CYBERINFRASTRUCTURE COMMITTEE REPORT

Executive Summary

The Cyberinfrastructure Committee was formed in October, 2004 largely as a response to the Access to the Sea report. The committee was tasked by Vice Presidents Luyten and Detrick to examine the state of cyberinfrastructure at WHOI and if deemed necessary to develop specific proposals for steps which should be taken to remedy areas of weakness. After a series of preliminary meetings, the committee formed three working groups focused upon the key topics which emerged as needing to be addressed. The three working groups were: Architecture and Infrastructure, Shipboard and Vehicle Data and Data Portal. Each working group produced a report and recommended action plan that is attached. The working groups included people beyond the Cyberinfrastructure Committee membership to increase representation and add specific areas of expertise. Overall over 20 people were involved not counting those simply interviewed for their perspective.

The Committee as a whole endorses the working group reports and urges that immediate steps be taken to begin implementation of the actions contained in each report. Even recognizing current budget constraints, it is recommended that funding be provided to begin the first year tasks outlined in each report. Delays in proceeding will only exacerbate the current problem due to the lack of a systematic approach to data management and access and place us even further behind our peers.

The federal agencies, particularly the National Science Foundation, are making data preservation an increasingly prominent issue. The recent report from the National Science Board entitled “Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century” (May, 2005) recognizes different classes of digital data with differing preservation requirements and notes that direct observational data of time dependent phenomena “are historical records that cannot be recollected” and “are usually archived indefinitely.” The impact of the recommendations of this report and others have yet to be fully incorporated into agency policies but it appears likely that more stringent data preservation requirements and enforcement of them is the future. WHOI needs to prepare its infrastructure to meet its Institutional obligations so as not to jeopardize future grant opportunities.

There is, at this moment, a unique opportunity to leverage any WHOI sponsored efforts with the work being done in association with the recent NSF Digital Archiving award (DIGARCH) that we have received in collaboration with SIO and SDSC. This grant, while small, provides funding to begin to address some of the recommendations contained in the Committee’s report, especially in the area of shipboard and vehicle data. There is substantial overlap between the goals of the DIGARCH project and the Committee’s proposals and every effort should be made to maximize commonality and to coordinate development.

The primary conclusions of each of the working groups are as follows:

Architecture and Infrastructure Working Group conclusions:

WHOI should immediately embark upon an effort to establish a coherent and pervasive cyberinfrastructure which is designed to support the scientific processes of data collection, data lifecycle management, analysis and modeling and presentation. This infrastructure must:

- facilitate the efficient discovery and use of resources both within and external to the Institution.
- be easy to incorporate into the experiments and computational tasks of researchers and unobtrusive to use for routine operations.
- remove barriers to cooperation both internally and globally while preserving security and controlled access to data and resources.
- support the scheduling of large complex calculations involving distributed computing and storage resources and complex workflows.
- provide mechanisms to support the permanent archiving of information in a retrievable manner
- be responsive to changing priorities (both long term and as triggered by naturally occurring events) and massive unexpected influxes of data
- be efficient to operate and support by the IT staff.

Software to support a computing environment of this type has been developed over the past few years with extensive NSF support and it has now matured to the point were it can and should be used by WHOI. The two major software components are the latest version of Globus Grid Toolkit which integrates grid computing with Web Services and the Storage Resource Broker (SRB) from the San Diego Supercomputer Center. SRB is a key element of the DIGARCH work so there is synergism between these recommendations and that project.

Shipboard and Vehicle Data Working Group conclusions:

Efforts need to be undertaken to insure that every WHOI cruise and vehicle lowering produces an organized, complete and documented collection of data and metadata; efforts need to be made to implement display and collection systems that are readily available and whose design is more science-driven, by integrating real-time displays of multiple sensors and using automatically generated metadata; and efforts need to be made to improve data search and access on shore (i.e.; by location, by cruise, by time, etc.) and provide a catalog of data in terms of what was collected and where did it go. Additionally, efforts need to be made in the areas of interoperability including developing and enhancing current systems and coordinating with the cyber committee portal and infrastructure efforts.

In order to achieve these goals, the working group recommends embarking on the following efforts in parallel:

- WHOI should begin development and documentation of a formal cruise data set for WHOI ships. The cruise data set would be defined as the package of information returned to WHOI upon completion of each leg, and would consist of an organized, complete, documented, and standards-based collection of data and metadata.
- WHOI should expand and enhance the GeoBrowser technology (currently used on WHOI ships, Alvin, and ROV Jason II) to provide science users with better tools to enrich at-sea productivity and data access/availability post-cruise. Such tools would include developing web-based application software for services such as Dredging, Coring, and CTD operations and provide scientists with easy to use forms, automatic metadata generation, and summary reports; improved real-time displays; scientific cruise summary reports; improved capability for searching and retrieving data; etc. At the same time, efforts should be undertaken to improve interoperability of the GeoBrowser systems with other systems such as Geographic Information Systems like Roger Goldsmith's systems, WHOI data portal efforts, Scripps SIO Explorer, etc. using standards-based protocols like WMS, XML, and Web Services.

Elements of both of these recommendations have begun to be worked on as part of the DIGARCH grant; however, the funding in it is insufficient to go beyond prototype and conceptual demonstration.

Data Portal Working Group conclusions:

The working group recommends six strategies that facilitate further development of a WHOI data portal. These recommendations include:

- Establish leadership and coordination for ocean informatics and cyber-infrastructure efforts at WHOI by identifying an individual or group to provide the guiding vision for information and data management at WHOI.
- Design and implement a pilot project to further investigate scientific need and infrastructure and processes required of a data portal
- Compile a complete data inventory of all data resources at WHOI, conduct user testing, evaluate current data portals, and survey WHOI scientists and students to assess their need for a data portal
- Identify metadata and interoperability trends
- Implement outreach, communication and staff development for long-term success of a data portal, and
- Provide long-term planning and evaluation of data portal activities

Some preliminary steps have already been taken in response to these recommendations. The Digital Data Center web page is partial collection of data resources at WHOI and the metadata efforts called for in the working group report overlaps with that in the other reports.

One year goals:

The three working group reports in total call for a relatively modest investment to lay down the foundation for a WHOI cyberinfrastructure. The combined reports call for about three person years of support (spread over several people) to accomplish the first year goals described in them. This estimate attempts to account for overlap between the reports. No equipment, travel, or other types of support is requested in any of the reports. If the WHOI effort is cotemporaneous with the DIGARCH work then the projects can be managed so that there is significant synergism with a corresponding cost reduction to WHOI. It is expected that with $150,000 of internal support that most of the key first year development goals could be met. This funding would be managed by a project team composed of both scientists and technical staff and led by Bob Detrick. It will used to cover salary time for members of the technical staff as they work on these tasks.

The primary tasks that would be accomplished with this level of support are:

1. The deployment and pilot use of key software components for a next generation computing infrastructure. This would be based upon the Globus grid version 4 software and the Storage Resource Broker from SDSC.
2. The design of a formal cruise data set, including data and metadata standards and defined pre- and post-cruise procedures, based upon a review of ship data collected over the past 3 years.
3. The extension of the Ship DataGrabber system to gather in real time information about a greater range of common shipboard operations such as dredging, coring, and CTD operations. In addition, the Ship DataGrabber will be enhanced to support links to other databases and have improved search capabilities.
4. The development of a pilot web data portal to provide access to selected data repositories with the ability to dynamically select and operate upon subsets of the data.

In addition to these specific results, a significant outcome of this effort will be the creation of a core group of WHOI staff, knowledgeable in WHOI's datasets, and skilled in emerging cyberinfrastructure techniques. All of these tasks require a substantial investment by the participants in learning new information technologies and the agreement upon an Institutional approach to data and metadata management. In terms of preparing for the future demands of large-scale data preservation, this outcome is as, if not more, important than the deployment of specific software technologies.

TABLE OF CONTENTS

1. Architecture and Infrastructure Report (A. Gaylord chair, R. Goldsmith)

1.1 Motivation for a WHOI CI initiative

Advances in information technology continue to have a dramatic impact upon the science of oceanography. WHOI scientists and engineers have always been at the forefront of adopting state-of-the-art electronic and computing technologies into specific research projects as evidenced by our leadership in the development of sophisticated underwater vehicles and instruments, pioneering of use computers on ships, and early adoption of the Internet. As an Institution, however, the infrastructure to support these project based efforts has often lagged. Over the past several years, WHOI has built up a physical network infrastructure and core set of services which are adequate for current and near-term future needs. The Institution has not, however, committed to the establishment of kind of modern cyberinfrastructure that will be necessary to continue to make effective use of information technology as tool to advance scientific discovery. This has already had an impact upon the Institution as demonstrated by our limited ability to access and share cruise data in a convenient and timely manner as well as the difficulty we have even cataloging all our data holdings. As ocean observatories and other automated observing systems are deployed the volume of data that needs to be managed will increase by several orders of magnitude over current quantities. This in turn will drive vastly more computationally demanding models and visualization techniques.

Several of recommendations of the Access to the Sea Task Force Report (July 2004) are concerned with responding to the impact of information technology upon the Institution's work and the need to improve our cyberinfrastructure. Recommendation 10 was the motivation for the formation of this Cyberinfrastructure Committee and in part states that

> "WHOI has fallen behind in its involvement with scientific data management and cyberinfrastructure developments, and is not well positioned to manage current and legacy data or to handle the vast volumes of data associated with planned ocean observing networks."

Recommendation 11 goes on to state that

> "WHOI should take immediate steps to facilitate the development of procedures and tools required for managing metadata and data at WHOI. This should include the initiation of one or more high-profile 'demonstration projects' that could serve the dual purpose of developing our internal expertise in scientific metadata/data management and stimulating broader interest in the development of cyberinfrastructure at WHOI."

While the Access to the Sea Task Force Report concentrates on the critical issues directly related to the collection, management, and storage of observational data collected from the ocean, there are many additional challenges that need to address within the WHOI information environment

It will become increasingly unlikely that individual investigators and even individual institutions will be able to meet computational and information storage needs of their research programs. Multi-organizational cooperation will become a necessity. It seems safe to assume that funding agencies will not adequately support isolated efforts of any significant scale in the future. Indeed this is implicit in the thinking behind NSF's strong support of grid technology and other shared infrastructure initiatives. Within NSF, the topic of cyberinfrastructure is so pervasive that CISE (Computer & Information Science & Engineering) recently reorganized to form a division of Shared Cyberinfrastructure

The 2004 report "Trends in Information Technology Infrastructure in the Ocean Sciences," by NSF's Ocean ITI Working group extensively discusses the challenges and emerging trends in this area. It highlights many of the same advanced cyberinfrastructure elements that are recommended in this report. The NSF report concludes its introductory section with the following statement:

> Adequate ITI [Information Technology Infrastructure] capabilities at each local research institution are essential for designing and implementing new scientific programs and new observing systems that are needed to address complex, interdisciplinary problems in ocean research. New ITI capabilities will support more-effective data management and dissemination, advanced analyses of complex data sets, and communication with the public and policymakers."

