

# **I. Interoperable Data Discovery, Access, and Archive**

## **Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems**

### **Part III. Appendices**

**Appendix 3. Data Archive and Access**  
*IOOS DMAC Data Archiving and Access Team*

**March 2005**



**The National Office for Integrated  
and Sustained Ocean Observations**  
**Ocean.US Publication No. 6**

# Contents

<b>Vision</b> .....	209
<b>The Archive System</b> .....	210
<b>Data Receipt</b> .....	213
<b>Data Preservation</b> .....	215
<b>Data Provision and Access</b> .....	220
<b>Data Policy</b> .....	224
<b>Interactions and Partnerships with Other Data Centers</b> .....	225
<b>Cost Estimates</b> .....	227
<b>Annex A. Additional Infrastructure Costs</b> .....	229
<b>Annex B. Glossary of Terms</b> .....	230

IOOS data archiving and access will be a distributed system of interconnected archive and data centers that function collaboratively to receive and preserve the data, and provide easy and efficient access to the data. Search and discovery of data and products will be easy and will directly support the seven IOOS goals.

Archive collections range greatly in size, complexity, and importance to public and scientific needs. Currently, diverse data service paradigms are used to support access to the archives. IOOS data transport methods, metadata standards, and data discovery interfaces will be implemented in the Archive System. The result will be a system that provides more uniform access across multiple centers and that can handle all collections consistently. The data discovery component will allow access by both humans and machine.

As the amount of IOOS data steadily increases, the old and new systems of access must remain compatible in order to maintain the high levels of service and allow users to fully discover the archived data.

# The Archive System

The Archive System will use coordinated methods for data collection, quality control, archiving, and user access. The system will consist of a distributed network of archive centers, regional data centers, modeling centers, and data-assembly centers, all interconnected to provide efficient flow of data into the IOOS archive and easy access to its data and products (Figure 1). Although data may flow from observing systems to any of the four types of centers, at least one copy of each observation desired by IOOS must ultimately reside in an IOOS archive center. For the purpose of IOOS, data will be considered in the Archive System if the following two conditions are met: (1) the data are held and access is provided by one of the System components, and (2) there are procedures in place to preserve the data at an archive center. Through this approach data will be under IOOS management early in its life cycle and thereby maximize the amount securely archived and uniformly accessible. The IOOS Archive System will take full advantage of the infrastructure, expertise, and historical reference data sets at existing data centers. It is probable and practical that more than one type of center may be physically collocated, for example, a data assembly center may be an entity at a national archive center. Additional resources (expertise, people, funds) will be needed to meet the expanding requirements of IOOS.

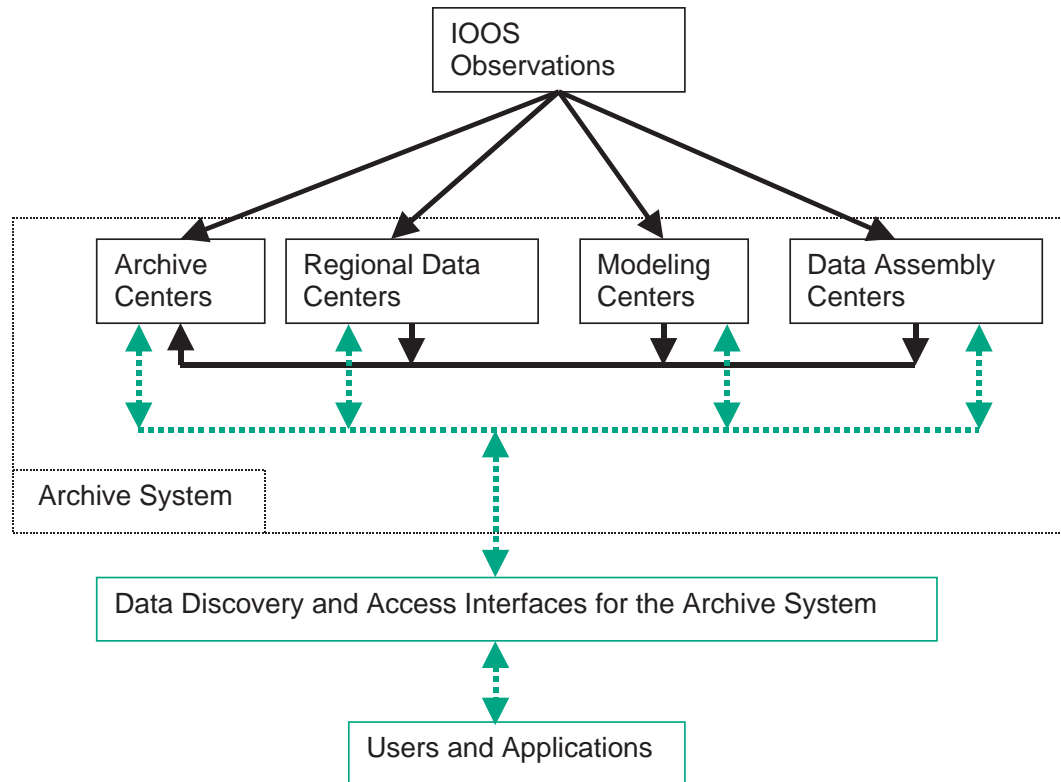


Figure 1. Primary archival (solid lines) and access (dashed lines) data flow within the DMAC Archive System of IOOS. Not shown are the secondary bi-directional archival data flow between all the centers and IOOS observations flowing directly to users and applications external to the Archive System.

