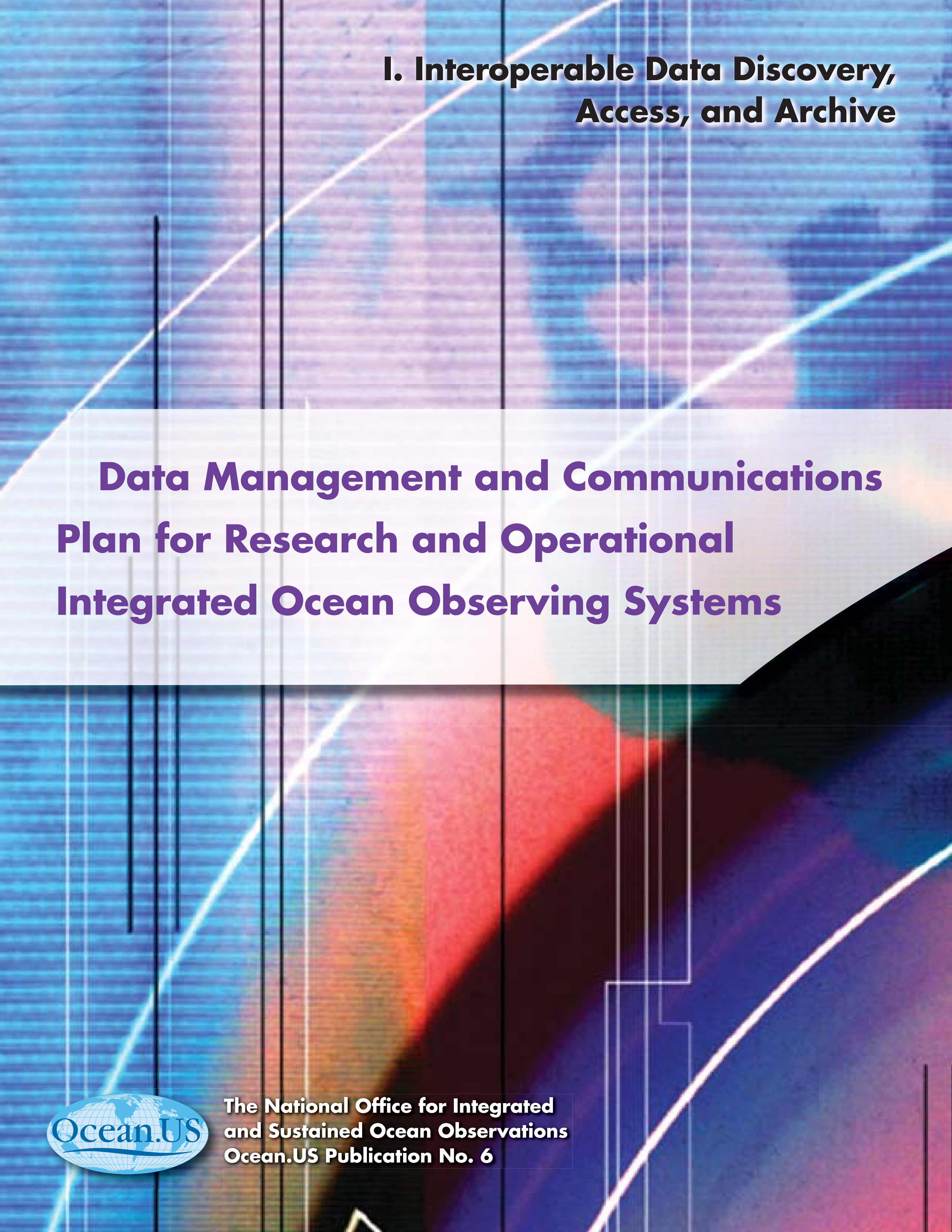


I. Interoperable Data Discovery, Access, and Archive

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

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

Editing
and design by
Geosciences Professional
Services, Inc.

We greatly appreciate and thank the reviewers for their many thoughtful and insightful comments.

Funding for the development of this Plan, and all related activities of the DMAC Steering Committee and its supporting teams, was provided by NASA, NOAA, and NSF. We gratefully acknowledge their support.

Suggested Citation: Hankin, S. and the DMAC Steering Committee, 2005, Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems: I. Interoperable Data Discovery, Access, and Archive, Ocean.US, Arlington, VA 304 pp. http://dmac.ocean.us/dacsc/imp_plan.jsp

I. Interoperable Data Discovery, Access, and Archive

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

March 2005

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

DMAC Steering Committee

Steven C. Hankin, Chairman, NOAA

Landry Bernard, NOAA

Peter Cornillon, University of Rhode Island

Fred Grassle, Rutgers University

Chuck Hakkarinen, Consultant

David Legler, U.S. CLIVAR Office

John Lever, Navy

Phil Mundy, Gulf of Alaska Ecosystem Monitoring and Research Program,
Exxon Valdez Oil Spill Trustee Council

Worth Nowlin, Texas A&M University

Susan Starke, NOAA

Steven J. Worley, NCAR

Ocean.US Staff

Roz Cohen

Contents

Table of Contents

Roadmap to the DMAC Plan

This detailed, phased Implementation Plan for the Data Management and Communications (DMAC) Subsystem of the Integrated Ocean Observing System (IOOS) was prepared at the request of Ocean.US, the IOOS National Office. The Plan is organized into three main parts:

- Part I is intended for a semi-technical audience, including potential IOOS partners and users. It presents an overview of the unique challenges and requirements of the DMAC Subsystem, and proposes strategies and a technical design for addressing them. Subsystem management, oversight and coordination are discussed, and a summary cost model (preliminary in this draft) is included. Part I concludes with a section outlining the highest priority activities for implementation.

- Part II is intended for a highly technical audience, including software engineers, information management specialists and IOOS program planners. It presents a detailed Phased Implementation Plan for DMAC using many of the formalisms of software engineering. It summarizes the formal requirements for all parts of the subsystem, and suggests a number of coordinated activities through which the DMAC Subsystem may be built.

- Part III includes appendices which provide in-depth discussions of: (1) Metadata and Data Discovery, (2) Data Transport, (3) Data Archive and Access, (4) User Outreach, (5) System Engineering, (6) Technology Refreshment and Maintenance, (and 7) Biological Data Considerations. Appendices 1-4 were prepared by Expert Teams appointed by the DMAC-SC. Each appendix is intended for readers with significant technical expertise in the topic addressed.

Table of Contents

Glossary

AGU	American Geophysical Union
AMS	American Meteorological Society
API	Application Programmer Interface
APEX	An autonomous drifting profiler used to measure subsurface currents and make profile measurements
Argo	A broad-scale global array of temperature/salinity profiling floats
ASPC	Assimilation, Synthesis, and Product Center
ASLO	American Society of Limnology and Oceanography
AVHRR	Advanced Very High Resolution Radiometer
BAA	Broad Agency Announcement
CalCOFI	California Oceanic Fisheries Investigation
CAOS	Coastal Alaska Observing System
CAS	Commercially Available Software
CLEANER	Collaborative Large-Scale Engineering Analysis Network for Environmental Research
CODAR	Coastal Ocean Dynamics Applications Radar
COP	Conference of Parties
COSMIC	Constellation Observing System for Meteorology, Ionosphere, and Climate
COTS	Commercial Off-The-Shelf
CSDGM	Content Standard for Digital Geospatial Metadata
CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science, Inc.
DIF	Directory Interchange Format
DMAC	Data Management and Communications
DMAC-SC	Data Management and Communications Steering Committee
DMAC-StC	Data Management and Communications Standing Committee
DNA	Designated National Agency
DODS	Distributed Ocean Data System
EarthScope	A program exploring the structure and evolution of the North American continent
EEZ	Exclusive Economic Zone
EXCOM	Executive Committee
FAQ	Frequently Asked Question
FGDC	Federal Geographic Data Committee
FNMOC	Fleet Numerical Meteorology and Oceanography Center
FTE	Full-Time Equivalent
FTP	File Transfer Protocol
FWIS	Future WMO Information System
FY	Fiscal Year
GBIF	Global Biodiversity Information Facility

Glossary

GCMD	NASA Global Change Master Directory
GCOS	Global Climate Observing System
GDACs	Argo Global Data Access Centers
GEOSS	Global Earth Observation System of Systems
GIS	Geographic Information System
GLOBEC	GLOBal ocean ECosystem dynamics
GML	Geography Markup Language
GODAE	Global Ocean Data Assimilation Experiment
GoMOOS	Gulf of Maine Ocean Observing System
GOOS	Global Ocean Observing System
GrADS	Grid Analysis and Display System
GTS	Global Telecommunications System of the WMO
HTTP	Hypertext Transfer Protocol
HUGO	Hawaii Undersea GeoObservatory
ICG	WMO Intercommission Coordinating Group
IOC	Intergovernmental Oceanographic Commission
IODE	International Oceanographic Data and Information Exchange
IOOS	Integrated Ocean Observing System
IOWG	Implementation Oversight Working Group
ISDR	International Strategy for Disaster Reduction
ISO	International Organization for Standardization
IT	Information Technology
ITIS	Integrated Taxonomic Information System
JASON1	Satellite that will provide altimetry data
JCOMM	Joint Commission for Oceanography and Marine Meteorology
JCOMM-ETDMP	JCOMM Expert Team on Data Management Practices
JGOFS	Joint Global Ocean Flux Study
LAS	Live Access Server
LEO 15	Long-term Environmental Observatory 15
LLC	Limited Liability Company
LOICZ	Land-Ocean Interactions in the Coastal Zone
MARS	Monterey Accelerated Research System
MBARI	Monterey Bay Aquarium Research Institute
MET	Meteorological Data
MOA	Memorandum of Agreement
MODIS	Moderate Resolution Imaging Spectroradiometer
NARA	National Archives and Records Administration
NASA	National Aeronautics and Space Administration

Glossary

NCDDC	National Coastal Data Development Center
NDBC	National Data Buoy Center
NEON	National Ecological Observatory Network
Neptune	A project seeking to establish a seafloor observatory in the Northeast Pacific Ocean
netCDF	Network Common Data Form
NFRA	National Federation of Regional Associations
NOAA	National Oceanic and Atmospheric Administration
NODC	National Oceanographic Data Center
NOPP	National Oceanographic Partnership Program
NOS	NOAA's National Ocean Service
NSF	National Science Foundation
NVODS	National Virtual Ocean Data System (NVODS)
OAI	Open Archive Initiative
OBIS	Ocean Biogeographic Information System
Ocean.US	The national office for integrated and sustained ocean observations
ODAP	Ocean Data Access Protocol
OGC	Open Geospatial Consortium
OIT	Ocean Information Technology
OITP	Ocean Information Technology Project
OMB	Office of Management and Budget
OOI	Ocean Observatory Initiative
OPeNDAP	Open source Project for a Network Data Access Protocol
ORION	Ocean Research Interactive Observatory Networks program
OSSE	Observing System Simulation Experiment
PODAAC	Physical Oceanography Distributed Active Archive Center
PDA	Primary Data Assembly
QA	Quality Assurance
QC	Quality Control
QuikSCAT	Satellite that has wind-measuring instrumentation
RA	Regional Association
RDBMS	Relational Database Management System
RNODC	Responsible National Oceanographic Data Center
R&D	Research & Development
SeaWiFS	Sea-viewing Wide Field-of-view Sensor Project
SeaWinds	A specialized radar on the QuikSCAT satellite that measures near-surface wind speed and direction at a 25 km resolution.
SMTP	Simple Mail Transfer Protocol
SQL	Structured Query Language

SW	Software
TAO	Tropical Atmosphere Ocean
TCP/IP	Transmission Control Protocol/Internet Protocol
TeAM	Technology Assessment and Management
T/P	TOPEX/POSEIDON satellite, providing ocean surface topography
TRP	Technology Refreshment Plan
UNFCCC	United Nations Framework Convention on Climate Change
URI	Universal Resource Identifier
USCOP	U.S. Commission on Ocean Policy
VOS	Volunteer Observing Ship
WCP	World Climate Program
WCS	Web Coverage Service
WDC	World Data Center
WFS	Web Feature Service
WMO	World Meteorological Organization
WOCE	World Ocean Circulation Experiment
WWW	World Wide Web
XBT	Expendable Bathythermograph
XML	Extensible Markup Language

I. Interoperable Data Discovery, Access, and Archive

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

Executive Summary

March 2005

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

Executive Summary

Congress has directed the U.S. marine science communities to come together to plan, design, and implement a sustained Integrated Ocean Observing System (IOOS). IOOS is envisioned as a network of regional, national, and global systems that rapidly and systematically acquires and disseminates data and data products to serve the critical and expanding societal needs to:

- Improve predictions of climate change and weather and their effects on coastal communities and the nation;
- Improve the safety and efficiency of maritime operations;
- More effectively mitigate the effects of natural hazards;
- Improve national and homeland security;
- Reduce public health risks;
- More effectively protect and restore healthy coastal ecosystems; and
- Enable the sustained use of ocean and coastal resources.

Internationally, IOOS will be the U.S. contribution to the Global Ocean Observing System (GOOS) and the Global Earth Observation System of Systems (GEOSS).

A coherent strategy that enables the integration of marine data streams across disciplines, institutions, time scales, and geographic regions is central to the success of IOOS and other regional, national, and international ocean and coastal observing systems. The system that must be developed, while challenging, is within the scope of current information technology (IT). It can be developed by building upon existing capabilities through relatively straightforward software engineering. *The greatest challenge to enhancing marine data integration is one of coordination and cooperation among the members of IOOS and its user communities.*

Ocean.US, the IOOS national office, established the Data Management and Communications Steering Committee in the spring of 2002 to develop a detailed, phased implementation plan that will lead to an effective data management and communications (DMAC) component of IOOS, and to provide oversight during its evolution. The DMAC Plan has undergone multiple levels of review by technical and scientific experts, as well as by the broader marine environmental data supplier and user communities. It is divided into three main parts. Part I, intended for general readers, provides an overview of requirements, strategies for addressing them, and technological considerations. Part II, intended for technical readers, presents a detailed DMAC Implementation Plan in outline form. Part III, the Appendices, provides in-depth analysis of key technical topics.

This DMAC Plan is the first in a series of documents that addresses IOOS data management and communications requirements, and those of other observing systems such as the National Science Foundation's (NSF) Ocean Research Interactive Observatory Networks (ORION). This Plan presents an overview of DMAC; provides a technical focus on the issues of interoperable data discovery, access, and archive; and provides a development time line with estimated costs. As one of several Subsystems of IOOS, DMAC will be developed, implemented, operated, and enhanced according to the planning and governance procedures described in the IOOS Development Plan (www.ocean.us).

INTERNATIONAL COOPERATION

Producing global assessments and predictions of coastal ecosystem health and sea-level change, as well as addressing the other IOOS goals, requires that IOOS observations and data products be fully integrated with other national and international Earth observation efforts. Coordinated and sustained cooperation is already well established within the weather community, and the World Meteorological Organization's (WMO) World Weather Watch demonstrates the value of this international collaboration. Coordination is less well established in the ocean, ice, land, water, and climate observation communities. Nevertheless, much important work has been accomplished on the international front, and IOOS is well positioned to contribute to these efforts, especially in the area of data management and communications.

Many of the contributors to the IOOS DMAC Plan are involved with international efforts addressing global ocean and coastal observing needs. As a result, the DMAC Plan is being examined by the WMO as an early model for standards and protocol development. Effective coordination among the relevant international programs is essential to realizing a truly interoperable, global and national coastal and ocean observation framework. Expanded coordination and more formal programmatic linkages by IOOS with GEOSS, GOOS, and the WMO are needed. The DMAC Plan therefore recommends that steps be taken to address this need.

TECHNICAL ANALYSIS OF THE DMAC SUBSYSTEM

IOOS will consist of three subsystems:

- **Observing Subsystem:** remotely sensed and *in situ* measurements and their transmission from regional and national backbone platforms;
- **Modeling and Analysis Subsystem:** evaluation and forecast of the state of the marine environment based upon assimilated measurements; and
- **Data Management and Communications Subsystem (DMAC):** information technology infrastructure such as national backbone data systems, regional data centers, and archive centers connected by the Internet, and using shared standards and protocols.

Figure 1 illustrates the flow of data from observation platforms to intermediate components (e.g., modeling centers and archive centers), to generators of information products, and finally to end users. The DMAC Subsystem is a framework for integration of large and small independent and heterogeneous data management and communications systems. Most of the planning and investment within the DMAC Subsystem for data management *per se* lies outside the scope of the DMAC Plan. The thousands of individual organizations that comprise IOOS will continue to manage their data in the manner they deem most appropriate to their individual missions, but through DMAC they can broaden the impact of their data, serve a larger community, and contribute to long-term data archives that will benefit generations to come.

IOOS Observing Subsystem elements are managed by regional, national, and international entities. Measurements made by these elements are highly heterogeneous. A wide range of data distribution and dissemination systems, including the WMO's Global Telecommunication System (GTS), are used to transfer data from the measurement platforms to and among the locations at which Primary Data Assembly and Quality Control (PDA&QC) occur. **The systems that convey data from sensors to primary data centers/sites lie outside the scope of the current DMAC Plan.** PDA&QC processes typically lie at the interface between the Observing Subsystem and the DMAC Subsystem. In general, some form of PDA&QC is required before ocean observations and measurements can be used.

The DMAC Subsystem will include a data and communications infrastructure that consists of a suite of components—standards, protocols, facilities, software, and supporting hardware systems. The design and planning for the DMAC framework will emphasize continual, smooth evolution. The components upon which the architecture is built will, themselves, be an evolving collection. New components will be introduced; recognized components will be advanced; obsolete components will be removed. A significant level of duplication of function among components will be tolerated as a necessary consequence of a continuously evolving system. The DMAC Standards Process will define the manner by which the level of maturity of components will be designated: R&D, pilot, pre-operational, and operational. At the outset, no formal DMAC standards process exists. Yet, there is an imperative to provide immediate guidance to would-be data providers. To address this need, the DMAC Plan includes preliminary recommendations for (1) the maturity designations of certain named components that are viewed as essential to the initial architecture and (2) a roadmap leading to rapid designation of other initial components by community-based working groups.

The DMAC Plan provides a roadmap to achieve the following functionality for the DMAC Subsystem: (1) IOOS-wide descriptions of data sets (**Metadata**); (2) the ability to search for and find data sets, products, and data manipulation capabilities of interest (**Data Discovery**); (3) the ability to access measurements and data products from computer applications across the Internet

IOOS Data Communications

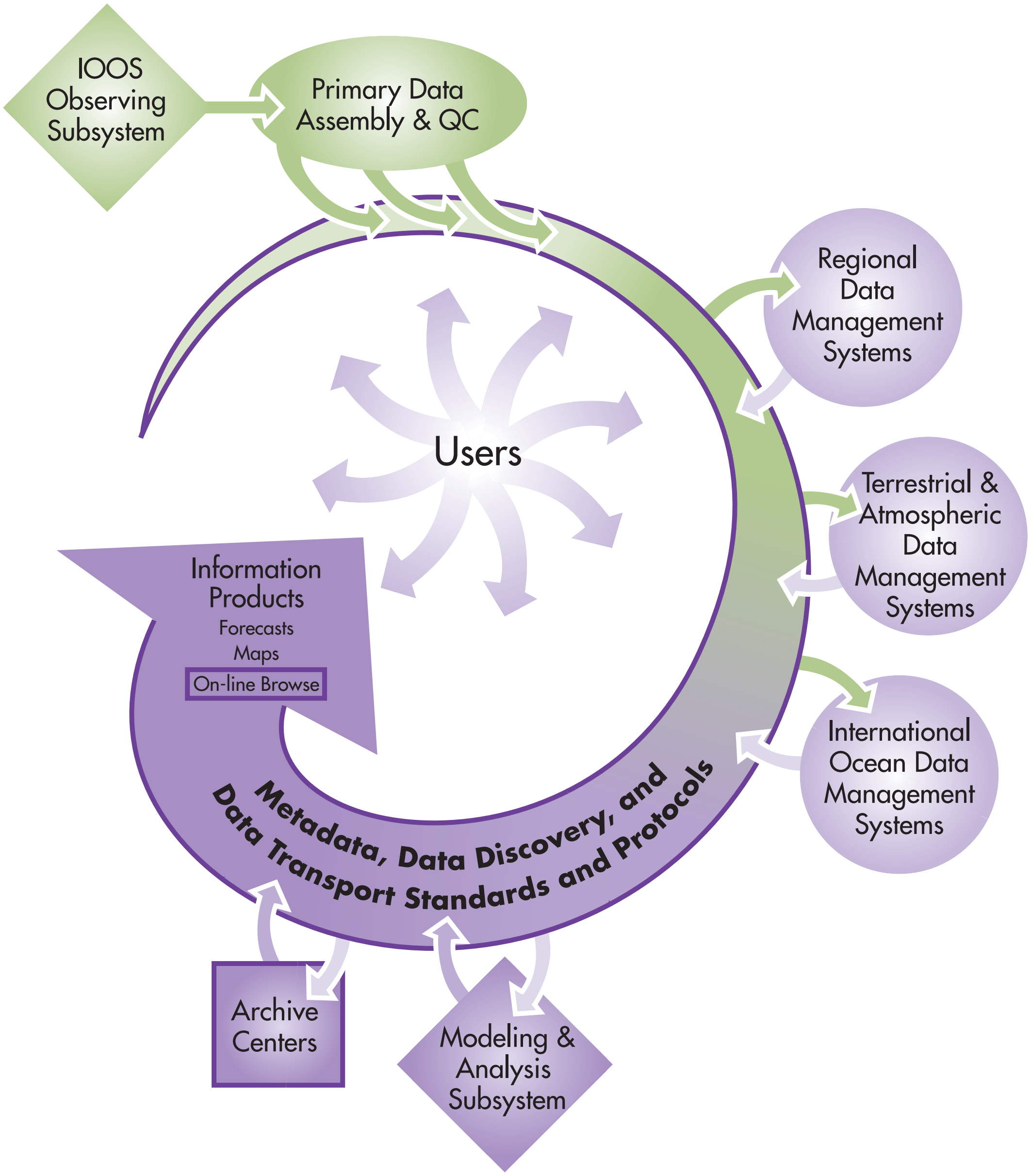

IOOS Observing Subsystem
Primary Data Assembly & QC
Regional Data Management Systems
Users
Terrestrial & Atmospheric Data Management Systems
Information Products
Forecasts
Maps
On-line Browse
International Ocean Data Management Systems
Metadata, Data Discovery, and Data Transport Standards and Protocols
Archive Centers
Modeling & Analysis Subsystem

Figure 1. Solid outlines indicate the elements of the IOOS Data Communications framework, which are detailed in the DMAC Plan. The arrows flowing outward from users indicate the feedback and control mechanisms through which users ultimately direct the functioning of all parts of the system. Note that the National Data Management Systems are included in the concept of Primary Data Assembly and Quality Control.

(**Data Transport**); (4) the ability to quickly evaluate the character of the data through commonly-available web browsers (**Uniform On-line Browse**); and (5) secure, long-term data storage (**Data Archive**).

DMAC Metadata will be based upon standard vocabularies and content specifications. The metadata specifications will support the publication of Federal Geographic Data Committee (FGDC)-compliant records. Agreement on vocabulary and content does not yet exist. The DMAC Plan recommends establishing an interdisciplinary Metadata Working Group to provide recommendations on a common vocabulary and initial metadata content for IOOS data. DMAC metadata will support data discovery capabilities that complement and extend the publicly accessible search capabilities that are available today through web search engines such as Google®. The DMAC Plan recommends that the **Data Discovery** architecture be determined by a working group that includes representatives from existing metadata management facilities and other metadata experts.

Underlying **DMAC Data Transport** is the unifying vision of DMAC web services. Through web services all types of client applications—for example, tools for end-users, modelers, and planners; and value-added marine information web sites—can access data from the broad range of IOOS data suppliers (servers). The methods by which client applications access web services remain uniform despite the servers being layered upon various legacy data management systems, developed with diverse programming languages, and run under different operating systems. DMAC web services will provide the means to connect IOOS to data management systems operated by international marine data partners and by partners in other disciplines such as meteorology.

Both the Open-source Project for a Network Data Access Protocol (OPeNDAP) and the Open Geospatial Consortium (OGC) support web services of relevance to DMAC Data Transport. OPeNDAP, the web service that underlies the National Virtual Ocean Data System (NVODS), is a discipline-neutral transport protocol that conveys data, metadata, and structure without regard to the scientific interpretation of the data. The DMAC Plan recommends the designation of OPeNDAP as an initial "operational" component for transport of gridded data, and recommends that a "pilot" activity be undertaken to explore the delivery of non-gridded data using OPeNDAP (See IOOS Development Plan at www.ocean.us for definitions of system component maturity). The DMAC Plan further recommends that two OGC web services, the Web Feature Service (WFS) and the Web Coverage Service (WCS) be examined for incorporation into the DMAC data transport suite.

The DMAC Plan anticipates that many IOOS data providers will host metadata-enabled, open source, or commercial on-line browse tools for end users. In addition, the DMAC Subsystem must provide a system-wide view of IOOS data—the ability to visualize and assess all IOOS data in a uniform manner. The **Uniform On-line Browse** capability of DMAC will use the Data Transport web services for access to IOOS data. The DMAC Plan recommends the designation of the Live Access Server (LAS), which provides browsing capabilities with NVODS, as an initial "pre-operational" component for system-wide Uniform On-line Browse. The DMAC Plan further recommends that OGC-compatible GIS web servers likewise be examined as candidates for DMAC Uniform On-line Browse clients.

The **DMAC Data Archive** component will be assembled from existing and new marine data archive facilities. The DMAC Plan recommends that, to be recognized as an official partner in the IOOS Data Archive enterprise, a facility must enter into a formal agreement(s) stipulating that they perform archive and access functions using DMAC standards and protocols and conform to IOOS Data Policy. The DMAC Plan further recommends that a community-based, interdisciplinary working group of archive specialists and advisors initiate an orderly strategy to determine DMAC Data Archive policies and procedures, and to ensure that designated archive facilities exist for all IOOS data.

The **IOOS Modeling and Analysis Subsystem** will provide numerical (digital) data products through computer modeling and analysis of real-time and historical data collections. Planning for the numerical data products that IOOS must produce lies outside the scope of the DMAC Plan. It will be handled elsewhere within the IOOS framework.

Information products, such as text and verbal forecasts, maps, and scientific plots, will be generated throughout the IOOS network. It is understood that the private sector will be a primary producer and distributor of value-added information products within IOOS, particularly to meet the specialized needs of targeted user groups.

IT SECURITY

IOOS is being deployed in a distributed, heterogeneous information technology (IT) environment with web services as the eventual target architecture. The IOOS therefore faces a number of security challenges that include:

- Participants joining the IOOS enterprise are accustomed to operating under diverse security guidelines and cultures that may not conform to required federal IT security practices.
- Agreement on and compliance with a common security policy must be reached across multiple heterogeneous systems.
- "Desktop-level administrators" must be able to understand and implement IOOS security policies and measures, and will often be in environments where IT management resources are limited.

- Many legacy applications will be incorporated into IOOS that will be web-enabled, but may not have been originally designed for exposure to the public Internet or for use in a structured IT security environment.

A thorough treatment of this topic is not possible at this time due to time and resource constraints. Therefore, it is recommended that a community-based working group on IOOS IT Security be established to develop an IOOS security policy, and to provide more specific guidelines for IOOS participants on implementation.

GOVERNANCE

The DMAC Subsystem operates within the context of the overall IOOS governance mechanisms described in the IOOS Development Plan. At the First Annual IOOS Implementation Conference (August/September 2004), the Ocean.US Executive Committee (EXCOM) agencies and the emerging Regional Associations (RAs) endorsed a DMAC governance strategy to ensure that the development and implementation of the DMAC Subsystem is coordinated closely with, and leverages upon, related activities in the federal agencies and other national, regional, and international Earth observing systems. This strategy includes: a community-based **DMAC Steering Team** (DMAC-ST) to coordinate and oversee the evolution of DMAC standards and best practices; **Expert Teams and Working Groups** to support the DMAC-ST; and a federal-government-only **Implementation Oversight Working Group** (IOWG) to coordinate DMAC implementation within the federal agencies. It is also recommended that the National Federation of Regional Associations (NFRA) establish a DMAC subcommittee to oversee and facilitate coordination, communications, and data and technology exchange at the regional level. IOOS stakeholders will be urged to participate in the DMAC planning and assessment activities to ensure that current and future community needs and priorities are addressed.

The DMAC Plan recognizes that data interoperability is largely a reflection of the ability of the marine community to successfully agree upon and use standards. Therefore, the DMAC Plan recommends that Ocean.US convene (or participate in) a working group to investigate and recommend a process for the development of future community data and metadata standards. The process should include guidelines that maximize the compatibility of new standards with pre-existing ones, and uniform review procedures for standards.

The DMAC Subsystem plays an essential role in IOOS user outreach by providing Internet portals for user feedback, and mechanisms for automated collection and analysis of system performance metrics.

COSTS

The DMAC Plan provides first-order estimates of those expenses associated with development and implementation of the policies, standards, protocols, and tools comprising the DMAC Subsystem. These cost estimates do not reflect any formal review or endorsement by the participants or agencies supporting IOOS deployment. Further, it should be noted that these estimates do not include costs resulting from growth in data services that would occur irrespective of IOOS development; nor do they include the costs of sensors, data telemetry, modeling and applications, and most product-development activities. These estimates also do not include those costs associated with the implementation of DMAC standards within the regions, or with anticipated capitalization and maintenance/operating costs likely to be incurred.

The DMAC Plan calls for the initiation of the full DMAC Subsystem over a five-year period at a cost of $82 M. The initiation costs include the development of core standards, protocols, and tools ($28 M); costs of hardware, software, networking capacity, data archiving center expansion, and systems integration labor ($37 M); and a budget for focused pilot projects to usher in and test the new technologies ($17 M). Out-year recurring costs over the following five years (to Year 10) total an additional $85 M. All cost estimates provided in the DMAC Plan include an inflation factor of 2.2 percent per year. Substantial new funding for IOOS is not anticipated until fiscal year 2007 (FY 07), **yet a minimally functioning DMAC Subsystem must already be in place to support the initial growth in IOOS** (and other ocean observing systems) measurements, modeling, and usage at that time. Thus, the DMAC Plan includes tasks totaling $2.1 M during FY 05 to FY 06 that are deemed to be very high priorities for immediate implementation to prepare for FY 07 demands on the Subsystem.