WHOI needs to heed this statement and immediately embark upon an effort to establish a coherent and pervasive cyberinfrastructure which is designed to support the scientific processes of data collection, data lifecycle management, analysis and modeling and presentation. This infrastructure must:

- facilitate the efficient discovery and use of resources both within and external to the Institution.
- be easy to incorporate into the experiments and computational tasks of researchers and unobtrusive to use for routine operations.
- remove barriers to cooperation both internally and globally while preserving security and controlled access to data and resources.
- support the scheduling of large complex calculations involving distributed computing and storage resources and complex workflows.
- provide mechanisms to support the permanent archiving of information in a retrievable manner
- be responsive to changing priorities(both long term and as triggered by naturally occurring events) and massive unexpected influxes of data
- be efficient to operate and support by the IT staff.

It is technically feasible to begin testing and deploying an environment just as this immediately, but it will take several years to complete the transition to it from our current environment. This is both because of the need for some elements of the envisioned cyberinfrastructure to mature, the need to maintain compatibility with existing research systems and applications and perhaps mostly critically the need for the Institution to culturally adapt to the new environment. It is expected that the change to the Institution will be comparable to that caused by the introduction of the Internet.

1.2 Current cyberinfrastructure at WHOI

WHOI already has a significant networking and computing infrastructure. Up until the recent appropriation of the term a few years ago, this would have been called our cyberinfrastructure. We currently have a solid network infrastructure which is capable of handling significantly more traffic than is placed upon it now and has the capacity to be expanded by at least a factor of 100 as the need arises.

The primary services offered on the network are largely targeted at individual users and are for the most part statically defined. These include Email, Web access, FTP, simple file sharing and printing. CIS does manage a modest amount (~15 TBs) of storage space to support these services but the vast majority of information storage is distributed throughout the Institution on project dedicated computer systems. CIS also operates a network-based backup service which provides short- term protection against information loss for all CIS operated servers and a few hundred project dedicated systems.

There are three primary, and disjoint, user authentication methods in use on the WHOI network. The first is authentication via a password stored in and LDAP directory which is primarily used for access to internal resources, such as Email and restricted Web pages. The second is RADIUS authentication which is used mostly for access from external locations using RAS dialup services, iPass or the VPN service. While both of these authentication methods are WHOI system- wide not specific to a given computer system, the third method is exactly that, an individual host-based authentication. This is widely used on both research and administrative systems.

The Institution has a number of data repositories, some of which are in formal database systems and others not. These efforts serve distinct target groups, both internal and external to WHOI and tend to have diverse data access procedures and policies.

Virtually all scientific computing at WHOI is done on computer systems owned and operated by individual research groups or small collaborations. There is no centrally supported computing resource. A number of researchers use national supercomputer facilities (NCAR, UCSD, Los Alamos, etc.) for their most extensive computations.

Currently an Institution-wide directory service based upon LDAP provides information about the people associated with the Institution. This is being enhanced by the ConnectWHOI effort which will maintain more information about people, integrate several current disparate data sources of personal information and give individuals the ability to update and add information about themselves and their research. This will make it easier to discover who is working on what and what researcher is being done at WHOI. It will not, however, designed to be a comprehensive directory of data repositories at WHOI nor is there any consistent approach to metadata describing the contents of these holdings.

1.3 The need to advance WHOI’s cyberinfrastructure

The current environment has been considered adequate for current needs and is very consistent with WHOI’s general philosophy of supporting project-based research. It is increasingly in danger of becoming insufficient to support the growing computational and data management needs of many researchers and can be anticipated to be seen as major Institutional weakness if not changed over the next few years. In particular, in the current cyberinfrastructure environment will be difficult to support complex automated computations across multiple computer systems, cross-disciplinary data mining, and perhaps most critically multi-institutional collaboration.

The lack of a uniform and consistent means of identifying individuals for controlled access to resources is a major barrier to any efforts to share information and resources. When only a small number of people, all known to each other, want to share it is possible although burdensome to establish user accounts on the computer systems involved. More general sharing relationships, especially those across organizations or those which are spontaneous, are very difficult, if not impossible to manage securely and efficiently. To address this issue, a computational environment which uses a standards based method of identification and can be used across organizations needs to be established. This would not necessarily supplant existing authentication methods at WHOI, but would be the basis for access to a new infrastructure.

The lack of any Institution-wide procedures for resource identification and location is similarly a major impediment to support advanced computing needs. Without a set of consistent and universally applied (within the scope of the infrastructure) procedures for naming and describing both the static and dynamic state of resources, it is extremely difficult to share information and use computational resources efficiently.

This does not mean that WHOI needs to move to a more centralized computing environment. Indeed that would not only be counter to the WHOI culture and very expensive, it is not likely to be technically possible. The needs of individual researchers at WHOI vary too widely to attempt to accommodate them all into a single computational structure or to be able equitably cost recover the associated expenses. Instead what is recommended is the adoption of a consistent distributed computing model which will enable the secure and controlled sharing of information and resources across project and institutional bounds. By establishing a minimal set of centrally managed services with well-defined functions and interfaces, it is possible to empower individuals to locate and offer/use resources of all types (data, instruments, storage, compute cycles) under conditions that are defined as precisely as desired by the resource owner. Resources that are desired to be publicly, or at least broadly, available can be exposed through portals and/or proxy access methods to avoid introducing any unnecessary complexity to the consumer. This type of distributed environment is one which can dynamically adjust to the changing needs of both project and Institutional needs. It is also one which can be progressively introduced over time in a manner that is minimally intrusive. That is new services supporting an advanced cyberinfrastructure can be installed and provide benefits to those you chose to use them without disrupting the current environment.

This recommendation is basically an acknowledgement that the everyday scientific computing environment is becoming more complex and that new mechanisms must be introduced to effectively provide adequate support services to operate it. This in many ways is analogous to the evolution of the networking environment that we have today and more recently the trends in building and maintaining web environments. The Internet works today because a number of standards and conventions were adopted and consistently applied to all who connect to it. Many of these standards are virtually invisible to the average person and others have become so widely used that they seem perfectly natural. In the early days of networks, there were some who needed them more than others. Some people strongly advocated for the deployment of a network infrastructure, while others resisted the concept as unnecessary or even disruptive in the context of their research. There is a growing consensus within the computing and networking community that the type of environment described in this document is the next stage of evolution for a new cyberinfrastructure that will become as ubiquitous as the Internet is today. As with the Internet, some of the more demanding science and engineering problems are leading the way, although this time commercial applications are not far behind.

1.4 Key components required for a modern Cyberinfrastructure

A minimal set of fundamental building blocks are required to support a cyberinfrastructure environment that facilitates the sharing of resources. These have been described in various terms over the years but are well described using the current terminology of the Grid community, so that terminology will be used here. The four core categories they use are:

- **Security services** used to identify people and other entities and to control access to resources.
- **Data Management** service to support data transfer and replication
- **Execution management** to enable the scheduling and allocation of resources
- **Information services** for resource discovery, naming and location

These services categories are not new and indeed exist in one form or another in almost all complex systems whether they are computer based or not.   What is different is that the scope and complexity of the services used to implement these functions has grown as the range of computing resources potentially available to researchers has become more diverse and more distributed. Many techniques that work well for individuals or small well-defined groups of collaborators do not scale well to larger, more amorphous communities. Creating a modern cyberinfrastructure will mean a transition from a network of autonomous devices individually managed to a more coherent shared environment. Participation in this new environment can be optional, but for the option to be available for those who need it, the key infrastructure components must be in place.

The next section describes each of the four categories in greater detail.

1.4.1 Security

- Authentication – identifying who a person or resource is
- Authorization, - identifies group membership. It is often not necessary or even desirable to make security decisions based upon an individual’s identification but rather his/her/its role.
- Access control – limits access to resources based upon authentication and authorization
- Credential management – maintenance of authentication and authorization info.

- Delegation – allowing actions to be done on your behalf – important for when a process or server is doing work for another process

1.4.2 Data Management

- Data transfer file transfer and copying services
- Replication – creation, deletion, and maintenance of multiple copies of data for redundancy and optimized access
- Location - discovery of where the data is. This is particularly important service in a large and changing environment where the physical location of the data may move for performance and reliability reasons.

1.4.3 Execution management

- Scheduling and multi-step workflow
- Resource allocation –allows for absolute and conditional reservations

1.4.4 Information services

- Resource discovery – finding out what is available and what it state is. This is critical is a dynamic, shared environment where resources can come and go. The grid allows for extremely details resource descriptions and requirements matching.
- Resource monitoring – gathering and reporting of state information about resources.

1.5 Grid computing and the Globus Toolkit

There has been a fair amount of hype and confusion around the term “Grids” which has led many to be skeptical about their usefulness in general and in an environment such as WHOI’s in particular. The mass media, even the technically oriented publications, have frequently misrepresented what is or is not grid computing and some computer companies have further confused things by rebranding some of their proprietary products as grids in an attempt to take advantage of the hype. Grid technology began by drawing upon the experiences of the previous decades of work on large-scale distributed computing environments, high performance computing and is now incorporating the most advanced concepts from the web and data management communities.

At least within the scientific and engineering research communities, and especially the Higher Education and government communities, grids are defined by the Global Grid Forum (GGF) standards and the implementations of those standards. The most commonly used implementation, and one directly supported by NSF, NASA, DoD and DOE, is the Globus Toolkit. The remainder of this section discusses the grid within the context of the GGF standards and Globus.

The Grid concept originally emerged in the high performance computing environment to meet the needs of people with very large computational requirements by better coordinating access to supercomputers and other relatively scarce large computing resources. First generation grid implementations were essentially distributed batch processing environments with the appropriate security mechanisms to allow sharing across organizational boundaries. At that time, it might have made sense for an individual WHOI researcher to install grid software on a workstation in order to join an established grid, but it would not have been appropriate to build a grid at WHOI.

The second and third versions of grid software generalized the basic concepts, expanded the scope of usefulness and improved ease of use. At this point, people began to speak of computational grids, data grids and sensor grids to describe different environments where grid technology could be used and a number of highly publicized large projects based upon grids where initiated (GEON, NEESgrid). NSF and other government agencies were the primary sponsors of these efforts and in Europe the e-Science programs were started. In 2001, NSF started their Middleware Initiative which resulted in a series (six so far) of the regular releases of tested and packaged grid software along with several additional scheduling and management tools. The third generation of grid software began to use web service and portals as access methods into the basic grid components to make it easier to develop and use complex grid applications. CIS investigated early releases of version 2, but decided that its deployment and use was still too complicated to be of general use at WHOI. Version 3 was not investigated because it became obvious that this would be an intermediate release and would be quickly replaced by another as there were significant technical and standards related difficulties with the way web services were being melded with the grid software.

The fourth version of the Globus toolkit, which is in beta testing now and scheduled for official release at the end of April, 2005, is a redesign of the grid architecture to fully unify grid and web services. It is compliant with standards of the much broader Web Services community. It also adds significant new services to support event notification (to people and other processes) and improves the management of grid services. A Primer on the design and use of this version of Globus can be found at http://www-unix.globus.org/toolkit/docs/development/3.9.5/key/GT4_Primer_0.6.pdf . This version of Globus grid software could be deployed and supported across the Institution and could act as the backbone of a new cyberinfrastructure. The original grid concept has been significantly generalized and the software used to implement a grid matured so that grids can now be viewed as generic computational environments to support the coordinated sharing of resources.