## Part III. Appendix 3: Data Archive and Access

**Archive centers** are the core of the Archive System. Their mission is to acquire, preserve, and provide access to IOOS data in perpetuity. High-priority objectives include integrity and completeness of the archives. Essential functions include constant monitoring of data streams, accounting for all files and records, and frequent checks of accuracy. Metadata are equally important since they ensure that the maximum information can be derived from the data. Archive centers must have maintenance strategies that protect the data as storage media and systems change. Data stewards must constantly guard against changes in formats and software that could make accessing the data more difficult, more costly, or even impossible. Since important collections are seldom static, a significant effort is required to integrate new metadata, add improvements and corrections to the data, and make additional related historical archives easier to access.

**Regional data centers** acquire and provide access to IOOS data collected in specific geographic regions. These centers often collect a variety of physical, biological, and chemical ocean data that are used to support scientific, public, and commercial interests in the area. Resident staff may also apply quality-control measures to data and derive specialized products. Regional data centers may support long-term archives if they meet the IOOS standards for integrity and stewardship or they will systematically transfer the data to an archive center.

**Modeling centers** procure and synthesize observational data to produce products such as analyses, predictions, or hindcasts that may span a wide range of spatial and temporal scales. These centers often provide access to their products, but their mission does not include long-term archiving. Model products that are essential to IOOS goals will be transferred and preserved at an appropriate archive center.

**Data assembly centers** also obtain IOOS data and provide access to it. They typically specialize in certain types of data, and often provide quality control and data products in their area of expertise. These centers may be permanent (e.g., NDBC) or exist only for limited periods (e.g., WOCE data assembly centers). They do not provide long-term archiving, but often provide access. Distributing data assembly centers is an efficient way to acquire and process data over a wide range of disciplines, with the assembled data and products then being submitted to archive centers for long-term storage and access.

Although IOOS data may flow into the archive centers over several pathways (Figure 1), at least one copy of each set will reside in a designated archive center. Some categories of data will require that multiple copies be stored securely. When data must be duplicated, a primary and secondary data steward will be designated. The primary data steward will typically be an archive center and will provide the highest level of access. The secondary steward need not maintain full access, but will

### Part III. Appendix 3: Data Archive and Access

maintain the data at the same level of integrity. Creating separate primary and secondary archives also provides two physically separated copies of irreplaceable data while avoiding the cost of full access at two locations.

Access services for IOOS users will be provided from most centers in the Archive System. For IOOS, archive centers will expand their access services beyond current levels providing more real-time services, and enhance data discovery by using the IOOS metadata standard and data discovery techniques. When regional, modeling, and data assembly centers provide access on schedules that meet the IOOS goals, duplication of this effort is not essential for the archive centers; however, the archive centers will ultimately receive the data, provide for its long-term preservation, and provide access to the full archived data set.

Success for the Archive System hinges on center-to-center collaboration. The modeling, assembly, and regional data centers can benefit by having a secure data repository at an archive center. Conversely, the archive centers can benefit by having high-quality, useful data streams developed at the modeling, assembly, and regional data centers.

The scientific community also has an important role. The System will enable scientific endeavors that make comparisons of model and observed data, develop analyses and reanalyses data products, and provide additional quality control on the data, thereby quality checking the observing systems. The Archive System will receive these additional data products, use the discoveries to augment data stewardship activities, and have mechanisms to inform the IOOS observation subsystem about data quality concerns.

## Data Receipt

Two types of data reach the IOOS Archive System: real-time and delayed-mode. Real-time data arrive in real time or near real time, with the goal of being made available with minimum delay. High-level quality control is not practical here. Delayed-mode data arrive later than real-time data, and sometimes much later. They may be research collections that have been improved through further processing, or simply raw data collected under circumstances where prompt transmission was not feasible or needed. The Archive System will receive sets of either type that address the seven IOOS goals. All appropriate metadata should arrive with the data.

High priorities for the IOOS Archive System include ensuring that all valuable data are sent and that an exact copy is received. The data may be transferred over networks or on hard digital media. The integrity of the data must be constantly checked. Acceptable tools and procedures include:

- Receipts and reconciliation reports for transfers over networks,
- Skilled staff to review metrics (e.g., how much of the expected data was received and how much of the data set was made available),
- Byte counts, inventories of data files, and checksums of records or files,
- Test files that can be confirmed against archived data and used to verify local software,
- Accuracy relative to other data sources (i.e., whether a set of data falls within acceptable ranges or compares acceptably with other data known to be correct).

Unfortunately, data transmissions can fail, or data can change unexpectedly. Because both can significantly degrade the value of the data, it is important to verify data as soon as possible after receipt. Detecting problems early will minimize their harm. Cooperative efforts between the data providers and the archiving centers are sometimes required to repair an archive. Having expert contact persons available is important in evaluating and resolving these problems.