HIGH-PRIORITY RECOMMENDATIONS

Implementation of IOOS and other regional, national, and international ocean and coastal observing systems has already begun, and will continue to accelerate over the coming years. To support these activities, the DMAC Subsystem must quickly achieve a useful minimum level of functionality. The DMAC Plan recommends the following steps as high priorities for implementation:

1. Initiate working groups and/or applied R&D activities to address: (a) development of metadata, including vocabulary, content, and discovery components; (b) assessment and selection/development of missing data transport components; and (c) community building and partnerships as outlined in Part II of the DMAC Plan.

2. Engage software engineering services to initiate development of well-organized documentation, centralized coordination of assistance to IOOS data suppliers and product generators, and software life-cycle planning for the critical components of the DMAC Subsystem.

3. Ocean.US should establish (1) a permanent DMAC Steering Team with representation from the full IOOS community responsible for technical planning and recommending standards and (2) an all- federal group responsible for allocation of resources to address DMAC priorities and implementation.

4. Data providers should examine and where possible implement the Part I, "Concrete Guidance to Data Providers." This section offers guidance to help coordinate the implementation of initial DMAC functionality, while ensuring compatibility with future IOOS standards. Highlights of that guidance include: (a) create FGDC-compliant metadata; (b) enable data discovery by sharing metadata with designated IOOS metadata facilities; (c) make gridded data accessible through the OPeNDAP data access protocol; (d) implement on-line browse solutions (using Live Access Server, GIS web servers, or others); and (e) ensure that designated IOOS archive centers have plans in place for long-term archiving of the contributed data.

I. Interoperable Data Discovery, Access, and Archive

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

Part I: Overview

March 2005

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

Contents

Note to Readers

We wish to thank the approximately 150 national and international organizations and individuals (representing the government, academic, public, and private sectors) who provided comments at each stage of the review process. New drafts of the Data Management and Communications (DMAC) Plan were produced to address the comments received at each stage. The process was as follows:

- February-March, 2003: Internal review by the DMAC-Steering Committee (DMAC-SC) and members of the six supporting teams[1];
- April-May, 2003: External review by over a dozen national and international technical and scientific experts;
- June 2003: Public review by over sixty policy-makers and technical experts who participated in a national workshop sponsored by the Gulf of Maine Ocean Observing System and Ocean.US in Portland, Maine;
- September-November, 2003: Formal public review process. The draft DMAC Plan was posted on the Ocean.US/DMAC web site and announcements of its availability appeared in several marine community newsletters. E-mail notifications were also sent to several hundred members of the marine community, including participants in the 2002 Ocean.US Community Workshop (Airlie House), the March 2003 Ocean.US Regional Summit, the National Ocean Research Leadership Council, the U.S. GOOS Steering Committee, and the EXCOM; and
- November-December 2004: Formal public review process. The same process was followed as outlined above for *September-November 2003*, and the Plan's availability for review was announced in a Federal Register Notice on November 10, 2004.

In addition, DMAC-SC members and Ocean.US staff presented numerous briefings, and received feedback on the Plan, at regional, national, and international conferences and meetings, including those sponsored by AGU, ASLO, AMS, JCOMM-ETDMP, IODE, NSF Cyberinfrastructure, ORION, CAOS, GoMOOS, NVODS, and GODAE. The DMAC-SC also provided briefings to staff within their own agencies and organizations.

Summary of Major Changes Compared to the May 2004 Draft Plan

This final version of the first DMAC Plan contains the following substantive changes from the May 2004 version, primarily addressing comments received during the most recent review period announced in the November 10, 2004 Federal Register:

1. Cover and report number: changed to conform to the Ocean.US report series specifications.
2. Executive Summary: Updated to correspond to changes made in the body of the Plan.

[1] The DMAC-SC was supported in their work by six teams Data Discovery and Metadata; Data Transport, Data Archive and Access; Applications and Products; Data Facilities; and User Outreach.

3. Part I: New section added (this section: Note to Readers) summarizing the Plan review process, and the major changes made as a result of the final public comment period.
4. Part I, Preface: Updated to include references to the linkage between IOOS and GEOSS, and the First Annual IOOS Development Plan.
5. Part I, Main Sections: Updated references to the linkage between IOOS and GEOSS, and the First Annual IOOS Development Plan. Minor edits occur throughout Part I, and the following sections have more extensive edits in response to comments:
   - International Cooperation
   - On-Line Browse (now "Uniform On-Line Browse")
   - Management, Oversight, and Coordination (now "Governance, Oversight, and Coordination")
   - Concrete Guidance to Data Providers
6. Part I, Main Sections: The following new sections were added in response to comments:
   - Scope and Evolution of the DMAC Subsystem
   - OGC (Open Geospatial Consortium)
   - IT Security
7. Part II, General Requirements Section: Expanded to include performance requirements, and the two new recommendations in Part I regarding OGC have been incorporated.
8. Part III: No changes were made to the appendices because they are final reports from teams formed to support the work of the original DMAC-SC.

In addition, please note that the following points included in the May 2004 version of the Preface already address concerns expressed by several reviewers:

- "This Plan is the first in a series of documents that will address Data Management and Communications (DMAC) requirements of the Integrated Ocean Observing system (IOOS), and other regional, national, and global observing systems."
- "The DMAC Plan focuses on enhancing the interoperability of existing IOOS components through development of a common Data Communications Infrastructure. The infrastructure will consist of standards and protocols for metadata, data discovery, transport, on-line browse, and long-term archive. Other important issues such as QA/QC, modeling and applications, security, data assembly, and telemetry will be addressed in future IOOS documents."
- " .... this document is a plan, not a specification. The cost model presented herein includes support for systems engineering services to conduct a formal design analysis leading to a formal specification. The specification will guide the planning and implementation decisions and application of resources."

Preface

This Plan is the first in a series of documents that address the Data Management and Communication (DMAC) requirements of the Integrated Ocean Observing System (IOOS), and other regional, national, and global observing systems. In the Preface, we provide some background information to assist the reader in placing the DMAC Subsystem into its proper context as a component of the larger IOOS. As such, the DMAC Subsystem will be developed, implemented, operated, and enhanced through the planning and governance structures described in the IOOS Development Plan (go to www.ocean.us to download a copy).

IOOS

There is strong support in the U.S. Congress, the Executive Branch, and the U.S. Commission on Ocean Policy (USCOP) for development of a sustained, integrated coastal and ocean observing system that will make better use of existing resources, new knowledge, and advances in technology to achieve the following seven related societal goals:

- Improve predictions of climate change and weather and their effects on coastal communities and the nation;
- Improve the safety and efficiency of maritime operations;
- More effectively mitigate the effects of natural hazards;
- Improve national and homeland security;
- Reduce public health risks;
- More effectively protect and restore healthy coastal ecosystems; and
- Enable the sustained use of ocean and coastal resources.

Congress directed the U.S. marine environmental communities to come together to plan, design, and implement this observing system. The National Oceanographic Partnership Program[2] (NOPP) established the Ocean.US Office through a Memorandum of Agreement (MOA) in 2000. Ocean.US is charged with coordinating the development of the U.S. Integrated Ocean Observing System, based on concepts developed by national and international experts over the past dozen years. Ocean.US is overseen by an Executive Committee (EXCOM) composed of representatives from those NOPP agencies that signed the MOA.

[2] NOPP was established by Congress in 1997 (P.L. 104-201) to (1) "promote the national goals of assuring national security, advancing economic development, protecting the quality of life, and strengthening the science and education through improved knowledge of the ocean" and (2) "coordinate and strengthen oceanographic efforts to achieve these goals by identifying and carrying out partnerships among Federal agencies, academia, industry, and other members of the oceanographic community in areas of data, resources, education, and communications."

Part I. Overview

Ocean.US prepared an IOOS Development Plan in cooperation with participating NOPP agencies. The Plan was developed in three parts: Part I—Structure and Governance; Part II—Fiscal Years 2005-2006 Integrating Existing Assets; and Part III—Improving the IOOS Through Enhancements and Research. Approval of the IOOS Plan by the NOPP National Ocean Research Leadership Council (NORLC) is anticipated in early 2005.

IOOS is envisioned as a coordinated national and international network of observations, data management, and analyses systems that rapidly and systematically acquires and disseminates marine environmental data and information on past, present, and future states of the oceans. The IOOS is being developed as two closely coordinated global and coastal components that encompass the broad range of scales required to assess, detect, and predict the effects of global climate change, weather, and human activities. The global component consists of an international partnership to improve forecasts and assessments of weather, climate, ocean state, and boundary conditions for regional observing systems. It is the U.S. contribution to the Global Earth Observation System of Systems (GEOSS) and the Global Ocean Observing System (GOOS). The coastal component blends national observations in the Exclusive Economic Zone (EEZ) with measurement networks that are managed regionally to improve assessments and predictions of the effects of weather, climate, and human activities on the state of the coastal ocean, its ecosystems and living resources, and the nation’s economy. The coastal component encompasses the nation’s EEZ, the Great Lakes, and the estuaries.

Existing and planned observing system elements that address both research and operational aspects of the seven IOOS goals will be integrated into the system. Evolution of an integrated system that is responsive to user needs will require an iterative process of selection, incorporation, evaluation, and improvement over time. Candidate technologies and capabilities may pass through a series of stages (research, pilot, pre-operational) prior to being incorporated into the operational IOOS, long-term research, or both. Detailed criteria for activities to successfully pass through each of these stages are presented in the IOOS Development Plan.

A four-year cycle of planning, programming, and budgeting for IOOS implementation and development is described in Part I of the IOOS Development Plan. Ocean.US, in cooperation with NOPP agencies and Regional Associations (RAs)[3], will specify priorities for implementation and

[3] RAs will be established, based on regional priorities, to design, implement, operate, and improve regional observing systems by increasing the resolution of the variables measured; supplementing the variables measured by the national backbone with additional variables; providing data and information tailored to the requirements of regional stakeholders; and implementing programs to improve public awareness and education. Regional observing systems are needed to provide data and information on phenomena that are more effectively detected or predicted on regional scales that go beyond the jurisdiction of individual states.

advancement of IOOS; formulate timetables; work within the federal budget process to determine costs; and capitalize on unplanned opportunities. Research and development projects may be funded competitively through the NOPP process, or through mechanisms established by individual agencies in cooperation with Ocean.US. Operational elements are funded for extended periods of time based on demonstrated utility and performance.

DMAC

For planning purposes, IOOS is considered to be composed of three subsystems: the OBSERVING SUBSYSTEM (remotely sensed and *in situ* environmental measurements and their transmission from regional and national backbone platforms); MODELING AND ANALYSIS SUBSYSTEM (evaluation and forecast of the state of the marine environment based on assimilated measurements); and the DATA MANAGEMENT AND COMMUNICATIONS SUBSYSTEM (DMAC—information technology infrastructure such as national backbone systems, regional data centers, and archive centers connected by the Internet, and using shared standards and protocols).

Central to the success of IOOS (and other regional, national, and international ocean and coastal observing systems) is the presence of a DMAC Subsystem capable of supporting the wide variety and large volumes of data, the reliability and integrity requirements of operational data delivery, and the many other needs of the IOOS user community. The DMAC Subsystem is the primary integrating element of IOOS, and will provide the linkages among other IOOS components, partner organizations, and systems in other disciplines (e.g., terrestrial, atmospheric). Because of the critical need for a basic data communications infrastructure to support existing and newly emerging IOOS observing systems, the DMAC effort was initiated early on in the IOOS planning process. In the spring of 2002, the Director of Ocean.US appointed the Data Management and Communications Steering Committee (DMAC-SC), including representatives from federal and state government agencies, academia, and the private sector. The DMAC-SC was tasked with developing a detailed phased implementation plan for this IOOS subsystem.

The DMAC Plan, this document, presents a coherent strategy for integrating marine data streams across disciplines, organizations, times scales, and geographic locations. It has been divided into three main parts: Part I provides an overview of the requirements and technological considerations, and the strategies for addressing them. Part II presents the detailed DMAC System Implementation Plan in outline form. Part III, the Appendices, provides in-depth discussion of key technical topics.

The DMAC Plan focuses on enhancing the interoperability of existing IOOS components through development of a common Data Communications Infrastructure. The infrastructure will consist of standards and protocols for metadata, data discovery, transport, on-line browse, and long-term archive. Other important issues such as QA/QC, modeling and applications, security, data assembly, and telemetry will be addressed in future IOOS documents. It should be noted that this document is a plan, not a specification. The cost model presented herein includes support for systems engineering services to conduct a formal design analysis leading to a formal specification. The specification will guide planning and implementation decisions such as application of resources.

The DMAC Subsystem is an integral part of IOOS, and it is being developed in close coordination with other IOOS components. The longer-term implementation priorities and recommendations articulated in the DMAC Plan are being incorporated into the overall IOOS planning and budgeting processes described in the IOOS Development Plan. The IOOS planning and budgeting processes will not have a direct influence on the federal budget process until FY 2007. Annual updates to the DMAC Plan, developed in collaboration with the participating NOPP agencies and RAs, and with recommendations from new DMAC planning and implementations committees, will feed into future planning and budgeting cycles beyond FY 2007. In the interim time period (FY 2005 to 2006), efforts are being made to obtain support from participating NOPP agencies, RAs, and other sources of opportunity for the immediate, shorter-term priorities of the DMAC Subsystem. Implementation of these immediate priorities will lead to development of an initial DMAC Subsystem to support existing and emerging observing system activities at the local, regional, and national levels.

Section 1. Overview

INTRODUCTION

At the present time, no coherent data management and communications strategy exists for effectively integrating the wide variety of complex marine environmental measurements and observations across disciplines, institutions, and temporal and spatial scales. As a result, U.S. society is denied important benefits that might otherwise be derived from these data, such as improved climate forecasts and more effective protection of coastal marine ecosystems. Data are obtained by diverse means: nets are dragged; traps are set; instruments are lowered from ships, set adrift, or moored on cables and platforms; satellites scan the oceans from space; and laboratories are constructed on the seafloor. Measurements are made for a wide variety of purposes by individuals and sensors supported by many different kinds of institutions, including private industry; federal, state, and local governments; and non-governmental organizations. These data come in many different forms, from a single variable measured at a single point (e.g., a species name) to multivariate, four-dimensional collections of data that may be millions of gigabytes in size. These considerations, among others, led Congress to direct the U.S. marine data communities to come together to plan, design, and implement a sustained Integrated Ocean Observing System (IOOS).

Central to the success of IOOS, and other regional, national, and international ocean and coastal observing systems, is the presence of a Data Management and Communications (DMAC) Subsystem capable of delivering: real-time and delayed-mode observations to modeling centers; model-generated forecasts to users; distributed biological measurements to scientists, educators, and planners; and all forms of data to and from secure archive facilities. The needs of end users must be a part of the implementation and operation of the subsystem, both as sources of specifications for subsystem design, and as agents of change to keep the delivery of products from IOOS relevant to national interests. At a minimum, the DMAC Subsystem will make data and products readily accessible, allow users to readily locate data and information products, and advise users on the specifications and limitations of data by providing essential metadata (descriptive information about the data) along with the data.

The information technology required to meet most of the needs of DMAC, while challenging, can be developed from existing capabilities through relatively straightforward software engineering. The greatest challenge facing DMAC is one of coordination and cooperation among IOOS partners and user communities. DMAC can succeed only if the participants actively use the data and metadata standards, communications protocols, software, and policies that will knit IOOS into an integrated whole. The creation of a successful IOOS DMAC will require a sustained effort, a commitment across the U.S. marine community, and continual coordination with our international counterparts.

THE VISION

IOOS is envisioned as a system of regional, national, and global elements that rapidly and systematically acquire and disseminate data and data products to serve the needs of government agencies, industries, scientists, educators, non-governmental organizations, and the public. The IOOS vision is one of cooperative integration. The member entities will continue in the independent pursuit of their missions, while participating in a well-ordered data and information infrastructure. If IOOS were a living being, the DMAC Subsystem would be its blood and circulatory system—the data used to produce the information products needed by IOOS are analogous to the oxygen and nutrients transported in the blood to feed the many highly specialized organs.

The following set of guiding principles addresses the DMAC vision:

Interoperability: DMAC will serve as a framework for interoperability among heterogeneous cooperating systems.[4] The cooperating systems will be free to evolve independently to address the needs of their target users. Software and standards needed to participate in DMAC will be available directly to partners, or provided through commercial and non-commercial sources. DMAC will also be interoperable with systems outside of the marine community that manage atmospheric and terrestrial data.

Open, easy access and discovery: DMAC will enable users from all over the globe to easily locate, access, and use the diverse distributed forms of marine data and their associated metadata and documentation in a variety of computer applications (e.g., Geographic Information Systems-GIS, and scientific analysis applications). Users will be unencumbered by traditional barriers such as data formats, volumes, and distributed locations. DMAC will integrate cooperating systems so that data discovery will be seamless, and multiple versions will be easily tracked. There will be a "free market" of ocean sciences information, including officially sanctioned IOOS data sets, as well as data and products from other sources.

Reliable, sustained, efficient operations: DMAC will provide high reliability with 24/7 delivery of real-time data streams from measurement subsystems to operational modeling centers and users with time-critical requirements. It will provide high reliability in the delivery of computer-generated forecasts, estimates of state, and delayed-mode and real-time data to end users.

[4] By "interoperable" we mean that systems can function cooperatively through seamless exchange of data, the blending of data derived by different methods/instruments, an ability for models to access all varieties of source data without detailed knowledge of their origins, or the ability of users to see derived information in a way that is not limited by knowledge of the data collection methods or processing.

DMAC will require sufficient bandwidth and adequate carrying capacity to support large exchanges of raw data and model outputs among high-volume users. DMAC will offer techniques that reduce the need for large data transfers, such as server-side subsetting and computation, to allow users with limited bandwidth to enjoy the benefits of IOOS. Feedback mechanisms will be built into the technical design of DMAC to ensure that problems are detected and rapidly addressed.

Effective user feedback: IOOS will provide a continuous, vigorous outreach process addressing all levels of users of marine data, emphasizing the benefits of participation in IOOS/DMAC, and helping to identify and remedy difficulties encountered by those who are participating. In addition, this process will identify and address changing user requirements that drive the development and growth of DMAC.

Open design and standards process: DMAC will commit to an open software design. All standards and protocol definitions will be openly published so that participating organizations may create functioning DMAC components based on these specifications. The standards development process will be open and inclusive, so that it fosters buy-in by all stakeholders. Existing information technology and scientific standards will be used in preference to development of new solutions, whenever suitable standards exist. The standards and protocols will be of sufficient breadth and quality to guarantee interoperability of all observations and products. Institutions participating in IOOS will ensure that the data they contribute comply with these standards and protocols.

Preservation of data and products: Irreplaceable observations, data products of lasting value, and associated metadata will be archived for posterity in an efficient and automated manner.

CHALLENGES

The DMAC design faces a trio of competing characteristics:

1. **Loosely federated organizations:** No top-down corporate management structure exists to effectively manage major shifts in data management strategy (and the resulting dislocations) in order to achieve interoperability.

2. **Physically distributed repositories:** Data must reside and be managed at many distinct locations (some of which contain vast volumes of data).

3. **Heterogeneous data:** Classes of data range from huge satellite track records, to multi-dimensional model outputs, to Lagrangian drifters, to polygonal geographic regions and point measurements. Variables are from diverse disciplines and are unevenly distributed in time and space.

Though challenging, the technical requirements for DMAC that are imposed by characteristics 2 and 3, alone, could be addressed by relatively straightforward software engineering. Solutions that do not adequately embrace the loosely federated structure of IOOS, however, cannot succeed. Community-building considerations must be central in the design of the DMAC Subsystem.

COMMUNITY BUILDING

For DMAC to succeed it must achieve acceptance and recognition by marine data providers and data user communities. Only "gentle" (non-coercive) tools of persuasion will be effective within the loosely federated structure of IOOS. Individuals working in pursuit of their organization-specific goals must perceive that participation in DMAC will lead to a net gain toward achieving those goals. Thus, the greatest challenges for initial acceptance of DMAC and for subsequent growth in its usage are in the areas of community outreach and organizational behavior (the factors that enable a community to agree upon and use standards) rather than in technology.

Organizational challenges exist at both management and technical levels of the system. The leadership of existing marine organizations and programs must understand that access to the benefits that will accrue from IOOS depends upon their willingness to commit their organizations to the development and use of the DMAC Subsystem. At times short-term inconveniences to their organizations may occur as a result. Technical staff involved in the development of information management systems will need to ensure that their systems conform to the interoperability standards set by DMAC. Sometimes duplication of software functionality that is available through in-house systems may be necessary.

INTERNATIONAL COOPERATION

Producing global assessments and predictions of coastal ecosystem health and sea-level change, as well as addressing the other IOOS goals, requires that IOOS observations and data products be fully integrated with other national and international Earth observation efforts. Coordinated and sustained cooperation is already well established within the weather community, and the World Meteorological Organization's (WMO) World Weather Watch demonstrates the value of this international collaboration. Coordination is less well established in the ocean, ice, land, water, and climate observation communities.

Nevertheless, much important work has been accomplished on the international front. Table 1 summarizes some of the significant international programs and activities relevant to IOOS and DMAC.

Each of the programs listed in Table 1, and IOOS, recognize that until recently, international efforts to capture ocean-related, observational data have been hindered by a number of factors, including:

1. A lack of access to data (especially by the developing world), and to the benefits resulting from the value-added data products and new knowledge developed from these data;
2. Inadequate data integration and interoperability;
3. Eroding technical infrastructure;
4. Large temporal gaps in specific data sets and in observing locations;
5. Inadequate processing systems to transform data into useful information;
6. Insufficient end-user involvement;
7. Uncertainty over continuity of observation programs; and
8. Inadequate attention to long-term data archiving.

Table 1. Summary of International Programs Relevant to IOOS DMAC

International Program or Activity	Objectives relating to IOOS
International Strategy for Disaster Reduction (ISDR)	Enhanced understanding of natural hazards
World Climate Program (WCP)	Improved understanding of climate patterns and variability
Global Climate Observing System (GCOS) and the Conference of Parties (COP) of the United Nations Framework Convention on Climate Change (UNFCCC)	Improved climate monitoring
Global Ocean Observing System (GOOS), and the coastal module of GOOS	Global ocean and coastal monitoring, modeling, and forecasting. Establishment of Global Coastal Network (GCN) and Regional Alliances
International Oceanographic Data and Information Exchange (IODE)	Exchange of oceanographic data and information
Joint Technical Commission for Oceanography and Marine Meteorology (JCOMM)—established by the WMO and the Intergovernmental Oceanographic Commission	Intergovernmental coordination, regulation, and improved operational oceanographic and marine meteorological observing, data management, and services
Global Earth Observation System of Systems (GEOSS)	Framework for achieving comprehensive, coordinated, and sustained Earth observations

IOOS is well positioned to contribute to these international efforts, especially in the area of data management and communications. Internationally, IOOS will be the U.S. contribution to GOOS and GEOSS. Many of the contributors to the IOOS DMAC Plan are involved with international efforts addressing global ocean and coastal observing needs. This peer-level interaction has greatly benefited both the U.S. and international efforts, and has resulted in consistency among the different activities and programs with respect to requirements, governing principles, and goals. For example, the key functional components of GEOSS are quite consistent with those of the observing, modeling, and data management and communications subsystems envisioned for IOOS.

The IOOS DMAC Plan outlines a process for developing community-based standards that will enhance interoperability across a global, distributed observing system of systems. This process is especially relevant to the challenges listed above. Effective coordination among the contributors to all these programs is essential to realizing a truly interoperable, global and national coastal and ocean observation framework.

More-structured coordination mechanisms are needed as these programs and activities progress beyond the stage of high-level, broad agreements to the planning and implementation phases. For example, the WMO has identified its emerging Future WMO Information System (FWIS) as a core contribution to implementation of the data exchange and dissemination[5] functions of GEOSS. Observational data will be distributed through an interoperable data exchange framework, enabling discovery of, and access to, all of WMO program data using existing dedicated communications avenues, as well as Internet web services. The IOOS DMAC Plan is being used by the WMO as an early model for standards and protocol development.

Coordinated development of internationally accepted data management and communications standards that meet both the needs of IOOS DMAC and WMO FWIS is strongly recommended. Recent discussions between IOOS DMAC and the WMO Intercommission Coordinating Group (ICG) on the FWIS indicate that the near-term, high-priority areas for FWIS and DMAC are both consistent and complementary, including metadata content and representation, metadata and data representation, and data exchange standards and recommended protocols.

More formal programmatic coordination and linkages between the WMO and IOOS DMAC are needed. The DMAC Plan therefore recommends taking steps to address this need.

[5] The WMO Global Telecommunications System distributes weather observations and meteorological information worldwide.

Data Management Needs For Research Observatory Networks

NSF's Ocean Research Interactive Observatory Networks (ORION) program will establish a research observing network (the Ocean Observatory Initiative, OOI) and develop instrumentation and mobile platforms needed to enable the broad range of research envisioned for this system. This integrated observatory network will provide the oceanographic research and education communities with a new mode of access to the ocean. The network has three elements: (1) a regional cabled network consisting of interconnected sites on the seafloor spanning several geological and oceanographic features and processes, (2) re-locatable deep-sea buoys that could also be deployed in harsh environments such as the Southern Ocean, and (3) construction or enhancements to existing facilities leading to an expanded network of coastal observatories. The scientific problems driving the need for this infrastructure encompass nearly every area of ocean science and will provide Earth and ocean scientists with the opportunity to study multiple, interrelated processes over time scales ranging from seconds to decades; to conduct comparative studies of regional processes and spatial characteristics; and to map whole-Earth and basin-scale structures.

In addition to the ORION program, there are a number of observing networks funded by or proposed to NSF whose goals are to investigate Earth structure (EarthScope), terrestrial ecology (NEON), hydrology (CUAHSI), environmental remediation (CLEANER), and atmospheric structure (COSMIC). To fully realize the benefits of these observing networks, it is important that data systems established for each program are interoperable. This interoperability will facilitate investigations leading towards a better understanding of the entire Earth system.

Observing networks established with basic research and education as the primary mandates share many of the same data management and communication requirements that drive the design of the IOOS DMAC, but will impose significant further constraints. To keep pace with research, the system must have sufficient flexibility to evolve in response to changing and innovative data collection techniques. It must address rapidly evolving user requirements in the classroom, field, and laboratory, and in data analysis and synthesis. Research demands will continually stretch the capabilities of the data system to ingest diverse data types at varying frequencies. To meet the needs of the research and education communities, such a data management and communications system should be embedded in an information infrastructure that simplifies the task of finding and using data from multiple sources, and that facilitates collaborative work. The research infrastructure must embrace a central role in creating facilities for data management that differs from the "framework for interoperability between independent systems" paradigm of IOOS DMAC. The research infrastructure must also include analysis and visualization tools, a range of numerical models and model-data fusion tools, workflow tools, and one or more digital libraries. This infrastructure will be integrated with a computational grid to provide the necessary resources to support analysis, visualization, and modeling tasks.

As an information infrastructure suitable for ocean science evolves, a number of specific elements are likely to be required. Many of these elements can be most effectively implemented only through close coordination with other agencies and with the efforts to establish a US IOOS. In some cases, coordination with international efforts is also required. NSF expects to receive specific advice and recommendations on the requirements of this information infrastructure from the ocean science community and from information technology experts, as part of the OOI implementation design phase.

Section 2. Technical Analysis

IOOS DATA COMMUNICATIONS

The relationship among the DMAC Subsystem and other IOOS components and partners is depicted in Figure 1. Data flow within IOOS begins with the Observing Subsystem. Raw measurements from the Observing Subsystem elements are processed at various Primary Data Assembly and Quality Control sites to make them available to the uniform DMAC Data Communications Infrastructure through a DMAC Data Entry Point. Every IOOS data stream must have a DMAC Data Entry Point. The infrastructure consists of standards and protocols to support: (1) IOOS-wide descriptions of data sets (Metadata); (2) the ability to search for and find data sets of interest (Data Discovery); (3) the ability to access the data in an interoperable manner from client applications (Data Transport); (4) the ability to evaluate the character of the data through common Web Browsers (Uniform On-line Browse); and (5) the ability to securely archive data and metadata and retrieve them on demand (Data Archive).

The DMAC Data Communications Infrastructure provides access to IOOS data for all IOOS components and partners. Bi-directional communications will exist between independent data management systems both internal and external to the IOOS framework—data management systems from regional and international entities and from distinct disciplines such as meteorology. The DMAC Data Communications Infrastructure also conveys data, metadata, and data products to users' applications (programs) and to those entities both inside and outside of IOOS who generate value-added information products. The information products and data address the interests of U.S. society through the advancement of the seven IOOS goals (see Preface). Ultimately, it is the needs of U.S. users that guide the selection of new measurements, infrastructure, procedures, and products through IOOS User Outreach mechanisms.

The DMAC Plan (this document) offers a detailed, phased implementation strategy specifically for the development of the DMAC Subsystem Data Communications Infrastructure and Archive.

THE OBSERVING SUBSYSTEM AND PRIMARY DATA ASSEMBLY/QUALITY CONTROL

IOOS Observing Subsystem elements are managed by regional, national, and international entities. The measurements are highly heterogeneous, originating from surveys (e.g., fish stock assessments), cruise measurements, laboratory measurements, satellites, and automated inputs from *in situ* and remotely sensed sources that include time series, profiles, swaths, grids, and other data structures. A wide range of telemetry systems, as well as the WMO's Global Telecommunication System (GTS), are used to transfer data from the measurement platforms to and among the locations at which Primary Data Assembly and Quality Control (PDA&QC) occur.