1.6 Web services

Web services have already been mentioned in the context of their incorporation into the latest grid toolkits, but this is just one of many ways in which they are used. It is important to realize that web services are not solely, or even primarily, about web access as one usually thinks about it, although dynamic web sites can be built from web services. Web services provide a means to execute programs remotely independent of the location or type of computer system on which the programs are running. A collection of standards and procedures for creating well defined, public interfaces to programs form the basis for interoperability and communications between web services. The mechanism for doing this is to define the interfaces using the Web Services Definition Language (WSDL) and to communicate between program units using SOAP messages which are typically, but not necessarily transmitted using HTTP and XML. The standards are largely independent of the operating system and programming language (so that Java, C, etc. can be used to write programs that use web services). Web services are designed to facilitate program to program interactions. Additional conventions define how web services interact with security and management services. Complex applications can be created from assemblages of simple web services as well as through wrappers around pre-existing applications and data sources. It is possible to dynamically locate, start if necessary and connect to components of Web service enabled programs.

1.7 Higher level toolkits and systems

Grids and Web services, although they provide useful core services, are too low level for most scientists to use directly other than in a rudimentary manner. These services are in many ways analogous to the basic network services such as DNS, LDAP, FTP and Email, which are now used routinely by the scientific community, often transparently. They are necessary but do not directly benefit scientific research.

Commodity Grid kits or CoGs have been developed to provide help bridge the gap between traditional computing environments and grids. They provide easier access to grid services from additional languages (including Java, Python and Perl), applications, web portals and from traditional desktops. They hide many of the complexities of the native interfaces to grid services and provide abstractions that are more familiar to more people. Perhaps most important to WHOI are two specific CoGs. One enables easy access to grid services from Matlab and the other provides grid enabled Message Passing Interface (MPI) support.

Additional more sophisticated and more discipline specific CoGs are being developed or ported from other distributed system projects to the grid environment. One example of such an effort is the ICENI project at the Imperial College's London e-Science Centre. The goal of this "is to provide high-level abstractions for e-science (scientific computing) which will allow users to construct and define their own applications through a graphical composition tool integrated with distributed component repositories and to deliver this environment across a range of platforms and devices." (from http://www.lesc.ic.ac.uk/iceni/index.html). Another project of the London e-Science Centre called GENIE (Grid Enabled Integrated Earth systems model) builds upon ICENI and grid technologies in an even more discipline specific manner and could prove to be a highly useful example for us.

The Community Grid kits and the ICENI project are works in progress that still need time to mature into refined tools. They demonstrate, however, the potential of providing tools that support the creation and execution of complex calculations in a distributed environment with similar ease to use of Matlab and Labview. As discussed in the next section, there are efforts under way to incorporate instruments and sensors into the grid and web service environments, which then allow the extension of these tools into the support of real-time environments.

Grid portals are the obvious extension of grid services to be web. This can be a very effective method of providing easy access to fixed computational and/or data services.

Many of the early grid-based, science projects, such as GEON, have taken this approach. With the adoption of the web services model in the Globus Toolkit version 4, the grid portal approach becomes even and more natural to support. Ironically it also becomes less necessary as use of web services makes the direct use of grid services and resources much easier.

1.8 Data Collection

Traditionally the data collection process has largely been as distinct from data analysis and modeling. Data is collected and stored in files or databases that are subsequently processed and analyzed. There have been, of course, exceptions to this but for the most part data gathering sensors and instruments are not integrated into the cyberinfrastructure to the same extent as computational and data storage devices are. There are several trends which are making this isolation of data collection as a separate part of the scientific process less and less satisfactory. Ocean observing systems, especially cabled observatories, as they become increasing complex and deliver larger and larger volumes of data, need to be closely coupled with both analysis systems and data storage management systems both to be responsive to real-time science requirements and to make them manageable. Similarly large-scale sensor networks become very difficult to manage without a well defined support environment. Another trend is the increasing tendency towards large scale multi-organizational projects with distributed collaborations. These projects often require shared control of expensive instrumentation and extensive data sharing. A third trend is that computational simulations are increasingly becoming able to incorporate real-time or near real-time information into their calculations. The NSF initiative is encouraging efforts along in the direction through its Dynamic Data Driven Application Systems initiative.

Sensors and instruments can be incorporated into a gird-based cyberinfrastructure using web service interfaces and thus become accessible and manageable in the same manner as computational and storage resources. Several projects are working on a Common Instrument Middleware Architecture and it is well described in a paper by Chiu et al. (http://www.cs.indiana.edu/~chiuk/pubs/CIMA_whitepaper.pdf). Within the oceanography community, this approach is planned as the implement of the NSF ITR funded LOOKING project (Laboratory for the Ocean Observatory Knowledge INtegration Grid).

The integration of instruments into a cyberinfrastructure can occur in many ways. Large complex entities such as a cable observatory like Neptune will require a substantial cyberinfrastructure just to maintain operational integrity and to respond dynamically to changing scientific observational requirements. At the other extreme, simple sensors cannot be expected to support any local cyberinfrastructure. These sensors can still, however, be virtually represented as a web service element in a unified infrastructure. The advantage of a consistent approach is that all the security, resource location, process management functions of the infrastructure are available to use as part of the data collection process. Additionally, it can make metadata collection easier and more systematic.

1.9 Data Storage

Currently most of the scientific data and derivative products are stored on the computers operated by individual researchers, although an increasing amount is being additionally saved in discipline centric repositories. Typically the information is stored in files using operating system specific semantics and filename paths that are meaningful only on the local machine upon which the files are stored. Data sharing is most often arranged through personal contacts between researchers and occurs via explicit data transfers (FTP or Email) or local file sharing mechanisms (NFS, Samba, etc.) While this approach has been considered adequate at WHOI for many years, it is becoming increasingly burdensome to maintain and is a significant barrier to broader information sharing.

Data repositories such as JGOFS and GLOBEC are extremely valuable for the management and sharing of well defined data sets that are intended to be broadly available. Access methods such as DODS and data portals can be used to hide many formatting and local storage details that inhibit data sharing. These approaches are designed to server as general purpose distributed file systems.

A robust modern cyberinfrastructure both requires and supports more flexible methods of information naming, location, and retrieval. These facilitate the sharing of information in ways that retain the data owner's control over the data while making it more accessible. By introducing information location services and the abstraction of a virtual file system into the cyberinfrastructure, it is possible to mask the details of where and how the data is physically stored (it can and most often will remain on researcher maintained computers) and to make it accessible via searches of descriptive metadata rather than obscure and inflexible file path names.

In order to take advantage of the resources available in a distributed environment, a file system must be able to store and access information independent of operating system semantics, honor security constraints, and support both replication and portioning of data. Applications should be able to access data independent of the storage mode; that is without knowledge of whether the information is coming from a file, database or real-time data stream, but still has the option of using optimized access via specific methods. Furthermore all this should happen as much as possible without preplanning, especially by humans.

It is important that data can generally be considered transportable and replicable, at least for the purpose of making transient copies to optimize computations. Both real-time data collection processes and large-scale distributed computations can benefit from features such as opportunistic data replication and transient storage allocation at the file system level.

1.9.1 Storage Resource Broker

A data storage system that matches these requirements is the Storage Resource Broker developed at the San Diego Supercomputer Center. As described on their web site

> "the SDSC Storage Resource Broker (SRB) is client-server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets. SRB, in conjunction with the Metadata Catalog (MCAT), provides a way to access data sets and resources

> based on their attributes and/or logical names rather than their names or physical locations." (from http://www.sdsc.edu/srb/)

SRB is an established application that has been developed and used extensively for many years at SDSC. As of 2004, they were using it to manage over 16 million files and 90 TB of data across several different projects. It is the leading data storage system for use in large distributed environments and with very large datasets. It has been used in a number of Grid and large-scale data environments. It is relatively easy to establish an initial pilot program and then grow it to institutional scale. There are existing interfaces to SRB from the Globus toolkit (version 3 currently), web services, Java, Perl, and Python. Additionally there is an interface to standard Unix I/O for compatibility with legacy applications. There is also a command line interface and an SRB windowing client.

1.10 Data Archiving

Data archiving refers to the procedures and tasks necessary to insure the long-term perseveration and recovery of information. This is distinctly different from the requirements of immediate data access and routine backup operations related to recovery from failures or accidents. A modern cyberinfrastructure provides most of the basic functionality required by data archiving operations, including data replication and migration capabilities, location services and a security framework. A successful data archiving program, however, requires additional planning, documentation, and stringent procedures above and beyond those found in standard operational environments.

Some of the specific issues that need to be considered when planning for data archiving include:

- The provisioning both local and remote archives to insure accessibility and survivability in the event of catastrophic failures.
- The preservation of ownership and access rights over time, even as people and projects come and go from the active security information databases.
- The creation and maintenance of a sufficiently rich set of metadata to document of data formats, calibration information, instrument/sensor ID, processing requirements and all other descriptive data necessary to insure that the observational and derived data remains useful.
- The preservation of computer programs and algorithms used to process data. For critically important data that is sensitive to processing details, the preservation of the actual computer systems, operating systems, and programs should be considered.
- Problems associated with the media upon which the archived data are stored including media longevity, maintenance and migration. Media should be regularly tested to insure that it remains readable.

1.11 Recommendations

1. Create a WHOI grid using the Globus Toolkit version 4
    a. CIS should deploy the necessary core services of version 4. It is estimated that this will take about 2 person months (320 hours) to install, test, and learn to operate in our environment.

    b. Establish a small pool of processor and storage resources that can be used by grid testers in trial and demonstration projects. Initially this can be done using existing CIS computers and contributions from the community.

    c. Solicit proposals for one to three computationally oriented applications that will demonstrate the use of the gird. WHOI should provide small grants (2- 4 weeks of time each) to support the labor involved in grid enabling the selected projects.

2. Begin a trial deployment of the Storage Resource Broker (SRB)
    a. CIS should deploy the core SRB services for trial use and make some of its storage resources and information available through SRB. Estimated about 2-3 person weeks of effort.

    b. CIS should test, and then make available the various clients that interface with SRB.

    c. Task a working group to look at the standard SRB metadata structure and determine what extensions may be needed for use at WHOI. Note the metadata structure is flexible and can be project specific, but if there is specific information that should be required at WHOI then the earlier this is established the better.

    d. Configure SRB to work with the WHOI grid. It is recommended that a WHOI use SRB and grid technology together in its cyberinfrastructure but this is not absolutely necessary.

    e. Solicit proposals for one to three data oriented applications that will demonstrate the use of the SRB. WHOI should provide small grants (2- 4 weeks of time each) to support the labor involved in SRB enabling the selected projects. Ideally at least one project would involve an established database collection and another real-time data stream.