The demand for timeliness implied in the IOOS goals means that data and metadata must be made available as soon as possible after they are received. Because metadata are harder to handle than bulk data, they need to be checked and standardized when they are received (and possibly supplemented with information garnered from reading the data), and the data catalogs updated. These steps will allow the metadata and data-discovery techniques to reveal the fullest and most current information to users.

Guidelines must be drafted so that providers developing new data streams can select formats and metadata that can be easily integrated into IOOS. Specifications should be set as part of the IOOS data-transport, metadata, and data-discovery components.

### Part III. Appendix 3: Data Archive and Access

Data in Archive Systems are commonly resubmitted and replaced. IOOS standards for metadata will allow different versions of the same data and metadata to be traced by means of information on lineage and version. The number of old versions of data to be preserved remains an open question, however. Managers of data centers need a formal procedure to help them resolve this difficult issue. It will be carefully considered, probably with representatives of the scientific community and possibly input from the public during the early implementation of IOOS.

The broad range of data to be included in IOOS (physical, biological, chemical) means that many different native data formats will be used. Data providers should use only established, fully documented formats, which the data-transport methods will handle and so make the format issue transparent for the user. Nonetheless, the data centers will need to accommodate native formats from numerous providers, especially in the beginning of the IOOS Archive System. Because these formats will be somewhat discipline specific, each center will not necessarily have to be proficient in every format.

In contrast to the diversity of data to be collected, metadata will all have to meet a common standard, or at least be interpretable through a filter as a standard, so that they can be accessed and interpreted by all of IOOS.

Archive centers will consider accepting data in all formats, with the following understandings:

- Unique specialized formats (such as occasionally found in research or field data) are significantly more expensive to manage. Standard formats are preferred.
- Proprietary formats (with undisclosed internal structure and typically with proprietary software) are unacceptable for long-term archiving and are explicitly discouraged because they would have to be converted to public formats accessible with open-source software. Such conversion is expensive and may corrupt the data.

Software for accessing each native format must be kept fully operational at the centers. Because the inevitable evolution of formats can quietly create discontinuities in data, even in time series from a single source, centers must track these changes and maintain software that will access all segments of data sets. This software will also provide further documentation of data sets and changes in their lineage.

Another serious consideration for the Archive System is data-compression software. File-compression techniques used for transferring IOOS data (or any other kind) should always use standard protocols with open documentation, such as GNU zip. File compression is important for efficiently transporting and storing data. Decompression is equally important because the long-term mission of the archive centers requires them to reproducibly decompress a data set over its entire lifetime.

# Data Preservation

All four component data centers of the IOOS Archive System will be responsible for acquiring and providing data, but only the archive centers will be primarily responsible for preserving data long term (i.e., much longer than the typical funding period of an oceanographic research project or the career of a principal investigator). To qualify as an archive center, a data center must be able to perform the following functions related to data preservation:

- Create and manage multiple copies of the data and metadata,
- Verify and generate metadata as well as preserve it with its associated data,
- Frequently check data integrity,
- Plan for evolution of technology.

Archive centers must be able to create and manage one or more copies of all IOOS data and metadata, both online and offline, according to the specified IOOS data category and according to NARA and other Federal guidelines. Initially, a working group, with balanced representation from the science and archive management communities, will categorize each extant IOOS data set. The IOOS categorization will become part of the standard metadata. As new data sets become available they will be categorized by the same criteria and requirements.

The selection of data category requires careful consideration, because it determines the minimum time period for preservation and the minimum number of copies that must be maintained. Table 1 summarizes the four data categories and the number of archival copies required to meet the minimum IOOS Archive System standards.

- Irreplaceable Data—Maintain two copies in separate archive centers in perpetuity.

Irreplaceable data have the most stringent maintenance requirement because these data are unique and impossible to retake. All satellite and *in situ* measurements and some difficult-to-reproduce data products (e.g., long-term global atmospheric reanalysis or primary productivity fields from blended *in situ* and satellite data) are in this category. Historically, irreplaceable data have not always been archived in perpetuity (e.g., to reduce data storage and prepare for subsequent calculations observed ocean profile data were discarded after they were reduced to estimates at standard levels). Modern technologies now allow for all observational data to be preserved so current and future researchers can derive products based on the original data.

The two copies of irreplaceable data will be preserved in separate facilities under independent data management. One facility will be designated as the primary archive center for a particular data set, and the other as the secondary archive center. The primary and secondary archive centers storing irreplaceable data may operate as mirror sites, both offering the same level of access,

### Part III. Appendix 3: Data Archive and Access

Table 1: IOOS Data Categories for Archiving and Access.

Data Category	Data Description	Examples	Minimum Number of Archival Copies
Irreplaceable	Observational and research-quality data that cannot be reproduced or easily regenerated	<ul style="list-style-type: none"> <li>• Raw, ancillary satellite observations</li> <li>• Instrumental measurements</li> <li>• Biological samples</li> <li>• Model reanalyses</li> <li>• Complex merged data analyses</li> </ul>	Two
Replaceable	Derived from irreplaceable data, can be regenerated through systematic processing	<ul style="list-style-type: none"> <li>• Calibrated satellite radiances</li> <li>• Simple composites or analyzed data</li> </ul>	One
Perishable	Real or near-real-time data; typically replaced by higher-quality data	<ul style="list-style-type: none"> <li>• Direct broadcast satellite data</li> <li>• Operational analyses</li> <li>• Quick-look analyses based on uncalibrated or incomplete data</li> </ul>	One
Virtual	Data provided through on-demand processing	<ul style="list-style-type: none"> <li>• Subsets from GUI</li> <li>• Analyses from a Live Access Server</li> </ul>	Two*

\* Original generation algorithms and documentation only.

or one as the exclusive access center and the other as a “deep” back-up center (e.g., a regional data center could serve as a secondary archive center). Mirrored sites will reduce the risk for archive down time and maximize data availability, but will increase the data management cost.