IOOS Data Communications

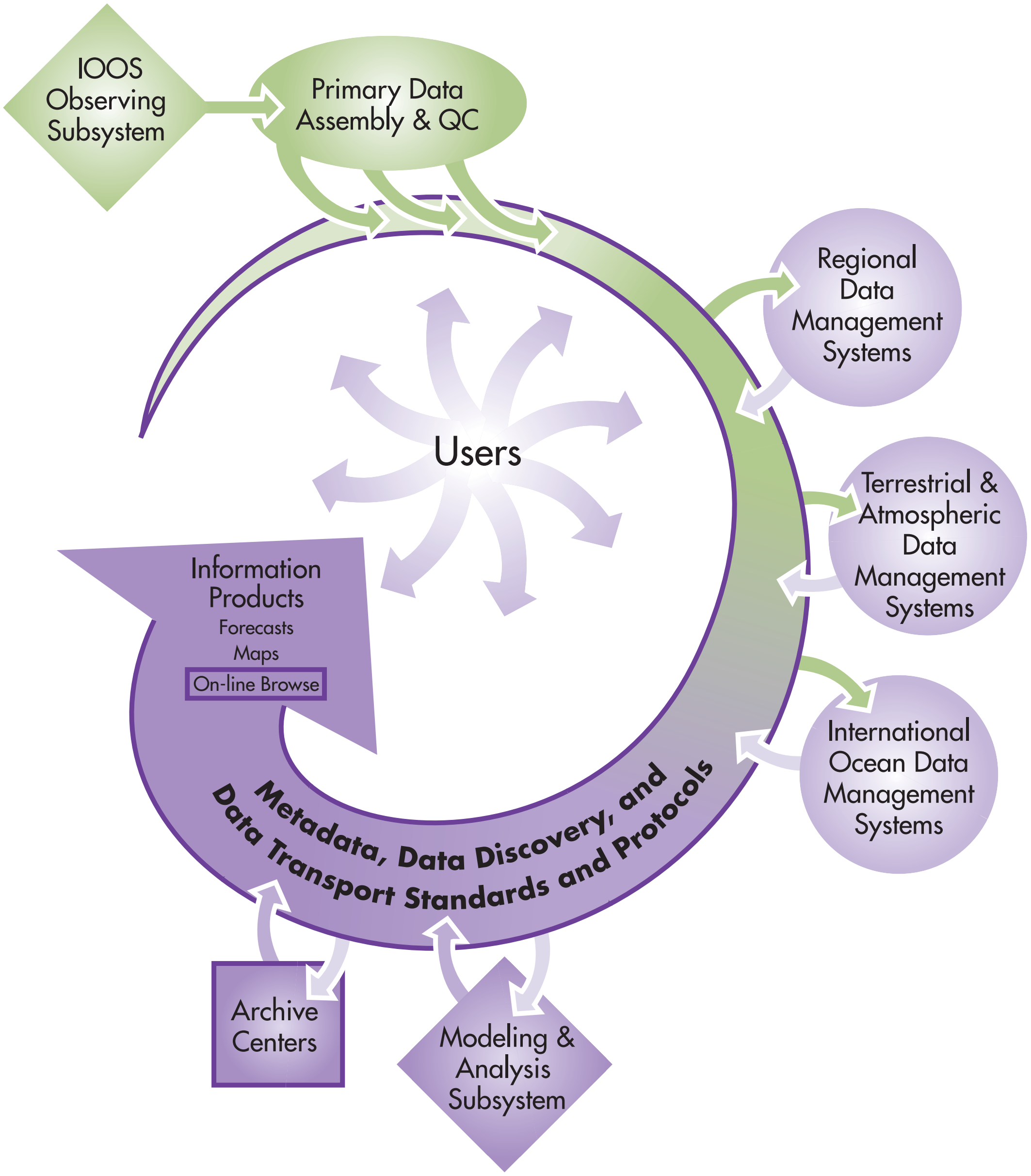

IOOS Observing Subsystem
Primary Data Assembly & QC
Regional Data Management Systems
Users
Terrestrial & Atmospheric Data Management Systems
Information Products
Forecasts
Maps
On-line Browse
International Ocean Data Management Systems
Metadata, Data Discovery, and
Data Transport Standards and Protocols
Archive Centers
Modeling & Analysis Subsystem

Figure 1. Solid outlines indicate the elements of the IOOS Data Communications framework, which are detailed in the DMAC Plan. The arrows flowing outward from users indicate the feedback and control mechanisms through which users ultimately direct the functioning of all parts of the system. Note that the National Data Management Systems are included in the concept of Primary Data Assembly and Quality Control.

PDA&QC processes lie at the interface between the IOOS Observing Subsystem and the DMAC Subsystem. In general, some form of PDA&QC is required before IOOS data can be used. The character of this activity varies greatly by data type. It includes such processes as hand entry of numbers from log sheets, conversion from raw instrument voltages to physical units, calibrations, and the quality assessment of measured values against neighboring measurements or climatological norms. It is in this activity that myriad individual measurements are assembled into "data sets" (named collections of data) that may be referenced and queried as a whole. The activity may be conducted as a part of a data management strategy for a particular measurement type, where examples include NOAA/NDBC management of mooring data and NASA/PODAAC management of ocean satellite data. Alternatively, it may be managed by an IOOS RA, for example assembling regional ecosystem measurements; or it may be found in conjunction with operational forecast modeling and state estimation, where a prime example is the U.S. GODAE Server (which operates in close association with the Navy's Fleet Numerical Meteorological and Oceanography Center).

Because of its location at the boundary between two subsystems, the responsibilities for PDA&QC are shared between the Observing Subsystem and the DMAC Subsystem. As a general rule, the DMAC Subsystem's responsibility for procedures and standards of scientific quality control (QC) are limited to providing mechanisms to ensure that QC flags are reliably associated with the corresponding measurements. The standards and procedures for quality control will be developed by the relevant marine science communities. The DMAC Plan (this document), however, also includes a management responsibility to ensure that all IOOS data streams undergo primary data assembly and quality-control processing to make them available using DMAC standards and protocols at a specified DMAC Data Entry Point. In some cases, this responsibility may involve support for specific facilities, such as the U.S. GODAE Server. In the future, we may see the growth of "intelligent" instruments that can perform data assembly and quality control functions at the instrument subsystem level[6].

The telemetry systems that convey data from sensors to primary data centers/ sites and the standards and procedures for data assembly and quality control lie outside the scope of this DMAC Plan. It is recognized, however, that these areas require careful planning within IOOS, and these topics will be addressed in future IOOS plans. The intent of the current DMAC Plan is to provide a foundation for these community building and planning activities.

[6] Delin, K.A., 2002, The Sensor Web: A Macro-Instrument for Coordinated Sensing, *Sensors*, 2, 270-285)

THE DMAC SUBSYSTEM – A DATA COMMUNICATIONS INFRASTRUCTURE

Scope and Evolution of the DMAC Subsystem

The DMAC Subsystem is a framework for the integration of independent and heterogeneous data management and communications systems, large and small. It builds upon and complements its member systems; it does not in any sense replace them. The member systems integrated through the DMAC framework will develop over time to meet the demands placed upon them by growing data volume, increasing data complexity, and the evolving user needs. For example, network bandwidths will be increased to meet growing data volumes, independently from DMAC integration. Similarly, processor speeds, mass storage capacities, and data base sophistication will increase. The scope of the DMAC Plan is largely limited to the evolving framework for the integration of its member systems. The DMAC Plan is not a roadmap to the evolution of the many individual systems that the DMAC Subsystem comprises (nor could it ever be).

The DMAC Subsystem will include a data management infrastructure that consists of a suite of components—standards, facilities, software, and supporting hardware systems. The design and planning for the DMAC framework will emphasize continual, smooth evolution. The components upon which the DMAC Subsystem is built will, themselves, be an evolving collection. New components will be introduced; recognized components will be advanced; obsolete components will be removed. A significant level of duplication of function between components will be tolerated as a necessary consequence of a continuously evolving system. The DMAC Standards Process (p. 49) will define the manner by which the level of maturity of components will be designated: R&D, pilot, pre-operational, and operational (see IOOS Development Plan [www.ocean.us] for details on these designations).

Particularly challenging is the process for designating the maturity level of components to be included in the initial DMAC Subsystem. At the outset, no formal DMAC standards process exists. Even if a suitable community standards process were broadly accepted today, experience has demonstrated that application of a standards process requires considerable time. Yet there is an imperative to provide immediate guidance to would-be data providers. To address this need, the DMAC Plan includes preliminary recommendations for: (1) the maturity level designation of certain named components that are viewed as essential to the initial subsystem and (2) a roadmap leading to rapid designation of other initial components by community-based working groups.

Metadata Management

(see Appendix 1 for a more detailed discussion)

Metadata are a critical component of the Data Communications Infrastructure, required for all key infrastructure functions: discovery, transport, uniform on-line browse, archive, and access. Metadata considerations are equally applicable to both data and information products. A sustained commitment to the creation and management of metadata is a requirement to support the ability of users to locate and use data. Sustaining this commitment will be a challenge to IOOS Leadership. Certain classes of metadata (e.g., variable names, units, coordinates) are indispensable to any utilization of the data, and must be tightly bound to data transport as an integral part of the data delivery protocols. We refer to this class of information as "use metadata." Other types of information, such as descriptions of measurement and analysis techniques, help to place the data in context and are essential to overall understanding and usefulness of the data. We refer to this class of information as "descriptive metadata."

Federal Executive Order[7] mandates, "each [federal] agency shall document all new geospatial data it collects or produces, either directly or indirectly, using the standard under development by the Federal Geographic Data Committee (FGDC)." While there are many IOOS members to whom this mandate does not directly apply, the breadth of participation found in FGDC makes it a natural initial foundation for DMAC[8] metadata. The FGDC developed the Content Standard for Digital Geospatial Metadata (CSDGM) that provides a common set of names and definitions of compound and individual data elements used to document digital geospatial data. The content of FGDC records encompasses the elements of many other metadata formats including most of the content contained in the Directory Interchange Format (DIF) metadata records in common use by international IOOS partners. The scope of FGDC, however, is far broader than marine data. A focused activity to determine the precise information that will define DMAC-standard metadata content, along with mechanisms for extensibility, are initial tasks identified within Part II of this phased DMAC Implementation Plan. Controlled keywords (standardized topic names) and controlled vocabularies (standardized technical terminology) need to be adopted or developed. The breadth of scientific disciplines that will participate in DMAC guarantees the existence of overlapping terminology, and therefore tools and techniques to perform translation among these controlled vocabularies are needed. "Parent-child" hierarchies of metadata must be supported, since marine data are often managed as collections of observations that require description both as inventories and as individual observations.

[7] Executive Order 12906 (April 11, 1994)

[8] It should be noted that the International Organization for Standardization (ISO) developed a standard for geospatial metadata. This standard, ISO 19115, was formally accepted in May 2003. It is anticipated that the next version of FGDC Content Standard for Geospatial Metadata (CSDGSM) will be in a form compatible with the international standard.

Data Discovery

Data Discovery will initially be implemented as a process of locating data and products of interest through searches of metadata. (The ability to search within the data itself—so-called "data mining"—will be incorporated into DMAC data discovery as the technologies for doing so mature.) The data-discovery capabilities provided by the DMAC Subsystem will complement and extend the search capabilities that are widely available today through commercial web search sites such as *Google*™ and *Yahoo*™. Typically, search parameters include geospatial location, temporal information, keywords, controlled vocabulary items, and in some cases "free text." The speed and reliability of searching are improved when the catalogs are centrally cached. Centralized caching greatly improves the system interoperability, and may be achieved through automated metadata harvesting such as that being developed by the Open Archive Initiative (OAI – http://www.openarchives.org) Protocol for Metadata Harvesting. This mechanism may serve as a tool or a template for DMAC because it provides cross-repository harvesting of metadata in XML format.

In marine data management today, metadata records are often managed independently from the data. As a practical matter, the initial solutions to Data Discovery within DMAC must support this architecture. The separation of metadata record management from the associated data management, however, leaves two significant challenges to be solved: (1) assuring that links between metadata records and points of data access (see Data Transport) remain valid over time and (2) assuring that changes made to data sets are reflected in corresponding changes to the metadata records. In addition, the issue of what constitutes a "data set" from the data discovery perspective when the data are assembled on-the-fly from distributed sources, must be addressed. These challenges are made more acute because a given data set may be replicated and made available at multiple data-providing organizations. Changes to the data (new versions) may be made independently by these organizations. Part II of the DMAC Plan lays out steps to investigate and resolve these technical problems as the DMAC Subsystem matures.

The data-discovery capabilities of the Data Communications Infrastructure will permit humans to formulate queries directly to the catalogs, and also support machine-to-machine queries. DMAC will provide at least one "portal"—web pages through which end users can search for IOOS data. In the mature phases of DMAC, search entry points will also exist at many alternative locations that relay their queries via machine-to-machine communications to the DMAC search service. Advanced data discovery and metadata management techniques, such as the "Semantic Web" (www.w3.org/2001/sw/), will be actively pursued and considered for incorporation into the system as they mature.

Data Transport

(see Appendix 2 for a more detailed discussion)

The concept of a "web service" is fundamental to the ability of DMAC to connect quasi-independent systems. The term "web service" is used in many contexts today; in the DMAC Plan we intend the term to mean reusable software components that provide a standardized means for computer systems to request data and data processing from one another, typically using messages expressed in the eXtensible Markup Language (XML) and conveyed using the ubiquitous communications protocol of the World Wide Web, HTTP (the Hyper Text Transfer Protocol). Web services make data and software capabilities available on one computer, accessible to other computers via the Internet through the familiar Universal Resource Identifiers (URIs) that begin with "http://".

DMAC will endorse a suite of web services to serve as a shared communications toolbox connecting systems that are operated by regional, state, and federal agencies; academic projects; international partners; and others. Data suppliers (including PDA&QC sites) will be responsible for making data accessible through DMAC web services tools and standards. Data users will find that in many cases the software applications upon which they depend for product generation and scientific analysis will be "DMAC ready" (possibly with some adaptation required), having been adapted to work directly with the DMAC web services. In this case, the applications will perform much as if the data existed on users' local hard drives. To provide a bridge from current practices, compatibility between DMAC and user applications may be achieved using formatted files that are made readily available as products through DMAC web services.

The DMAC data transport framework will designate a suite of freely available software components adequate to meet typical needs. The goal in doing so is to minimize the barriers to participation in the DMAC. The uniformity provided by the DMAC web services standards will permit all the components related to data transport to be interoperable at the machine level (i.e., data can be moved from one component of the system to another, retaining complete syntactic[9] and semantic[10] meaning without human interaction).

[9] Syntactic Meaning: refers to the syntax of a data set—the atomic data types in the data set (e.g., binary, ASCII, real), the dimensionality of data arrays (P is a 90 by 180 by 25 by 12 element array), the relationship between variables in the data set (lat is a map vector for the first dimension of P), etc.

[10] Semantic Meaning: refers to the semantics of the data contained in the data set—the meaning of variables (P represents phytoplankton abundance), the units used to express variables (multiply P by 8 to obtain number of specimens per cubic meter), special value flags (a value of –1 means missing data, 0 land,...), descriptions of the processing or instrumentation used to obtain the data values, etc.

Web services exist in the context of the web. Data transport on the web involves protocols at multiple levels. The foundation of transport on the Internet is TCP/IP, which handles the routing of "packets" of information between source and destination hosts. Layered upon TCP/IP are a variety of protocols, for example, FTP, HTTP[11], and SMTP. These protocols are supported on a very wide range of computers and operating systems, and all of them will be used to move various types of data over the network as part of IOOS. There is, however, no uniform syntactic and semantic meaning that is guaranteed for data communicated via these transfers, and therefore no guarantee of immediate interoperability among computer applications. This function is the role of the DMAC web services standards.

Several solutions currently exist for the syntactic description and transport of binary data, however none is universally accepted. The most broadly tested and accepted of these solutions within oceanography are the OPeNDAP[12] data access protocol and the Open Geospatial Consortium, Inc. (OGC) (http://www.opengeospatial.org/) data access protocols.

OPeNDAP

OPeNDAP underlies the National Virtual Ocean Data System (NVODS[13]). OPeNDAP has been serving the marine community since 1995. OPeNDAP provides the very general approach to data management that is needed in support of research and modeling. OPeNDAP also supports server-side subsetting of data, which greatly reduces the volumes of data that need be transferred across the Internet in many cases. This capability is vital when considering the large volumes of data that will be produced in the near future by observing platforms and modeling activities. Tables 2 and 3 provide estimates of near-term data flow for the U.S. IOOS using selected data streams as examples; the lists are not intended to include all observing system data types. The DMAC Steering Committee recommends the designation of OPeNDAP protocol as an initial "operational"[14] component for Data Transport of gridded data, and a "pilot"[15] component for the delivery of non-gridded data.

[11] When using a web browser, most images and text are delivered via http.

[12] The Open Source Project for a Network Data Access Protocol (OPeNDAP) is a non-profit corporation formed to develop and maintain the middleware formerly known as the Distributed Ocean Data System (DODS).

[13] NVODS was created in response to a Broad Agency Announcement (BAA) issued by the National Oceanographic Partnership Program in 2000.

[14] "Operational" is stage four of a four-level classification scheme for the maturity of system components within IOOS: R&D, pilot, pre-operational, operational. See IOOS Development Plan (www.ocean.us).

[15] "Pilot" is stage two of a four-level classification scheme for the maturity of system components within IOOS: R&D, pilot, pre-operational, operational. See IOOS Development Plan (www.ocean.us).

Table 2. Near-term Data Flow Estimate for U.S. IOOS[16] (NASA & NOAA sources excluding cabled observatories)

Data Source Class	Data Source Type	Annual Volume (MB)	Totals Annual Volume by Class (MB)
Direct Observation Systems: Buoys	Moored buoys - NDBC, TAO, MBARI, etc.	2,000	
	Drifting buoys - Surface, APEX, etc	100	2,100
Direct Observation Systems: Ships	NOS Charting/Resurvey	1,800,000	
	Other Ship Data - VOS MET, XBT, CALCOFI, etc.	12,000	1,812,000
Remote Sensing Systems	Surface currents-CODAR	13,000	
	Sea Surface Temperature - AVHRR, MODIS	500,000	
	Sea Surface Height – T/P, JASON1	120,000	
	Ocean Vector Winds - QuikSCAT, SeaWinds	130,000	
	Ocean color-MODIS, SeaWIFs	400,000	1,163,000
Total Near-Term Annual Data Flow			**2,977,100**

Table 3. Near-term Data Flow Estimate for U.S. IOOS: Cabled Observatories[14] (NASA & NOAA sources)

Data Source	Annual (GB)
MBARI/MARS	13,000
HUGO	5,000
LEO 15	5,000
Neptune (approval pending)	177,000
Total cabled Observations	**200,000**

[16] Science Applications International Corporation. October 18, 2002. "Consolidated Data Flow Estimates for the Integrated Ocean Observing System (IOOS)." Submitted to the National Ocean Service, National Oceanic and Atmospheric Administration, Department of Commerce. Note that this was a preliminary study. Further investigation is needed.

Biological Data Considerations

Management and stewardship of biological data present special challenges, which historically have often been neglected. Biological data management requires that special consideration be given to metadata (see Part III, Appendix 7 for a more detailed discussion). For example, the basic units for biological data are species. New species are continually being discovered and named, and names of recognized species are sometimes changed. The hierarchical tree of evolutionary relationships among species, and the associated hierarchical nomenclature, must continually be revised to incorporate new information. Biological data systems require name translators that provide currently recognized scientific names from synonymous scientific names and common names. The taxonomic authority for each major group of organisms maintains the accepted list of species, with oversight from the Global Biodiversity Information Facility (GBIF), Catalogue of Life, and organizations such as the Integrated Taxonomic Information System (ITIS)/ Species 2000, and the Ocean Biogeographic Information System (OBIS). Protocols for using DNA sequence data as a "bar code of life" have been proposed as an aid to taxonomic identification and evolutionary relationships.

OPeNDAP uses a discipline-neutral approach to the encapsulation of data for transport. Discipline neutrality is seen as a key element of the data transport protocol for a system such as IOOS that covers such a broad range of data types and data users—for example, four-dimensional geospatial grids, time series, vertical profiles, and species type and abundance. The protocol ensures that the structure, numeric values, and metadata attributes of the data are preserved between server and client. It does not, however, impose a particular geospatial data model. For example, OPeNDAP does not "understand" what a time series is, nor does it "know" the significance of "phytoplankton_abundance" as a variable name. When transmitting a simple time series, OPeNDAP merely knows that it is conveying a one-dimensional array of values with attributes such as units = "seconds" and title = "Phytoplankton Abundance" attached to it. Such an approach greatly lowers the barrier to initial participation by data suppliers, since nearly all data holdings can easily be encapsulated in this fashion and sent over the Internet. It also ensures the level of generality needed to provide semantically aware data transport for the very broad range of ocean data classes.

To achieve the desired level of interoperability, the mature DMAC will require that all data are delivered in a consistent geospatial data model (or family of models). In this phased DMAC Plan, the development of a rich and comprehensive data model occurs in parallel with the pilot deployment of OPeNDAP data servers and clients. The comprehensive data model(s) must harmonize with ongoing work in several communities, such as GIS and forecast and climate modeling. It must standardize controlled vocabularies, include the encoding of ocean biological data and taxonomies such as those demonstrated in OBIS[17], and describe a broad range of data structures, including for example, spectral and finite element models, arbitrary curvilinear coordinate systems, and multilevel hierarchies of *in situ* measurements. The parallel progress made by deploying OPeNDAP servers and clients, while simultaneously designing a comprehensive data model and community-wide metadata standards, is a key element of the phased DMAC Plan. This element will enable rapid progress both in capacity building and in broad community standards building. It is anticipated that the design of the data model may necessitate changes or additions to OPeNDAP.

OGC

The Open Geospatial Consortium, Inc. (OGC) is a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services. OGC works with government, private industry, and academia to create open and extensible software application programming interfaces for geographic information systems (GIS) and other mainstream technologies. Adopted specifications are available for the public's use at no cost.

The OpenGIS Web Mapping Services is a family of specifications that enable servers to dynamically query, access, process, and combine different types of spatial information over the web. Two of the Web Mapping Service specifications are relevant to data transport (as the term is used in this document): (1) OpenGIS Web Feature Service (WFS) and (2) OpenGIS Web Coverage Service (WCS).

The WFS specification proposes interfaces that allow a client to retrieve geospatial data encoded in the XML-based Geography Markup Language (GML). Geographic "features" are described by a set of properties, where each property can be thought of as a {name, type, value] tuple. The geometries of geographic features are restricted to what OGC calls simple geometries. A simple geometry is one for which coordinates are defined in two dimensions, and the delineation of a curve is subject to linear interpolation. The traditional zero-, one-, and two-dimensional geometries defined in a two-dimensional spatial reference system are represented by points, line strings and polygons. In

[17] OBIS is an on-line, open-access, globally distributed network of systematic, ecological, and environmental information systems. Collectively, these systems operate as a dynamic, global digital atlas to communicate biological information about the ocean and serve as a platform for further study of biogeographical relationships in the marine environment (http://iobis.org/).

addition, the OGC geometry model allows for geometries that are collections of other geometries, for example multiple points. Currently WFS and GML directly address only restricted classes of marine data—those that can be described as two-dimensional features. However, the potential exists for applying WFS to further marine data structures (e.g., time series and vertical profiles). The DMAC-SC recommends that WFS be examined for incorporation into the DMAC data transport suite and that community working groups be formed to consider extensions to the protocol (WFS and/or GML) that may permit it to more completely address marine data transport requirements.

The WCS SPECIFICATION supports electronic interchange of geospatial data as "coverages, "that is, digital geospatial information representing space-varying phenomena. Like WFS, WCS allows clients to choose portions of a server's information holdings based on spatial constraints and other criteria. Unlike WFS, which returns discrete geospatial features, the WCS returns representations of space-varying phenomena that relate a spatio-temporal domain to a (possibly multi-dimensional) range of properties. WCS provides coverage data (that is, values or properties of a set of geographic locations), bundled in so-called "well-known" data format. The DMAC-SC recommends that WCS be examined for incorporation into the DMAC data transport suite, and that community working groups be formed to consider extensions to WCS that may permit it to smoothly interoperate with multi-dimensional data formats that are in common use in the marine data community, such as netCDF.

Uniform On-line Browse

The DMAC Plan anticipates that many IOOS data providers will host metadata-enabled, open-source, or commercial on-line browse tools for end users. In addition, effective management of the DMAC Subsystem requires a system-wide view of IOOS data—the ability to visualize and assess all IOOS data in a uniform manner. The Uniform On-Line Browse capability of DMAC must provide geo- and time-referenced graphics and data in human-readable tables. It will use the DMAC Data Transport services for access to IOOS data. It must be accessible through standard web browsers. The DMAC Subsystem must provide a seamless segue from Data Discovery to Uniform On-line Browse.

The Uniform On-line Browse capability is a form of information product (see Information Products and Applications). As such it must be an effective informational tool for its target user groups, namely the marine data specialists across the IOOS community who have responsibilities for managing elements of IOOS. These users may be assumed to share a high level of technical training, but they represent diverse professions, including scientists, computer specialists, engineers, and technical managers. (Although the Uniform On-Line Browse capability must be designed principally to address the needs of these users, it will also be accessible to the general public and will doubtless prove useful to many groups of users.)

The DMAC Subsystem may support multiple user interfaces for Uniform On-Line Browse, as needed to address the range of users who have responsibilities for managing elements of IOOS. The Uniform On-Line Browse capability must be accessible through a computer-to-computer "web service" interface, enabling the browse products that are provided to be incorporated into a range of applications. The Uniform On-line Browse architecture must be designed to scale with the growth of IOOS, and must be sufficiently flexible that new and modified browse products can readily be added to meet evolving user needs.

Within the National Virtual Ocean Data System, the Live Access Server (LAS; http://www.ferret.noaa.gov/LAS) has been effectively used for several years in the delivery of on-line browse capabilities across a broad range of marine data types. The DMAC-SC recommends the designation of the LAS as an initial pre-operational component for Uniform On-line Browse. Recently, web visualization tools based upon the OGC protocols have achieved prominence. The DMAC Plan further recommends that OGC-compatible GIS web services be examined as candidates for DMAC Uniform On-Line Browse.

Data Archive and Access

(See Appendix 3 for more detailed discussion)

The Data Archive System will receive and provide access to both real-time and delayed-mode data and metadata, serving the needs of real-time assessment and prediction, scientific research, and all others who require access to archived IOOS data. It will be a high priority for the Archive System to ensure that all valuable data are sent and that an exact copy is received. The Archive System will be designed to detect and correct failures using a combination of technological backup and expert oversight that will check the integrity of the data streams. In addition, during the phased implementation, a comprehensive process will be undertaken to ensure that all critical data streams and existing historical archives are inventoried and are scheduled to enter into the system.

The Data Archive System will consist of a designated set of existing and new facilities. Initially, existing centers will be the basis for the System; operating principles, requirements for additional facilities, and cross coordination among facilities will all be defined during the early planning phase. The Archive System will include distributed archive centers, regional data centers, modeling centers, and data assembly centers (Figure 2). Although data may flow from observing subsystem components to any of the four types of centers, at least one copy of each observation will ultimately re-

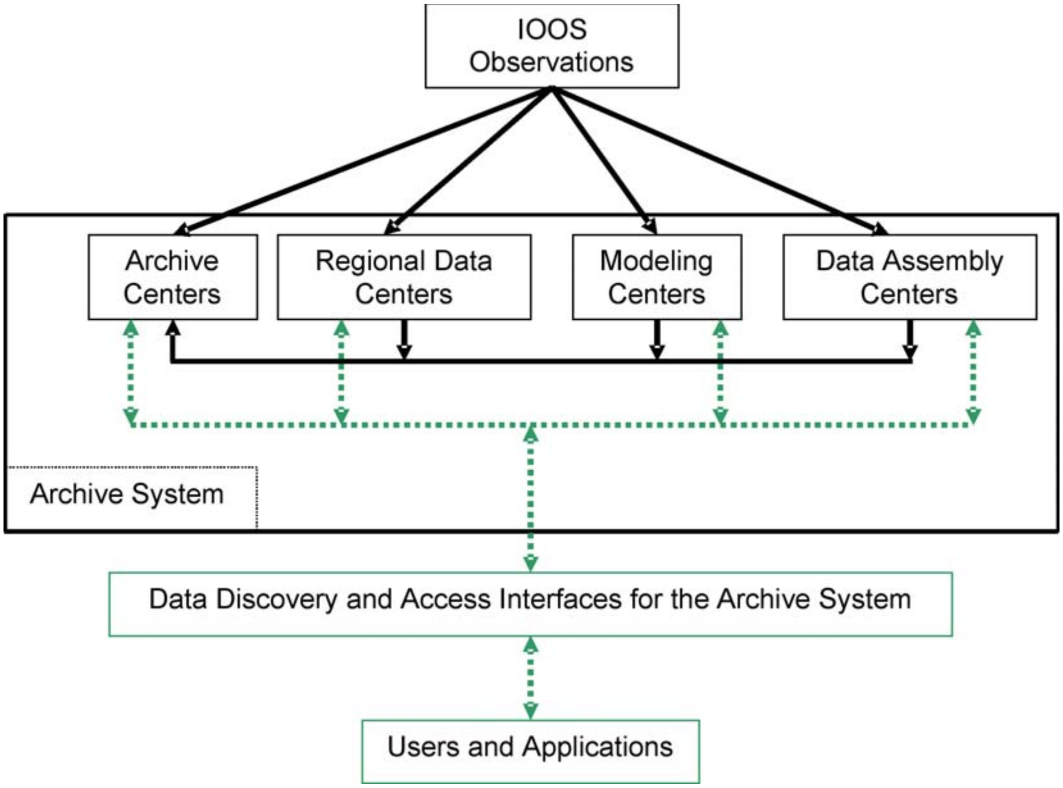


Figure 2. The Archive System represents an alternative view of those DMAC Subsystem elements that are involved with archiving data. Primary data archiving (solid lines) and access (dashed lines) show data flow. Not shown are other data flows that are essential to IOOS but not directly pertinent to the Archive System.

side in an archive center. Data will be considered as officially in the Archive System if the following two conditions are met: (1) the data are held and access is provided by one of the Archive System centers and (2) there are established procedures in place to preserve the data at an archive center. Through this approach, data will be under IOOS management early on in their life cycle, thereby maximizing the amount of securely archived and uniformly accessible data. It is probable and practical that more than one type of center may be physically collocated, for example, a data center may be an entity at a national archive center.