3. Establish a GEON grid node in cooperation with USGS
    a. This is a fast way to get experience in an established, multi-organizational grid environment, although it is based upon earlier versions of the Globus grid toolkit.

    b. USGS has expressed interest in this and will help fund the effort through the Cooperative agreement.

4. Offer training in grid technology and Web Services
    a. Extensive training is necessary at all levels from general concepts to specialized programming techniques.

   b. Invite speakers from Globus, UCSD, W3C to give general overviews and preliminary training. Much of this can be done at minimal cost (travel expenses or less).

   c. Establish a WHOI internal seminar series for self education and sharing experiences.

5. Task a working group to examine use of higher level grid based toolkits such as the Matlab CoG and Iceni.

2. Shipboard Data Report (R. Detrick (ex-officio), P. Lemmond co-chair, S. Lerner co-chair, A. Maffei, D. McGillicuddy, D. Smith, M. Tivey, B. Walden)

Report Summary

This group was tasked with examining how to advance the cyber-infrastructure on ships and vehicles operated by WHOI. Within the context of the working group, "cyber-infrastructure" was broadly defined as the collection, display, and dissemination of the scientific data originating on WHOI ships and vehicles. We formed a sub-committee workgroup[1] consisting of scientists and engineers integrally knowledgeable of WHOI's ships, vehicles, and sensor/data systems. The goal of this sub-committee was to identify areas that need specific improvement, develop a plan to implement short-term (1-3yr) pilot projects to address these, and at the same time, identify longer-term issues and potential funding avenues. The group identified these broad areas that could use improvement: metadata documentation and collection, real-time data displays, and data accessibility/availability for both real-time and post-cruise users.

Within these areas, efforts need to be undertaken to insure that every WHOI cruise and vehicle lowering produces an organized, complete and documented collection of data and metadata; efforts need to be made to implement display and collection systems that are readily available and whose design is more science-driven, by integrating real-time displays of multiple sensors and using automatically generated metadata; and efforts need to be made to improve data search and access on shore (i.e.; by location, by cruise, by time, etc.) and provide a catalog of data in terms of what was collected and where did it go. Additionally, efforts need to be made in the areas of interoperability including developing and enhancing current systems and coordinating with the cyber committee portal and infrastructure efforts.

In order to achieve these goals, the committee recommends embarking on the following efforts in parallel:

1. WHOI should begin development and documentation of a formal cruise data set for WHOI ships. The cruise data set would be defined as the package of information returned to WHOI upon completion of each leg, and would consist of an organized, complete, documented, and standards-based collection of data and metadata.
2. WHOI should expand and enhance the GeoBrowser technology (currently used on WHOI ships, Alvin, and ROV Jason II) to provide science users with better tools to enrich at-sea productivity and data access/availability post-cruise. Such tools would include developing web-based application software for services such as Dredging, Coring, and CTD operations and provide scientists with easy to use forms, automatic metadata generation, and summary reports; improved real-time displays; scientific cruise summary reports; improved capability for searching and retrieving data; etc. At the same time, efforts should be undertaken to improve interoperability of the GeoBrowser systems with other systems such as Geographic Information Systems like Roger Goldsmith's systems, WHOI data portal efforts, Scripps SIO Explorer, etc. using standards-based protocols like WMS, XML, and Web services.

2.1 Introduction

The purpose of this working group was to determine how to advance the cyber-infrastructure on ships and vehicles operated by WHOI. Within the context of the working group, “cyber-infrastructure” was broadly defined as the collection, display, and dissemination of the scientific data originating on WHOI ships and vehicles. The general consensus of the working group was that:

- Data collection is being done well on-board WHOI ships and vehicles.
- The display of and interaction with data at-sea lags data collection efforts. Present capabilities are geared toward individual systems and sensors (predominantly engineering-driven); better integration could enrich operations (more science-driven).
- Dissemination of data after a cruise needs substantial improvement. Duplicating the data set from the ship to shore is inadequate.

Probably the most serious challenge for advancing shipboard cyber-infrastructure lies in the wide diversity of projects undertaken with WHOI ships and vehicles. A different mix of science participants in the working group would almost certainly skew implementation priorities, especially in the area of at-sea data displays and interaction. Another major challenge, maybe equal to the first, is in the tradeoff between short-term, high visibility solutions versus longer-term, infrastructure and process-oriented solutions.

2.2 Discussions

The members of the working group discussed a wide variety of cyber-infrastructure related topics, technologies, applications, and goals. As an organizational method, it seems best to classify these in a somewhat chronological order.

2.2.1. Issues During a Cruise

a) Real-time displays

Current displays generally present useful information as it pertains to a single, specific instrument. However, linking multiple, real-time data sets as they are generated would be a useful enhancement, both for scientific interpretation and real-time decision-making. Possible improvements would include:

- Expand available *DataGrabber* and viewers to link additional real-time data sets.
- Expand *EventLogging* for more types of science operations.
- Provide science users with a real-time navigation and operational display (where are we now, where have we been, where are we going, how far are we from the last core site, etc.).

In this category, the major challenge lies in hitting multiple, moving targets. Shipboard systems change over the years, change cruise-to-cruise, some systems used frequently, others rarely. Setting priorities will be difficult, especially without an established data architecture.

b) Daily Shipboard Operations and Activities

Science users of WHOI ships use a variety of methods to track the overall daily progress of cruise operations. Some use hand written logbooks, some keep on-line journals, some use pre-printed forms, and some don't bother with anything. Shipboard technical personnel also maintain such records. Regardless of the method, this information most often constitutes a wealth of knowledge regarding the science aspects of a cruise, and thus needs to be preserved along with system and sensor generated data files. Some of this information will form the basis of parts of the various cruise metadata files.

c) Quality Control

Ensuring that high standards of quality control are met is a goal no one can argue against. Quality control is almost always an issue that applies to a single instrument or system, and is most likely handled best on a case-by-case basis, within the existing shipboard technical support structure. One could reasonably expect that new initiatives for real-time displays and the implementation of data and metadata standards will advance quality control efforts.

d) Communications & Network Access

Real-time and near real-time access to shipboard data will likely migrate more to a network model than a direct connection or file-based model. We do appear to be on the cusp of, or at least approaching the cusp of, ships at sea being continuously connected to the Internet. Ships will also be connected to other ships, either via the Internet or via a private, direct network. Users accessing shipboard data may be sitting in their stateroom on the ship, on the bridge of another, close-by ship, or in an office on shore. Some of the issues that might arise include:

- How to provide a path for technical support issues at sea.
- Development of fleet-wide standards for data transmission and messaging.
- How to insure network security.
- How to balance network performance issues.

2.2.2. Issues at the End of a Cruise

When a cruise is ending, there is always a flurry of activity assembling reports, summaries, and data products. Onboard technical personnel put the current cruise behind them to gear up for the next, and the science party scatters to various places to resume other pursuits. Within a short time, the intimate knowledge about cruise activities (was the weather bad during one day in a survey, what are these empty file in the XBT directory, why is there a gap in navigation, etc.) begins to fade. Much of this information is not captured in the actual data itself, but is vital to using the data in the coming years. Preserving this information will be useful to both immediate science needs, short and long-term technical support, and to the broader community.

The creation of more content-rich, standardized cruise data sets has been a goal of science users, both from informal discussions and formal committees and organized workshops. To accomplish this within WHOI shipboard infrastructure, the following would need to be accomplished:

a) Insure that cruise data is complete and organized at the end of a cruise.

b) Insure that complete, standards-based metadata accompanies all data leaving the ship.

c) Identify the ancillary data that needs to be included.

d) Create a mechanism to insure that data and metadata, when received at WHOI, can be verified and ingested into a larger data environment.

A unique challenge for this type of effort would be the lack of technical advances and expertise needed to accomplish it. Rather, this effort would be a process-oriented and management challenge.

2.2.3 Issues After a Cruise

When shipboard data returns to WHOI, it has traditionally, and will likely in the future, leave the domain of shipboard technical services and become “someone else’s problem.” Whether such data ends up in a massive, on-line, web-accessible, relational, GIS-based database system, or is bundled up in a cardboard box and put on a shelf in a warehouse, two requirements need to be satisfied prior to such actions:

a) Shipboard data must be in a form to insure interoperability within the Institution. Some of this relates to the efforts of the WHOI Cyber-Infrastructure Data Portal Group, and also to in-place archive and dissemination efforts (WHOI Data Archive, Multibeam Data Archive, etc).

b) Shipboard data must also be in a form to insure interoperability with peer institutions and federal agencies efforts to provide data for the broader community.

2.3 Applicability Beyond WHOI Shipboard Data

2.3.1 NDSF Vehicles

Data infrastructure issues on NDSF vehicles more or less mimic those of WHOI ships. The status of data collection (good), display and interaction (somewhat lagging), and dissemination (needs improvement) within the NDSF would all benefit if the methodologies implemented to support shipboard operations were adopted. There appear to be two major challenges. First, data operations on NDSF vehicles are generally faster and more voluminous than shipboard systems. At-sea vehicle turnaround is faster than ship in-ports. Vehicles are more likely to produce video, imagery, and acoustic data of scientific interest, in addition to important, engineering, performance data. The second major challenge is that NDSF vehicles are generally subjected to more experimental, one-of-a-kind improvements and upgrades, making standardization more difficult.

2.3.2 Observatories

Observatory systems also share the same elements with WHOI ships as NDSF vehicles. Probably the major difference, however, is that observatory system experiments will be conducted over much longer time periods, so that the distribution of data will occur continuously. With observatories in their infancy, having methodologies and expertise in data infrastructure in place prior to full-scale implementation may prove to be very useful.

2.4 1930 to 2004?

WHOI has assembled a vast quantity of shipboard data during its first seventy-five years. At some point in the future, this data will either be lost to current practitioners (not on-line, cannot read media, cannot determine usage, etc) or will be available in much the same form as data currently being collected. Determining the effort needed for data rescue will be a project in itself; actually rescuing data will certainly be a much larger undertaking. While neither of these tasks is necessarily part of the current shipboard data infrastructure efforts, current and future efforts must acknowledge past data holdings and be prepared for, and compatible with, any future data rescue efforts.

2.5 Recommendations

The following recommendations reflect the general consensus of the group's participants.

The shipboard data working group members discussed and identified these broad areas that could use improvement including metadata documentation and collection, real-time data displays, and data accessibility/availability for both ship and shore based systems. Within these areas, efforts need to be undertaken to insure that every WHOI cruise and vehicle lowering produces an organized, complete and documented collection of data and metadata; efforts need to be made to implement display and collection systems that are readily available and more science-driven, by integrating real-time displays of multiple sensors and using automatically generated metadata; and efforts need to be made to improve the capability to search and retrieve data on shore (e.g.; by location, by cruise, by time, etc.) and provide a catalog of data in terms of what was collected and where did it go.