- Replaceable Data—Maintain one copy (residence time in the archive will vary with replacement cycle).

Replaceable data are directly derived from irreplaceable data and are often more readily useful (e.g., weekly gridded SST from AVHRR satellite measurements). Only a single data copy is required because replaceable data may be systematically regenerated. However, having several copies at multiple centers will enable greater accessibility, which is especially critical for generating data products that are necessary for timely decision-making.

### Part III. Appendix 3: Data Archive and Access

- Perishable Data—Maintain one copy until higher-quality data are available.

Most perishable category data are real-time data derived from uncalibrated measurements or products provided at reduced spatial and temporal resolution. Perishable data are undoubtedly valuable data in the near term (e.g., quick-look analyses and forecasts based on incomplete and uncalibrated, *in situ* measurements), but they lose value when quality-controlled measurements and full-resolution products become available. When decision-critical data products are derived from data in this category, and it is necessary to reproduce the data product, the perishable data may inherit an extended term for data preservation that is not obvious for the original data alone.

- Virtual Data—No copies of the data are necessary, but an archive center and the virtual data provider should maintain separate copies of generation software and documentation.

Virtual data are those derived from the other data categories by “on demand” systems. The systems may include data subsetting, data analysis, and format conversion capability. Automated data access for applications through IOOS data discovery and transport methods are in this data category. These data products need not be preserved in the Archive System. However, the complete algorithm and documentation, including source code, should be saved by the providing center and must be saved by an archive center for future reference. Data analysis algorithms, format conversion standards, and the source data identification must be determinable long after a user generates the virtual data and even after the software has changed and may no longer be operable.

Metadata come in many forms, including: use metadata (the semantic and syntactic information about a data set); discovery metadata (standard structured information describing a data set); and documentation metadata (bibliographic information about documentation associated with a data set). The capability to discover and accurately use data, in the long term, relies heavily on the available metadata of all three forms. As such, metadata collections throughout the Archive System are critically important.

Documentation metadata have been commonly collected in the past and will continue to be significant for IOOS. New potential to improve data management, user discovery and access, and application access is possible through the forthcoming IOOS standard for metadata. Representatives from the Archive System will participate in the metadata development for IOOS and work to transition current systems to the new standards that will make data retrieval more effective. For example, IOOS-wide data catalogs will enhance data discovery (by both humans and machines) across

### Part III. Appendix 3: Data Archive and Access

data centers, and data service catalogs (see description in the Data Provision and Access section) will identify where the data are available and how they can be accessed. Some Discovery metadata elements that are particularly important for managing the Archive System are:

- Data set lineage history (e.g., which irreplaceable data set was used to create this current data set),
- Data category specification, which determines the storage requirements,
- Release date, which is the date to remove temporary restricted access,
- Version number and description of the version number,
- Description of the file naming convention,
- Unique IOOS-wide data set name or identification,
- Mechanisms for correct publication citation and reference tracking.

Because some archived data sets go through numerous incremental updates, modifications, corrections, and occasionally, full replacements, the metadata strategy must be dynamic so the centers can easily maintain accurate information and so the users have complete ancillary information. Furthermore, as data are referenced in publications it is desirable to have bibliography tracking capability. This would provide an end-to-end lineage record, starting with the measurements or computation through the change and modification history and eventually to established scientific or public knowledge. Consequently, the data set could be properly cited in the literature and the IOOS program would gain another metric to measure success.

A lapse in data security could quickly result in the loss of irreplaceable data. The Archive System will guard against unrecoverable data loss by making data integrity (or security) a primary objective. As with data received from each provider, byte counts and checksums will be calculated and used to verify that the data are uncorrupted when transmitted between data centers. These quantities will again be calculated after every internal process at the archive centers, and then recalculated periodically on all archived data to protect against such problems as hard disk failures, media degeneration, incomplete file transfers, and malicious hacking. Virus checks will be performed on the data before archiving, then periodically on all data kept online.

Long-term preservation requires that all archive centers have a plan to address evolving mass storage technology. The plan must include strategies for storage media migration. Current systems are based on magnetic tape cartridges, which typically have a three- to five-year life cycle, and are approaching a petabyte in size. Under IOOS these systems will grow and the rate of increase will accelerate. This growth can be accommodated in the Archive System, but will require increases in facilities infrastructure and support.

### **Part III. Appendix 3: Data Archive and Access**

The Archive System will be a cohesive set of centers that interoperate by using metadata standards and data transport methods in a system of computers, software, and networks.

Undoubtedly, the future will bring new technologies in networks, computing systems, and evolutions in software. In order to take advantage of the new technologies and software, and not disrupt the interoperability, a coordinated plan is required for handling system-wide technology infusion.

