Woods Hole Oceanographic Institution
Information Technology Advisory Committee
Data Management Working Group

August 2004

Working Group Members

Michael Caruso, Chair (Department of Physical Oceanography)
Cynthia Chandler (Department of Marine Chemistry and Geochemistry)
Lori Dolby (Communications)
Arthur Gaylord (Computer and Information Services)
Melissa Lamont (Communications)
Andrew Maffei (Computer and Information Services)
Ralph Stephen (Department of Geology and Geophysics)

Objective

Our goals were to assess the need for data management at WHOI, review current capabilities and provide an implementation plan for fulfilling the data management needs at WHOI. In this document, the term “Data Management” is the storage of data and information in a manner for efficient and meaningful retrieval. The term “Data” is a generic term that refers to any digital object.

Summary of Recommendations

Our principal recommendation is that WHOI designate an interdepartmental data archivist who reports to the Director of Research. The data archivist should develop an initial sustainable archiving project related to ocean observing systems and create a plan for preserving new data and adding historical data. In essence, it is imperative that WHOI provide archive capabilities for new data sets immediately and then extend the archive with historical data.

Motivation:

Data and metadata have become an integral part of scientific research. Advances in technology are providing researchers with unprecedented amounts of data from in situ measurements, models and satellites. To effectively analyze, visualize and share information, data needs to be accessible and contextual. The management of this data for analysis, information sharing and funding agency requirements is time consuming for individual researchers. Common data management techniques should be used to reduce the individual efforts that are currently implemented.

Introduction

Research at WHOI is dependent upon data. Researchers gather in situ data, generate data with numerical simulations and acquire data through collaborative research. Anyone who works with data has data management requirements. For most researchers, this creates two main challenges. Data generated by WHOI research needs to be accessible to external researchers and externally generated data needs to be accessible internally. For some researchers, this may be as simple as e-mailing a single data file. For other users, it may require locating, retrieving and reformatting data from multiple archives. Given the spectrum of data management needs at WHOI, is it feasible to provide institution resources to manage internal data? Should the scope of the Institution's management involvement simply be to assist researchers migrating data to national archives? Or should the Institution create a facility to archive and distribute data?

Additionally, effective data management will facilitate participation in multiple institution programs by giving researchers the tools to share data reliably. Although it is possible to defer data management to external sources, WHOI involvement is necessary to ensure the data context is maintained. The future value of WHOI data is reliant on maintaining the context of the data **– Who** generated the data? **What** is the data?; **Where** was the data generated? **Why** was the data generated? And **How** was it created?

The diversity of research conducted at WHOI makes a unified data management scheme impractical. However, there are clear overlaps in research that make a common management capability desirable. This document is intended to compliment the report being prepared by the Access to the Sea taskforce. The Data Management and Visualization working group have indicated to us that their recommendations will focus on requirements in the 5 – 15 year range. The recommendations of this working group should be implemented within 5 years.

During the preparation of this document, it became clear that the ad hoc Scientific Data Advisory Committee's (SDAC) report from 1999 was very thorough and is still relevant. Therefore, we must reiterate much of the SDAC report while we provide updated recommendations. It is also beyond the scope of this report to provide an extensive assessment of oceanographic data management. The intent of this report is to reprise the need for WHOI data management and provide guidance for implementing a data management strategy over the next 5 years.

Summary of SDAC Report

The SDAC report is an excellent assessment of the data archive needs at WHOI. Most of the report is still relevant today. It describes general considerations for an archive and defines a specific application. Parts of the executive report are repeated here with our comments in italics.

