. provide data to community as quickly as possible, encourage use of the data and provide a feedback mechanism for questions and comments regarding the data collection
9. management of ocean science data should always involve a cooperative and collaborative relationship with scientists, preferably those also responsible for data collection
10. choose the system design that best serves the scientific needs of the PI/project/community you expect to serve
11. the data system should be simple, reliable and easy to maintain; it should be dynamic and modular, permitting modification as new enabling technologies become available
12. keep the ultimate goals of the project in mind when making data management decisions; synthesis products are impossible to generate without sufficient metadata

What next ?

Summary

While the lessons were learned from experience managing data for a large, coordinated program, the best practice guidelines can apply to individual investigators as well as future programs. Good data management practices have always been a part of well designed scientific research projects. The trans-disciplinary nature of current ocean science research projects necessitates an even more rigorous dedication to data quality, metadata reporting and data dissemination.

Data management similar to the effort invested by U.S. JGOFS costs about 5-10% of the overall project budget. This is a sound investment considering that the data generated by a project are an important component of that project's legacy. Our community must adopt a philosophy of data stewardship, collaborating with investigators to ensure comprehensive data management from proposal to preservation. Well documented data has lasting value as it can be used in new ways to generate additional synthesis products, aid scientific discovery and yield new knowledge.

metadata metadata metadata metadata metadata

Recommendations for the Future

The ocean science research community must continue to invest in development of the infrastructure required to improve data management practices in support of scientific excellence. Some of the key areas in which we should focus our efforts include: metadata; identification, development and adoption of standards; effective integration of enabling technologies and development of interoperable data delivery systems.

The process of recording sample collection metadata should be standardized and automated as much as possible. Procedures should be developed wherein data is tagged with the critical metadata during acquisition with continued recording of metadata through subsequent analysis and processing phases.

Data managers should maintain awareness of emerging standards and strive for compliance or work with members of the research community to modify and adapt existing standards to meet their needs. Use of common standards by the community, including metadata and parameter vocabularies, will facilitate automated data aggregation and interpretation in the future.

Data managers will need to successfully manage the tension between existing (familiar and functional) and new technologies (perhaps enabling - perhaps just new) in designing their comprehensive data management schema.

Our ultimate goal should be the incorporation of the previous recommendations in the development of interoperable data delivery systems the design of which is science-driven with a goal of open access to well documented, high quality data.

Much of this information has been published in a recent paper:
Glover, D.M., C. L. Chandler, S. C. Doney, K. O. Buesseler, G. Heimerdinger, J.K.B. Bishop and G. R. Flierl. 2006. The US JGOFS data management experience, *Deep Sea Research II*, **53** (5-7), 793-802.

References Cited:
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.

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

Acknowledgements:
Over the years of service to the U.S. JGOFS project, the DMO also included Jeff Dusenberry, Christine Hammond, George Heimerdinger, Dave Schneider and Raymond Slagle and the project office was staffed by Ken Buesseler and Mary Zawoysky.