IOOS will instantiate new and parallel data sets that will augment the extant digital historical collections now at the archive centers. Focused efforts at the archive centers will be necessary to maintain continuity across related data sets while IOOS evolves. The goal is to have the broadest reference data sets possible through smooth integration of historical digital data and the new IOOS data sets.

# Data Provision and Access

Data can be accessed from any suitable component of the IOOS Archive System (Figure 1). By querying the system with its data-discovery interface, users or applications can discover what data are available. The data may then be pulled automatically with the OPeNDAP protocol and data transport methods, or by the user from a GUI that displays the various options.

Using the OPeNDAP protocol for transporting data will allow the Archive System to provide a host of services beyond current-day simple file downloads. They include real-time subsetting, on-line analysis, reformatting, and support for GIS applications.

Although the designation of IOOS data sets is yet to be determined, the March 2002 Ocean.US workshop defined the most important variables in various disciplines. Relevant, extant data sets will be identified and potential new data sets and products determined and prioritized during the early phases of implementation.

Not all access requirements fit all data sets. As the IOOS grows, its services will evolve. To accommodate this evolution and to provide service to the expected broad IOOS user community, access services will be tailored to data sets. This can be illustrated conceptually as a matrix of data sets and services (Table 2).

Table 2. Conceptual matrix of data access services for different data sets at different components of the IOOS Archive system. Note that data set 3 is offered at two centers, but with different services.

Center	Data set	Core Services			Extended IOOS Services						
		FTP	HTTP	OPeNDAP	Spatial Subset	Parameter Subset	Temporal Subset	Temporal Aggregation	OpenGIS Map	Online Analysis	Online Ordering
Center 1	Data set 1		X	X	X						
	Data set 2		X	X	X	X		X	X	LAS	X
	Data set 3		X	X							X
Center 2	Data set 3	X	X	X						GrADS	
	Data set 4		X	X			X				

## Part III. Appendix 3: Data Archive and Access

The core protocols include FTP, HTTP, and OPeNDAP. Most IOOS data sets are expected to be available in at least one, and ideally two or three, of these protocols. As the IOOS standard transport protocol, OPeNDAP should be used whenever possible. The characteristics for each of these core services are:

- FTP—Direct downloads of data files, unrestricted public access, and no application support,
- HTTP—Direct downloads of data files, restricted or unrestricted access, and no application support,
- OPeNDAP—Application-layer protocol that supports a number of data storage formats and allows a number of client applications to access data transparently. Importantly, it can allow additional extended services.

As data sets increase in size and complexity, extended services will offer users helpful options for accessing data. Although setting up OPeNDAP for accessing data sets will take more effort initially, it will be cheaper to maintain in the long run. It is most advantageous for data sets that are accessed frequently. Data centers can use OPeNDAP to offer the following extended services:

- Spatial subsetting—Extracting spatial sub regions from data sets for larger geographic areas,
- Parameter subsetting—Extracting one or more variables from data sets containing many variables,
- Temporal subsetting—Extracting short periods from data sets covering longer periods,
- Temporal aggregation—Creating a longer time series from data files for shorter periods,
- GIS products—Depicting data projected, interpolated, and rendered onto a map with GIS protocols,
- Online analysis—Analyzing online by using tools on the data server such as the Grid Analysis and Display System (GrADS) or the Live Access Server (LAS). The resulting data or graphics can then be downloaded.

There will always be some data sets stored offline, typically those that are too large or accessed too infrequently to justify the cost of storing them on line. Nevertheless, they will still be kept accessible and discoverable through the data-discovery interfaces. This access to off-line data will likely be initiated by on-line ordering. On-line ordering, which is an extended service, is a mechanism by which data are ordered and then picked up or delivered later. Normally a WWW GUI is presented to the user, who then specifies the data needed. This service deviates somewhat from the IOOS objective in that it is neither standardized nor transparent.

### Part III. Appendix 3: Data Archive and Access

The IOOS DMAC methods of transporting and discovering data and metadata will evolve during its early years. They will eventually set the foundation for increased data usage through “data mining,” which is currently a research endeavor focused on accessing data and automatically searching out suitably described patterns in the largest data sets.

Data latency is a requirement that links the users’ needs to the archiving costs. For IOOS, access latency is defined as the time between the earliest primary observation (not counting ancillary data) in a data file and the availability of that file to users. For example, a field of monthly mean SST has a minimum latency of one month, whereas broadcast satellite data and buoy observations used in operational modeling could have a latency of only minutes. Affordability is a factor here because low latencies are expensive. Requirements for low latency often come with requirements for high availability, which together imply around-the-clock staffing and special redundancies in hardware. For IOOS data users, latency requirements need to be assessed and suitably defined in the metadata.

Unrestricted access is a first principle for non-commercial IOOS data sets. Restricting access goes against this principle and is not encouraged. A policy on this issue will have to be established when IOOS begins. There are circumstances where access may have to be temporarily restricted, however, typically beginning when the data are collected. Such circumstances include:

- Proprietary embargo—Data are available only for sale from commercial companies (e.g., the initial two-week embargo on SeaWiFS data),
- National security—Data are available only for defense purposes,
- Calibration and validation—Data are available only to the science team while they calibrate or validate instruments, data, or models,
- Non-commercial use only—Data are available for government applications and academic research, but not for resale.