In order to achieve these goals, the committee recommends embarking on the following efforts in parallel:

1. WHOI should begin development and documentation of a formal cruise data set for WHOI ships. The cruise data set would be defined as the package of information returned to WHOI upon completion of each leg or vehicle lowering, and would consist of an organized, complete, documented, and standards-based collection of data and metadata. A strategy for this development is shown in Appendix 3.
2. WHOI should expand and enhance the GeoBrowser technology (currently used on WHOI ships, Alvin, and ROV Jason II) to provide science users with better tools to enrich at-sea productivity and data access/availability post-cruise. Such tools would include developing web-based application software for services such as Dredging, Coring, and CTD operations and provide scientists with easy to use forms, automatic metadata generation, and summary reports; improved real-time displays; scientific cruise summary reports; improved capability for searching and retrieving data; etc. At the same time, efforts should be undertaken to improve

interoperability of the GeoBrowser systems with other systems such as Geographic Information Systems like Roger Goldsmith's systems, WHOI data portal efforts, Scripps SIO explorer, etc. using standards-based protocols like WMS, XML, and Web services. A strategy to develop and implement these improvements is shown in Appendix 4.

A timeline is shown in Appendix 5 outlining these efforts. By the end of year1, recent ship cruise datasets will be reviewed, data standards identified, and procedures defined. At the same time, new enhancements will be made to the GeoBrowser technology allowing rapid development of new applications with greater interoperability capabilities. The Ship DataGrabber system will be expanded to include features that assist in operations such as dredging and coring. Greater data access and availability will be achieved via improved capabilities for searching and retrieving data on-shore, and greater interoperability with systems internal and external to WHOI will be implemented. At the end of year2, the cruise dataset procedures will be deployed and integrated in with existing systems on the ships, and new composite real-time displays will be developed and deployed. Year3 will be for operations and maintenance support with additional efforts for any newly needed scientific tools and displays as well as investigating getting historical cruise datasets organized and available on-line.

3. Data Portal Report (K. Bice, C. Chandler co-chair, D. Fino co-chair, J. Fredricks, N. Galbraith, R. Groman, M. Lamont, M. Rioux, A. Shepherd)

Report Summary

Oceanography has entered a new era of e-Science in which research is increasingly done through distributed, Internet enabled, global collaborations that use very large data collections, tera-scale computing resources and high performance visualization. Currently at Woods Hole Oceanographic Institution (WHOI), there is no institution-wide system for cataloging, organizing or accessing Institution data repositories. The repositories that do exist are often difficult to locate and if located, require navigation of a potentially unfamiliar system for each repository. Further, the lack of a controlled vocabulary and metadata standards hinders the Institution's ability to archive and share data.

As stated in the July 2004 Access to the Sea Task Force Report, "WHOI has fallen behind in its involvement with scientific data management and cyber-infrastructure developments, and it is not well positioned to manage current or legacy data or handle the vast volumes of data associated with planned ocean observing networks".

To prepare for increasing funding agency, scientific, educational and public demands to access data and to increase prestige, WHOI must develop a comprehensive data and information management vision—one that includes a unified data interface as well as recommended best-practices for data management—especially with regard to metadata. To develop this vision, the Ocean Informatics and Cyber-infrastructure Steering Committee was formed and consequently created three subgroups to concentrate on specific data and information management needs.

The Data Portal Working Group (DPWG) was charged by the Cyberinfrastructure Committee to recommend processes and infrastructure needed to develop a data portal — a uniform, interactive gateway to scientific data — to facilitate access to oceanographic data with special focus on WHOI data.

This report recommends six actions to facilitate development and success of a WHOI data portal system. These actions include:

1) Establish Leadership and Coordination
2) Design and Implement a Data Portal Pilot Project
3) Compile Institution Data Inventory
4) Identify Metadata and Interoperability Trends
5) Implement Outreach, Communication and Staff Development
6) Provide Project Evaluation and Long-term Planning

3.1 Introduction

Earth science researchers have begun to make use of coordinated resource sharing through increasingly powerful information technology. Theory, experimentation and teaching are being facilitated by vast amounts of sensor observations and model output, for example. Ultimately, however, the utility of large complex datasets in oceanographic research depends on the existence of a flexible, secure, coordinated infrastructure within institutions such as WHOI that produce or archive such data. Effective science data management has therefore become a mandate of funding agencies and scientific, educational and public audiences.

The Access to the Sea Report (July 2004) concluded that WHOI has fallen behind in scientific data management. In response, the Cyberinfrastructure Committee was formed in October 2004. The committee identified several components, based on working knowledge and previous reports (see Appendix 2), that are required for success of an Institution-wide data management effort. One major component was the development of a data portal, an interactive gateway that facilitates access to data repositories. To address the complexity of building a data portal, the Cyberinfrastructure Committee created the Data Portal Working Group (DPWG).

3.1.1 What is a Data Portal?

A data portal is a single location where a user can access data repositories that exist in a variety of locations. Depending on the accessibility of known repositories the portal can provide the user an easy interface to view, extract and analyze data and model results. Appendix 6 provides a list of example data portals.

Figure 1 illustrates some of the basic components needed for a robust data portal implementation.

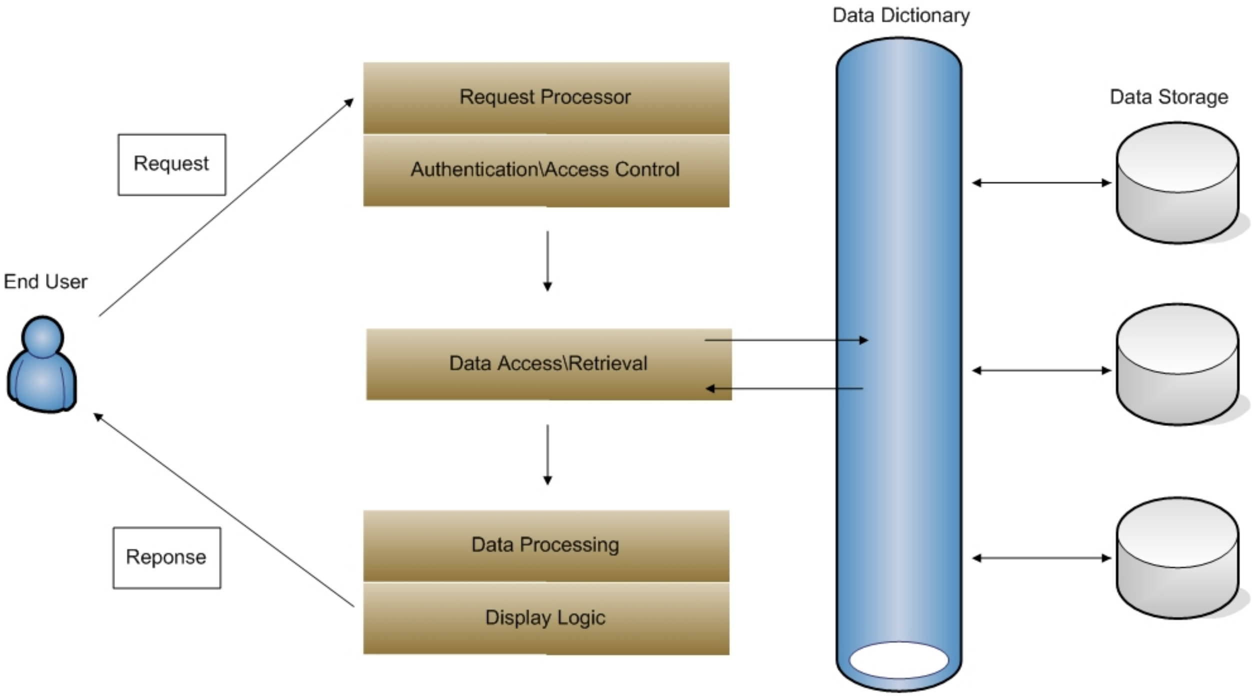


Figure 1. **Portal Request Life Cycle.** All requests made through the portal are first processed to determine what data is being requested. Once the data has been identified, the portal determines if the user has the privileges to access the data. If so, the next step is to determine who owns the data and where it is stored, and retrieve the desired result set to prepare it for processing. Next, the data is translated from the retrieved format at the Data Processor into something that can be displayed by the portal. After the data is

formulated, the portal wraps the data in display logic that specifies visual formatting of data. Finally, this visual display of the data is passed back to the user.

Two major considerations in the portal design are system interoperability and scalability. Interoperability is defined as the ability of two or more systems or components to exchange and subsequently use information[1]. The ability to access, utilize and integrate different data sets both within and outside of the Institution is a major attribute of the data portal and distinguishes it from a simple list of data collections. Interoperability is often achieved by opting for standards-based solutions wherever possible.

Scalability, the capability of a system to maintain performance under an increased load when resources are added, will allow the portal to remain effective as the number and diversity of data sets grows over time. Scalability will allow for our recommended pilot project to serve as the beginning of a long-term portal that can handle almost any dataset generated at WHOI. Other important considerations include data governance and data distribution policies, including user authentication and data use constraints.

At present WHOI.edu maintains a simple data portal, the "Digital Data Center"—a page launched in November 2004 that allows users to link to individual existing data repositories. While the data center page offers limited functionality, it has proven a demand for access to WHOI data repositories from a variety of audiences. From November 2004 to January 2005 there were approximately 3,100 visits to the page with 15% of these visits from WHOI scientists and staff. Figure 2 shows a breakdown of visits by domain.

To build on this initial step the Institution must develop a long-term strategy for organizing and providing access to its data repositories.

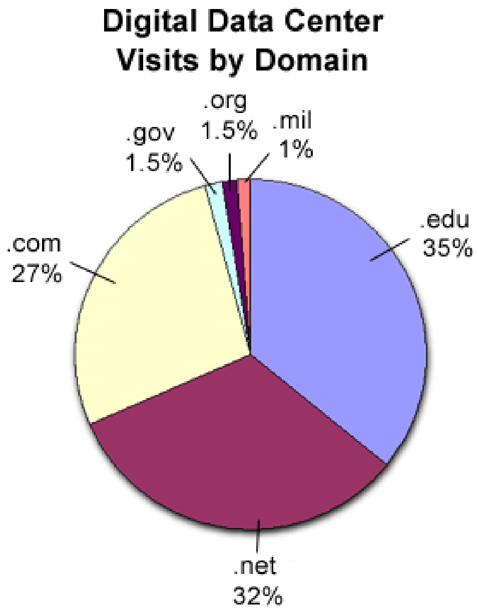


Figure 2. From November 2004 to January 2005, the top visitors to the Digital Data Center page were from educational (.edu) institutions. Users entering on network (.net) and commercial (.com) domains comprised 59% of visits—these often represent users visiting from home or from commercial offices. A small percentage of visits were from government (.gov), organizational (.org) and military (.mil) domains.

3.1.2 Recommendations for Data Portal Development

The DPWG recommends six strategies that facilitate further development of a WHOI data portal. These recommendations include:

- Establish leadership and coordination for ocean informatics and cyber-infrastructure efforts at WHOI
- Design and implement a pilot project to further investigate scientific need and infrastructure and processes required of a data portal

[1] The Institute of Electrical and Electronics Engineers Standard Computer Dictionary

- Compile a complete data inventory of all data resources at WHOI, conduct user testing, evaluate current data portals, and survey WHOI scientists and students to assess their need for a data portal
- Identify metadata and interoperability trends
- Implement outreach, communication and staff development for long-term success of a data portal, and
- Provide long-term planning and evaluation of data portal activities

3.1.3 Challenges

Funding: The current ocean science research funding mechanisms do not readily support institution-wide data management efforts. Therefore, one of the main challenges will be to identify funding sources for this initiative. It will be important to monitor announcements from funding agencies regarding cyber-infrastructure and to anticipate funding agency mandates pertaining to data access and "ocean informatics".