- WHOI has a fundamental responsibility to collect, archive, manage, and distribute important scientific data. *This is still fundamental*.
- WHOI should institute a proactive policy for archiving scientific data that are acquired by WHOI scientists and by WHOI ships and deep submergence vehicles. *A proactive policy will ensure that WHOI remains competitive with other institutions.*
- To the extent reasonable and possible, scientific data should be archived with national data centers and other established archives. Other scientific data, including those data desirable for local access, should be archived at WHOI, following a priority list based on

the scientific value of the data. NDSF data are at the head of this list. *The location of the archives is not as important as the ability to access the data. It does not matter whether the data resides at WHOI or a remote national data center, we need to have the ability to discover and utilized the information.*

- Implementation of this policy will require personnel, internal adjustments, and physical resources as follows:
    - A Scientific Archivist to track, acquire, and help manage scientific data for both the Data Library and the Seafloor Samples Laboratory. *We see this as more of a data liaison position. The archivist will work with scientists and national data centers to ensure that data is properly identified and archived.*
    - A Cataloger, a Clerical Assistant, and a Mixed-Media Preservationist to manage and rescue existing data, and to manage future acquisitions within the Institution Archive. *The actual staffing will probably vary depending on the data being processed. A database programmer will be useful for integrating and configuration of software.*
    - Establishment of a permanent Scientific Data Advisory Committee to deal with continually evolving issues of acquisition, archiving protocols, management, and dissemination of scientific data. Establishment of mechanisms routinely to cull unwanted data from the archives and to identify data that need to be rescued or migrated. Provision of support to staff who contribute to these functions. *An active committee can ensure that the archivist is aware of current and pending research programs.*
    - Construction and use of an efficient, WWW-accessible metadata (data about data) system (hardware and software) to manage scientific data and make it accessible to the scientific community. Incorporation of appropriate web links to other archives and digital databases. *The adoption of standard metadata definitions is a key requirement to any data system and will be a requirement for participation in national and international programs. One of the primary functions of the archivist is to relieve WHOI scientists of the burden of metadata requirements.*
    - Upgrade of DSOG hardware and procedures to allow complete duplication and archiving of NDSF data. *Separate data management hardware should be used to provide access to the archive for multiple projects.*
    - Development of a mechanism similar to the Independent Study and Sr. Technical Staff Awards to award internal, proposal-based grants dedicated to enhancing the Scientific Archives. *Institution support would help seed proposals to the federal agencies.*
    - Investigation of mass-store systems for large volumes of high-priority digital data. *A high capacity storage device could be used to store metadata and stage data being transferred to or from a remote national archive.*

Additional Considerations

Although there are many arguments for and against developing and maintaining an archive, it is clear that the ability to distribute and retrieve data is playing a greater role in oceanographic research. This is evident by the growing number of data archive initiatives at WHOI, nationally and internationally. There are significant efforts at the national (e.g. http://dmac.ocean.us) and international (e.g. http://www.iode.org) levels to develop mechanisms for distributing data. These efforts indicate that ocean observing systems will need to distribute data in near real-time using standardized methods. It is imperative that WHOI becomes prepared to interoperate with these efforts if we want to participate in Global Ocean Observing Systems programs. The fact that the Access to the Sea committee has created a data management subcommittee demonstrates the importance of managing data in the future of research at WHOI. The most compelling argument against a WHOI data archive is that WHOI researchers are not specialists in managing data. Data is an integral part of research, so we have become reluctant data managers. As an individual and often ad hoc process, this often results in duplicate, yet incompatible efforts.

There are many approaches WHOI can take to manage data in the coming decades from simple policy statements to full data vault development. The most realistic approach is to provide tools and services to interoperate with external archives. The key to this approach is to recognize the interdisciplinary nature of WHOI research and the varied requirements within each discipline and research group. The institution will need to provide resources to coordinate the efforts of individual projects and to integrate with external archive centers. At a minimum, the institution should provide guidance to researchers on metadata standards, data formats and short to long-term storage requirements. Although many archive centers will take almost any form of metadata, the archival quality of data is reliant on the integrity of the metadata. Additional effort should be made to provide tools and assistance for migrating data from measurement to long-term archive. The ability to generate useful metadata and provide migration paths for data is necessary regardless of the location of the archive.

A summary of the SDAC report with our comments italicized is given in the next section. Some additional considerations for managing data are then given followed by our detailed recommendations. The rapid changes occurring in oceanographic data management could outdate any specific implantation details so the finer details have been left intentionally vague.