These periods are envisioned to be temporary. Cost and efficiency make it useful to enter data into the archives during the restricted period, however, while they are still fresh. Any archive center that supports temporary restrictions must be able to authenticate and properly authorize users so as to shield the data from general public use. The opportunities for restricted access, data security, and metadata and data discovery support offered by the IOOS Archive System are an asset, previously unavailable, for the research science community.

No archive system is complete without user services and use metrics. On-line documentation and knowledgeable staff will provide assistance and advice on both access and content. Additional background information will be available through references and citations in the metadata. Broad

### Part III. Appendix 3: Data Archive and Access

use metrics are required to evaluate the system effectiveness and gain a sense of how to improve it. Ideally, they would measure the impact of the data, for example, the number of scientific articles written based on IOOS data. Although such metrics are currently outside today's capability, new techniques for metadata could be used to capture and hold this information. Some metrics will be furnished by the DMAC data transport mechanism. Others include:

- Number of “users”—The anonymous nature of much of the access prevents the true number of users from being collected. Unique Internet addresses are the closest proxy to this number that can be collected, and are useful for evaluating trends as well as access by well-constrained domains such as .gov, .mil, .edu, and international domains.
- Number of accesses—This is the number of files downloaded or otherwise accessed through the various services. Note that volume of data is not used here; a cornerstone of DMAC data access is to provide subsets, GIS maps, on-line analyses—in short, only the information required by the user. This renders “data volume distributed” a relatively meaningless metric (although it is useful for system performance). The data access metric should also be broken down by data set and service method.
- System performance statistics—This includes use of disks and computers as well as work performed (i.e., services executed and volume accessed). While not useful for measuring use of data, it is needed for planning systems.

In addition to numeric metrics, measurements of qualitative access are also useful. Specifically, all archive systems should have a means of soliciting and capturing user feedback on services and data sets. One way is to include voluntary user registration, which has the added benefit of supporting the transmission of newsletters, information on data products, and updates. Voluntary user surveys are also useful for this purpose, but must be approved by OMB for federal data centers. Obtaining clearance for such surveys throughout IOOS could be a useful function of the IOOS program.

# Data Policy

IOOS data policies will be developed in an early phase of implementation. The policies will include all applicable Federal policies. Recommendations for the Archive System follow.

The IOOS Archive System data policy will be consistent with the GOOS design principles, the IOC/IODE Data Exchange Policy, adopted in 1993 (Meeting of the Ad Hoc Working Group on Oceanographic Data Exchange Policy IOC/INF-1144rev, 4 July 2000), and the policy for free exchange of meteorological and related marine data of the WMO (WMO Resolution 40, Publication WMO – No. 837). Accordingly, the IOOS data center policies will be based on the following guidelines:

- Full and open sharing of non-commercial IOOS data and products.
- Coordination and cooperation between IOOS Archive System centers and the international GOOS data centers.
- Preservation of all data according to the IOOS defined categories. Federal standards for data preservations will apply to the Archive System.
- IOOS metadata standards or software to interpret metadata to the IOOS standard. Federal standards for metadata will apply to the Archive System.
- Data sets reprocessed will be managed under version control. Previous versions will be retained as subject to IOOS data polices.
- The IOOS Archive System will provide access to the data
  - to the greatest extent practical data will be made accessible online at no cost to the users;
  - data from offline sources will be available at no more than the cost of providing the service.
- All data collected and prepared under IOOS funding shall be submitted to the IOOS Archive System.
- Restricted access, if any, will be in accordance with IOOS data policy.

## Interactions and Partnerships with Other Data Centers

IOOS DMAC will operate as a federation among cooperating groups that share IOOS objectives. Forming and maintaining effective partnerships over time is essential to implementing and sustaining the system. The near-term challenge is to identify and approach the groups most likely to share IOOS objectives. This challenge will be addressed in the early phases of implementation. Identifying such potential partners requires searching both national and international ocean communities—among governmental and non-governmental bodies—keeping in mind the full scope of IOOS objectives. For example, in terms of archive and access to IOOS-relevant data, valuable partners may be found among groups that specialize in socio-economic studies or public health statistics as well as among the ocean operations and research communities.

There is a need to develop and maintain a list, in a systematic manner, of potential interactions and partnerships. Interaction with the Oceans Commission is a good starting point because it has attracted many participants likely to share IOOS interests. IOOS should request that the Commission provide a list of these participants. Another source is the NOPP federal agencies themselves. IOOS should request that each agency compile a list of their own ocean programs and external groups that those programs serve. The federal agencies already have tabulated their major ocean programs for the Oceans Commission. With that base, adding information about users and partners in those programs could start a systematic listing of potential IOOS partners and users.

International organizations and programs are another source of potential partners. The international GOOS program is an obvious example. But, there are many more within the structures of the World Meteorological Organization (WMO), the Intergovernmental Oceanographic Commission (IOC), the International Council of Scientific Unions (ICSU), and similar bodies. As U.S. participation in IOOS begins, representatives from these organizations will be tasked to identify other potential international partners.