Community Education*:* It will be important to accurately determine what features WHOI investigators want in a portal and to demonstrate that a data portal will serve their needs.. Success will depend on continuing education of the WHOI community about best practices in information technology and their effective integration into ocean science research.

Evolving field*:* Data management and information technology are evolving quickly. Using community-wide standards and, whenever possible, open source solutions will help keep the data portal project from becoming outdated.

3.2 Recommendations

In the pages that follow the following six recommendations are presented:

1) Establish Leadership and Coordination
2) Design and Implement a Data Portal Pilot Project
3) Compile Institution Data Inventory
4) Identify Metadata and Interoperability Trends
5) Implement Outreach, Communication and Staff Development
6) Provide Project Evaluation and Long-term Planning

Recommendation 1: Establish Leadership and Coordination

Data portal design is a complex, cross-disciplinary and relatively new concept, and success will require innovative solutions originated by a diverse team of contributors. The requisite technical expertise already resides at WHOI in staff members with training in graphic design, data and information management, library and computer science and with years of practical experience utilizing that expertise to facilitate ocean science research.

The working group recommends that WHOI identify an individual or group to provide the guiding vision for information and data management at WHOI. It will also be important for that individual to identify appropriate funding resources and coordinate the various information management activities at WHOI. Information architecture will continue to happen at WHOI even without a guiding vision, but we strongly believe leadership and proper coordination will yield a more desirable result.

Recommendation 2: Design and Implement a Data Portal Pilot Project

We propose that a small pilot project be undertaken with institution funding to fully investigate a WHOI Data Portal. Using currently evolving standards, the working group recommends creating a web-based system that could potentially provide access to a wide variety of data created or used by WHOI. The pilot portal project should consist of an online resource discovery tool providing: access to data repositories and information environments through a variety of interface methods; the ability to dynamically interact with the data, create custom subsets and visual products; download capability; and tools for analyzing and manipulating data.

Objectives

- Provide a proof-of-concept for a WHOI data portal
- Demonstrate the value of a data portal
- Identify desired functionality
- Provide leverage for obtaining external funding
- Stimulate new research directions

Recommended Actions

- **Fund the pilot project with Institution resources.** This project should move forward with Institution funds because it would be difficult and time consuming to apply for funding from outside agencies for this task. It is imperative that sufficient funds be allocated to permit an adequate, on-going evaluation of the pilot program.
- **Select the pilot project team.** Key team roles include: Project Manager, User Interface Designer, Application Developer, Dataset Evaluator and data managers.
- **Identify and prioritize potential datasets.** Select a few datasets that are already available on WHOI web servers, preferably in several departments, that represent the broad spectrum of types and formats of data in use at WHOI. The datasets selected for the pilot study would ideally have complete metadata, quality control information, and be in the public domain.
- **Identify desired data portal features**. The pilot project itself should provide a good starting point for this identification process and could then be supplemented by other methods, such as a web-based survey, interviews with specific target groups, researchers and support staff. Identifying features of the data portal will develop the functional specification document for a larger project. Existing data

portals should be reviewed and feedback from surveys of and interviews with WHOI personnel should be documented.

- **Design and build the pilot portal.**
  - Identify criteria that define a successful pilot project
  - Investigate interoperability trends and existing software and systems
  - Draft an evaluation plan that identifies milestones at which the pilot project is reviewed and critiqued
  - Build the portal using an iterative process of design, implementation, testing, and evaluation
  - Integrate the results of the pilot portal into a functional specification document for the long-term portal
- **Evaluate progress.** The iterative evaluation process should continue until a successful scalable pilot project has been completed. The results of this process should be integrated into the functional specification document for a persistent WHOI Data Portal.
  - Aside from assessing user satisfaction within the Institution, it will be important to evaluate the ability of the portal to interact with similar projects being developed around the world.

Recommendation 3: Compile Institution Data Inventory

In order to fully implement a WHOI data portal it will be necessary to determine what historical and new data should be accessible via WHOI's data portal, gain familiarity with existing data management operations and develop an understanding of state-of-the-practice data servers. This evaluation of types and magnitude of data as well as current data management systems' operations will provide guidelines for the design, development, and implementation of the portal. Also, by meeting with WHOI scientists and staff to discuss their data and its use, we begin to educate them about the potential benefits of a coordinated, institution-wide data management effort.

Objectives

- Catalog current WHOI data repositories, identifying types, volume and complexity
- Evaluate internal and external data management systems
- Educate the WHOI community about data management effort underway
- Identify and prioritize need and requirements for a data portal system

Recommended Actions

- Review reports from recent committees (Appendix 2)
- Locate internal data repositories and evaluate need for a data portal via web-based survey, e-mail request and personal interviews

- Identify and evaluate external data repositories to which WHOI PIs have been required to contribute data and those frequently accessed by WHOI Investigators
    - See Appendix 6: Data Portal Examples
    - See Appendix 7: Sample Data Repository Evaluation Form
- Identify and evaluate new data portal applications

Recommendation 4: Identify Metadata and Interoperability Trends

Funding agencies are increasingly stressing the importance of making one's data accessible to other researchers, managers, educators, students and the community at large. In order to facilitate this and the goal of interoperability, it is necessary that information about the data (metadata) be prepared and made available in both human and machine-readable form.

A metadata schema based upon existing and emerging oceanographic metadata standards should be developed for the WHOI Data Portal. The schema should be extended as needed to support WHOI data and features to be included in the portal.

Objectives

- Develop metadata schema for data portal use based on existing and emerging standards
- Increase awareness at WHOI of metadata standards and concept of interoperability
- Identify technologies to support interoperability

Recommended Actions

- Research metadata and interoperability trends (see Appendix 8: Recommended Resources)
- Increase collaborations with partner institutions, designed to foster data system interoperability (MOU, November 2004)
- Define and adopt a standards-compliant Marine Metadata schema

Recommendation 5: Implement Outreach, Communication and Staff Development

As funding agencies demand more effective systems for archiving, publishing and distributing data, the Institution should act to support the changing needs of staff and students. The data portal can be an important tool to meet these needs. To involve WHOI investigators and staff in the development and success of the data portal, the working group recommends ongoing outreach and communication, and support of staff development.

Objectives

- To foster interest and expertise in data management issues, data availability requirements, and new technologies
- To encourage investigators, students, and staff to participate in the data portal project

- To encourage interaction with outside institutions working in data management
- To ensure the data portal project remains an effective tool for WHOI
- To communicate the use of the data portal as a teaching and learning tool

Recommended Activities

- Support the pilot project and data inventory process
    - The data inventory process and the pilot project will provide a good beginning for educating WHOI investigators about data management solutions.
    - The data gathering process will allow feedback from WHOI investigators and staff on their opinions and preferences, and can be used to educate them about the issues at hand.
    - The pilot project will show by example the importance of using existing standards.
- Sponsor in-house workshops and short courses on data management, metadata standards, and tools being developed to increase interoperability
- Identify and sponsor in-house venues for outreach and staff development, including ITAC presentations, WHIT seminars, departmental meetings, IEG, and Buoy Lunch talks.
- Develop technical libraries stocked with up to date references

Recommendation 6: Provide Project Evaluation and Long-term Planning

The working group recommends ongoing and final evaluation of the pilot project, data inventory phase, and communication and staff development activities. We also recommend the development of a long-term plan that documents steps and resources needed for the pilot project's further development.

Objectives

- To bring together all the accumulated knowledge and experience
- To determine if and why the project succeeded or failed
- To better assess the potential impact of a data portal on WHOI staff, students and external users
- To recommend implementation and design choices most likely to lead to a successful long-term WHOI data portal
- To communicate to the WHOI administration requirements for further development of the data portal

Recommended Actions

- Before pilot project implementation, document criteria it should meet to deem it a success. These criteria may evolve over time, but recording the pilot study's initial goals will be a useful exercise.

- Conduct ongoing evaluation of the pilot project, data gathered, and communication, outreach, and staff development activities
- After completion of the pilot project and data inventory, complete a final evaluation by interviewing, user testing, and surveying data contributors and data users
- Based on the ongoing and final evaluation develop a long-term plan that recommends steps and resources needed for the long-term development of WHOI's data portal

Specific questions that should be answered by the evaluation include:

- Are staff receptive to the idea of a long-term WHOI data portal?
- Are WHOI staff likely to voluntarily contribute to the portal?
- Do staff and students see a WHOI data portal as something likely to save them time or increase productivity? … something likely to provide access to new resources?
- Are the potential positive impacts of a WHOI data portal primarily internal to the Institution, or will external users see substantial benefits as well?
- Is a data portal likely to add to the prestige of the Institution?
- What is the estimated cost for development of a full-scale data portal at WHOI?
- What is the estimated fixed cost of maintaining a full-scale portal?

APPENDICES

Appendix 1: Committee and Working Group Members

Cyberinfrastructure Committee

Cyndy Chandler, MCG
Robert Detrick- ex officio, Marine Ops
Danielle Fino, Communications
Arthur Gaylord, chair, CIS
Roger Goldsmith, CIS
Melissa Lamont, Communications
Peter Lemmond, G & G
Steven Lerner, AOPE/DSL
Andy Maffei, CIS
Ralph Stephen, G & G

Architecture and Infrastructure Working Group

Arthur Gaylord, chair, CIS
Roger Goldsmith, CIS

Shipboard Data Working Group

Robert Detrick, ex-officio, Marine Ops
Peter Lemmond, co-chair , G&G
Steven Lerner, co-chair, AOPE/DSL
Andy Maffei, CIS
Dennis McGillicuddy, AOPE
Debbie Smith, G&G
Maurice Tivey, G&G
Barrie Walden, AOPE/OSS

Data Portal Working Group

Karen Bice, G&G
Cyndy Chandler, co-chair, MCG
Danielle Fino, co-chair, Communications
Janet Fredricks, AOPE
Nan Galbraith, PO
Bob Groman, Biology
Melissa Lamont, Communications
Maggie Rioux, Library
Adam Shepherd, CIS

Appendix 2: Prior Data Management Committee Reports

There have been several excellent reports prepared recently by various Ad Hoc WHOI committees which address the topic of data management. Many of the recommendations listed in the earliest report from May 1999 are reflected in the latter reports and are still valid today.