Recommendations

Since the needs of the Institution and the capabilities of national and international archives are in a state of flux, we are recommending a phased implementation. The selection of a tractable project and adopting metadata standards will provide the foundation for adding larger and more diverse data. This will also give WHOI the ability to participate in setting the direction of national archives with regards to metadata standards and functionality. Specifically, we suggest the following steps for instituting data management at WHOI:

1. Designate a data archivist/liaison.
2. Coordinate with national and international data centers.
3. Develop an initial archive project.
4. Develop a long-term relationship with a university computer science group.

5. Prepare a formal data distribution policy.
6. Develop metadata standards and tools.
7. Create a plan for adding historical data.
8. Provide Access to the Sea synergy.

Beyond the appointment of an archivist, the recommendations are not necessarily listed in order of importance. Many of the recommendations will need to be implemented in parallel to be effective.

Data archivist:
We would like to reiterate the recommendations of the SDAC committee and suggest that a data archivist/liaison should be hired with the intent of initializing data management at WHOI. An archivist would facilitate the remaining recommendations by leveraging existing capabilities such as JGOFS and GLOBEC as well as utilizing our oceanographic expertise to influence national standards. Ultimately, this position will incorporate recommendations made by the Access ToThe Sea Committee. Initial commitment over the next five years will require at least a part time data entry person and part time database administrator/programmer.

Distribution policy:
A formal distribution policy is needed that adheres to standards such as WMO, NSF and ONR. Each agency has a policy describing data distribution rights. This policy will also need to address issues such as version control and ownership and copyright. The archivist/liaison will need to work with a wide range of groups at WHOI including all of the departments, institutes, library and Development to ensure compliance and consistency.

Initial project:
The initial project should leverage the Martha's Vineyard Coastal Observatory to test necessary infrastructure to store, forward and retrieve data. Although the SDAC report recommended the National Deep Submergence Lab, an MVCO project would be more likely to produce efficient feedback.

The first step here includes developing metadata consistent with external standards such as the Federal Geophysical Data Center (FGDC, http://www.fgdc.gov) and MarineXML (http://ioc.unesco.org/marinexml/). Although there are numerous "standards" being developed and there is risk in selecting an appropriate standard, waiting for a clear winner will be riskier. If WHOI is not a participant in the definition of metadata standards for oceanographic research, we will be required to incorporate incomplete or inaccurate externally defined standards.

The next step would be to work with Integrated Ocean Observing System (IOOS, http://www.ocean.us/) to develop the mechanisms to distribute real time and archive data. This will provide the insight and experience to add additional data capabilities such as those outlined in the SDAC report. The goal is to use the MVCO as the starting point and commit to maintaining data into the future, while adding additional data (*e.g.* NDSL data) based upon cost/benefit of managing the data.

On-line browse of data should be a separate concern and implemented after metadata and data distribution issues are resolved. Initial browse capability should be used to test data archive capabilities. This is not to imply that access to the data is a second priority. Although the access capability is dependent upon managing the data and the context of the data, it must be separated to allow the development of application or project specific interfaces.

Coordinate with external agencies:
The requirements for an archive – managing multiple copies of data and metadata, creating and verifying metadata, insuring the integrity of the data and preserving the data – are necessary regardless of the data location. These capabilities will give WHOI researchers the ability to be more effective analyzing data. Time will not be wasted locating and analyzing incorrect data.

In addition to national data centers, smaller regional data centers, discipline specific data centers or project specific data centers will be created. The mechanisms for submitting and retrieving data from these archives will help fulfill grant requirements for sharing research results. The ability to coordinate with external agencies will become a requirement for any data WHOI produces regardless of the physical location of the archive.

Historical data:
Once a system is in place, historical data should be added to the archive as funding and time permits. It is not economically feasible to try to archive historic data while developing a management scheme. However, an assessment of the historical data should be made to ensure compatibility with older data.

Synergy:
Although the Access to the Sea report is likely to be completed before an archivist can be hired, we would expect that archivist would refine the findings of the data management section of the report. The archivist would also provide continuity for the Data Library and Archives, MVCO, NDSL, JGOFS, GLOBEC and provide a unified front for data management proposals.