Another community to consider is the commercial, value-added information providers. Environmental engineers and consultants, publishers, and forecasting services are some examples. While this category would likely be users of IOOS data, IOOS should carefully coordinate its level of information services to the public with the capabilities of the value-added vendor community. There must be sensitivity to encroaching on the capabilities of commercial vendors. IOOS should identify and approach such organizations early in the implementation to clarify respective roles in providing information products to the public.

### Part III. Appendix 3: Data Archive and Access

In parallel with identifying potential partnerships, IOOS should develop a standard briefing package to use in approaching these groups. The briefing should explain the IOOS objectives and options for their participation as a data center in the IOOS Archive System. A second version, intended for potential user groups, would assist in building support for the effort.

A second parallel effort should be started to develop partnership tools—standard Memorandum of Understanding, grant/contract clauses, etc.—that convey IOOS requirements. Having these tools pre-approved by the appropriate legal and administrative authorities will avoid delays in implementing partnership arrangements later. An additional benefit of starting these early in the implementation phase is that the approval process will uncover any obstacles to our partnership strategy, and allow more time to obtain any necessary exemptions or revisions to the regulations.

# Cost Estimates

Estimates for the cost of managing data sets in the Archive System are shown in Table 3. These cost estimates are only for data set work and storage media; other necessary supporting infrastructure is not accounted for here. Some supporting infrastructure costs are given in Appendix B.

The estimates are largely controlled by the costs of:

- On line and Off line Storage Costs—Marginal hardware costs to add new data to preexisting infrastructure,
- Data set Adaptation Costs—The cost to bring a non-compliant data set up to IOOS metadata and data transport standards.

The first year startup cost and annual costs thereafter are approximated with consideration given to the yearly data volume and number of years of data held, how much data are stored on line and off line, the number of data-set copies, and the following parameters:

- Structure Factor: How closely the data and metadata adhere to access standards,
- External Provision: The amount of cost that is assumed by an organization or agency outside of IOOS,
- Staff Implementation:  $(\text{Data set Adaptation Cost}) \times (1 - \text{Structure Factor})$ ,
- Annual Maintenance:  $15\% \times (\text{Hardware Cost} + \text{Staff Implementation})$ ,
- IOOS Start Up: All first year costs for IOOS data sets,
- IOOS Maintenance Per Year:  $(\text{Annual Maintenance}) \times (1 - \text{External Provision})$ .

The costs are scaled for two example data sets and three archive centers.

Example Data set 1:

- 10 TB per year with a 3 year total (30 TB)
- The data stream feed is 80% IOOS compliant (structure factor = 0.8)
- 90% of the costs are covered by external programs
- 1 TB is maintained on line

Example Data set 2:

- 100 GB per year with a 10-year total (1 TB)
- The data stream is 10% IOOS compliant
- No cost sharing with external programs
- 1 TB maintained on line

### Part III. Appendix 3: Data Archive and Access

Table 3. Data sets management cost estimates. Values are in dollars unless otherwise noted.

Storage (TB)	NASA	NCAR	NODC	NCDC	EPA	MIL	ORNL					
Off line	400	325		2100								
On line	6000	10000		9500								
Data set Adaptation Cost	75000	80000		75000								
Data set	Yrly Vol. (TB)	Years (#)	Struct. Factor (0-1)	External Provision (0-1)	On line Storage (TB)	Off line Storage (TB)	Archive Copies (#)	Hard- ware	Staff Imple.	Annual Maint.	IOOS Start Up	IOOS Cost per year
Data set 1 @ NASA	10	3	0.8	0.9	1	30	1	18000	15000	4950	3300	495
Data set 1 @ NCAR	10	3	0.8	0.9	1	30	1	19750	16000	5363	3575	536
Data set 1 @ NCDC	10	3	0.8	0.9	1	30	1	72500	15000	13125	8750	1313
Data set 2 @ NASA	0.1	10	0.1	0	1	1	2	6800	67500	11145	74300	11145
Data set 2 @ NCAR	0.1	10	0.1	0	1	1	2	10650	72000	12398	82650	12398
Data set 2 @ NCDC	0.1	10	0.1	0	1	1	2	13700	67500	12180	81200	12180

# Annex A. Additional Infrastructure Costs

## Infrastructure costs at NCAR

- The STK 9940B cartridge tapes now hold 200 GB each. Migration to this media has begun (08/2002). Each tape costs \$65. Previous STK storage was 60 GB/tape.
- For reliable on-line storage, RAID configured disks are used.
- A STK storage Silo holds roughly 6000 tapes. New cost is \$400K, and can be purchased used for \$150K. The high-capacity tapes are creating a healthy used-Silo market.
- Infrastructure costs (heating, cooling, system and operation staff, servers, networks, fiber connections, maintenance fees for hardware and software licenses) for a static system that moves 2TB/day is \$1–3M/year.
- Additional infrastructure costs for a growing system, approximately 2 TB/day, is about \$1 M/year
- The start-up costs for facilities are two to three times greater than the operational costs. Hardware vendors require most of the money up front.
- Media migration is a constant effort. Very little technology lasts for more than five years.

## Annex B. Glossary of Terms

**Archive** (noun)—A repository for preserved data and metadata. Analog and digital information is stored with identification tags, computer integrity measures, and descriptive data for reference. A deep archive contains the original data, plus Archive System derived products in an off-network environment. A working archive contains the same data as maintained in the deep archive, plus other data and products in an on-network environment for internal and external access.