May 1999, Final Report of the Ad Hoc Scientific Data Advisory Committee

July 2004, Access to the Sea Task Force Report

(Available from Ruth Goldsmith, rugoldsmith@whoi.edu)

August 2004, WHOI Information Technology and Advisory Committee, Data Management Working Group
http://www.whoi.edu/committees/ITAC/internal/pdf/Data_Management_Update.pdf

Appendix 3: Development of Cruise Data Set

This project would seek as its goal the development of a formal cruise data set for WHOI ships. The cruise data set would be defined as the package of information returned to WHOI upon completion of each cruise leg, and would consist of an organized, documented, and standards-based collection of data and metadata. The development process for a cruise data set would best be structured as a typical, multi-phase software development project, beginning with a Requirements and Specifications phase, then a Design and Implementation phase, followed by a Deployment and Testing phase, and ending with a Maintenance phase.

Initial Project

The initial work for this project will consist chiefly of determining the specific requirements for a WHOI cruise data set. The requirements would define what data and metadata needs to be included, how will data and metadata be organized and stored, and how a formal cruise data set will meld with current practices. The end product of this phase of the project would consist of a documented data description of a WHOI cruise data set, detailing how and what data and metadata is to be produced. If possible, a small number of sample cruise data sets (maybe two for each ship?) could be produced. The following are estimates of the tasks required the resources needed:

1. Review ship data from past 3 years.

Cruises from 2002 through 2004 from Atlantis, Knorr, and Oceanus will be reviewed to make an assessment of what data is collected on WHOI ships. At roughly ten cruises per year per ship, this should provide an adequate sample size. For each cruise leg, will catalog:

- What data and metadata was collected.
- What data and metadata was returned to WHOI.
- What ancillary data and metadata is available.
- What non-digital data and metadata was collected (cores, water samples, etc) that needs to be referenced in a cruise data set.
- What other cruise-related activities would need to be included in the cruise data set (Science cruise report, proposal numbers, etc).

A very rough estimate of the time required to complete this task could be based on one day per cruise leg, for a total of 90 days (three ships, three years, ten cruises per year per ship). This estimate would also need to include additional support (0.5 months) from current shipboard operations and data archiving personnel

Estimated Time: 5 months

2. Identify data and metadata standards to be used.

For each of the data entities identified in Step 1, a data format with accompanying metadata would be identified. These would need to be an appropriate balance of:

- Adherence to known data and metadata standards

- Interoperability with peers and peer institutions.
- Ease of implementation

This task should be relatively straightforward, and flow naturally from the results obtained in Step 1. Consultation with peers within and outside the Institution would be done chiefly via email. All of the documentation concerning existing standards and requirements are readily available on-line.

Estimated Time: 0.75 months

3. Identify pre- and post-cruise procedures needed.

The creation of a cruise data set will begin prior to a cruise, and must be of value after a cruise. The following needs to be specified:

- What information needs to be assembled prior to a cruise that will ultimately be part of a formal cruise data set.
- What are the WHOI (and non-WHOI) sources of pre-cruise information.
- What will happen to a cruise data set when returned to WHOI. The answer to this will also impact Step 2.

This task will consist chiefly of consultations within WHOI concerning availability and accessibility of cruise data set related information.

Estimated Time: 0.75 months

Beyond the Initial Phase

Following the completion of the initial phase of development, it would be expected that subsequent phases would be needed to successfully complete this overall project.

Phase 2 – Design and Implementation

This phase will identify the procedures needed to fulfill Phase 1, determine how these procedures will actually work, and then develop the software, forms, and documentation needed.

Estimated Time: 6 months

Phase 3 – Deployment and Testing

This phase will involve the deployment of software, forms, and documentation to WHOI ships and to pre- and post-cruise data facilities. The purpose of this phase will be to test if items developed under Phase 2 meet the requirements specified in Phase 1.

Estimated Time: 6 months

Phase 4 – Maintenance and Beyond

This phase will involve operational use of the items developed and tested from Phases 1 through 3. As new shipboard systems come online, they could be incorporated into the data infrastructure. This phase will be where data previously collected would begin to be brought up to current practices.

Estimated Time: 3 to ? months

Appendix 4: Roadmap for Improvements in Shipboard Data Handling

The Shipboard data committee members discussed and identified several areas that need specific improvement including metadata documentation and collection, real-time data displays, and improved search and retrieval of data for both ship and shore based systems. A plan or roadmap as to how to implement these improvements via pilot projects was discussed and is outlined below.

In developing and implementing this plan, the following “Guiding Principles” are recommended:

1) Do what we can in real-time and automate whenever possible
2) Make it easy to get data in & out of the system (ASCII whenever possible)
3) Minimize impact to ship operations & staff
4) Provide catalog of data – i.e.; what was collected and where did it go
5) Find funding for continual support for ship & shore data management systems and equipment.

Items with an asterisk in the roadmap outlined below can be done within the 1st year, and these will provide a high-level of capabilities with a minimal amount of effort. Items such as 2b & 3b (eg; integrating the seabeam or ctd equipment) could be started in the first year and then expanded into a multi-year effort for additional capabilities. A timeline for these efforts are shown in Appendix 5.

Real-Time ShipData Roadmap

1) Improve Metadata Documentation and Collection

a) Develop forms for SSSG to fill-out and expand the Ship DataGrabber system to address “What was collected”. Currently we have a very short "End of Cruise" form that could be expanded to include a checklist (methods to automate this should also be investigated). An enhanced form may also provide a means for quick QA assessment for instruments that have problems.

b) Expand the Ship DataGrabber system utilization of EventLogging systems by developing custom web-based application/forms for requested services such as Dredging, Coring, and CTD operations. As part of the development process, meet with scientific experts like Henry Dick for dredging and Jim Broda for coring to identify necessary metadata to collect and to help develop user-interface forms.

c) As part of the Ship DataGrabber system and in conjunction with the applications developed in (b) above, automatically generate cruise data summary reports such as table of dredges, cores, number of CTDs, etc. and include ship time and position for each entry.

d) Coordinate efforts for what metadata should be collected for shipboard data systems and for increased interoperability within WHOI and peer institutions.

2) **Improve Real-Time Data Monitor Displays**

a) Identify sensors/instruments where real-time QA displays can be improved.

b) Develop a series of R/T displays for shipboard investigators either web-based or via video distribution system (need to be constantly available and automatically updating).

c) A point was raised that some systems are fragile and dependent on multiple computer systems. Identify these and develop method of improvement.

d) Provide interactive viewing and plotting of ship data.

3) Improve Data Access and Availability

a) Current access to ship cruise data is by the cruise id. Expand search criteria to by location, by time, by data collected, etc. Also provide a capability to interface to a GIS system and cataloging systems.

b) Integrate external instrument/images/data with the Ship DataGrabber system, e.g., seabeam maps, GMT maps, CTD plots, etc.

c) Identify what amount of data needs to be on-line verses having pointers to the data, i.e.; 1-minute IMET data or snapshots of Alvin video sufficient, or is full resolution needed?

d) Identify the need for video streaming capability, e.g.; is there a need for video highlight snippets on-line and does WHOI's cyberinfrastructure support video streaming?

e) Include pointers to other databases like Rick's Alvin database, Scripps, Lamont, etc.

f) Include pointers to science cruise reports if available on-line

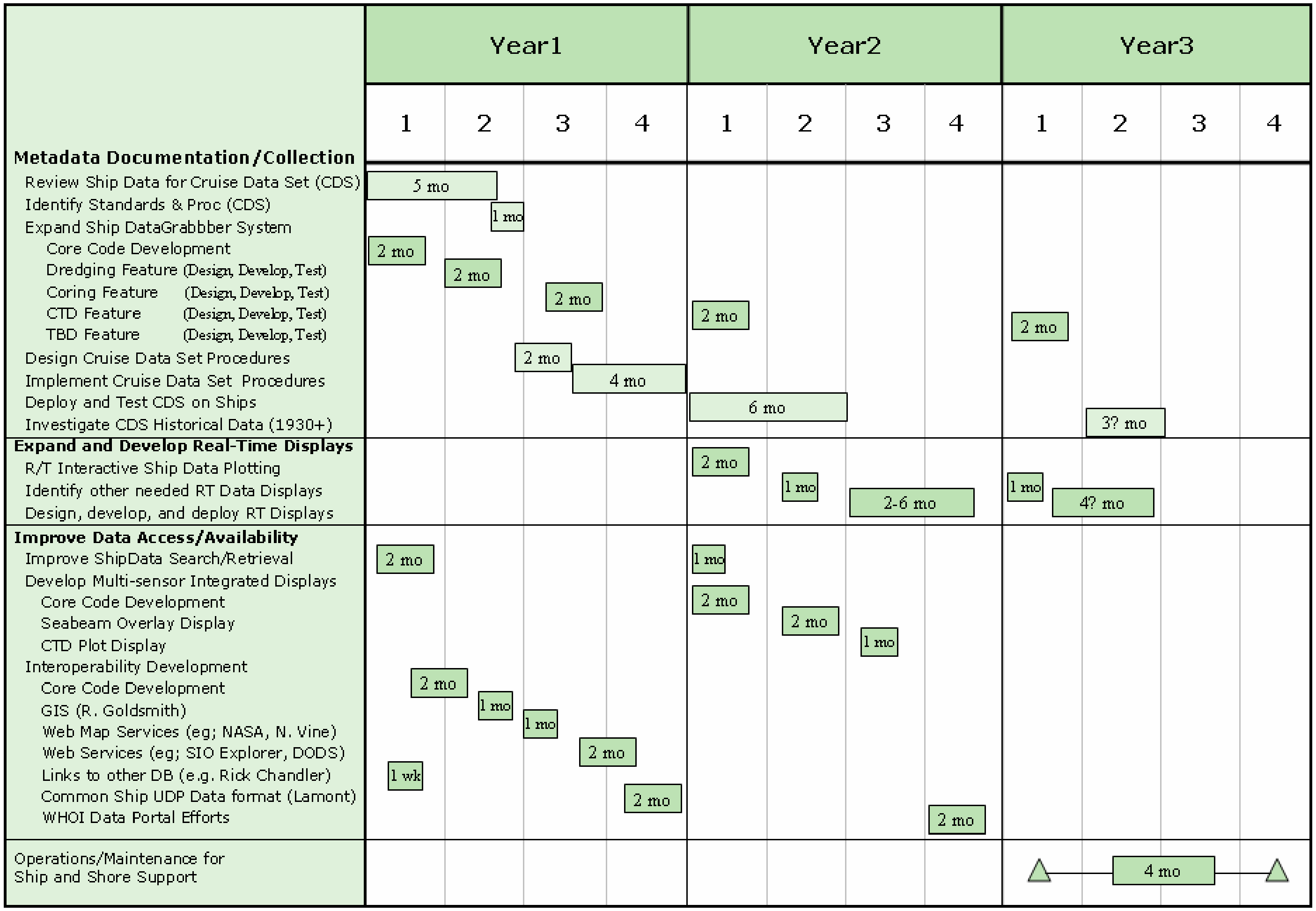
g) Develop a standard network-based UDP real-time data stream format for underway ship data. This will allow developers and users to build real-time monitoring, acquisition, and display systems.

h) Increase interoperability of the GeoBrowser systems with other systems such as Geographic Information Systems like Roger Goldsmith's, WHOI data portal efforts, Scripps SIO Explorer, etc. using standards-based protocols like WMS, XML, and Web services.