All centers in the Archive System may be responsible for acquiring and providing data, but (by definition) it is the archive centers that will have primary responsibility for preserving data for the long term (Figure 2). To qualify as an archive center, a data center must be able to manage multiple copies of the data and metadata, create and verify the metadata, frequently check data integrity, and have plans to evolve systems and media through generations of technology. Data will be preserved according to data categories, which will be developed by an Archive Working Group during phased implementation of the DMAC, and according to U.S. National Archives and Records Administration (NARA) and other federal guidelines.

Technologies and standards developed by metadata management and data transport activities will be an integral part of the Archive System. In the mature DMAC Subsystem, which is required to deliver data in a timely manner, it is anticipated that data and metadata will be received by the archive centers and redistributed through the DMAC transport mechanisms in standardized formats, eliminating many of the delays and difficulties of the non-standard and diverse methods that have burdened the systems in the past. Evolution to this state will be stepwise so that current data services are not interrupted and users can make a smooth transition. The result will be a system that provides uniform access across multiple centers, and provides data discovery and access by both humans and machines. Furthermore, all irreplaceable observational and research-quality data that are difficult to regenerate will be maintained and managed in perpetuity.

Unrestricted data access is a primary principle for all IOOS data, however, circumstances may arise where temporary restrictions are permitted. These instances are envisioned to be short term, where the burden of managing data set authorization and authentication can be offset by the reduced cost and increased efficiency of archiving the data at an early stage. The opportunities for limited restricted access, data security, and metadata and data discovery support offered by the Archive System are an asset, previously unavailable, and are intended to encourage broad participation from the scientific community.

The scientific community will add value to the data that will become part of the Archive System. The System will enable scientific endeavors that make comparisons of model and observed data, develop analyzed and reanalyzed data products, provide additional data quality control, and thereby quality checks on 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 Archive Policy

All facilities that participate as official archive centers in the Archive System will agree to adhere to data archiving guidelines that will be established in the phased implementation of the DMAC. A few key points that will be part of the guidelines are:

- Data distribution policies will follow the international recommendations of the IOC and WMO. Generally, the policy will call for full and open sharing of data and products. As a possible extension, the ability to provide restricted access for limited periods of time may be provided in certain cases;
- Data will be made accessible, to the greatest extent practical, on line and at no cost to the user. Data from off-line sources will similarly be available at no more than the cost of providing the service;
- Centers in the Archive System will make the data and metadata available using the DMAC transport protocols, metadata standards, and data discovery interfaces. The details of the transition from existing access systems to systems using the DMAC standards remain to be determined;
- The archive centers in the Archive System will have a data and metadata migration plan to accommodate media and system evolution and assure long-term preservation of irreplaceable data;
- All data collected and prepared under IOOS funding shall be submitted (or, in appropriate cases, notification of its availability shall be submitted) to the IOOS Archive System;
- As new versions (upgraded or changed) of a data set become available the versions will be distinguishable through standard metadata. Old versions can be deleted only under restrictive circumstances—when all relevant IOOS data policies and federal regulations are met.

MODELING AND ANALYSIS SUBSYSTEM

The IOOS Modeling and Analysis Subsystem has responsibility for the generation of numerical (digital) data products through: (1) computer modeling and the assimilation of marine observations, which provide estimations of the current state of the marine environment and forecasts of its future state and (2) analysis of data collections, which incorporate late-arriving observations and apply further quality controls in order to ensure that the historical record of marine observations is as complete and accurate as possible. The activities of the Modeling and Analysis Subsystem are carried out at many distinct centers. The term "center" may refer to an organization that has a specific focus on numerical modeling or data analysis, or it may refer to an individual project embedded within a university, state government, or other organization. It is the responsibility of IOOS Governance and User Outreach mechanisms to ensure that the Modeling and Analysis Subsystem meets the needs of the full range marine stakeholders.

Planning for the Modeling and Analysis Subsystem and the particular numerical data products that IOOS must produce lies outside the scope of the DMAC Plan (this document). It is recognized, however, that these topics require careful planning within IOOS, and they will be addressed in future IOOS plans. The intent of the DMAC Plan is to provide a foundation for these community-building and planning activities.

INFORMATION PRODUCTS AND APPLICATIONS

The ultimate goal of DMAC is to provide information about the marine environment to end users in a manner that permits them to advance the seven goals of IOOS (see Preface). Information can be provided to users through the generation of Information Products, or through ingestion of data into DMAC-ready applications. Information Products include text documents, such as printed assessments of fish stocks; verbal reports, such as wave height announcements on marine radio; maps and charts, including GIS layers; and graphics, animations, 3D visualizations and other media generated by computers to assist with the communication of information. The term "DMAC-ready Applications" refers to those applications that can directly access data and information through DMAC standards and protocols. This designation will (in the mature DMAC) include common GIS applications, common scientific analysis and visualization applications, educational software, and common business tools, such as spreadsheets and word processors.

Because IOOS is a user-driven system, the IOOS User Outreach and Governance mechanisms must ensure that the users' needs for Information Products and DMAC-ready applications are continually assessed and accommodations made within IOOS to meet them. Information Products will be generated at all levels of IOOS: by Primary Data Assembly Centers/sites (e.g., the Argo GDACS); by the DMAC Communications Infrastructure (see Uniform On-line Browse); by DMAC Modeling Centers and Archive Centers; and by the regional, international, and discipline-specific data management systems that interoperate with DMAC data.

It is understood, however, that these products alone will not be sufficient to meet the needs of all user groups. There is a vital role for private sector IOOS partners in providing users with specialized value-added Information Products and DMAC-ready applications. Production and sale of value-added products by the private sector is to be encouraged, and data providers and users from both private and public sectors should be able to contribute to and use IOOS data and information. This policy is consistent with policies in the Paperwork Reduction Act of 1995 (44 U.S.C. §§ 3501 et seq.) and Office of Management and Budget (OMB) Circular No. A-130.

INTEROPERABILITY WITH OTHER DATA MANAGEMENT SYSTEMS

The DMAC Data Communications Infrastructure provides for interoperable communications between DMAC components and other data management systems containing data that are of interest to IOOS users. These entities include data management systems operated by disciplines lying outside of marine sciences (e.g., public health), data management systems operated by other nations or international bodies, and specialized data management systems that may be operated by RAs within IOOS. Like the DMAC Subsystem these systems may be highly complex and involve multiple distributed partners. They may use data communications infrastructures that are distinct from the DMAC Standards and Protocols. To achieve interoperability with these entities, software "gateways" must be built that will translate between standards and protocols used within the DMAC Subsystem and those used in the other data systems. Similar considerations apply to commercial organizations that have adopted custom "enterprise" solutions[18], as well as frameworks adopted by communities that share specialized computing needs, such as the high performance computing community's DataGrid[19] and the GIS community's geospatial data systems. Thus, as IOOS matures, it should be viewed as a system of systems in which the DMAC Data Communications Infrastructure provides a uniform language for communications.

IT SECURITY

IT Security IOOS is being deployed in a distributed, heterogeneous information technology (IT) environment with web services as the eventual target architecture. This implementation environment will likely be characterized by a decentralized architecture, decentralized administration, multiple server and client components, and open access to and from the public Internet. Migrating to a web services architecture from the present HTML-based architecture promises to offer significant savings in the cost and development time. It will also significantly enhance data and information services. However, this approach requires that careful attention be devoted to the enhanced protection required to address the additional security vulnerabilities introduced. This protection will be in addition to the customary methods now employed: firewalls, intrusion detection, and other approaches for guarding against security breaches. The IOOS therefore faces a number of security challenges that include:

- Participants joining the IOOS enterprise are accustomed to operating under diverse security guidelines and cultures that may not conform to required federal IT security practices.
- Agreement on and compliance with a common security policy must be reached across multiple heterogeneous systems.
- "Desktop-level administrators" must be able to understand and implement IOOS security policies and measures, and will often be in environments where IT management resources are limited.
- Many legacy applications will be incorporated into IOOS that will be web-enabled, but may not have been originally designed for exposure to the public Internet or for use in a structured IT security environment.

[18] "Enterprise" solutions are typically commercial interoperability frameworks that operate on secure network connections.

[19] http://www.globus.org/datagrid

Some of the IOOS data will be used in time-sensitive environmental forecasts that affect the protection of life and property. In these instances, threats from denial of data service attacks and intentional corruption of data will be particularly critical. Conventional web traffic involves HTML-based exchanges (typically HTML pages and tables) between client and server. Web services traffic involves the use of application program interfaces (APIs) for exchanging data using a variety of standards (e.g., HTTP, HTTPS, SMTP). This architecture complicates the security picture because each web service application interface may invoke multiple operations that are potentially susceptible to new and difficult-to-detect security breaches.

A common ground must be found in developing solutions to this challenge that can satisfy both the federal and non-federal partners, while not impeding the present rapid pace of IOOS development and implementation. The DMAC-SC recommends that a community-based working group on IOOS IT Security be established to develop an IOOS security policy, and to provide more specific guidelines for IOOS participants on implementation.

Section 3. Governance, Oversight, and Coordination

GOVERNANCE

The governance of the DMAC Subsystem is being designed to operate within the context of the IOOS governance mechanisms described in the IOOS Development Plan.

At the First Annual IOOS Implementation Conference (August/September 2004), the Ocean.US Executive Committee (EXCOM) agencies and the emerging RAs endorsed the following strategy:

- **DMAC Steering Team**: Ocean.US will establish an IOOS DMAC Steering Team drawn from government, industry, academia, public, and non-profit communities to: (1) coordinate and oversee the evolution of DMAC standards; (2) identify and provide recommendations regarding gaps in needed standards; and (3) help ensure that the DMAC standards process is conducted in an open, objective, and balanced manner.

- **DMAC Expert Teams and Working Groups**: Ocean.US will organize expert teams and working groups to address key IT standards areas as identified in the DMAC Plan. Experts from the emerging GEOSS and relevant international data management standards activities will be invited to participate.

- **Interagency Coordination**: The EXCOM agencies have agreed to establish a government-only IOOS DMAC Implementation Oversight Working Group (IOWG). Consistent with the governance guidelines outlined in Part I of the First Annual IOOS Development Plan, the IOWG will coordinate DMAC implementation within the federal agencies. Specifically, the IOWG will provide oversight of federal implementation of relevant IOOS DMAC standards and best practices recommended by the Steering and Expert Teams; recommend to the agencies actions relating to inter-agency adoption and/or development of common standards, protocols, and shared communications software; and serve as an information resource in DMAC planning efforts.

To aid in the implementation of these processes, it is recommended that the National Federation of Regional Associations (NFRA) establish a DMAC subcommittee to oversee and facilitate coordination, communications, and data and technology exchange at the regional level. This NFRA subcommittee will also serve as a major contact point facilitating national and regional DMAC coordination, similar to the relationship between the NFRA and IOOS.

The above strategy ensures that the development and implementation of the DMAC Subsystem is coordinated closely with, and leverages upon, related activities in the federal agencies and other national, regional, and international earth observing systems (e.g., GEOSS, Joint Technical Commission for Oceanography and Marine Meteorology [JCOMM], and Ocean Research Interactive Observatory Networks [ORION]). The DMAC Subsystem will be planned, developed, maintained, and enhanced in a systematic, coordinated, cost-effective, interoperable manner, with support from professional systems engineering services. IOOS stakeholders will be urged to participate in the DMAC planning and assessment activities to ensure that current and future community needs and priorities are addressed.

IOOS DATA POLICY

IOOS data policy is under development, and will be put into effect at an early stage of IOOS implementation. It will be consistent with:

- U.S. federal data policies;
- International GOOS Design Principle 7, "GOOS contributors are responsible for full, open, and timely sharing and exchange of GOOS-relevant data and products for non-commercial activities," (IOC, 1998);
- 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 updates adopted at the 22nd session of the IOC Assembly in July 2003 - Resolution XXII-6 & 7;
- The WMO policy of free exchange of meteorological and related marine data (WMO Resolution 40, Publication WMO—No. 837);
- Production and sale of value-added products by the private sector is to be encouraged, and data providers and users from both private and public sectors should be able to contribute to and use IOOS data and information. This policy is consistent with policies in the Paperwork Reduction Act of 1995 (44 U.S.C. §§ 3501 et seq.) and Office of Management and Budget (OMB) Circular No. A-130 (First Annual IOOS Development Plan).

All data collected and prepared under IOOS funding shall be subject to the IOOS data policy. Generally, the policy will call for full and open sharing of non-proprietary data and metadata, products, model code, and related information. It will also call for adherence to data, metadata, and data products standards promulgated by IOOS. Specific requirements for each of the DMAC Subsystem elements are discussed in the Plan sections on Metadata, Data Discovery, Data Transport, Uniform On-Line Browse, and Data Archive and Access. Archive facilities that participate as official IOOS Archive Centers will further agree to adhere to data archiving guidelines that will be established in the Archive phased implementation of DMAC.

IOOS will not compete with the private sector because it will not distribute commercial data, products, or services produced by commercial enterprises.

IOOS/DMAC STANDARDS PROCESS

The marine sciences community has made significant progress toward data integration during the past two decades. Community-wide programs such as GLOBEC, OBIS, LOICZ, WOCE, JGOFS, and, recently, Argo have done much to establish new traditions in marine data management. These new traditions recognize the importance of data standards and the value of shared data access within individual disciplines. The community is now facing a new generation of standards-related problems: to be effective, operational oceanography will require integration of data and product streams from many distinct disciplines. Marine data standards that have been narrowly focused, though documented and well supported, are often not sufficiently interoperable to address this requirement. With the development of GEOSS and GOOS, similar standards issues are also being addressed by JCOMM, the IODE, and the OITP.

Defining data standards is a slow and expensive process. Typically a group of technical experts must meet repeatedly over a period of years to develop and agree upon a data standard of modest scope. Thus the DMAC development cannot wait upon a systematic redesign of marine data standards, alone, to achieve the required level of interoperability. Rather, the focus of this Plan is on the use of protocols and translators that can achieve an acceptable level of interoperability building upon standards that exist today. This approach is discussed in greater detail in Part II and in the Data Transport Appendix.

Adopting this approach represents a compromise. The level of interoperability that can be achieved among differing standards is often limited by mismatches in the information content of the standards, or differences in the semantic data models that underlie them. In the long term, achieving the desired level of data interoperability will require that the community develop and use fewer standards that are of greater breadth.

In parallel with building the interoperable Data Communications Infrastructure, this phased DMAC Implementation Plan recommends that DMAC begin work to foster an improved standards process. The DMAC standards process must be open so that it represents community consensus. It must be highly visible so that the standards are broadly used, and it must carry official stature so that the standards will be respected and used appropriately. It must also build on existing standards and standards processes whenever possible. To be fully successful, IOOS must foster the adoption of community standards that encompass quality control, scientific analysis, data-set versioning, metadata, products and services, data discovery, network data transport, file formats, and data archiving.

USER OUTREACH

(see Appendix 4 for complete User Outreach Team Report)

The recognition and incorporation of users' needs is essential to the success of IOOS and all other components of GOOS. This effort extends well beyond the boundaries of the DMAC Subsystem. Indeed, from its inception IOOS is envisioned as an "end-to-end" system tailored to address the needs of the end user. True end users are generally not information technology specialists, but professionals who rely on information that has been developed from data by other professionals. Examples of end users include the commercial fishery manager, the oil spill response team leader, the U.S. Coast Guard watch officer, and the harbor master.

User Outreach refers to the structures within IOOS governance that will recognize and address the needs of end users on an on-going basis. The goals of User Outreach are to: (1) identify the needs of users, (2) influence IOOS to meet those needs, (3) assess how well those needs are being met, and (4) report the results of 1 and 3 to all parts of IOOS. User Outreach also seeks to establish and enhance the societal relevance of IOOS through informing the concerned public about existing products and services.

There are at least three classes of users: (1) the end users mentioned above, (2) users who process (e.g., assimilate and analyze) data and provide information in the form of forecasts, ocean state estimations and hindcasts, and (3) users who provide specialized software and other services. Geographic and institutional structures add further complexity to identifying the needs of the various user groups. The needs of users are generally structured according to the seven phenomena of interest to IOOS: marine operations, natural hazards, national defense, public health, climate change, healthy ecosystems, and sustainable marine resources. For an in-depth description of user needs from the viewpoint of a national panel of experts, please see Appendix 4 (the complete User Outreach Team Report).

User needs are translated into software requirements and then into software product design through the spiral design cycle described below in the System Engineering Approach. On-going input from users will be solicited at each stage of the design cycle. This input will shape the standards-generating and system implementation tasks that are outlined in Part II of the Plan.

SYSTEM ENGINEERING APPROACH

(see Appendices 5 and 6 for more detailed discussions)

This document describes a wide variety of requirements addressing the needs of a diverse group of stakeholders. Because of the resulting complexity, the success of the DMAC Subsystem requires a formalized system engineering process. A brief description of three system engineering process models is presented in Part III (Appendix 5), as well as recommendations for the approach that should be used for the DMAC development and integration.

Based on a review of the subsystem requirements and a comparison of the features of several common models, it is recommended that the Spiral Model for systems engineering be selected for DMAC implementation. The Spiral Model accommodates a "task-oriented," highly structured approach, while allowing rapid prototyping and risk-analysis to be performed at juncture points of the project. In the Spiral Model, selected requirements are chosen for development to an operational level. Then, more requirements are added, and the development process is repeated through this "spiral" until all requirements are accomplished. The phases can be executed using a waterfall-like process (i.e., with requirements specification [or updates], analysis and design, system development, and verification performed for each phase). Each phase (sometimes referred to as an effectivity), would then represent a complete end-to-end execution of a subset of the requirements. A phased approach adapted to fit the DMAC purpose is presented in Appendix 5.

To ensure that the DMAC Subsystem meets the program goals, it is critical that the technology stay current and operational through adherence to a concrete plan for maintenance and refreshment. Appendix 6 discusses the life cycle maintenance and refreshment of the technology components of the DMAC.

Much of this document describes systems capabilities that are reliant on technology, whether hardware or software. A Price Systems LLC[20] paper defines technology refreshment as "the periodic replacement of commercial off-the-shelf (COTS) components; e.g., processors, displays, computer operating systems, commercially available software (CAS) within larger ... systems to assure continued supportability of that system through an indefinite service life." We would add communications capabilities and storage media to this list. Systems are being acquired that are ever larger and more complex, constructed out of components designed and built by commercial third parties. If the DMAC is viewed as a "system of systems," this situation still applies; whether "component"

[20]Technology Refreshment - A Management/Acquisition Perspective" (2001), available at http://www.pricesystems.com/downloads/pdf/technology%20refresh.pdf

in this context means a server or an application software suite, it still represents an item that must be evaluated and, if appropriate, refreshed periodically in order for the overall system to meet its evolving mission requirements.

For the DMAC Subsystem to remain current, the technology might be refreshed during the system life for a number of reasons, including the following:

- The existing system component has malfunctioned and either cannot be repaired, or the repair costs exceed the replacement costs,
- The existing system component has reached its life expectancy,
- The surrounding technical infrastructure has evolved and is incompatible with the existing component under consideration,
- Evolving requirements have made an alternative technology more cost-effective,
- Newer technology has come to market that provides more capability for the same or lower total cost of ownership.

The DMAC Technology Refreshment Plan (TRP) (discussed in Part II) will to the greatest extent practical draw upon established protocols such as the Navy's Technology Assessment and Management Methodology (TeAM)[21] or the Technology Refreshment Cost Estimating and Planning Model[22].

DMAC COST MODEL

The provision of marine data over the Internet is experiencing a period of rapid growth in both the volume and breadth of services offered. The aggregate cost of these activities is likely to increase over time irrespective of the existence of IOOS. The cost estimates for the DMAC Subsystem that are presented in Tables 4 and 5 reflect only those expenses that will be incurred over and above this background of growth. Expressed another way, the cost estimates in the DMAC Plan represent only those expenses that explicitly address the tasks of data integration achieved through the development of and participation in the standards, protocols, tools, and policies of the DMAC Subsystem.

Table 5 shows the cost model for the DMAC Subsystem for a ten-year period. The Plan calls for the initiation of the full DMAC Subsystem over a five-year period at a cost of $82 M in new funding. The initiation costs include the development of core standards, protocols, and tools ($28 M); costs

[21] Technology Assessment and Management Methodology – An Approach to Legacy System Sustainment Dynamics" (1998), available at http://smaplab.ri.uah.edu/dmsms98/papers/samuelson.pdf

[22] Technology Refreshment Cost Estimating and Planning Model: User s Guide (2001), available at http://www.its.berkeley.edu/nextor/pubs/RR-00-5.pdf

Table 4. DMAC Overall Program (thousands of $)

	Y-2	Y-1	Y1	Y2	Y3	Y4	Y5
Pilot Projects							
Data Discovery			$1,000	$1,022	$783	$534	$0
Access/Infrastructure			$500	$511	$522	$0	$0
Data Transport			$1,000	$511	$522	$534	$545
Archive			$500	$409	$418	$801	$1,091
Information Assurance			$500	$204	$104	$534	$545
Innovative Architectures			$300	$307	$522	$1,601	$818
Total Pilots			**$3,800**	**$2,964**	**$2,871**	**$4,004**	**$2,999**
Program Initiation Labor							
Program Management Activities	$36	$72	$726	$742	$752	$769	$792
Metadata and Data Discovery	$335	$271	$2,480	$2,412	$1036	$463	$975
Data Archive and Access	$235	$335	$1,612	$3,076	$1,799	$1,497	$1,139
Data Transport	$450	$348	$2,234	$2,980	$699	$740	$693
Total	**$1,056**	**$1,026**	**$7,052**	**$9,210**	**$4,286**	**$3,469**	**$3,599**
Program Initial Fixed/Maintenance Costs (Inflation-adjusted costs shown)							
Communication/Infrastructure			$1,460	$1,594	$1,734	$747	$764
Servers at Centers			$2,400	$2,821	$3,259	$1,153	$1,178
Archive Capacity						$4,163	$2,836
Engineering/Integration			$3,000	$3,475	$3,342	$1,921	$1,746
Total Initial Fixed/Maintenance			**$6,860**	**$7,890**	**$8,335**	**$7,984**	**$6,524**

Table 5. Total Program Costs (thousands of $)

Program Initiation Costs							Outyear Recurring Costs				
Preparatory		Core Program									
Y-2	Y-1	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
$1,056	$1,026	$17,712	$20,064	$15,492	$15,457	$13,122	$17,645	$18,033	$18,430	$15,265	$15,600

Grand Total, Preparatory	**$2,082**
Grand Total, Initiation	**$81,847**
Grand Total, 10 Years	**$166,819**

of hardware, software, networking capacity, data archiving center expansion and systems integration labor ($37 M); and a budget for focused pilot projects to usher in and test the new technologies ($17 M). Out-year recurring costs over the following five years (Years 6 through 10) total an additional $85 M. Substantial new funding for IOOS is not anticipated until fiscal year 2007 (FY 07), yet a minimally functioning DMAC Subsystem must already be in place to support the initial growth in IOOS measurements, modeling, and usage at that time. The DMAC Plan includes tasks and associated costs totaling $2.1 M during FY 05 and 06 that are deemed to be very high priorities for immediate implementation in order to prepare for FY 07 demands on the Subsystem.

The assumptions driving the cost model, and additional details about the model components are presented below:

- **General Assumptions**
  - The plan assumes that the program cost reflects new efforts above and beyond existing program elements. It further assumes the existence of data collection and processing components. Therefore, program costs reflect new services, hardware, software, and infrastructure costs, and do not include costs that are already budgeted for ocean observing systems.
  - The plan assumes that funds will be appropriated to Executive Agencies as a part of the normal budget process, and that the funds will be targeted for DMAC. The exact mechanism for appropriation and allocation to DMAC is outside the scope of this plan (for details, refer to the IOOS Development Plan).
  - It is assumed that agencies will apply DMAC funds to the execution of the activities in this Plan based upon recommendations from the DMAC planning and implementation activities described earlier in the Governance section.
  - The cost model is designed to support the Data Management and Communications Subsystem of the IOOS only; costs for sensors, data ingest, modeling, and end user capabilities are not included. Specifically, the DMAC does not include the costs of the observing subsystem nor the modeling and analysis subsystem, although those components are part of the IOOS.
  - The cost model is transparent with respect to the origin of the observation type; for example, remotely sensed data would be within scope, but only from the point at which the data enter the IOOS, not including data providers' sites.
  - The overall costs of the plan were not derived by comparison to like systems; indeed, it is not believed that a comparable system of the magnitude of IOOS exists today. Rather, the cost estimates were derived by a combination of level of effort analysis and estimates of the magnitude of system capacity required. It is fully anticipated that the planned Systems Engineering Tasks will provide more refined estimates of acquisition and integration costs.

- **Pilot Projects** - The goals of the Pilot Projects are to provide proof-of-concept demonstrations and advanced technology development and integration. Given these goals, it should be noted that the cost estimation for Pilot Projects is much less precise than that for the labor costs of the three subsystems. Pilots are intended to move components from the developmental to the operational stage.
  - Pilot Projects include both design and demonstration, depending on the level of maturity of the underlying technology. The program assumes an approximate division of 60 percent for design and 40 percent for demonstration.

- **Labor by Subsystem** – This category reflects the labor costs to develop system capabilities and reach a Full Operating Capability by Year 5.
  - "Program Integration Labor" consists of the total cost of labor for the development and integration activities of the three segments: Metadata and Data Discovery (including Uniform On-line Browse), Data Archive and Access, and Data Transport. Labor costs are based on a standard contractor rate of $180,000 per annum, except for staffing of the Archive and Access, at $100,000 per annum. The exact distribution of the labor costs can be determined from the Phased Implementation Plan, Part II of this document. The DMAC Subsystem has been conceived to minimally impact the freedom of individual data providers to make independent data management choices. Yet providers of data—for example, federal, regional, and commercial—will nonetheless incur significant expenses creating metadata that conforms to DMAC (FGDC) standards and configuring the Data Transport software components that will make their data accessible to others. These costs are included under "Program Integration Labor."

- **Program Initial Fixed/Maintenance Costs** – This category refers primarily to non-labor costs of initiating the program in the first five years. These costs include the following:
  - Communications infrastructure including communications hardware at 30 sites (or equivalent). These sites contribute to essential DMAC infrastructure (e.g., archive centers and primary data assembly centers). The majority of sites that provide data to IOOS will not fall into this category.
  - Communications lease for the entire infrastructure.
  - Servers at a total of 30 sites, including hardware and software, and hardware maintenance after the year of installation. It is understood that participation in DMAC may in some cases require additional dedicated hardware to address issues of security and elevated performance demands. Notably the DMAC Subsystem will require one or more primary data assembly centers as fundamental infrastructure.
  - Server hardware and software, including media at five archive center sites.

- Professional services of a Systems Integrator to (1) coordinate and manage the total hardware, software, and infrastructure definition, design, procurement, installation, integration, and maintenance and (2) oversee **Capacity Building**, the effort in providing labor and services to data providers to enable them to reach and maintain the level at which they can participate. This effort might include technical training, system administration or help with preparing metadata, for example.

- **Outyear Recurring Costs** – These costs reflect a hardware recapitalization beginning in Year 6. "Maintenance of Custom SW (software)" reflects software maintenance and upgrades (i.e., custom programming) for the totality of deliverables at Initial Operational Capability. Systems Engineering/Integration will be provided for the life cycle maintenance of the hardware/software systems, including integration of the custom software. Archive operations consist of two staff per center per year at a cost of $100,000 per staff member. Outyear Recurring Costs also include
  - **Capacity Building** – (as defined under Program Initial Fixed/Maintenance Costs),
  - **Outreach** – referring to the effort to acquire new data providers, to educate the pool of possible data providers on the IOOS process and requirements, and to perform similar services as **Capacity Building**, on a limited basis.

Inflation costs are calculated based on the Real Discount Rates, published by the Office of Management and Budget. Published rates for five, seven, and nine years are 1.9 percent, 2.2 percent, and 2.5 percent, respectively. For this model, a rate of 2.2 percent was used.

Section 4. Immediate Priorities for Implementation

Implementation of IOOS has already begun. Once the IOOS Plan has been approved, participation will accelerate. Developments will be initiated at local, regional, and national levels. To support these activities, DMAC must quickly achieve a useful minimum level of functionality. The DMAC outcomes listed below are required to reach that point. They include (1) community-building and planning, (2) standards-generating, (3) the development of critical (currently missing) technology components, and (4) initiating a technical oversight function by a professional systems engineering service. The activities that accompany these outcomes are described in detail in Part II, Section 2 of this phased Implementation Plan. The activities will be overseen by the DMAC planning and implementation committees, which will also have initial responsibilities for coordination of data management and communications among national backbone components, regional observing systems, and international collaborators.

METADATA/DATA DISCOVERY/DATA LOCATION

Outcome 1: The development of (i) initial metadata standards that will guide IOOS/DMAC data providers in the creation of metadata and (ii) agreement upon the initial organizations that will participate in metadata management for IOOS:

- Task 1: Convene a community-based working group of metadata and archive experts to agree upon and document: (i) interim metadata standards, and (ii) an initial list of IOOS member organizations that will provide metadata catalog management services.

Outcome 2: The development of initial data discovery services needed for IOOS data users to identify data sets of interest.

- Task 1: Convene a community-based working group of data discovery experts to agree upon the Data Discovery architecture.

- Task 2: Implement a testbed based upon existing data discovery services leading to the development of an interim distributed metadata search capability.

Outcome 3: Agreement upon a technical solution to create bi-directional linkages between metadata-based DMAC Data Discovery services and (i) DMAC Data Transport and (ii) Uniform On-Line Browse (and other information products).

- Task 1: Conduct an applied R&D activity that will explore existing technologies capable of linking metadata-based searches to points of data and product access. This effort must address the dynamic (changeable) nature of the access points and the problems of multiple (replicated) copies of data sets that are available at different points.

DATA TRANSPORT

Outcome 4: The development of an initial semantic data model (for a restricted class of marine data) that (i) is capable of demonstrating machine-to-machine interoperability with semantic meaning and (ii) will form the foundation for further semantic data modeling leading (in the mature DMAC) to a comprehensive marine data model. This work will be conducted in close coordination with the Metadata Working Group activities.

- Task 1: Convene a community-based working group of data management experts from a broad range of marine disciplines to (i) develop and document a (restricted) interim semantic data model and (ii) initiate development of a comprehensive semantic data model.