**Archive** (verb)—To place original digital data and information files into the working archive area, where those files are preserved and maintained according to the processes defined by the Archive Center.

**Archive Center**—An organization that has a mission to procure, preserve, and provide access to data in perpetuity. An Archive Center maintains multiple archive repositories. Data archives and data services are explicitly part of their function. General responsibilities include:

- Acquiring and accepting data and metadata from many different individuals and organizations and in many different formats,
- Ensuring data integrity,
- Ensuring that back-up copies of data are made and that metadata are preserved with the data,
- Storing data either in original form or in a form from which all the original data and metadata can be recovered,
- Refreshing or updating the medium on which the data and metadata are stored so that both are readable in the future,
- Providing the data and all supporting metadata to users on request, free of charge or at a cost no more than the cost of reproduction or transmission.

**Catalog**—A directory, plus a guide and/or inventories, integrated with support mechanisms that provide metadata access and answers to inquiries. Capabilities include browsing and data searches, and it may be integrated with data retrieval capabilities.

**Checksum**—An error-detection scheme that uses a numerical value based on the number of set bits in a file. Using the same formula for computing checksums at later times makes it possible to identify digital files that have been truncated or corrupted.

**Data Assembly Center**—An organization that has a mission to procure and provide access to data. These data centers specialize in one or more data types—providing quality control and data products in their area of expertise. These centers may be permanent (e.g., NDBC) or exist only for limited periods of time (e.g., WOCE Data Assembly Centers). They do not provide long-term archival services. Distributing Data Assembly Centers is an efficient way to acquire and process data over a wide range of disciplines, with the assembled data and products then being submitted to Archive Centers for long-term storage and access.

## Part III. Appendix 3: Data Archive and Access

**Data Category**—The arrangement of data into groups by their distinct archiving requirements. These requirements include the minimum retention time of the data in the archive and the minimum number of data copies that must be archived. There are four IOOS data categories.

- Irreplaceable data are observational and research quality data that cannot be reproduced or easily regenerated, such as raw satellite and *in situ* measurements.
- Replaceable data are derived from irreplaceable data and can be regenerated through systematic processing. Such data include calibrated satellite radiance.
- Perishable data are low-resolution or uncalibrated real or near real-time data that are replaced by higher-quality data, such as XBT data broadcast over the Global Telecommunications System as part of the Ship-of-Opportunity Program.
- Virtual data are data provided through on-demand processing, such as analyzed data generated with the Live Access Server software on the Internet.

**Data Discovery Tool**—Software used to search through metadata to find data sets of interest.

**Data Product**—A data set derived from original data.

**Data Security**—Measures taken to guard against computer viruses and other forms of data corruption. Also known as data integrity.

**Inventory**—A list of archive objects that includes some information meant to aid a user in selecting and obtaining a group of archive objects. Inventories may include temporal and spatial coverage, status indicators, and physical storage information.

**Latency**—The time between the earliest observation in a data set and the availability of that data set to a customer.

**Lineage**—Information about the events, parameters, and source data that constructed a data set and information about the parties responsible for that data set (adapted from FGDC CSDGM).

**Lineage Control**—A method for tracking the lineage of a data set (contrast with Version Control).

**Media Migration**—Act of moving data from one type of archive media to another usually in response to changing technology (e.g., 9-track to 3490 cartridge tape).

## Part III. Appendix 3: Data Archive and Access

**Metadata**—The several types of information, which may be analog as well as digital, created and maintained to describe and manage a data set or archive object (i.e., “data about data”). The metadata types relevant to the IOOS Archive System are Use, Discovery, Documentation, and Administrative.

- Use metadata are the semantic and syntactic information about the contents of an archive object (e.g., descriptions of measured parameters, data collection methods, and file formats).
- Discovery metadata are the standard structured information that is designed to help find a data set (e.g., IOOS-wide data-set name and data-set version).
- Documentation metadata are the information about documents that refer to an archive object (e.g., the title, author and date of publication of a cruise report).
- Administrative metadata are the information used to manage an archive object within a data center and do not change or affect the description of the archive object (e.g., file location, file size, and checksum values). These metadata are created by a data center as the data are archived.

**Modeling Center**—An organization that synthesizes observational data to produce analyses, predictions and hindcasts of ocean conditions. Modeling centers often provide access to their products, but typically are not long-term archives.

**Pull**—To download data from a server.

**Push**—To upload data to a server or to send data to a customer (e.g., via e-mail).

**Quality Assurance**—To assess the quality of data collected via a particular method and then provide feedback to the data collectors so as to improve the data-collection method.

**Quality Control**—To assess the quality of data collected and then correct or flag the bad data.

**Regional Data Center**—An organization that has a mission to procure and provide data from a specific geographic region (e.g., Gulf of Mexico) and that provides quality control and data products in their area of expertise. These organizations may, also, serve as secondary IOOS data Archive Centers.

**Server**—Location on the Internet where data are available to be downloaded via protocols such as FTP, HTTP, and OPeNDAP.

**Version**—An instance of a data set in which some part of the content of the data has been changed.

**Version Control**—A method for tracking the version of a data set (contrast with Lineage Control).