4) **Operations and maintenance**

a) Provide support for operations, maintenance, and equipment for ship and shore-based data management systems.

Appendix 5: Ship Data Working Group Timeline



	Year1				Year2				Year3			
	1	2	3	4	1	2	3	4	1	2	3	4
Metadata Documentation/Collection												
Review Ship Data for Cruise Data Set (CDS)	5 mo											
Identify Standards & Proc (CDS)		1 mo										
Expand Ship DataGrabbber System												
Core Code Development	2 mo											
Dredging Feature (Design, Develop, Test)		2 mo										
Coring Feature (Design, Develop, Test)			2 mo									
CTD Feature (Design, Develop, Test)					2 mo							
TBD Feature (Design, Develop, Test)									2 mo			
Design Cruise Data Set Procedures			2 mo									
Implement Cruise Data Set Procedures			4 mo									
Deploy and Test CDS on Ships					6 mo							
Investigate CDS Historical Data (1930+)										3? mo		
Expand and Develop Real-Time Displays												
R/T Interactive Ship Data Plotting					2 mo							
Identify other needed RT Data Displays						1 mo			1 mo			
Design, develop, and deploy RT Displays							2-6 mo		4? mo			
Improve Data Access/Availability												
Improve ShipData Search/Retrieval	2 mo				1 mo							
Develop Multi-sensor Integrated Displays												
Core Code Development					2 mo							
Seabeam Overlay Display						2 mo						
CTD Plot Display							1 mo					
Interoperability Development												
Core Code Development	2 mo											
GIS (R. Goldsmith)		1 mo										
Web Map Services (eg; NASA, N. Vine)			1 mo									
Web Services (eg; SIO Explorer, DODS)			2 mo									
Links to other DB (e.g. Rick Chandler)	1 wk											
Common Ship UDP Data format (Lamont)				2 mo								
WHOI Data Portal Efforts								2 mo				
Operations/Maintenance for Ship and Shore Support									△	4 mo		△

Appendix 6: Data Portal Examples

Below are some examples of existing web portals to earth science data. Where known, the names of the funding and supporting agencies and institutions are given. This list was compiled 1/27/05 and last updated 2/14/2005. For the most up to date version go to http://www.whoi.edu/science/GG/people/kbice/portals.htm

Earth Reference Data and Models

http://www.earthref.org/

The EarthRef.org portal provides access to data describing the geochemical make-up of all earth reservoirs. It serves, for example, databases compiled as part of the Geochemical Earth Reference Model (GERM) project. The portal is coordinated and hosted by researchers at Scripps Institution of Oceanography (SIO), Lawrence Livermore National Laboratory (LLNL), and San Diego Supercomputer Center (SDSC). (No mention is made of funding source on the site.)

MARGINS and RIDGE 2000 Data Portals

http://www.marine-geo.org/margins/
http://www.marine-geo.org/ridge2000/

These data portals provide access to cruise information and data collected during MARGINS- and Ridge 2000-funded projects. They include mapping and gridding utilities. Access to these data portals is through the Marine Geoscience Data Management System at Lamont Doherty Earth Observatory (LDEO). Funding is provided by NSF. (Note: On their web site, Ridge 2000 refers to their data portal as the "Ridge 2000 Open Data Exchange System," or RODES.)

Marine Environmental Data Inventory (MEDI)

http://ioc.unesco.org/medi/

MEDI contains an inventory of world-wide marine-related datasets within the UNESCO Intergovernmental Oceanographic Commission's International Oceanographic Data and Information Exchange program (IODE, http://iode.org/).

National Geospatial Data Clearinghouse

http://clearinghouse.esri.com

This portal allows access to 100 spatial data servers for digital geographic data for use in Geographic Information Systems (GIS), image processing systems, and other software. This is an initiative of the Federal Geographic Data Committee, a 19-member U.S. government interagency organization.

NOAAServer

http://www.joss.ucar.edu/NOAAServer/index.html

NOAAServer provides access to NOAA's nationally distributed environmental information databases.

NOAA/PMEL Live Access Server (LAS)

http://ferret.pmel.noaa.gov/Ferret/LAS/ferret_LAS.html

The Live Access Server is a portal that provides access to a growing number of geo-referenced climate data systems, including NODC's World Ocean Data Base, the National Virtual Ocean Data System, Pacific Fisheries Environmental Laboratory, the NGDC coastal bathymetry and topography data, and NOAA, NASA, Navy, DOE, LDEO, NERC, CNRS and CSIRO model output. The LAS is funded by NOAA and is operated primarily by personnel at NOAA's Pacific Marine Environmental Laboratory (PMEL, http://www.pmel.noaa.gov/).

Ocean Biogeographical Information System (OBIS)

http://www.iobis.org/

OBIS provides taxonomically- and geographically- resolved data on marine life and oceanography, access to physical oceanographic data at regional and global scales and software tools for biogeographic analysis. OBIS is supported by The Alfred P. Sloan Foundation, NSF, ONR and the U.S. National Oceanographic Partnership Program (http://www.nopp.org/). It is hosted by the Rutgers University Institute of Marine and Coastal Sciences (http://marine.rutgers.edu/).

Petrological Database of the Ocean Floor (PetDB)

http://beta.www.petdb.org/

PetDB, a Ridge2000-sponsored program, provides access to geochemical and petrological data of igneous and metamorphic rocks from the ocean floor generated at spreading centers. Funding is provided by NSF. The portal is operated by LDEO and the database is hosted by Columbia University's Center for International Earth Science Information Network (CIESIN, http://ciesin.columbia.edu/).

SIO Explorer

http://nsdl.sdsc.edu/

This portal provides access to data, documents and images from "822 expeditions of the Scripps Institute of Oceanography (SIO) since 1903." Funding is provided through the NSF National Science Digital Library program (http://www.nsdl.org/) and the NSF Information Technology Research program (http://www.nsf.gov/home/crssprgm/itr/). The portal is a collaborative effort among SIO researchers, computer scientists from SDSC, and archivists and librarians from the University of California San Diego Library.

U.S. Global Change Data and Information System

http://globalchange.gov/

The GCDIS web site provides access to data, news releases, publications and educational resources from the U.S. Global Change Research Program.

The U.S. National Data Centers and the World Data Center System

A growing number of the data deposited in the National and World Data Centers are accessed by portals such as the NOAAServer, NOAA/PMEL Live Access Server, GCDIS and PetDB. Other data may be accessible for now only through the individual data center web sites.

National Data Centers

Carbon Dioxide Information Analysis Center (CDIAC): http://cdiac.ornl.gov
Center for International Earth Science Information Network (CIESIN): http://ciesin.org
Earth Resources Observation Systems (EROS) Data Center: http://edcwww.cr.usgs.gov
National Climatic Data Center (NCDC): http://lwf.ncdc.noaa.gov

National Earthquake Information Center: http://neic.usgs.gov
National Geophysical Data Center (NGDC): http://www.ngdc.noaa.gov
National Oceanographic Data Center (NODC): http://www.nodc.noaa.gov
National Snow & Ice Data Center (NSIDC): http://nsidc.org
National Space Science Data Center (NSSDC): http://nssdc.gsfc.nasa.gov

World Data Centers

There are more than 50 centers in the World Data Center System, each funded and maintained by the host country. The World Data Center System web site is maintained by NOAA's National Geophysical Data Center. For links to individual centers and a central data search utility, go to http://www.ngdc.noaa.gov/wdc/.

Appendix 7: Sample Data Repository Evaluation Form

I. General

1) What is the scope of the repository? What data and data formats are accepted for inclusion?
2) Any quality control checks performed or is data "as is."
3) Who is responsible for the day-to-day operations of the repository?

II. Schema checking

1) How accessible is the data?
    a) Can the data be retrieved through our soon to be designed systems?
        i) Must we search the repository to retrieve a result set or would we receive full-scale data dumps?
        ii) Can we assimilate the data into our systems? [Could we provide the supplier with extra functionality that would be beneficial via assimilation?]
    b) What are the import/export capabilities?
2) Is the data model relational or organized in such a way that individual records can be indented by some key?
    a) Has the data been described at all?
        i) If the data has been described, how? Dictionary files? Relational data schema? Data tags (XML, schematic tags)?
        ii) Are the descriptions portable? If not, how difficult would they be to replicate
    b) If not, can we describe it? How much of our description capability depends on the retrieval process?
3) How well does the data model scale?
    a) If we were to implement this model for our systems, how would it perform under pressure?
    b) (RDMS vs. flat files) If RDMS is/can it be normalized?
    c) What is the total cost of ownership for the repository and its maintenance? in manpower? in dollars? in hardware?
    d) How was the repository funded? How likely is future funding?

III. Internal and External Issues

1) How is the data viewed?
    a) Do you need a special program?
    b) Is it web accessible or available over a common protocol (with or without auth)?
    c) Was the interface developed in-house or purchased?
    d) Have user tests been conducted and if so what were the results?

e) Could a non-expert user understand the interface and successfully retrieve data?

2) How does the search engine work? Does the engine search within the data or simply the metadata?

3) What type of authentication schemes are necessary?

    a) Who owns the data? Who is allowed to access the data?

    b) Is authentication a time sensitive issue? [ex. Does the data become available to a wider audience after 2 years]

    c) What type of authentication scheme should our system implement?

    d) Are there explicitly defined roles as to what is accessible? [ex. Administrator, Viewer, Time-Sensitive Viewer (2 week access -like a library card) etc.]

4) How are data obtained, described, stored?

    a) Are data accessions actively solicited or is the system completely voluntary?

    b) What volume of data (monthly, yearly, # of datasets, amount of storage, etc.) are accepted?

    c) Who develops metadata? Is a metadata standard in use?

    d) Is an archiving service associated with the repository?

        i) Are data migrated?

        ii) What is the long-term storage system?

5) How open are external sources to forming a working partnership to build/share systems?

    a) Can it be done via grid computing?

    b) What are the capabilities of Internet II and our inherent partnerships with other institutions b/c of Internet II?

    c) Where are the data stored? Who is responsible for back ups?

    d) Is it better to store data in-house? Outsource data storage with a conglomerate group of WHOI and other data partners?

Appendix 8: Recommended Resources

In the process of writing this report, the working group found these publications to be especially useful.

Information Architecture for the World Wide Web, 2nd Edition, Designing Large-Scale Web Sites. Louis Rosenfeld, Peter Morville. O'Reilly & Associates, August 2002. 486 pages. ISBN: 0-596-00035-9

Designing with Web Standards, 1st edition. Jeffrey Zeldman. New Riders Press, May 2003. 456 pages. ISBN: 0-735-71201-8

Metadata resources *(*Recommendation 4)

MBARI: Monterey Bay Area Workshop on Data Management and Visualization
http://www.mbari.org/iag/workshops/dmv/index.html

UNESCO/IOC Marine Mark-Up Language (MML) specification aligned with the W3C XML standard (http://marinexml.net/)

Project evaluation resources *(*Recommendation 6)

Information Use Management & Policy Institute

http://www.ii.fsu.edu/

Program Development and Evaluation, University of Wisconsin

http://www.uwex.edu/ces/pdande/