Outcome 5: The existence of three critical infrastructure components that will be needed to support Data Transport for vital classes of data: biological, GIS, and "generic" (non-standards conformant).

- Task 1: Conduct a software development activity to develop a linkage from biological data accessed through the OBIS system to the DMAC Data Transport component.

- Task 2: Conduct a software development activity to develop a "generic" server that would access such data as ASCII tab-delimited files.

- Task 3: Conduct a software development activity to develop a linkage from data accessible through the DMAC Data Transport component into common GIS applications.

DATA ARCHIVE

Outcome 6: The existence of community-wide agreements that will establish a framework for cooperation among IOOS archive centers.

- Task 1: Convene a community-based working group of archive center representatives to agree upon an initial list of IOOS partner organizations that will provide permanent archive services.

- Task 2: Archive working group to agree upon an initial framework to inventory and assess the state of marine data archiving in order to ensure that all irreplaceable marine data are associated with a responsible archive center.

Outcome 7: A demonstrated capability of IOOS Archive Centers to provide Data Discovery and Data Transport services using DMAC standards and protocols.

- Task 1: Continue development of existing pilot projects at NODC that use DMAC standards and protocols for Data Transport and Metadata to deliver near-real-time and real-time data sets (Global Temperature-Salinity Profile Program and Shipboard Environmental Acquisition System).

- Task 2: Modernize access to data currently received in real-time at Archive Centers by deploying pilots that rapidly deliver data to users employing DMAC standards and protocols.

SYSTEM ENGINEERING APPROACH

Outcome 8: The development of well-organized documentation of the DMAC Subsystem and its initial participants needed by IOOS participants and planners.

- Task 1: Engage the services of a software engineer to provide these services to IOOS.

CONCRETE GUIDANCE TO DATA PROVIDERS

(Technical guidance for data managers)

The DMAC Plan is not intended as a guide to marine data management; rather it is a plan for interoperability among independent data management systems. Furthermore, at the time of this writing (March 2005) the DMAC Subsystem remains very incomplete—important aspects of interoperability remain to be addressed through community processes described in Part II of the DMAC Plan. Yet it is possible to recommend some concrete actions that may be taken by data and metadata managers and product producers to coordinate early implementations and streamline their future compatibility with IOOS. Integration of data into IOOS implies that would-be users will more quickly and efficiently be able to (1) discover data through a comprehensive search, (2) browse and visualize data through standard web browsers, and (3) access data from many common computer applications. It is understood that any recommendations made at this time are subject to change as the DMAC Subsystem evolves.

There are two classes of solutions for sharing data that will ensure consistency with the emerging DMAC standards, protocols, and tools:

1. Providers of certain types of data may delegate responsibility for managing these data to another entity. For example, a data provider may be able to enter into an arrangement with the NOAA National Data Buoy Center (NDBC[23]) to perform quality-control and distribute mooring data or with the U.S. GODAE Server[24] to distribute operational model outputs.

2. Providers of all types of data can make choices to manage the data in a manner that is consistent with the emerging DMAC standards, protocols, and tools. The following approaches to managing data and metadata will help ensure compatibility with emerging DMAC:

   - **Metadata management and data discovery**
     It is recommended that all data providers:
     - Create metadata that are compliant with Federal Geographic Data Committee (FGDC[25]) standards for both current and legacy data holdings and inventories.
     - Submit metadata to the NASA Global Change Master Directory (GCMD[26]) and/or the NOAA Coastal Data Development Center (NCDDC[27]) so that data sets may be easily found through an open data discovery process.
     - Participate in the DMAC Metadata Working Group[28] (see Part II, Section 2, Metadata/Data Discovery Activity 1) to ensure that the special characteristics of their data will be thoroughly considered during the formulation of DMAC metadata standards.

   - **Data Transport**
     Depending upon the nature of the data to be provided, it is recommended that providers of:
     - *Gridded data* -- install servers providing access to their data through OPeNDAP[29] data access protocol.

[23] contact MarineObs@noaa.gov for details.

[24] contact the U.S. Global Ocean Data Assimilation Experiment (GODAE) Server at http://www.usgodae.fnmoc.navy.mil

[25] Information on FGDC is available at http://www.fgdc.gov/metadata/metadata.html. Contact either the NOAA Coastal Data Development Center (http://www.ncddc.noaa.gov/Metadata) or the NOAA Coastal Services Center (http://www.csc.noaa.gov) for direct assistance with creating FGDC metadata

[26] see http://gcmd.gsfc.nasa.gov

[27] see http://www.ncddc.noaa.gov/Metadata

[28] contact DMAC@ocean.us.net

[29] http://www.unidata.ucar.edu/packages/dods

- *Complex data collections in a relational data base (SQL)* -- make data accessible to DMAC by participating in data transport pilot activities to either (i) use OPeNDAP relational data base server or (ii) use enterprise GIS protocols. Full operational support for complex data collections in relational databases will be developed early in the evolution of DMAC.
- *Large collections of individual files that comprise a single (logical) data set* -- if OPeNDAP servers exist for the file types install these servers to provide access to the individual files. Participate in pilot activities to develop "aggregation" capabilities that will provide a higher-level (more ordered) view of the collections.

It is recommended that all data providers:

- Participate in the DMAC Transport (Semantic Data Model) Working Group[30] (see Part II, Section 2, Data Transport, Activity 1) to ensure that the special characteristics of their data (if any) will be thoroughly considered during the formulation of DMAC data transport standards.

Uniform On-Line Browse to all IOOS data

It is recommended that all data providers:

- install metadata-enabled, open source or commercial On-Line Browse tools for end users. For gridded data, the LAS[31] is a recommended pre-operational Uniform On-Line Browse component of IOOS. For complex data collections (non-gridded), participate in pilot activities to utilize commercial or open source GIS web clients or the LAS.

Archive

It is recommended that all data providers:

- Review their current data holdings to ensure that irreplaceable data are archived at a responsible entity.
- Contact the archive entity that is responsible for their classes of data and make arrangements for archiving the data.

[30] http://www.ocean.us/documents/docs/part_1_for_web_comment_022403.pdf

[31] http://www.ferret.noaa.gov/LAS/

I. Interoperable Data Discovery, Access, and Archive

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

Part II: Phased Implementation Plan for DMAC

March 2005

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

Contents

Section 1. Functional Requirements

This document states high-level functional requirements for the Data Management and Communications (DMAC) Subsystem of the Integrated Ocean Observing System (IOOS). The intended use of this document is to supplement the white papers that describe the DMAC subsystems (Part III of this document) and to rephrase the notions into the formal terminology of a requirements specification. The use of the term "requirements" is consistent with that used by the International Council on Systems Engineering.[1] As described in the Handbook, this section also is intended "to establish a database of baseline system requirements derived from the source, to serve as a foundation for later refinement and/or revision by subsequent functions in the Systems Engineering process and for a non-ambiguous and traceable flow down of source requirements to the system segments."[2] This section includes requirements of the following types:

- Program requirements
- Mission requirements
- Customer specified constraints
- Functional requirements
- System requirements
- Interface, environmental, and non-functional requirements
- Unclear issues discovered in the requirements analysis process

This section is written at a level that primarily addresses mission and customer requirements, along with high-level functional and system requirements. It is intended that this section serve as an asset to be used in subsequent engineering efforts to articulate the detailed system and interface requirements that are required to develop a DMAC.

GENERAL REQUIREMENTS

1. IOOS DMAC Vision

1.1. The DMAC Subsystem of the IOOS will knit together the distributed components of IOOS into a nationwide whole that functions as a unified component within the international GOOS framework. The vision for the DMAC Subsystem is not limited to the ingesting and archiving of observations; it includes the data and communications components needed to move data among systems and users in a distributed environment. The DMAC Subsystem will be required to link observations collected from a broad range of platforms: buoys, drifters, autonomous vehicles, ships, aircraft, satellites, and cabled instruments on the sea floor. Observations

[1] Systems Engineering Handbook, International Council on Systems Engineering, 2000.

[2] Ibid.

may be point measurements, continuous measurements, or imagery and variables may be biological, geological, chemical, physical, or abstract. The many millions of individual measurements anticipated to be obtained daily by the sensor networks will be transmitted (in real-time, near-real-time, and delayed modes) directly to end users, as well as to the applications and data-assimilating models that process these measurements into maps, plots, forecasts, and other useful forms of information.

1.2. While the DMAC vision recognizes that data products, rather than raw data, are typically required by users, the development of most data products will be the responsibility of the Applications, Modeling, and Product Services Subsystem of the IOOS. The requirements of the DMAC with respect to product generation are as follows:

1.2.1. to ensure that the needs of product generators are met for timely delivery of quality-controlled data;

1.2.2. to provide accurate and thorough metadata accompanying the data;

1.2.3. to provide a uniform guaranteed minimum level of geo- and time- referenced graphical browse capability for all classes of data.

1.3. The guarantee of assured data discovery and minimal browsing capability depend upon descriptive metadata, ensuring that the data are readily intelligible to users.

2. DMAC Overall Functional Requirements

2.1. Data transport. The DMAC shall provide capability for the collection/transmission of data from sensor subsystems at entry points where the data become available using DMAC standards and protocols either on the Internet or a supplied IOOS backbone, to assembly centers, users, and archive centers in real time and delayed mode, for operational, research, and product generation applications.

2.2. Quality control. The DMAC shall provide a mechanism for assuring that data are of known, documented quality. QC operations are a partnership among data observation/collection components, processors, analysts, other users, and the DMAC.

2.3. Data assembly. The DMAC shall provide mechanisms for aggregation and buffering of data streams over useful spans of time and space. Data assembly allows users to more easily exploit real-time data, especially data from distributed sensor arrays.

2.4. Product generation. Products include data products such as assimilation-friendly, real-time measurements, model nowcasts and forecasts, GIS layers and climatological reference fields; graphical information products such as scientific plots and maps; and text information products such as written forecasts and numerical tables. The DMAC will provide a minimal level product-generation capability, only—the guarantee of a uniform, interactive, geo- and time-referenced browse capability suitable for quick evaluation of data by IOOS scientists. Most product generation is the responsibility of the IOOS Modeling, Data Assimilation Subsystem and the value-added information product producers that will address the needs of specialized end-user groups.

2.5. Metadata management. The DMAC shall provide simple, clear guidelines and extensible standards for metadata; ensure that the linkages between data and metadata are maintained with great reliability; provide for communication of metadata among components of the system; provide training and tools to increase end users' and data providers' capacity in metadata generation and management.

2.6. Data archeology. The DMAC shall directly or indirectly facilitate activities to rescue, digitize, and provide access to legacy/historical data sets; retrieve data in danger of loss due to deteriorating media, out-of-date software, not in digital format, etc.

2.7. Data archival. The DMAC shall provide for the long-term archive and stewardship of IOOS data sets; conform to national archive standards, as well as IOOS standards and user requirements.

2.8. Data discovery. The DMAC shall provide a means for determining what data are available within the IOOS based upon queries that may be issued by users or by other machines. Data Discovery shall be seamlessly integrated with data and metadata access functions provided by the Data Transport and Metadata Management components, respectively.

2.9. Administrative functions. The DMAC shall provide oversight mechanisms to ensure the proper functioning and smooth evolution of IOOS. These include fault detection and correction, security, monitoring and evaluation of system performance, providing for system extensibility, establishing and publicizing policies for data availability, soliciting and responding to user feedback, and establishing and maintaining international linkages.

3. Participating Activities

3.1. The DMAC shall provide transport, access, and archival capabilities for TBD Regional Data Centers.

3.2. The DMAC shall provide transport, access, minimum browse, and archival capabilities for TBD Data Assembly Centers.

3.3. The DMAC shall provide transport, access, minimum browse, and archival capabilities for TBD Modeling Centers

3.4. The DMAC shall provide transport and access capabilities for TBD Archive Centers.

3.5. The DMAC shall provide transport, access, and minimum browse capabilities for TBD value-added product generators.

4. Infrastructure/Communications

4.1. Infrastructure

4.1.1. The DMAC shall leverage existing or deploy dedicated IOOS data servers at TBD locations, including up to all of the following: Regional Data Centers, Data Assembly Centers, Modeling Centers, and Archive Centers.

4.1.2. The DMAC shall leverage existing or provide aggregate storage as follows:

4.1.2.1. Regional Data Centers

4.1.2.1.1. Online - TBD

4.1.2.1.2. Near-line (e.g., online tape silo) - TBD

4.1.2.1.3. Offline - TBD

4.1.2.2. Data Assembly Centers

4.1.2.2.1. Online - TBD

4.1.2.2.2. Near-line (e.g., online tape silo) - TBD

4.1.2.2.3. Offline - TBD

4.1.2.3. Modeling Centers

4.1.2.3.1. Online - TBD

4.1.2.3.2. Near-line (e.g., online tape silo) - TBD

4.1.2.3.3. Offline - TBD

4.1.2.4. Archive Centers

4.1.2.4.1. Online - TBD

4.1.2.4.2. Near-line (e.g., online tape silo) - TBD

4.1.2.4.3. Offline - TBD

4.2. Communications

4.2.1. The DMAC shall leverage existing communications capabilities or provide dedicated broadband networks between/among the Regional Data Centers, Data Assembly Centers, Modeling Centers, and Archive Centers.

4.2.2. Data Providers for the IOOS will use existing Internet capacity to push data holdings to the Regional and National Backbone Data Centers.

5. Technology Infusion

5.1. The DMAC shall develop a plan to address technology infusion. The plan shall include mechanisms for member-provided technology infusion, as well as that which is centrally funded and maintained. The emphasis will be on integration, compatibility, and interoperability among all parties participating in the IOOS.

5.2. The plan shall include evolving mass storage technology.

5.2.1. The plan shall include strategies for storage media migration.

5.2.1.1. Current archive systems are based on magnetic tape cartridges, which typically have a three to five year life cycle, and are approaching a petabyte in size. 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.

5.3. The plan shall consider new technologies in networks, computing systems, and evolutions in software.

5.4. The plan shall account for the following categories of technology changes.

5.4.1. Technology Upgrades – A change that incorporates the next generation product or product upgrade to an existing technology or component which improves overall system functionality.

5.4.2. Technology Refreshers – A change that incorporates a new product to avoid an ensuring end of life or product/COTS obsolescence, or to correct a problem identified via a customer.

5.4.3. Technology Insertion – A change that incorporates a new product or function capability which is a result of industry growth or advanced development.[3]

6. Other General System Requirements

6.1.1. The DMAC as a whole shall be extensible in terms of function, volume, capacity, and throughput.

6.1.2. The DMAC shall provide access to data in a manner that is (largely) transparent to the user.

6.1.3. The DMAC shall not adversely impact existing data access methods or systems of the data providers.

6.1.4. It is a goal that the DMAC will not require data repositories to reformat their holdings to tie into the system.

6.1.5. Interfaces to data repositories may reside at any location that has network connectivity with the application and the data repository.

6.1.6. The DMAC shall provide a backward-compatible, version-controlled software environment.

6.1.7. The DMAC shall provide for the generic treatment of data sources isolating the requesting client from specific representations, unique request semantics, and protocols.

6.1.8. The DMAC shall make data available in multiple forms including the data's native form.

[3] "Technology Refreshment - A Management/Acquisition Perspective," available at http://www.pricesystems.com/downloads/pdf/technology%20refresh.pdf

6.1.9. The DMAC shall offer a cross-language and cross-platform data access mechanism that is independent of the data repository.

6.1.10. The DMAC shall enable the abstraction of encoding and transmission mechanisms and allow transparent distributed access to data using multiple protocols.

6.1.11. The DMAC shall provide access to all types of data: physical, chemical, biological, and geological.

6.1.12. The DMAC shall support system synchronization to permit multiple users access to the same database simultaneously.

6.1.13. Performance. The DMAC shall be developed to conform to minimum performance requirements. The following TBD notional performance requirements apply:

6.1.13.1. Minimum storage at Regional Data Centers, Data Assembly Centers, Modeling Centers, Archive Centers

6.1.13.2. Minimum aggregate computing capacity (ops/s) at Regional Data Centers, Data Assembly Centers, Modeling Centers, Archive Centers

6.1.13.3. Minimum communications bandwidth among Regional Data Centers, Data Assembly Centers, Modeling Centers, Archive Centers.

6.1.13.4. Maximum latency from data request to return to requesting user for simple data requests.

6.1.13.5. Maximum latency from data request to return to multiple simultaneous requesting users for simple data requests.

6.1.13.6. Maximum latency from data request to return to requesting user for complex data requests, including data aggregation, subsetting.

6.1.13.7. Maximum latency from data request to return to multiple simultaneous requesting users for complex data requests, including data aggregation, subsetting.

6.1.13.8. Minimum data volume rate of sustained delivery of volumes of data to a single user.

6.1.13.9. Minimum data volume rate of sustained delivery of volumes of data to multiple users simultaneously.

DATA COMMUNICATIONS INFRASTRUCTURE AND ARCHIVAL

The DMAC Subsystem is envisioned to consist of a Data Communications Infrastructure (standards, protocols, and tools for Metadata, Data Discovery, Data Transport, and On-line Browse) and an Archival capability. Figure 1 shows the interfaces through which Data Discovery functionality is achieved within the DMAC Data Communications Infrastructure. Numbers in italics refer to the sections of the requirements that reflect the drawing portion.

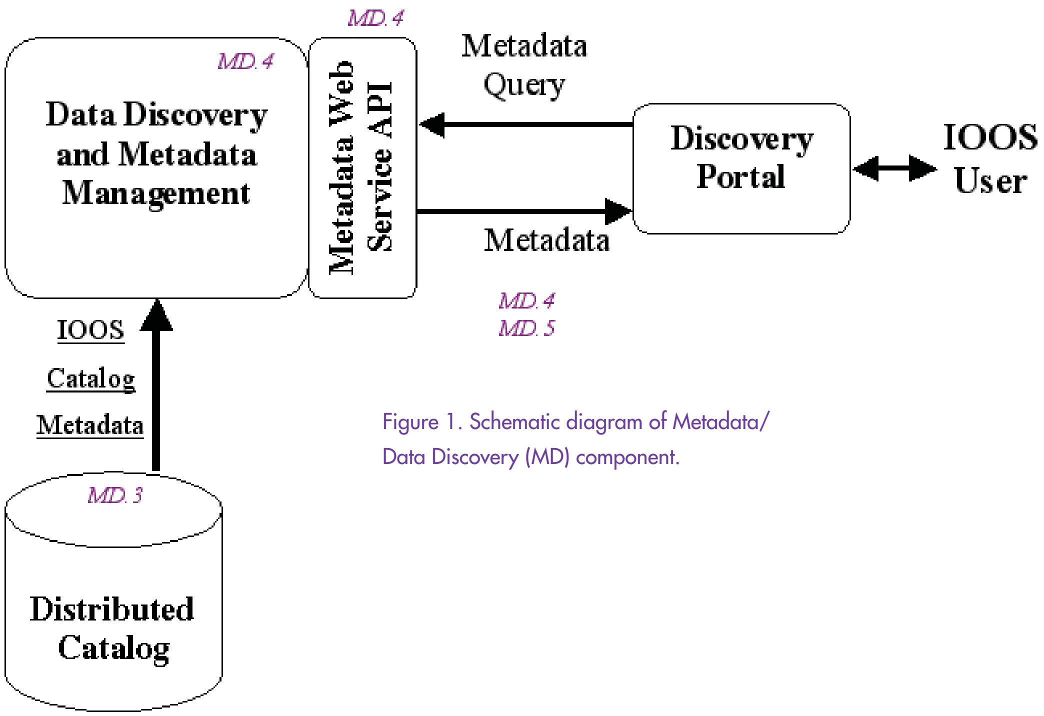


Figure 1. Schematic diagram of Metadata/ Data Discovery (MD) component.

1. Metadata/Data Discovery (MD) Requirements

(MD – Metadata; MMS – Metadata Management System; MC – Metadata Catalog)

1. Nature of Metadata

1.1. The IOOS MD shall be supplied using the guidelines established by the Federal Geographic Data Committee (FGDC) augmented by any applicable supplemental profiles.

1.2. The DMAC shall provide the capability to deliver metadata along with data delivery.

1.3. The MMS shall provide a mechanism to ensure that metadata found during data discovery are up to date, consistent, and understandable.

1.4. The MMS shall provide mechanisms for extensibility of the metadata.

1.5. The MD shall provide a framework for data versioning, data lineage tracking, and information citations.

1.6. The MD shall provide a framework for both semantic and syntactic metadata.

1.7. The MC shall provide a metadata query mechanism that supports access through a programming interface to any/all metadata fields.

1.8. The MMS shall support multiple standards that exist today and be able to extend beyond those to include expected future metadata standards.

1.8.1. Existing standards: FGDC; Biological Profile; Shoreline Profile, TBD

1.8.2. Possible future standards: TBD

2. Metadata Management System. The IOOS will include a master metadata management system.

2.1. The MMS shall be implemented as a distributed system that connects to all DMAC-compliant metadata holdings within the ocean community.

2.2. The MMS shall provide the capability for data providers to manage their metadata within a local system or through a centralized system via remote-access capabilities.

2.2.1. The MMS shall not require the data provider to maintain duplicate copies of metadata in two or more systems.

2.2.2. The MMS shall support a linkage between data discovery and data access that an application may utilize transparently to access both remote and local data via the DMAC Data Transport (DT).

2.3. The MMS shall include mechanisms to generate, validate and maintain metadata.

2.4. The MMS shall include a set of TBD controlled vocabularies for items such as keywords, entities and attributes, units, and other items to be determined.

2.5. The MMS shall provide support for parent/child metadata.

2.6. The MMS shall provide a mechanism for validation and approval of metadata.

2.7. The MMS shall include an automated metadata maintenance capability for checking URL links and any additional information within the metadata record that can be automated.

2.8. The MMS shall include mechanisms to facilitate the generation of metadata as close as possible to the collection and/or generation of the source data.

2.9. The MMS shall provide automated tools for versioning and configuration management of metadata.

2.10. The MMS shall provide a mechanism to access existing metadata servers to promote harvesting metadata.

3. Metadata Catalog. The MMS shall include a metadata catalog.

3.1. The implementation of the metadata catalog is TBD, but it is a requirement that the collective holdings of metadata shall comprise a distributed catalog. The implementation shall provide for integration of all such distributed sub-catalogs.

3.2. The catalog shall provide a capability to generate metadata records from self-describing data sources in which metadata and data have been integrated.

3.3. The catalog contents shall include items that will be used for discovery.

3.3.1. The catalog shall provide access control of metadata records, for maintenance and for viewing and searching on those records.

3.3.2. The catalog shall allow a catalog search from public search engines.

4. Search/Query Mechanism. The IOOS shall include a search/query mechanism.

4.1. The search interface shall search the MC for records that meet user-defined criteria.

4.2. End users and data providers can search for specific data sets.

4.3. End users and data providers can browse metadata about IOOS data holdings.

4.4. Automated agents can search for data.

4.5. The MC shall include a stable, documented-defined application programmers' interface (API) and a defined access protocol.

4.6. The search system shall support TBD types of actual data searches along with metadata searches.

4.7. The search system shall provide the following types of searches:

- 4.7.1. Full text and fielded searches
  - 4.7.1.1. Controlled vocabulary
  - 4.7.1.2. Free-text searches
    - 4.7.1.2.1. Single or multiple word searches
    - 4.7.1.2.2. Boolean operators on multiple words
    - 4.7.1.2.3. Thesauri to support text searches
- 4.7.2. Geospatial search
- 4.7.3. Temporal search
- 4.7.4. Thematic search
- 4.7.5. Parameter search
- 4.7.6. Taxonomic information
- 4.7.7. Browsing by thematic areas
- 4.7.8. Iterative/refinement searches

4.8. The system must be extensible to support other specific searches as required by the system, such as search by data quality or native format

4.9. Select - The Select functionality refers to those capabilities that allow an end user or data provider to examine data sets revealed from the data search and then choose sets of interest for downloading, on-line browse, or access via the DMAC Data Transport mechanism.

- 4.9.1. There shall be a user interface for allowing the user to select items for downloading. This will be referred to as the selection interface.
- 4.9.2. The selection interface shall display and accept selection requests for all data sets from the catalog software that meet the search criteria specified by the user in the search interface.
- 4.9.3. For each data set returned from a search of the catalog, the selection interface shall display the data set title and relevant metadata including spatial and temporal coverages and a method for viewing the metadata from that data set.
- 4.9.4. The selection interface shall provide a graphical means of viewing a thumbnail of each data set received from the catalog search.

4.9.5. The selection interface shall allow the user the ability to select from the items returned from the search and/or perform subsequent subsetting searches of the returned items.

4.10. Data set metadata shall be obtainable in multiple formats including both machine-readable XML and human-readable text.

5. Portal

5.1. Access to data and metadata shall be provided through the Internet via a portal. A portal is an Internet presence (e.g., web site) that redirects the user (possibly transparently) to a larger set of access points.

5.2. It is a system goal that graphical user interfaces (GUIs) shall be simple to use for a broad spectrum of users.

5.3. The IOOS portal will consist of an entry point (a Web "home page"), hierarchically lower level entries (other pages), and links to areas or functions within the IOOS.

5.4. There shall be simplified versions of the portal suitable for incorporation into non-IOOS Web sites for purposes of offering the capability to search IOOS data.

5.5. The portal shall provide mechanisms for accessing web-services enabled functions of the IOOS.

5.6. The portal shall conform to all Federal guidelines on Internet presence.

5.7. The portal shall be designed to be Section 508 compliant (see http://www.section508.gov/).

5.8. The portal shall provide all necessary policy statements and legal disclaimers.

5.9. The IOOS shall support web browsers Netscape and Microsoft Internet Explorer and others TBD.

5.10. The IOOS shall provide tools for remote content management of the portal structure.

5.11. The portal shall provide links to relevant information such as tools available for generation of the metadata required for this specific system.

5.12. The portal shall provide information on requirements for IOOS data providers.

5.13. The portal shall provide links to the supporting organizations.

5.14. The portal shall be easily modified to a new look and feel.

5.15. The portal shall provide FAQs.

5.16. The portal shall provide on-line documentation.

5.17. The IOOS shall provide a mechanism to solicit and receive user feedback concerning the operation of the system, data quality, portal content, and other issues.

5.17.1. User comments on data sets shall be accessible to IOOS staff for review.

5.17.2. The user feedback mechanism shall provide a "Help" function.

5.17.3. The user feedback mechanism shall provide a mechanism for Usage Tracking.

2. Data Transport (DT) Requirements

1. Overall requirement

1.1. The DT shall support machine-to-machine interoperability with semantic meaning, i.e., the DT shall incorporate some collection of methodologies that promote the scripted exchange of data between computers, with all computers involved in a transaction capable of determining both the syntax and the semantics of the exchanged data without human intervention.

1.2. The DT shall include an access method that is consistent with that which is referred to as "Web services" in the literature.

1.2.1. "A Web service is a software system identified by a Universal Resource Identifier[4], whose public interfaces and bindings are defined and described using XML[5]."

1.3. Other Web services requirements TBD.

2. Representational requirements

2.1. The DT shall support metadata as described below:

2.1.1. Syntactic metadata are information about the data types and structures at the computer level, the syntax of the data. For example, variable D represents a floating point array measuring 20 by 40 elements.

2.1.2. Semantic metadata are information about the contents of the data set.

2.2. DT shall be able to transmit all relevant semantic metadata, that is translational use, descriptive use, and search metadata. They must be available in both human-readable and machine-readable forms.

2.3. DT shall be able to express the structure of the numeric data it will encounter in oceanographic data repositories.

2.4. DT shall be able to transmit the numerical data themselves without corruption or loss of precision.

2.5. The following simple and compound types shall be provided:

2.5.1. Simple types:

2.5.1.1. Integers (signed 16, 32, 64-bit; unsigned 8, 16, 32, 64-bit); floating point (32, 64-bit);

2.5.1.2. Strings

2.5.1.3. Pointers to types

2.5.2. Compound

2.5.2.1. Structures

[4] Uniform Resource Identifiers (URI): Generic Syntax, IETF RFC 2396, T. Berners-Lee, R. Fielding, L. Masinter, August 1998 (See http://www.ietf.org/rfc/rfc2396.txt)

[5] Web Services Glossary W3C Working Draft 14 November 2002 (See http://www.w3.org/TR/ws-gloss/)

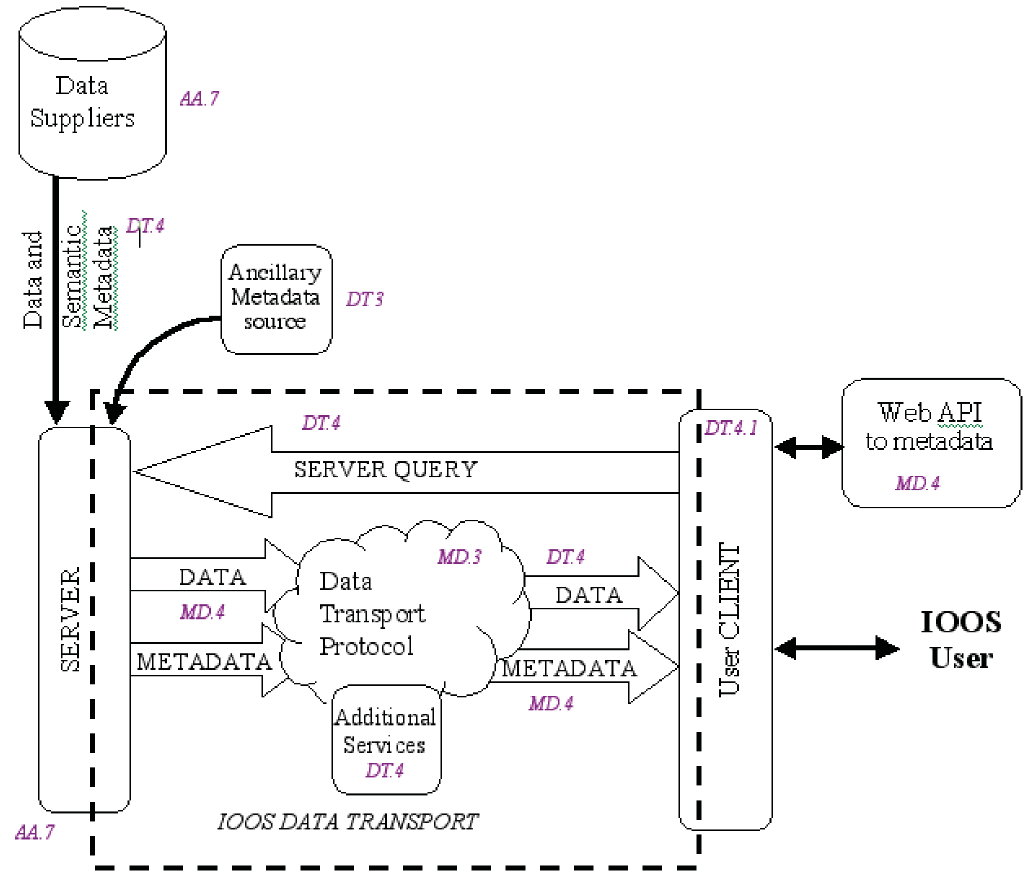


Figure 2. Schematic diagram of DMAC Data Transport (DT)

2.5.2.2. Arrays

2.5.2.3. others TBD

2.6. The DT shall be capable of accessing data in a variety of formats.

2.7. The DT shall be capable of delivering data of a given data type in a structurally consistent form across all data sets in the system.

2.8. The DT shall provide the metadata needed to transform the data to a consistent semantic form, or it must be capable of delivering the data in a consistent semantic form.

2.9. The DT shall provide access to metadata in a variety of forms, including the standard FGDC forms of the metadata, to take advantage of the metadata developed by different communities. The DT shall be capable of providing access to metadata from a site other than that of the data server.

2.10. The DT shall be capable of providing access to metadata from a site other than that of the data server together with the data. The DT shall be capable of binding these metadata to a data request where appropriate.

3. Modular approach

3.1. The DT shall be constructed in a modular fashion that allows capabilities to be utilized over alternative transport protocols to HTTP.

3.2. **Format** –The DT shall be capable of moving data from a site in which they may be stored in one format to a client application that may require them in another format.

3.2.1. Transport between sites will be implemented via an intermediate format, referred to as the DT syntactic data model.

3.2.1.1. The data model shall be discipline-neutral.

3.3. **Structure** - The system shall provide the capability of delivering data to clients in a structurally consistent form where appropriate. In this context, structure means the way that the data are organized, for example, grid, array, etc.

3.3.1. The structure layer protocol will define the organization of like data objects in a data set.

3.3.2. Operations and the associated modules in the structure layer that can be performed in a discipline-neutral fashion shall be logically separated from those that require a semantic understanding of the data.

3.4. **Semantic** – The DT shall provide a semantic data model, defined as the semantics implicit in the structural transformations that the system provides and the semantic information transported in the data access protocol.

3.4.1. The core semantic data model shall include the set of translational use metadata.

4. Functional requirements

4.1. The DT shall be capable of providing direct access to data via a variety of client programs, communicating directly with the program without the need to create data files.

4.1.1. The mechanism for user access to IOOS data can be either through a DMAC-enabled browser, or through user-supplied application software implementing DMAC access routines. In either case, DT shall provide the requisite software capability.

4.2. The DT shall support access to real-time data as well as access to retrospective (non-real-time) data.

4.2.1. The DT shall provide a push data delivery service.

4.2.2. The DT shall support “informed pull” of data.

4.3. The DT shall provide for online acquisition of data into legacy applications and new applications packages through the syntactic data model.

4.3.1. The DT shall allow users to obtain data subsets as formatted files (formats TBD) and human-readable ASCII numeric values via a standard Internet browser, possibly implemented via plug-ins.

4.3.2. The DT shall be designed and developed to accommodate the following considerations.

4.3.2.1. Data will be heterogeneous in type and storage format.

4.3.2.2. Data storage will be distributed.

4.3.2.3. Data will often, but not always, reside with the data collector.

4.3.2.4. The system to be developed will be a client-server system.

4.4. The DT shall provide a mechanism for subsetting data sets for retrieval, by parameter, by area, by time window, and by other criteria TBD.

4.4.1. When subsetting data the DT shall provide appropriate metadata.

4.5. The DT shall provide mechanisms for aggregating data.

4.5.1. Data of the same type and from same provider.

4.5.2. Data from different sources that do not or cannot share a single parent metadata record (e.g., observational data from different sources/systems).

4.5.3. When multiple data sets are aggregated, the DT shall provide a mechanism for providing appropriate aggregate metadata.

4.6. The DT shall display, for each data set it contains, the approximate size of the data set selected.

4.7. The DT shall support data restructuring, i.e., any process that ingests a data set described by one data model and maps that data set into another data set described by another data model.

4.8. The DT shall support data manipulation, including (but not limited to) such manipulations as

4.8.1. Re-projection—for example, Platte-Care to Mercator

4.8.2. Re-gridding—for example, same projection, different resolution

4.8.3. Averaging

4.8.4. Summing

4.8.5. Scaling of values such that they are delivered in a consistent system of units—for example, # specimens/m^3, m/s, °C

4.8.6. Conversion of time to different representations

4.8.7. Conversion of latitude and longitude

4.8.8. Conversion of depth

4.8.9. Conversion of missing values

4.9. The DT shall support access-restricted, secure transmission of data.

4.10. The DT shall support fault detection and localization within the DMAC.

4.11. The DT shall support the gathering of performance and usage metrics within the DMAC.

5. Design Constraints

5.1. The DT shall be designed to work cooperatively with other systems. For example, if a repository already uses a system that depends upon a particular data storage format, that site should not be forced to abandon its system in order to adopt IOOS.

5.2. The DT shall be designed to operate with minimum reliance on proprietary software.

5.2.1. The DT specifications shall be fully and openly accessible to the public.

5.2.1.1. There is a stated preference for software licensed under the General Purpose License.

5.3. The scheme that the DT adopts for generating syntactic and semantic data models shall be flexible and extensible so that any IOOS server can find a way to express its archive's storage format in an IOOS data model.

5.4. The DT shall support interoperability between Geographic Information Systems and Scientific Information Systems.

5.5. The DT shall support interoperability with other systems developed within other disciplines.

3. Data Archiving and Access (AA) Requirements

1. Vision

1.1. IOOS Data Archiving and Access (AA) will be a distributed system of interconnected archive and data centers that functions 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.

1.2. 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 shall be implemented in the Archive System. The result will be a system that provides more uniform access across multiple archive centers and that can handle all collections consistently. The data discovery component will allow access by both humans and machines.

1.3. 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.

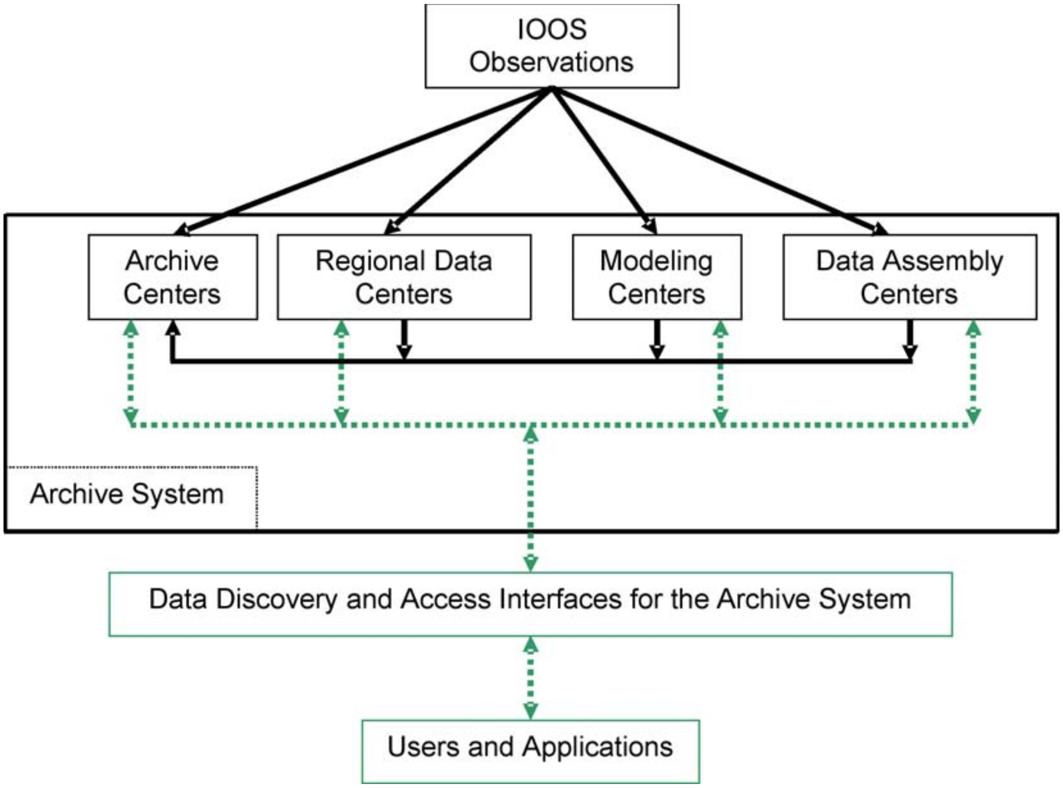


Figure 3. The Archive System represents an alternative view of those DMAC Subsystem elements that are involved with data archival. Primary archival (solid lines) and access (dashed lines) show data flow. Not shown are other data flows that are essential to IOOS but not directly pertinent to the Archive System.

2. The Archive System

2.1. The Archive System shall use coordinated methods for data collection, quality control, archiving, and user access.

2.2. The system shall 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 3, Data Archiving and Access Requirements).

2.3. 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.

2.4. 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.

2.5. Archive centers

2.5.1. Archive centers shall acquire, preserve, and provide access to IOOS data in perpetuity.

2.5.2. Archive centers shall implement mechanisms to ensure integrity and completeness of the archives.

2.5.3. Essential functions include constant monitoring of data streams, accounting for all files and records, and frequent checks of accuracy.

2.5.4. Archive centers shall provide for the archival of metadata.

2.5.5. Archive centers shall have maintenance strategies that protect the data as storage media and systems change.

2.5.6. Data stewards at the archive centers shall maintain constancy in formats and software to prevent conditions that could make accessing the data more difficult, more costly, or impossible.

2.6. Regional data centers

2.6.1. Regional data centers shall acquire and provide access to IOOS data collected in a specific geographic region.

2.6.2. Regional centers may collect a variety of physical, biological, and chemical ocean data that are used to support scientific, public, and commercial interests in the area.

2.6.3. Regional data centers shall apply quality control measures to data and derive specialized products.

2.6.4. Regional data centers shall fulfill the long-term archive obligation if they meet the IOOS standards for data integrity and stewardship or if they systematically transfer the data to an archive center.

2.7. Modeling centers

2.7.1. Modeling centers shall 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.

2.7.2. Modeling centers may provide access to their products, but their mission does not include long-term archiving.

2.7.3. Model products that are essential to IOOS goals shall be transferred and preserved at an appropriate archive center.

2.8. Data assembly centers

2.8.1. Data assembly centers shall obtain IOOS data and provide access to it.

2.8.2. Data assembly centers will typically specialize in certain types of data, and often provide quality control and data products in their area of expertise.

2.8.3. Data assembly centers may be permanent (e.g., NDBC) or exist only for limited periods (e.g., WOCE data assembly centers).

2.8.4. Data assembly centers do not provide long-term archiving, but often provide access.

2.8.5. Data assembly centers may gather distributed data 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.

3. Data Management

3.1. Although IOOS data may flow into the archive centers over several pathways (Figure 3, Data Archiving and Access Requirements), at least one copy of each set shall reside in a designated archive center.

3.2. Some categories of data will require that multiple copies be stored securely at separate locations under independent data management.

3.3. When data must be duplicated, a primary and secondary data steward shall be designated.

3.4. The primary data steward shall typically be an archive center and shall provide the highest level of access.

3.5. The secondary steward need not maintain full access, but shall maintain the data at the same level of integrity.

4. Access

4.1. Access services for IOOS users shall be provided from most centers in the Archive System.

4.2. Archive centers shall provide some real-time services, and enhance data discovery by using the IOOS metadata standards and data discovery techniques.

4.3. When regional, modeling, and data assembly centers provide access on schedules that meet the IOOS goals, duplication of this effort is not a requirement for the archive centers; however, the archive centers will ultimately receive the data, provide for their long-term preservation, and provide access to full archived data set.

5. Data Receipt

5.1. Modes of data receipt

5.1.1. Real-time-mode data arrive in real-time or near real-time, with the goal of being made available with minimum delay.

5.1.1.1. High-level quality control is not a requirement for real-time-mode data.

5.1.2. Delayed-mode data arrive later than real-time-mode 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.

5.1.3. The Archive System shall receive and archive sets of either type that address the seven IOOS goals.

5.1.4. All appropriate metadata should arrive with the data.

5.2. Integrity/Consistency

5.2.1. The Archive System shall implement mechanisms to ensure that all valuable data are sent and that an exact copy is received. The IOOS data transport system shall provide sufficient mechanisms to ensure accurate transfers over the networks.

5.2.2. Acceptable tools and procedures include:

5.2.2.1. Receipts and reconciliation reports for transfers over networks.

5.2.2.2. Skilled staff to review metrics (e.g., how much of the expected data were received and how much of the data set was made available).

5.2.2.3. Byte counts, inventories of data files, and checksums of records or files.

5.2.2.4. Test files that can be confirmed against archived data and used to verify local software.

5.2.2.5. Accuracy relative to other data sources (i.e., whether a set of data falls within acceptable ranges or compare acceptably with other data known to be correct).

5.2.3. The Archive System shall provide a failover mechanism for failed data transmissions.

5.2.4. The Archive System shall guard against unrecoverable data loss by making data integrity (or security) a primary objective.

5.2.4.1. Byte counts and checksums shall be calculated and used to verify that the data are uncorrupted when transmitted between data centers.

5.2.4.1.1. These quantities shall again be calculated after every internal process at the archive centers, and then re-calculated periodically on all archived data to protect against such problems as hard disk failures, media degeneration, incomplete file transfers, and malicious hacking.

5.2.5. Virus checks shall be performed on the data before archiving, then periodically on all data kept online.

5.3. The AA shall include guidelines to enable providers developing new data streams to select formats and metadata that can be easily integrated into IOOS. Specifications shall traverse the IOOS data-transport, metadata, and data-discovery components.

5.4. IOOS standards for metadata shall allow different versions of the same data and metadata to be traced by means of information on lineage and version.

5.4.1. The number of old versions of data to be preserved is TBD.

5.5. Data Formats

5.5.1. The AA shall process a broad range of data to be included in IOOS (physical, biological, chemical, geological, fisheries, socio-economic) encompassing many different native data formats.

5.5.1.1. Data providers shall use only established, fully documented formats, (TBD) which the data-transport methods handle. The data transport methods shall be robust and handle many established common formats.

5.5.1.2. It is not a requirement that each center be proficient in every format.

5.5.1.3. All metadata shall meet a common standard defined in Section 1, Metadata/Data Discovery Requirements.

5.5.1.4. If metadata do not meet the common standard, then the provider shall provide a mechanism that accepts the non-standard metadata as input and creates standard metadata as output.

5.5.1.4.1. Archive centers will consider accepting data in all formats.

5.5.1.4.1.1. Unique specialized formats (such as occasionally found in research or field data) are discouraged.

5.5.2. Proprietary formats (with undisclosed internal structure and typically with proprietary software) are unacceptable for long-term archiving and are prohibited.

5.5.3. Each center shall provide and maintain software for accessing each native format.

5.5.3.1. Centers shall maintain configuration management of the software in order to maintain currency with changing data formats.

5.5.3.2. This software shall also provide further documentation of data sets and changes in their lineage.

5.6. File-compression techniques used for transferring IOOS data shall use standard protocols with open documentation, such as GNU zip.

6. Data Preservation

6.1. All four component data centers of the AA will be responsible for acquiring and providing data, but only the archive centers will be primarily responsible for preserving data long term.

6.1.1. Long term is defined as much longer than the typical funding period of an oceanographic research project or the career of a principal investigator.

6.1.2. To qualify as an archive center, a data center shall be able to perform the following functions related to data preservation:

6.1.2.1. Create and manage multiple copies of the data and metadata;

6.1.2.2. Verify and generate metadata as well as preserve them with their associated data;

6.1.2.3. Frequently check data integrity;

6.1.2.4. Plan for evolution of technology.

6.2. Archive centers shall 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. Table 1 summarizes the four data categories and the number of archival copies required to meet the minimum IOOS Archive System standards.

6.3. Data Categories

6.3.1. Irreplaceable Data

6.3.1.1. The AA shall maintain two copies in separate archive centers in perpetuity.

6.3.1.2. The two copies of irreplaceable data shall be preserved in separate facilities under independent data management.

Table 1. IOOS Data Classes for Archiving and Access

Data Category	Data Description	Examples	Minimum Number of Archival Copies

Irreplaceable	Observational and research quality data that can not be reproduced or easily regenerated	Raw, ancillary satellite observations;
		Instrumental measurements;
		Biological samples;
		Model reanalyses;
		Complex merged data analyses
			Two

Replaceable	Derived from irreplaceable data, can be regenerated through systematic processing	Calibrated satellite radiances; Simple composites or analyzed data	One
Perishable	Real or near real-time data; typically replaced by higher quality data.	Direct broadcast satellite data; Operational analyses; Quick-look analyses based on uncalibrated or incomplete data	One
Virtual	Data provided through on-demand processing	Subsets from GUI; Analyses from a Live Access Server	Two*

* Original generation algorithms and documentation only.

6.3.1.3. One facility will be designated as the primary archive center for a particular data set, and the other as the secondary archive center.

6.3.1.4. The primary and secondary archive centers storing irreplaceable data may operate as mirror sites, both offering the same level of access, 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).

6.3.2. Replaceable Data

6.3.2.1. The AA shall maintain one copy (residence time in the archive will vary with replacement cycle).

6.3.3. Perishable Data

6.3.3.1. The AA shall maintain one copy until higher quality data are available.

6.3.3.2. When decision-critical data products are derived from data in this class, 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.

6.3.4. Virtual Data

6.3.4.1. 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.

6.4. Metadata. The AA shall maintain metadata as defined in Section 1, Metadata/Data Discovery Requirements, and will include the following types of metadata:

6.4.1. Use metadata (the semantic and syntactic information about a data set);

6.4.2. Discovery metadata (standard structured information describing a data set).

6.4.2.1. Data set lineage history (e.g., which irreplaceable data set was used to create this current data set);

6.4.2.2. Data category specification, which determines the storage requirements;

6.4.2.3. Release date, which is the date to remove temporary restricted access;

6.4.2.4. Version number and description of the version number;

6.4.2.5. Description of the file naming convention;

6.4.2.6. Unique IOOS-wide data set name or identification;

6.4.2.7. Mechanisms for correct publication citation and reference tracking.

6.4.3. Documentation metadata (bibliographic information about documentation associated with a data set).

6.4.4. Metadata shall be dynamic to accommodate through numerous incremental updates, modifications, corrections, and occasionally, full replacements.

6.4.5. Metadata shall be inclusive of sufficient information to 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.

7. Data Provision and Access

7.1. The AA shall accommodate data access from any suitable component of the IOOS Archive System (Figure 3, Data Archiving and Access Requirements).

7.1.1. By querying the system with the DMAC data-discovery interface, users or applications can discover what data are available. The data may then be pulled automatically with the data transport methods, or by the user from a GUI that displays the various options.

7.2. The AA shall implement the protocol for transporting data defined in Section 1, Data Transport Requirements.

7.3. The AA shall provide for access services tailored to data sets as provided in Table 2.

7.4. The core protocols shall include FTP, HTTP, and the IOOS DT protocols. Most IOOS data sets will be available in at least one, and preferably 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:

7.4.1. FTP – Direct downloads of data files, unrestricted public access, and no application support.

Table 2.

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				

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

7.4.2. HTTP – Direct downloads of data files, restricted or unrestricted access, and no application support.

7.4.3. OPeNDAP – Application-layer protocol that supports a number of data storage formats and allows a number of client applications to access data transparently.

7.5. The AA shall use the IOOS DT protocols to offer the following extended services:

7.5.1. Spatial subsetting – Extracting spatial sub regions from data sets for larger geographic areas.

7.5.2. Parameter subsetting – Extracting one or more variables from data sets containing many variables.

7.5.3. Temporal subsetting – Extracting short periods from data sets covering longer periods.

7.5.4. Temporal aggregation – Creating a longer time series from data files for shorter periods.

7.5.5. GIS products – Depicting data projected, interpolated, and rendered onto a map with GIS protocols.

7.5.6. On-line 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.

7.6. Data sets that are stored offline shall be kept accessible and discoverable through the data-discovery interfaces.

7.6.1. This access to offline data may be initiated by online ordering. Online ordering is a mechanism by which data are ordered and then picked up or delivered later.

7.7. The AA shall accommodate maximum latency periods as defined in the metadata.

7.7.1. 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.

7.7.2. Latency requirements shall be assessed and suitably defined in the metadata.

7.8. The AA shall provide unrestricted access under normal circumstances

7.8.1. The AA shall restrict access under special circumstances including:

7.8.1.1. Proprietary embargo – Data are available only for sale from commercial companies.

7.8.1.2. National security – Data are available only for defense purposes.

7.8.1.3. Calibration and validation – Data are available only to the science team while they calibrate or validate instruments, data, or models.

7.8.1.4. Non-commercial use only – Data are available for government applications and academic research, but not for resale.

7.9. The AA shall provide user services and use metrics.

7.9.1. Online documentation and knowledgeable staff shall be available to provide assistance and advice on both access and content.

7.9.2. Additional background information will be available through references and citations in the metadata.

7.9.3. The AA shall provide a facility to collect broad use metrics to evaluate the system effectiveness and gain a sense of how to improve it. Metrics shall include the following as a minimum:

7.9.3.1. 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.

7.9.3.2. 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, online analyses - in short, only the information required by the user. The data access metric shall also be broken down by data set and service method.

7.9.3.3. System performance statistics – This includes use of disks and computers as well as work performed (i.e., services executed and volume accessed).

7.9.3.4. In addition to numeric metrics, the AA shall provide for measurements of qualitative access.

7.9.3.4.1. Specifically, all archive systems shall have a means of soliciting and capturing user feedback on services and data sets.

Section 2. Phased Implementation Plan

METADATA/DATA DISCOVERY ACTIVITIES AND SCHEDULE

(see Figure 4)

1. Activity: Metadata: Determine IOOS Metadata Content and Format Standards

- **Description**: Determine the metadata contents and format for all IOOS metadata. Metadata will be FGDC CSDGM compliant but may require additional elements not in that standard.
- **DMAC Component:** Metadata
- **Milestone 1:** Compile IOOS metadata standards
- **Estimated Resources:**
- **Schedule:** Start beginning of Year 1 and continue for 3 years
- **Sequencing:**
- **Partnerships:** DMAC expert teams with additional input as needed.
  - **Task 1:** Convene an IOOS Metadata Standards Working Group with metadata representatives from all IOOS data disciplines to do a comprehensive assessment of metadata standards.
    - The make up of the W/G must:
      - represent the interests of all major ocean metadata holders;
      - be able to represent the broadest range of ocean data types: *in situ*, satellite-derived, biological data, sonar, various model output types, etc.;
      - liaisons to US environmental metadata standards activities should be identified (e.g., FGDC);
      - liaisons to international ocean metadata standards should be identified.
    - The W/G should consist of core standard group and a number of specializations to address special data types. The core group will have full responsibility for "format" issues as well as for content that is in common to all data types. The core group could be given responsibility to appoint specialist groups as it sees fit.
    - Also form expert subcommittees to address discipline specific issues. Identify need for extended elements to standard format. Evaluate developing a Standard Profile under the FGDC Content Standard.
    - After the initial work to produce the Interim Standards, some form of the committee will become a standing standards committee.
    - Consider whether to support standards other than FGDC CSDGM, e.g. Dublin Core, MARC21.
    - Category: Committee work

- **Task 2:** Develop Preliminary IOOS Metadata Standard
  - Category: Committee work
  - Deliverables: Initial IOOS Metadata Standards
- **Task 3:** Establish Liaison with metadata standards groups like FGDC.
  - Category: Committee work
- **Task 4:** Make interim list of keywords and a data dictionary
  - Category: Committee work
- **Task 5:** Study use of thesauri to enable machine-to-machine interoperability with semantic meaning.
  - Note: If thesauri are used, they will have to be maintained through a fair sized effort considering the variety of data and the complexity of language.
  - Category: Committee work
- **Task 6:** W/G incorporates input from expert subcommittees, results of studies, R&D, pilots, etc. into subsequent standards updates up to the release of the Interim Standards. Circulates draft interim standards for community comment. Specification will include guidelines on Data Quality issues along with lineage issues.
  - Category: Committee work
- **Task 7:** Develop policy for granularity of metadata.
  - Category: Committee work
  - Deliverables: IOOS Metadata Standards, keywords and data dictionary plus any updates or interim releases as needed
- **Task 8:** Develop guidance on metadata for subsetting and aggregation.
  - Category: Committee work
  - Description: This assumes that related metadata would be delivered along with transported data. How should the metadata change when the transported data are not identical to the source data? This must have strong representation from the Data Transport team and a joint sub-group should be considered. This is an activity overseen by the core standards group.
  - Task 8-1: Determine metadata modification for subsetted data.
    - Includes temporal, spatial, and parameter subsets
  - Task 8-2: Determine metadata modification for aggregated data.
    - Case 1: Data are from the same provider and same data type.
    - Case 2: Data are from different sources, may or may not be same data type.
  - Task 8-3: Determine metadata modification requirements for products or merged data.
    - Assumption is data products will have own unique metadata.
    - Key question: Should source data metadata be also delivered with product?
  - Deliverables: Documented guidance on metadata for subsetting and aggregation
  - Partnerships: Data Transport, Data Products

2. Activity: Develop Tools and Procedures to Support Metadata Providers

- **Description:** Develop or acquire procedures, practices and tools to aid developers and IOOS in designing, producing, verifying, and maintaining metadata.
- **Milestone 1:** Develop and maintain metadata.
  - **Task 1:** Select or develop a master metadata management system.
  - **Task 2:** Develop/acquire tools for metadata generation, validation, maintenance.
  - **Task 3:** Provide developer support to users in addition to tools and User Guide Training, support networks, consulting and help desk.
  - **Task 4:** Plan regular reviews of exiting metadata by data providers. Update or add information as data set circumstances change. Update perishable information such as contact info. This is assumed to fit within the standard metadata framework, otherwise make change recommendations to Standards Committee.
  - **Task 5:** Develop User Guide for Metadata.
- **Deliverables:** User Guide for Metadata
- **Estimated Resources:**
- **Schedule:** Year 2
- **Sequencing:**
- **Partnerships:** User Support

- **Milestone 2:** Develop tools to modify metadata as appropriate to data accessed.
- **Deliverables:** Tools to modify metadata appropriately
- **Estimated Resources:**
- **Schedule:** Years 3-4
- **Sequencing:** Following release of guidance on how to modify metadata for each of the conditions
- **Partnerships:** User Support

3. Activity: Discovery: Select or Develop and Maintain Catalog and Search Capability

- **Description:** The catalog(s) are the information source used for data discovery. Search capability is the prime purpose of the catalog.
- **Milestone 0:** (for DMAC Steering Committee): Designate an initial list of pre-operational Metadata Cataloging Centers. Initial list presumably to include NASA/GCMD and NOAA/NCD-DC. Additionally, this committee should look at the OBIS system for inclusion of biological information.
- **Milestone 1:** Convene a Catalog Architecture Working Group.

  - Membership should be of people expert in metadata searching and distributed metadata management. It must have broad representation from the data supplier community as well as DMAC teams.
  - **Task 1:** Design Catalog Architecture.
  - **Sequencing:** Can be concurrent with Activity 1.
- **Milestone 2:** Determine search/browse capabilities needed.
  - **Task 1:** Determine the level of search/browse needed and hence the composition of the catalog.
    - Minimum metadata vs. all metadata or in-between
    - If full text search required, then full metadata record will be required.
    - Search candidates or features:
      - Spatial Search
      - Temporal Search
      - Thematic Search – text searching esp. important
      - Taxonomic Search – biological data
      - Parameter Search
      - Additional Search Parameters
      - Browse Option
      - Results listing and Search refinement
    - Category: Committee work
  - **Task 2:** Get user feedback on search/browse needs.
  - **Task 3:** Choose and schedule search capabilities to be implemented.
- **Milestone 3:** Determine metadata loading and update procedures.
  - **Task 1:** Write Catalog Management Plan.
    - How is catalog managed and maintained?
    - How are catalogs kept up-to-date as data source metadata changes?
      - Option 1: Catalog harvests metadata from data source.
      - Option 2: Catalog metadata maintained by data source.
      - Option 3: Allow both (committee recommendation)
  - **Deliverables:** Design recommendations
- **Milestone 4:** Plan catalog security (cross-discipline with all DMAC)
  - Plan security tools, procedures, and practices to protect all participating systems from inappropriate access, intentional, or accidental.
  - **Task 1:** Determine catalog security needs with DMAC.
    - Category: Committee work
  - **Task 2:** Determine catalog access needs.
    - Category: Committee work
  - **Task 3:** Write Catalog Security Plan.

  - º Category: Committee work
- **Deliverables:** Catalog Security Plan
- **Partnerships:** Archive team, Data Transport team

- **Milestone 5:** Build initial capability (pilots are needed, TBD).
  - Multiple pilots and pre-operational tasks are acceptable if they can co-exist with the catalog structure including the OBIS development efforts for distributed biological data search.
  - **Task 1:** Build catalog.
  - **Task 2:** Populate Catalog.
    - º Populate DMAC catalog based on current archive centers.
    - º Accept or Harvest Metadata from archive centers.
  - **Task 3:** Build initial user interface (may be web portal).
  - **Schedule:** Beginning Year 2 and continuing
  - **Sequencing:** Will need to be repeated as archive centers join.
  - **Partnerships:** All archive centers currently part of IOOS.
- **Milestone 6:** Transition to pre-operational and operational systems.
  - **Task 1:** Transition to Pre-Operational.
  - **Task 2:** Transition to Operational.

4. Activity: Discovery: Develop Discovery Interface for Archive System

5. Activity: Discovery: Design Discovery Portal

- **Description:** Decide if a data portal is desirable and recommend functionality especially in search capabilities. Initiate pilot task(s) followed by a pre-operational task based on lessons learned and user feedback.
- **Milestone 1:**
  - **Task 1:** Design overall architecture.
    - º Single vs. many, governance?
    - º Category: Committee work
  - **Task 2:** Search content and scope
    - º Subscriptions
    - º Event association with parameters
    - º Broad vs. narrow search
    - º Note: These depend on and feed back to metadata standards and design. Functionality will be added in stages, probably from simple to more complex.
    - º Category: Committee work
  - **Task 3:** Solicit and incorporate user feedback.
    - º Category: IOOS Standards Process
  - **Task 4:** Determine access given to public search engines, for example, Google.

  - o Category: Committee recommendations => DMAC policy (governance)
  - **Task 5:** Pilot Data Portal - NOAA NCDDC
    - o Provide for a catalog and data access portal at NOAA's National Coastal Data Development Center and include discovery of NDBC data (hub of 70 moored buoys and 60 C-MAN shore sites – transporting hourly observations).
    - o Category: Contract(s)
  - **Task 6:** Pre-operational Data Portal
    - o Category: Contract(s)
  - **Task 7:** Operational Data Portal
    - o Category: Contract(s)
- **Deliverables:** Design recommendations
- **Estimated Resources:**
- **Schedule:** Start right away using existing data portals for study.
- **Sequencing:**
- **Partnerships:** User Support, user community, Data Transport, portal experts

6. Activity: Discovery: Study Alternate Discovery Approaches

- **Description:** Study alternate discovery approaches like Web Services and Semantic Web to address feasibility of this type of approach. This study could reveal methodologies to translate among multiple ontologies and allow the user to search among multiple controlled keywords and thesauri.
- **Milestone 1:** Study an implementation language neutral approach.
  - **Task 1:** Study, report alternatives like web service.
- **Milestone 2:** Study feasibility of using semantic web.
  - Description: Determine feasibility of using semantic web and ontologies in connection with IOOS metadata and for interoperability with other data catalogs.
- **Milestone 3:** Convene a working group to study semantic web feasibility
  - **Task 1:** How is this broken up if at all?
    - o If initial decision is "No," shouldn't it be revisited in a few years as the technology matures?
    - o If decision is "Yes," initiate a R&D or pre-operational task?
    - o Category: Committee work
- **Deliverables:** Feasibility Report
- **Estimated Resources:**
- **Schedule:** Year 1
- **Sequencing:**
- **Partnerships:** Interim Standards Committee, User Support team, semantic web experts.
- **Milestone 2:** R&D or Pilot study of alternate approaches in DMAC
  - **Task 1:** R&D or Pilot study of alternate approaches

  - **Category:** Committee work
- **Deliverables:** Functioning demonstration for review.
- **Estimated Resources:**
- **Schedule:** Year 2
- **Sequencing:**
- **Partnerships:** Interim Standards Committee, User Support team, semantic web experts.

7. Activity: Design and Implement Data Location Service

- **Description:** The end result of the Data Discovery process should segue seamlessly into the Data Transport (access) process—either the: (1) DMAC middleware connection, (2) on-line browse (visualization and subsetting), or (3) web file transfer (e.g., FTP). On-line, distributed data are dynamic in the sense that the point of access for data may move, and fine granularity information about the data sets may not be available in the catalog (e.g., the catalog cannot have the ability to perform GIS-style proximity queries about available data). The Data Location Service will be a standard machine-to-machine interface which enables the seamless segue from Data Discovery to Data Access.
- **Milestone 1:** Determine specifications for data location service.
  - **Task 1:** Determine specification for finding path to the requested data (e.g., the directory on the file server in the data archive).
    - Category: Contract(s) or RFP
  - **Task 2:** Determine specification for finding individual granules of requested data (e.g. the individual files which satisfy user request)
    - Category: Contract(s) or RFP
- **Milestone 2:** Data Location Service pilot projects
  - **Category:** Pilot projects with collaborating data suppliers and catalog services
- **Milestone 3:** Broad Deployment of Data Location capability
  - **Cross cut:** jointly with “population” of DMAC middleware solution
- **Deliverables:**
- **Estimated Resources:**
- **Schedule:** Year 2 or 3
- **Sequencing:**
- **Partnerships:** User Services

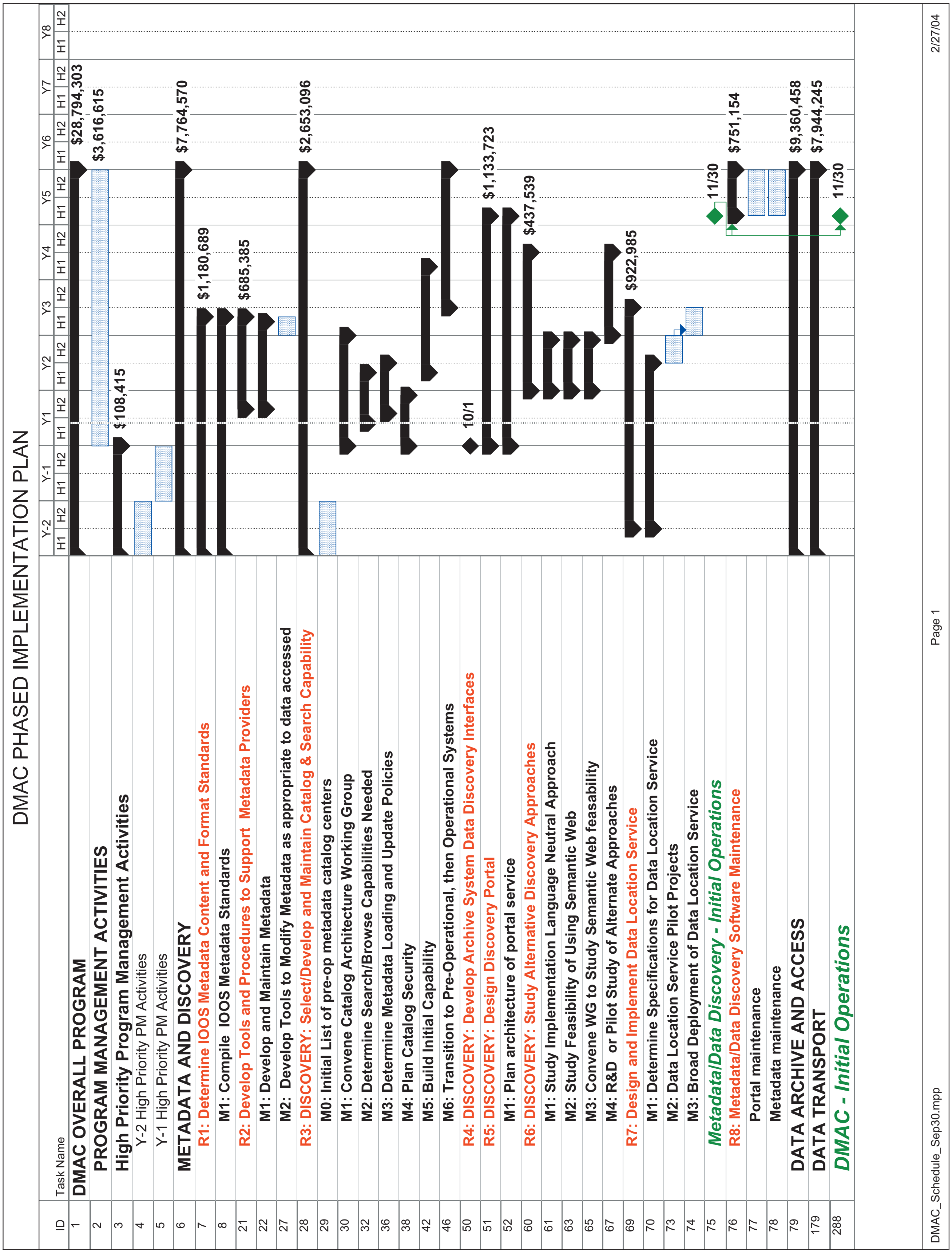


Figure 4. Metadata and Data Discovery Gantt Chart

DATA TRANSPORT ACTIVITIES AND SCHEDULE

(SEE FIGURE 5)

1. Activity: Develop Comprehensive IOOS Data Model

- **Description:** Accessed data will be moved into the data model for transport between server and client.
- **Comment:** There is no strict requirement for a single comprehensive model, but there should generally be only a single, correct representation for any given data class. If multiple representations of the same data class do exist, a reliable procedure to translate between them must be documented
- **Milestone 1:** Develop Comprehensive Data Model - Complete from the perspective of both syntactic and semantic elements – To provide the maximum flexibility in system design, the data model should be divided into a syntactic portion that is discipline neutral and a semantic portion that contains the discipline specific characteristics of the data.
  - **Task 1:** Develop a syntactic data model[6] for the system.
    - Approach: Adopt the OPeNDAP data model as the initial, provisional fast-track solution[7].
    - Special Considerations: The primary focus of this task should be the augmentation (if needed) of the OPeNDAP data model to accommodate data types within the OBIS and GIS data models.
    - Level of Effort: FTE[8] (TBD)
    - Duration: 6 months. The task outlined here is only for the initial development of the semantic data model. A subsequent task addresses the evolution of the model. This is true of Task 2 also.
    - Start: Immediately[9].
    - Category: Committee, OPeNDAP, OBIS, or the Data Model Working Group.
  - **Task 2:** Develop a semantic data model for the system.
    - Approach: Consider existing semantic data models such as that being promoted by OGIS, the DEI data model and others that may already exist. It is imperative that the semantic data model be kept as simple as possible to ensure the maximum compliance within the ocean community.

[6] The resulting data model may in fact consist of several data models. If so, they will collectively be referred to as the data model in this work plan.

[7] The OPeNDAP data model has been developed explicitly for this purpose, has already been vetted within the ocean community as part of the NOPP funded NVODS effort and is now in operational use by NVODS (as well as other non-oceanographic communities).

[8] The levels of effort identified in this work plan are supported FTEs (full time equivalent years). This support could be provided by subcontract to one of the groups indicated or provide funding to the Data Model Working Group for this purpose. Addition community participation is anticipated on many of these efforts through committee work. The level of effort for such committee work is not included here.

[9] Immediately means as soon as practical. These are absolutely essential components of the system, components on which much of the rest of the data management system rests.

- o Level of Effort: FTE (TBD)
- o Duration: A candidate semantic data model should be developed within 6 months of project initiation. There will likely be additions/modifications to the semantic data model over the next two years as it becomes more heavily exercised and as a result of some of the pilot efforts discussed below.
- o Start: Immediately.
- o Category: Subcontract, Metadata Standards[10], OBIS, OGIS, OPeNDAP. The group that puts this model and the various components together must include the data representation issues community very broadly. It should include representatives from each of the oceanographic sub-disciplines (biology, physics, chemistry, and geology), from GIS community (OGIS, ESRI, EaSY, etc.), from the ocean modeling community (GCM, coastal and finite element), and from the data collection communities—satellite (projections, swath), and *in situ* (hydrographic, moorings, floats).

- **Task 3:** Develop a controlled vocabulary for system contents.
  - o Approach: Adopt as a starting point the controlled vocabulary developed by the Marine XML consortium.
  - o Level of Effort: FTE (TBD)
  - o Duration: 6 months.
  - o Start: Immediately.
  - o Category: Subcontract, Metadata Standards, OBIS, OPeNDAP, and national and international metadata standards. The group that assembles the controlled vocabulary must include broad representation from the community of ocean data users: biologists, chemists, geologists, physicists, community planners, etc.
- **Task 4:** Synthesize the work of Tasks 1 and 2 into a complete data model.
  - o Special Considerations:
    - This task involves the assembly of the syntactic and semantic data models into a complete data model; i.e., linkages between the data model components in the two must be established. This work may require the addition of data types to the syntactic data model.
    - Pilot implementations designed to exercise the data model are discussed under Milestone 2.
  - o Level of Effort: FTE (TBD)
  - o Duration: 9 months.
  - o Start: 6 months.
  - o Category: Subcontract, OPeNDAP, OBIS.
- **Task 5:** Publish draft data model—follow IOOS Standards Process for Review.
  - o Subtasks:
    - Devise a plan to obtain community feedback.

[10] Metadata Standards refers to the IOOS Metadata Standards group.

        - Publish and circulate the draft data model.
        - Obtain feedback from the community on the data model.
    - Level of Effort: FTE (TBD)
    - Start: 1 _ year
    - Duration: 6 months
    - Category: Data Model Working Group
  - **Task 6:** Pilot implementations of data model
    - Brief pilot task using data model for non-gridded data, such as remote sensing "swath" data.
      - Restructuring and aggregation of this sort of "sequence" data are important due to the large amount of data in that format.
    - Pilots should include network transport utilizing the "fast-track" transport mechanism at a minimum
    - Category: Contract
  - **Task 7:** Broad testing of data model by distinct ocean data communities
    - including data observed from biological/laboratory sampling, cruises, moorings, floats, satellites, … and produced by the broadest range of models
    - Category: Community participation activity
    - Sequencing: Must follow pilot testing
- **Deliverables:** Comprehensive IOOS Data Model Standard
- **Estimated Resources:**
- **Schedule:** Year 1 and 2
- **Sequencing:** IOOS Standards process must be completed before Task 6
- **Partnerships:** oceanographic data communities, OPeNDAP, national and international metadata standards

2. Activity: Deliver time-critical (real-time) data to data assembly and operational modeling sites

- **Description:** IOOS sites that have regular, repeated need of time critical observations may best be served by a subscription-based "data push" service.
- **Milestone 1:** Provide operational support for time-critical data.
  - **Fast-track note:** An effective implementation of DMAC real-time delivery is singularly important in commencing the integration of IOOS operational observations with modeling activities. If a suitable candidate is available it will significantly advance the IOOS toward implementing a fast-track solution to this component. The Plan must include procedures to evaluate the effectiveness of the solution adopted and either make adjustments to it or abandon it, if it proves unsuitable.
  - **Task 0:** Characterize the need for real-time data.

    - Category: The mode in which real-time data should be received has not been clearly articulated. A careful examination of the issues related to this must be undertaken. Such an examination would include the number of sites that will want access to real time data, the number of originating data sites, the type of data that is required, etc.
  - **Task 1:** Adopt Unidata IDD as initial, provisional, fast-track transport solution.
    - Category: Unidata IDD is IOOS-pre-operational for operational, real-time delivery of formatted files to IOOS modeling sites. It is IOOS-pilot for format conversions of the data.
  - **Task 2:** DMAC evaluation and review of Unidata IDD, including data carry capacity, data integrity, and data delivery assurance.
    - Category: DMAC Governance Committee
  - **Task 3:** Evaluate the state of real-time data access to current and potential modeling operational sites.
    - Work cooperatively to minimize unproductive duplication and maximize timely, reliable availability of quality-controlled, real-time observations for modeling.
    - Category: Committee work
  - **Task 4:** Identify IOOS partner sites to serve as real-time data assembly and distribution centers
    - Adopt the US GODAE server (collocated with USN FNMOC) as a provisional, pre-operational IOOS real-time data assembly and distribution point (identification of other sites to follow).
    - Category: IOOS Governance
  - **Task 5:** As needed, initiate complementary and/or alternative real-time delivery solutions as R&D activities, Pilots, or Pre-operational solutions.
    - Category: DMAC Governance Committee
  - **Task 6:** Explore blended Push/Pull delivery in which data are pushed only to data assembly centers. All others use Pull delivery (typically middleware, FTP, or HTTP transfers).
    - Category: Contract
- **Deliverables:**
- **Estimated Resources:**
- **Schedule:**
- **Sequencing:**
- **Partnerships:** Modeling centers, data assembly centers

3. Activity: Develop DMAC Middleware

- **Description:** The middleware solution embodies four components: (1) the Ocean Data Access Protocol (ODAP) – the format-neutral method of requesting and receiving data and metadata over an Internet connection, (2) translating data from legacy data management systems (formatted files, RDBMS, etc.) into the ODAP, and (3) ingesting data from the ODAP into legacy and new client applications. The system will be capable of restricting data delivery based on data volume to be delivered. These must be changeable at the discretion of the DT team or DMAC Governance.

- **Milestone 1:** Determine the breadth of data management solutions in use by IOOS data suppliers, which must be supported by the middleware.
  - **Task 1:** Survey the IOOS participants to determine the current data management solutions in usage, the file formats and data management systems that must be addressed by the middleware, and the particular data sets which depend upon each management approach. Evaluate the subsetting needs that attend each data management system.
    - Category: contract
  - **Task 2:** Prioritize the server-side requirements based on IOOS theme priorities and critical data streams.
    - Category: DMAC Governance Committee
  - **Task 3:** Initiate development of server-side solutions based upon priorities
    - Sequencing: must follow adoption of initial transport protocol
    - Category: contract
- **Milestone 2:** Determine the breadth of legacy and new client applications that should be supported. Similarly survey and prioritize requirements for delivery of formatted subsets to users. Priorities should reflect the seven IOOS themes
  - **Task 1:** Survey user groups (and potential ocean information product suppliers) to access application and formatted file needs.
    - Category: IOOS User Outreach Committee activity
  - **Task 2:** Prioritize the client-side requirements based on IOOS theme priorities and critical data streams.
    - Category: DMAC Governance Committee
  - **Task 3:** Initiate development of application solutions and downloadable formats based upon priorities.
    - Sequencing: must follow adoption of initial transport protocol
    - Category: contract
- **Milestone 3:** Determine the specification for the ODAP.
  - **Fast-track note:** An effective implementation of the DMAC middleware component is singularly important to the ability to begin the integration of IOOS, as it allows data suppliers and users to bypass traditional barriers of file format, size, and locality. If a suitable candidate is available, it will significantly advance the IOOS to implement a fast-track solution to this component. The Plan must include procedures to evaluate the effectiveness of the solution adopted and either make adjustments to it or abandon it, if it proves unsuitable.
  - **Task 1:** Adopt OPeNDAP as initial, provisional fast-track transport solution.
    - Category: OPeNDAP is IOOS-Operational for gridded data; IOOS-Pilot for all other classes of marine data

- **Task 2:** Publish draft OPeNDAP specification document. Request comments.
  - Category: contract
- **Task 3:** DMAC evaluation and review for OPeNDAP
  - Category: DMAC Governance Committee
- **Task 4:** As needed initiate complimentary and/or alternative ODAP solutions as R&D activities, Pilots, or Pre-operational solutions
  - Category: DMAC Governance Committee
- **Task 5:** Develop an ancillary information framework allowing OPeNDAP servers to convert the native structure and attributes of a data set into the standard DMAC data model
  - Category: Contract
- **Task 6:** Develop detailed requirements and software specifications, followed by design and implementation of the aggregation servers required for *in situ* data collections.
  - Category: Contract
- **Task 7:** Identify (or adopt) procedures for developing consistent semantic use metadata for all IOOS data sets.
  - Category: Contract

- **Milestone 4:** Implement Server-side Middleware Tools
  - **Task 1:** Survey IOOS participants to determine current data management solutions
  - **Task 2:** Prioritize server-side requirements
  - **Task 3:** Initiate development of server-side solutions based upon priorities
- **Milestone 5:** Adapt or develop Client Software for Initial Transport Protocol
  - **Task 1:** Survey and recommend priority for applications to be supported by DMAC
  - **Task 2:** Initiate development of application and format solutions
- **Deliverables:** Functioning IOOS middleware component
- **Estimated Resources:**
- **Schedule:**
- **Sequencing:**
- **Partnerships:** Archive and Access, Metadata and Discovery, OPeNDAP, OpenGIS, OBIS, etc.

4. Activity: Make data available using IOOS middleware solution

- **Description:** Work with suppliers of data to make data available through the DMAC middleware solution.
- **Milestone 1:** Ensure that IOOS data suppliers make data available through middleware.
  - **Task 1:** Train middleware installers/troubleshooters/trainers.
    - Category: Contract
  - **Task 2:** Install middleware adaptors and other software as needed at suppliers' sites and train local personnel in configuration and management of the software.
    - Category: Contract
- **Deliverables:**
- **Estimated Resources:**

- **Schedule:** Start Year 1. Continuing.
- **Sequencing:**
- **Partnerships:** IOOS community, DMAC teams, governance, representatives from data providers in the disciplines.

5. Activity: Data Manipulation Services

- **Description:** Add optional functionality which may be so commonly required that great efficiencies and additional levels of integration are achieved through adding them as core DMAC services.
- **Milestone 1:** Prioritize and implement Data Manipulation Services.
  - **Task 1:** Prioritize Data Manipulation Services, including aggregation, regridding, and simple transforms such as averages and extrema.
    - Category: Governance Committee
  - **Task 2:** Develop specifications and implement services
    - Category: Contract
- **Deliverables:** Services descriptions and schedule for implementation
- **Estimated Resources:**
- **Schedule:** Year 2 and 3
- **Sequencing:** After initial DT work is done
- **Partnerships:**

6. Activity: Develop Metrics and Implement Performance Monitoring

- **Description:** Metrics and performance monitoring are necessary during development to monitor efficiencies, track user activity and inform further work. They are necessary in operations for reporting, monitoring for problems and to direct further improvements. Monitoring and metrics will evolve with time and should be reviewed periodically.
- **Milestone 1:** Determine specifications for Metrics and Performance Monitoring.
  - **Task 1:** Determine metrics to be used in DMAC and requirements for performance monitoring
    - Category: Committee work
  - **Task 2:** Implement performance monitoring in DMAC systems
    - Category: Contract/Pilot
  - **Task 3:** Pre-operational
    - Category: Pre-operations
  - **Task 3:** Operational
    - Category: Ongoing operations
- **Deliverables:**
- **Estimated Resources:**
- **Schedule:** Year 2, activity will continue

- **Sequencing:**
- **Partnerships:**

7. Activity: Implement Middleware Security (Cross-discipline effort with all DMAC)

- **Description:** Implement Security tools, procedures and practices to protect all participating systems from inappropriate access, intentional or accidental. Data providers must be able to configure compute and network resource limits. Selected data streams may have restricted access.
- **Milestone 1:** Create middleware Security Plan
  - **Task 1:** Determine Data Transport Security needs with DMAC
    - Category: Committee work
- **Milestone 2:** Develop and deploy middleware Security Plan
    - Category: Contract
- **Deliverables:** Middleware Security Plan
- **Estimated Resources:**
- **Schedule:** Year 3
- **Sequencing:**
- **Partnerships:** Archive team, Metadata and Discovery team

8. Activity: Provide guaranteed geo-temporal-referenced browse for all IOOS data

- **Description:** Ensure that all IOOS data are viewable through common web browsers (e.g., Netscape®, Internet Explorer®) in the form of intelligible "working graphics" on demand and human-readable numeric tables, and that they may be subsetted and downloaded in a range of common file formats. This will provide a necessary overview tool for IOOS scientists and a common entry point for all users who wish to explore IOOS.
- **Milestone 1:** Guaranteed minimum geo-temporally referenced graphics and numeric listings for viewing, and formatted file subsets for downloading.
  - **Fast-track note:** An implementation of a guaranteed minimal uniform DMAC data visualization (graphics) and access mechanism is vital to the integration of IOOS. It is fundamental to the ability to make quick evaluations of the effectiveness of all aspects of the DMAC and the quality of the data. If a suitable candidate is available, it will significantly advance the IOOS to implement a fast-track solution to this component. The Plan must include procedures to evaluate the effectiveness of the solution adopted and either make adjustments to it or abandon it, if it proves unsuitable. This effort should include consideration and/or implementation of Open GIS Consortium standards, including Web Feature Service, Web Coverage Service, Web Mapping Service and Geography Markup Language.
  - **Task 1:** Adopt the NVODS Live Access Server (LAS) as initial, provisional, fast-track guaranteed minimum browse solution.
    - Category: LAS is IOOS-Pre-operational for gridded data sets and IOOS-Pilot for others.

- **Task 2:** DMAC evaluation and review for OPeNDAP
  - Category: DMAC Governance Committee
- **Task 3:** As needed, initiate complementary and/or alternative real-time delivery solutions as R&D activities, Pilots, or Pre-operational solutions.
  - Category: DMAC Governance Committee
- **Task 4:** Determine the suite of visualization styles and file formats and data comparison capabilities that should be guaranteed by IOOS.
  - Category: Governance committee
- **Task 5:** Implement additional visualization styles and file formats and data comparison capabilities as determined.
  - Category: Contract

- **Deliverables:** Guaranteed minimum visualization and download capabilities
- **Estimated Resources:**
- **Schedule:** Year 1-4
- **Sequencing:**
- **Partnerships:** Archive and Access, Metadata & Discovery, GIS experts, User Support

9. Activity: OPeNDAP-OBIS Integration

- **Description:** OBIS is a globally distributed network of systematic, ecological, and environmental information systems. Data held in associated archives will be seamlessly integrated with those data accessible via the OPeNDAP.
- **Milestone 1:**
  - **Task 1:** Develop design for integration of the systems.
    - Category: Contract
  - **Task 2:** Implement design
    - Category: Contract
- **Deliverables:** Documented software
- **Estimated Resources:** 0.5 FTE
- **Schedule:** Year 1
- **Sequencing:** Immediately
- **Partnerships:** OPeNDAP and OBIS project

10. Activity: Aggregation of unstructured (a.k.a. vector, point, sequence, or profile) data

- **Description:** Design and implement a server or suite of servers capable of aggregating unstructured data. This server(s) will be capable of aggregating data across sites as well as within a site.
- **Milestone 1:** Design and implement a server for mooring data – fixed location variable in depth and time.
  - **Task 1:** Develop a consistent data model for mooring data.
    - Category: Contract

- **Task 2:** Convene workshop of mooring data providers to evaluate the model.
  - Category: Contract
- **Task 3:** Implement test and document design.
  - Category: Contract

- **Milestone 2:** Design and implement a server for hydrographic data – fixed location and time variable in depth.
  - **Task 1:** Develop a consistent data model for hydrographic data.
    - Category: Contract
  - **Task 1:** Convene workshop of hydrographic data providers to evaluate the model.
    - Category: Contract
  - **Task 1:** Implement test and document design.
    - Category: Contract
- **Milestone 3:** Design and implement a server for underway data – variable in space, depth and time (ship and drifter data).
  - **Task 1:** Develop a consistent data model for underway data.
    - Category: Contract
  - **Task 2:** Convene workshop of underway data providers to evaluate the model.
    - Category: Contract
  - **Task 3:** Implement test and document design.
    - Category: Contract
- **Milestone 4:** Design and implement a general server for unstructured data.
  - **Task 1:** Evaluate feasibility of a general aggregation server for unstructured data.
    - Category: DMAC
  - **Task 2:** If appropriate, design a general aggregation server for unstructured data.
    - Category: Contract
  - **Task 3:** Implement test and document design.
    - Category: Contract
- **Deliverables:** Documented software
- **Estimated Resources:** 2.0 FTE plus 3 workshops of ~20 attendees.
- **Schedule:** Year 1 and 2
- **Sequencing:** Immediately
- **Partnerships:** Data archivists for unstructured data, software developer

11. Activity: Develop a generic OPeNDAP server for unsupported data formats

- **Description:** A significant fraction of data to be made available by IOOS participants will not be in standard data formats, formats for which OPeNDAP servers already exist. In this project, a configurable server will be developed that may be used with a wide range of data formats.
- **Milestone 1:** Develop a generic OPeNDAP server for unsupported data formats.
  - **Task 1:** Design the server.

    - Category: Contract
  - **Task 2** Implement, test, and document design.
    - Category: Contract
- **Deliverables:** Documented server
- **Estimated Resources:** 2.0 FTE
- **Schedule:** Year 1
- **Sequencing:** Immediate
- **Partnerships:**

12. Activity: OPeNDAP-GIS client and GIS-OPeNDAP server

- **Description:** A significant fraction of IOOS users are expected to be GIS users while much of the data generated as part of IOOS will not generally be accessible from GISs. In this project, OPeNDAP client(s) will be developed for one or more commonly used GIS systems and OPeNDAP servers will be developed for commonly used GIS data formats.
- **Milestone 1:** Develop an OPeNDAP server for commonly used GIS data formats such as GeoTIFF.
  - **Task 1:** Design OPeNDAP server for GeoTIFF.
    - Category: Contract
  - **Task 2:** Implement, test, and document design.
    - Category: Contract
- **Milestone 2:** Develop a GIS OPeNDAP client.
  - **Task 1:** Delineate the issues, which GISs should be targeted and the level of support (access) that is appropriate.
    - Category: DMAC
  - **Task 2:** Design OPeNDAP GIS client
    - Category: Contract
  - **Task 3:** Implement, test and document OPeNDAP GIS client
    - Category: Contract
- **Deliverables:** Documented software
- **Estimated Resources:** 4.0 FTE
- **Schedule:** Year 1 and 2
- **Sequencing:** Immediate
- **Partnerships:** GIS, DMAC

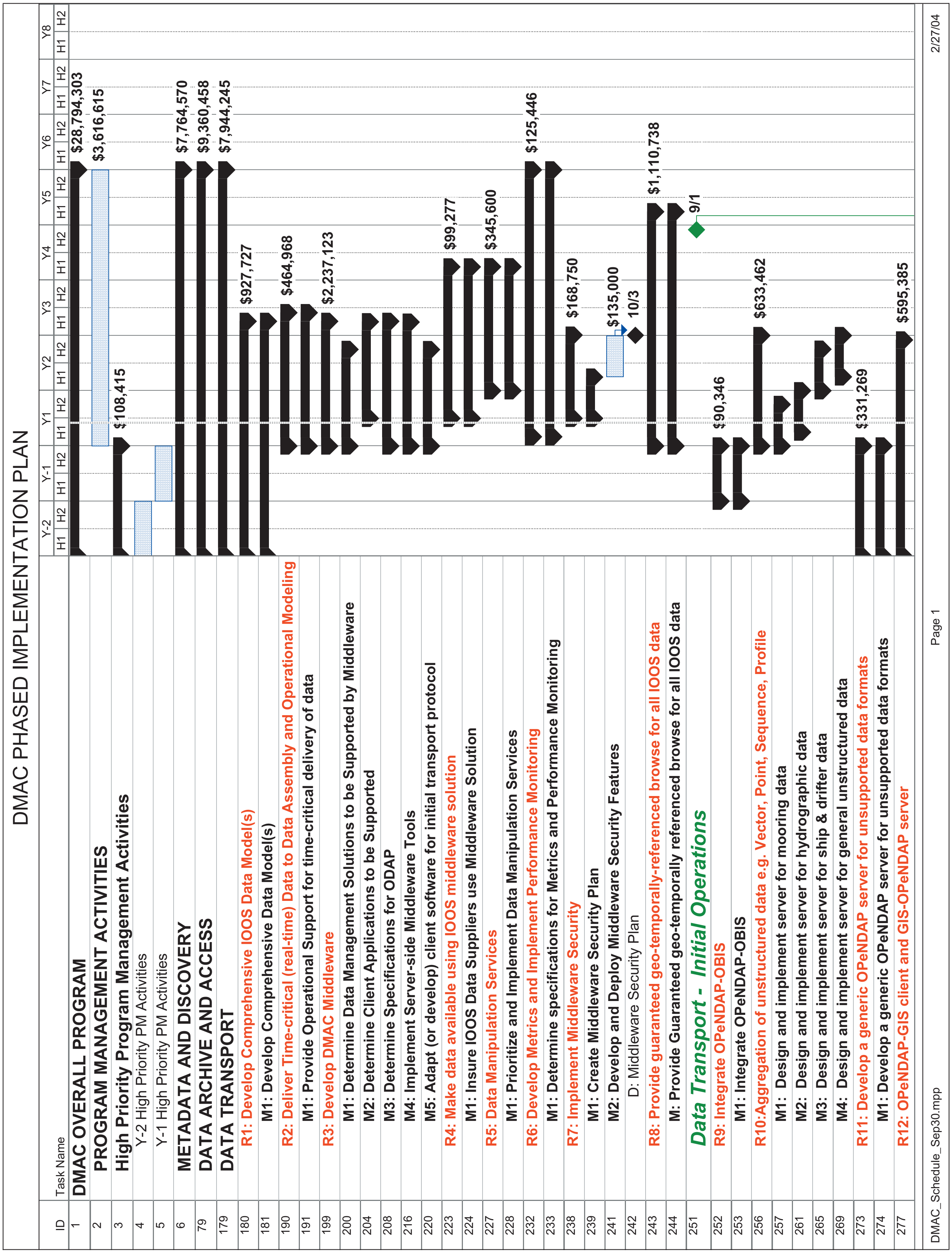
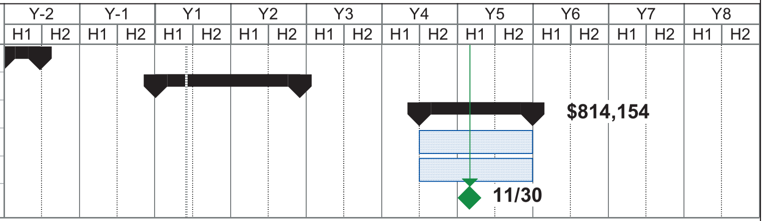


Figure 5. Data Transport Gantt Chart

DMAC PHASED IMPLEMENTATION PLAN

ID	Task Name
278	**M1: Develop an OPeNDAP server for common GIS data formats, e.g. GeoTIFF**
281	**M2: Develop GIS OPeNDAP Client**
285	**Data Transport Maintenance**
286	**Data Model Maintenance**
287	**Middleware Maintenance**
288	***DMAC - Initial Operations***



DATA ARCHIVE AND ACCESS ACTIVITIES AND SCHEDULE

(see Figure 6)

1. Activity: Current archive and access assessment

- **Description:** A comprehensive assessment of the current archive holdings and access methods is needed. A tabulation of data set name, content (temporal and spatial coverage), variables, format, storage location (online, offline, hardcopy), volume, resident center, and available access methods will form the starting benchmark for the IOOS. It will also uncover gaps in either the archive or access that need to be addressed and will point to archiving efforts where center-to-center collaborations would be beneficial.
- **Milestone 1:** Publish a current archive and access assessment report covering all U.S. centers holding IOOS-relevant data sets.
  - **Task 1:** Participate in metadata working group to define core metadata standards. Ensure that core development team has archive and scientific representatives from all IOOS data disciplines.
  - **Task 2:** Establish the set of data set descriptive parameters to be standard in the Archive System.
    - Description: To be effective, the set of descriptive parameters must be uniform across all participating data centers. There should be a two-way information exchange with the DMAC metadata development effort during this process. There are possibilities that this work could form the basis for the IOOS Discovery and Documentation metadata.
  - **Task 3:** Merge, tabulate, and evaluate the current status for all data sets and publish the findings.
  - **Category:** Tasks are committee work.
  - **Deliverables:** A published report on the starting benchmark for IOOS data archiving and access.
- **Estimated Resources:** Funding to cover costs for two or three meetings of eight to 10 people and publication.
- **Schedule:** During Year 1.
- **Sequencing:** Done prior to allocating or mapping of new data streams onto the existing set of data centers.
- **Partnerships:** Includes all U.S. IOOS data centers and is placed in the context of parallelisms, overlaps, and collaborations with GOOS as they apply.

2. Activity: Determine data set priorities for all IOOS data disciplines

- **Description:** The available and forthcoming data sets need to be ranked according to IOOS users' needs. Data set ranking is determined in conference with IOOS scientific representation, the DMAC User Outreach representatives, archive experts, and with reference to documentation on the most important variables as determined at the IOOS workshops. These lists set the priority for development efforts in the Archive System, and will lead to projects for improving archiving practices and access to both real-time and historical data sets.
- **Milestone 1:** Develop criteria for ranking data sets. Establish separate priority lists for each IOOS data discipline.
  - **Task 1:** Participate in the Transport "Data Population" group to identify and set priorities for ensuring accessibility of specific IOOS data sets.
  - **Task 2:** With an expert team and scientific representation prioritize extant archives according to IOOS needs.
  - **Category:** Tasks are committee work
  - **Deliverables:** A priority list for IOOS data sets.
- **Milestone 2:** Map the unfulfilled IOOS archiving needs onto the set of participating centers in the Archive System (see also the activity, Recruit Centers for the IOOS Archive System).
  - **Task 1:** Through a working group develop a plan to ensure all unarchived IOOS critical data become part of the Archive System.
  - **Deliverables:** A plan that maps IOOS data onto the Archive System.
- **Milestone 3:** List the products that are unavailable, but could be developed.
  - **Task 1:** Form a ranked list of data sets that could be developed. During this process a two-way sharing of information with the User Outreach component of the DMAC is required.
  - **Task 2:** In conjunction with the mapping exercise, assess the infrastructure capabilities at the centers that are to absorb additional archiving work. Document new infrastructure needs of IOOS support.
  - **Category:** Tasks are committee work.
  - **Deliverables:** A list of required products and estimated costs (by organization).
- **Estimated Resources:** Funding to cover costs for meetings.
- **Schedule:** Beginning 4th quarter during Year 1.
- **Sequencing:** Done immediately after (or possibly overlapping) the activity Current archive and access assessment.
- **Partnerships:** Inside IOOS (Archiving and Access (A&A) representatives, and Facilities Outreach members); Outside IOOS (scientific representation).

3. Activity: Determine IOOS data set categorization

- **Description:** Categorize IOOS data sets as irreplaceable, replaceable, perishable, and virtual. The data set storage strategy and retention period are determined by the categorization. In addition, adherence to Federal regulations at some centers in the Archive System is mandatory.
- **Milestone 1:** Assign data set category to all IOOS data sets.
  - **Task 1:** Convene a working group with experience in data archiving, metadata development, user outreach, and Federal regulations.
  - **Task 2:** Establish a process to review the categories set out in this document and that results in category assignments for all IOOS data sets.
  - **Task 3:** Assure irreplaceable data security.
    - It is planned that all irreplaceable data have two copies stored at separate locations and under independent data management. Irreplaceable data preserved below this standard are to be clearly documented, and archive centers need to immediately seek collaborations and support that resolve any deficiency.
  - **Category:** Tasks are committee work.
- **Deliverables:** A published report that categorizes all IOOS data sets.
- **Estimated Resources:** Funding to cover costs for meetings and publication.
- **Schedule:** Beginning 4th quarter in year one.
- **Sequencing:** Done immediately after the activity: Current archive and access assessment and possibly in conjunction with the activity. Establish data set priorities for all IOOS disciplines.
- **Partnerships:** Inside IOOS (A&A expert team, User Outreach Team, IOOS scientific representation, Metadata and Data Discovery Team)

4. Activity: Recruit centers for the IOOS Archive System and form partnerships

- **Description:** Effectiveness of IOOS will be achieved only by broad participation of the U.S. centers in the Archive System. Recruitment strategies need to be developed. Integration and cooperation with international programs are also critical. Global sharing of data will yield the maximum benefit to all programs, so the international contacts must be identified and actively engaged. IOOS must also be sensitive to the commercial 'value-added' data providers. This group will have objectives that overlap IOOS goals. IOOS should nurture partnerships based on open understanding and collaborations with this business sector.
- **Milestone 1:** Establish a set of guidelines for IOOS Archive System centers.
  - **Task 1:** Bring together the relevant IOOS governance and data policies into a set of guidelines so that interested centers can quickly know:
    - What is required to become part of the IOOS Archive System.
    - What are the benefits, for example, data sharing, backup, and archive.
    - What funding potential might exist for their centers.

The IOOS data policy document may not be in final form at this time, but a draft version could be used during the initial organization work.

- **Category:** DMAC policy (governance)
- **Deliverables:** Guideline document for Archive System centers
- **Estimated Resources:** Funding to cover meeting costs, publications, and potentially travel for a facilities management outreach liaison.
- **Schedule:** Beginning 3rd quarter Year 1 and onward
- **Sequencing:** Done after implementation is approved and data policies are in draft form.
- **Partnerships:** Inside IOOS (led by Facilities Management Outreach Team with assistance from A&A expert team, User Outreach Team, and IOOS governance)

- **Milestone 2:** Build international partnerships.
  - **Task 1:** Establish and devise a way to maintain a list of international centers and programs that could be collaborating partners for IOOS.
  - **Task 2:** Identify contacts, share understanding, and promote cross program partnerships and support.
  - **Category:** Planning and outreach
  - **Deliverables:** Document identifying relevant international partnerships and contacts
  - **Estimated Resources:** Funding to cover meeting costs, publications, and potentially travel for outreach liaison to international meetings and program offices
  - **Schedule:** Year 2 and onward
  - **Sequencing:** Done in parallel with IOOS developments following initial U.S. organization efforts.
  - **Partnerships:** To be defined
- **Milestone 3:** Evaluate and plan for commercial overlaps.
  - **Task 1:** Survey the commercial data business and identify overlaps with the IOOS goals. Assess and suggest way to integrate the business efforts with IOOS development so as to best serve the public needs.
  - **Category:** Planning and outreach
  - **Deliverables:** Document identifying relevant commercial interest overlaps.
  - **Estimated Resources:** Funding to cover meeting costs and publications.
  - **Schedule:** Year 2 and onward
  - **Sequencing:** Done in parallel with IOOS developments following initial U.S. organization efforts.
  - **Partnerships:** Inside IOOS (IOOS governance and access experts); Outside IOOS (commercial interests)

5. Activity: Develop archive critical metadata

- **Description:** To aid and ensure systematic (human and machine) access and data management across IOOS some archive specific metadata are critical. As a limited and brief example, some important elements are:
  - Unique data set identification code;
  - Expiration date;
  - Data set lineage and version history;
  - Points of access and available access methods;
  - Data set citation and references;
  - Data set latency specification.
- **Milestone 1:** Develop a comprehensive list of archive critical metadata and organize a plan that will lead to a system-wide metadata standard that is easy to implement and maintain.
  - **Task 1:** Convene a working group with experienced representation for data archiving and metadata development to prepare the list.
  - **Task 2:** Interact with the DMAC Metadata and Discovery Data and Data Transport development teams and working groups to ensure archive needs are accommodated.
  - **Category:** Tasks are committee work.
  - **Deliverables:** Archive specific metadata will appear in IOOS metadata standards.
  - **Estimated Resources:** Funding to cover committee work and collaboration with the DMAC Metadata and Discovery Data team.
  - **Schedule:** Beginning in Year 1 and continuing.
  - **Sequencing:** Done prior to completion of IOOS metadata standards work.
  - **Partnerships:** Inside IOOS (A&A expert team, and Metadata and Data Discovery Team, Data Transport Team).

6. Activity: Define IOOS archive and access data policy

- **Description:** The policies for contributing data and using data from IOOS need to be formally documented. A few key policy issues from the archive and access perspective are:
  - Full and open data sharing as per IOC and WMO policy and at no cost or minimum cost for reproduction. The conditions and authoritative protocol for allowing restricted access need to be discussed and defined, if required.
  - Data collected or prepared with IOOS funding must be placed in the Archive System. If possible, actions to be taken for non-compliance should be articulated.
  - Agreement that the four data categorizations are suitable to determine IOOS data preservation requirements.
  - Full cooperation with GOOS.

- **Milestone 1:** Develop a draft IOOS data policy.
  - **Task 1:** Form a committee with experienced representation for data archiving and IOOS management to draft the policy.
- **Milestone 2:** Receive community comment on the draft policy.
  - **Task 1:** Circulate the policy to interested parties and make it widely known that IOOS has a data policy.
- **Milestone 3:** Create a final draft.
  - **Task 1:** Resolve problems raised by the IOOS user community and ensure the data policies can be applied satisfactorily within the limits of standing Federal regulations.
- **Category:** Milestones and all Tasks are committee and IOOS management work
- **Deliverables:** IOOS data policy
- **Estimated Resources:** Funding to cover committee work.
- **Schedule:** Year 1
- **Sequencing:** Done before or at the same time as the activities of Develop archive critical metadata and Current archive and access assessment. A draft of the data policies is necessary for the data center recruitment and partnership requirement.
- **Partnerships:** Inside IOOS (A&A expert team, Metadata and Data Discovery Team, and IOOS management); Outside IOOS (done with consultation to GOOS).

7. Activity: Establish IOOS data stream developers guidelines

- **Description:** A document defining the IOOS data stream guidelines must be available for data providers. It will include at least:
  - IOOS archive and access data policy;
  - IOOS metadata and data discovery standards and recommended ways to easily develop the metadata;
  - IOOS recommended formats;
  - IOOS data transport standards and recommended ways to implement them and get support.
- **Milestone 1:** Publish a guideline document that is updated as progress and evolution in the DMAC take place.
  - **Category:** Milestone is committee and IOOS management work.
  - **Deliverables:** IOOS data stream developers' guidelines
- **Category:** Archive committee work
- **Estimated Resources:** Funding to cover committee work
- **Schedule:** Year 2
- **Sequencing:** Done after the data policy, data transport, and metadata and discovery data standards are in beta release form.
- **Partnerships:** Inside IOOS (A&A expert team, Metadata and Data Discovery Team, and IOOS management).

8. Activity: Develop Archive System data discovery interfaces

- **Description:** The IOOS Archive System will be distributed. Data discovery (both machine and human) must work across all centers, all data sets, and all methods for access.
- **Milestone 1:** Define the human data discovery interface
  - **Task 1:** Have archive component representatives work with the Metadata and Data Discovery component during the interface development. Some necessary interface features are:
    - It is dynamic, formed based upon user query;
    - It shows which data centers hold the data;
    - It shows the data set titles and unique IOOS data set identification;
    - It shows available core services (OPeNDAP, HTTP, FTP);
    - It shows available extended services (subsetting, aggregation, OpenGIS Map, online analysis, online ordering, etc.);
    - It provides users with links to all services that are available.
  - **Category:** R&D
  - **Task 2:** Dynamically harvest metadata from the Archive System and build IOOS metadata databases to serve the discovery interfaces.
  - **Category:** R&D
  - **Task 3-n:** Other tasks to be defined as standards and methods are developed by the Metadata and Data Discovery and Data Transport components. Following R&D, pilot projects will be necessary.
- **Milestone 2:** Define the machine data discovery interface.
  - **Task 1:** The machine, or application, interface will be specified by collaborations between the Data Transport, and Metadata and Data Discovery components. They are not described here.
- **Deliverables:** Data discovery interfaces for the Archive System.
- **Estimated Resources:** To be determined by other experts.
- **Schedule:** Beginning 1st quarter Year 2 and continuing.
- **Sequencing:** Done following the development of metadata standards and data transport methods.
- **Partnerships:** Inside IOOS (all components); Outside IOOS (other organizations that are attempting to do the same thing).

9. Activity: Receive and provide more data in real time

- **Description:** To meet the IOOS goals the Archive System must receive and provide more data to users in real time. Many IOOS goals have time critical schedules requiring prompt access to observed data and data products. Note: The provision of real-time data will come primarily from the modeling centers, regional centers, and data assembly centers in the Archive System. More limited real-time access will be the norm at the archive centers.

- **Milestone 1:** Put in place pilot projects and pre-operational real-time data systems based on DMAC data transport mechanism and existing data delivery infrastructure.
  - **Task 1:** Based on the activities of "current archive and access assessment" and "determine data set priorities for all IOOS disciplines," select several real-time data streams that are most important.
  - **Task 2:** Design pilot projects within the Archive System that must include the components for real-time data receipt and immediate public access.
  - **Task 3:** Depending on the advances in DMAC Data Transport, Metadata and Data Discovery, and Archive System collaborations, the following could also be components within the projects:
    - Data receipt and delivery through DMAC data transport methods;
    - Metadata records and catalogs in the DMAC standard;
    - Archive System data transfer to an archive center if initially received at a regional, assembly, or modeling center;
    - Value added data development (QC checks, and data merging, provision for server side subsetting), and product development (data analysis, maps, and data formatted for GIS ingest).
    - Archive System backup at a second archive center for irreplaceable data
  - **Category:** Tasks 1-3 are pilot study leading to pre-operational systems
  - **Deliverables:** Real-time test systems
  - **Estimated Resources:** Staff and facilities infrastructure commensurate with data source volume, complexity, and access service.
  - **Schedule:** During Year 2.
  - **Sequencing:** Done following the activities of "current archive and access assessment" and "establish data set priorities for all IOOS disciplines."
  - **Partnerships:** Internal to the IOOS DMAC and probably in collaboration with developments in IOOS measurement component.
  - **Task 4:** Pilot project to serve near-real-time GTSPP data.

    As a pilot project, near real-time profile data, which are harvested from the Global Telecommunications System (GTS) and delivered three times per week from the Marine Environmental Data Service (MEDS) of Canada to the U.S. National Oceanographic Data Center (NODC) as part of the Global Temperature Salinity Profile Program (GTSPP), will be posted by NODC on a DODS server (a precursor to the IOOS data transport protocol) for use by operational oceanographic data customers. For each data set delivered, NODC will create metadata, archive the data set according to its data category requirement, subject the data set to preliminary quality control, and then post the data set on a DODS server. After these data have been available via the DODS server for several months, customer feedback will be evaluated to determine subsequent improvements in this data system and whether the metadata are adequate for effective use of the data.

This pilot system will provide practical information for the DMAC developments in data transport, metadata, and data discovery. It would also be well positioned to transition to the IOOS standards as they become available.

- **Category:** Pilot
- **Deliverables: (**1) Near-real-time profile data available on a DODS server, (2) compilation report of customer feedback, and (3) data system improvement plan.
- **Estimated Resources:** Since this is an extension of an existing base funded project at NODC, no additional resources are requested for this pilot project. Transition to IOOS functionality may require additional support.
- **Schedule: (**1) March 2003, (2) August 2003, and (3) October 2003.
- **Sequencing:** Independent of other tasks.
- **Partnerships:** Internal to the IOOS DMAC, and in collaboration with MEDS and near real-time profile data customers such as the Naval Oceanographic Office.

10. Activity: Establish a protocol to report and resolve data and data flow problems

- **Description:** Inevitably, there will be problems with the data flows and data set integrity. These problems will have wide impact. Irregularities and changes will affect the data providers, the Archive System, the metadata, and most importantly, the data users.
- **Milestone 1:** Establish a protocol for reporting and resolving problems.
  - **Task 1:** Establish a method to post or broadcast problems to users as soon as possible after they are discovered. Use the same strategy to publicize when corrections to the real-time data stream are completed.
  - **Task 2:** Establish a method to publicize, in delayed mode, the analysis of problems and identify the impacted data in the Archive System.
  - **Task 3:** Establish a method to publicize, in delayed mode, substantive corrections that have been applied to the archive data.
  - **Category:** Committee work
- **Milestone 2:** Determine effective ways to solve chronic problems.
  - **Task 1:** Investigate if IOOS management through Facilities Management Outreach can assist in resolving chronic problems that are not addressed in a reasonable time period.
  - **Category:** Committee work
- **Milestone 3:** Implement a DMAC system for data problems.
  - **Task 1:** Pilot study leading to pre-operational then operational reporting system.

- **Deliverables:** A system and protocol through which problems can be reported, resolutions can be sought, and data providers and users can be informed.
- **Estimated Resources:** Staff and facilities infrastructure commensurate with the pilot project and services provided during the pre-operational and operational phases.
- **Schedule:** Beginning 4th quarter of Year 1.
- **Sequencing:** Done after the IOOS Archive System activity of Current archive and access assessment and when the IOOS data streams are coming online in the DMAC.
- **Partnerships:** Internal to the DMAC, and in collaboration with IOOS management and Facilities Management Outreach.

11. Activity: Broaden the base for user services

- **Description:** The IOOS is required to provide new services to many new users. In particular, decision- and policy-makers need rapid access to suitable ocean information. Plans are needed to determine which services are missing and most critical, and how to provide those services from the Archive System.
- **Milestone 1:** Accommodate new user services in the Archive System.
  - **Task 1:** Participate in the plans developed by the User Outreach component and determine how to improve extant and expand user services.
  - **Category:** Committee work (with IOOS User Outreach), R&D, and pilot.
- **Deliverables:** New services to meet the needs for the broad IOOS user base.
- **Estimated Resources:** Commensurate with the data products and systems needed for development.
- **Schedule:** Beginning 4th quarter in Year 2.
- **Sequencing:** Done following the gathering of advice from the User Outreach component and after completion of the activities, “current archive and access assessment” and “establish data set priorities for all IOOS disciplines.”
- **Partnerships:** Inside IOOS (Archive and Access, User Outreach).

12. Activity: Verify data security requirement for irreplaceable data sets

- **Description:** It is intended that irreplaceable data be stored in two separate locations, preferably under independent data management.
- **Milestone 1:** Run live tests to verify the security of irreplaceable IOOS data.
  - **Task 1:** Devise a DMAC pilot project to randomly check primary and secondary archive copies between centers for data in the irreplaceable category.
  - Category: Pilot project
  - **Task 2:** Execute data and inventory checks, and cross system archive file recovery.
  - **Category:** Pilot project

- **Milestone 2:** Establish regular verification procedures.
  - **Task 1:** Document and form cross-agency agreements to perform archive checks.
  - **Category:** Pre-operational and operational
  - **Deliverables:** Systematic methods to verify irreplaceable data security.
  - **Estimated Resources:** Commensurate with the pilot projects proposed.
  - **Schedule:** During Year 3.
  - **Sequencing:** Done following the activity to Determine IOOS data set categorization, and after data transport capabilities are available.
  - **Partnerships:** Inside IOOS (A&A expert team, IOOS management, and Facilities Outreach)

13. Activity: Establish procedures to document the Archive System metrics

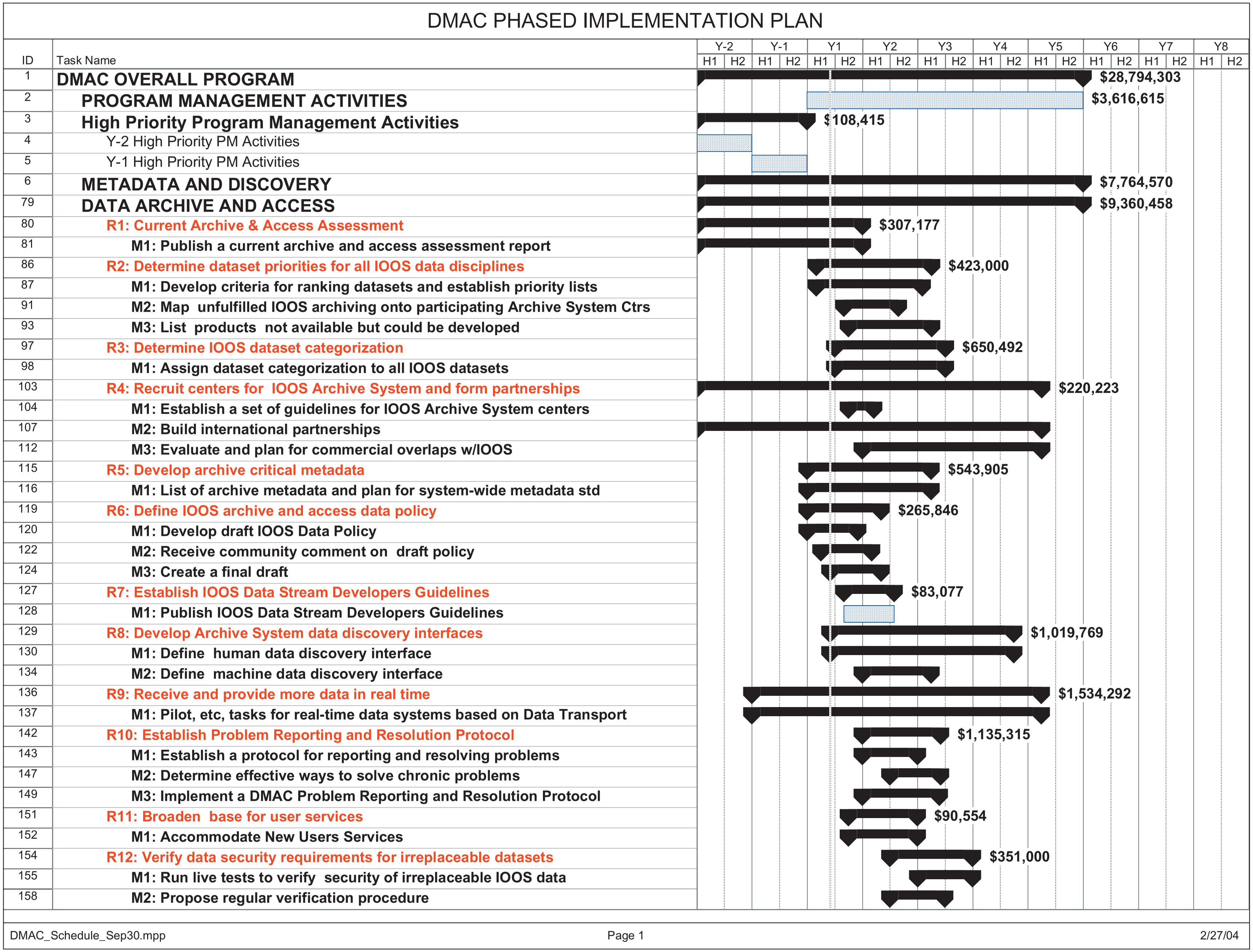
- **Description:** From the onset, measurement metrics for the Archive System are required. Tracking of incoming and outgoing data must be thorough, complete, and measured on equivalent scales system wide.
- **Milestone 1:** Compile a test annual report for the DMAC Archive System.
  - **Task 1:** As a pilot project, collect annual metric data from all Archive System centers. Work to develop an annual report. Through the process identify problems and needs to achieve accurate reporting. Use these results to refine the metric requirements for the system. General elements for consideration, review and refinement are:
    - media receipt and delivery;
    - network receipt and delivery;
    - DMAC data transport receipt and delivery;
    - error discovery and correction.
  - **Category:** Pilot project
- **Milestone 2:** Include the Archive System metric requirements in the guideline document for IOOS data stream developers.
  - **Category:** Committee work
- **Deliverables:** DMAC system wide data movement metric report.
- **Estimated Resources:** Funding for meetings, and pilot project work.
- **Schedule:** Beginning 2nd quarter in Year 2.
- **Sequencing:** Done after the data transport mechanism is established, data sets are categorized, and suitable metadata are available.
- **Partnerships:** Internal to IOOS

14. Activity: Improved efficiency for archive growth and access

- **Description:** A concerted effort to implement DMAC Data Transport methods and Metadata and Data Discovery standards is required and will result in improved efficiency within the Archive System.
- **Milestone 1:** Implement DMAC Data Transport receipt methods in parallel with extant methods.
  - **Task 1:** Pilot and Pre-operational, critically compare data received for six months to a year. Assure all aspects of data integrity, track system performance, and problems.
- **Milestone 2:** Implement DMAC Data Transport and Data Discovery in parallel with extant delivery systems.
  - **Task 1:** Pilot and Pre-operational, offer data access through extant interfaces and procedures, and through DMAC Data Discovery interfaces and Data Transport protocols. Track all user metrics, problems, and success. The pilot and pre-operational systems should run for at least one year each.
- **Category:** Pilot and pre-operational
- **Deliverables:** Archive System tests of DMAC functionality.
- **Estimated Resources:** Commensurate with Pilot and Pre-operational plans.
- **Sequencing:** Following beta tests on DMAC Transport and development and testing of Metadata and Data Discovery standards. Probably Year 2 and onward.

15. Activity: Procedure to resolve data retention issues

- **Description:** Data in archive systems are commonly resubmitted and replaced. The number of old versions of data to be preserved remains an open question. Managers of data centers need a formal procedure to help them resolve this difficult issue.
- **Milestone 1:** Reach a consensus agreement on how to resolve data retention issues and questions for data sets with multiple versions.
  - **Task 1:** Form a diverse expert team with representatives from archive management and the scientific user community.
    - Discuss and draft a set of procedures on data retention that possibly include:
      - An annual scientific review panel to consider data retention questions;
      - Input from IOOS management;
      - An adequately long public review period before any action is taken.
  - **Category:** Committee work (with IOOS policy implications)
  - **Task 2:** Develop schedule for removal of data sets identified for removal.
    - This should require approval from science, governances and parent agency.
    - This affects Data Discovery and Data Transport as they will need to make alterations to remove reference and support for that data.



DMAC PHASED IMPLEMENTATION PLAN

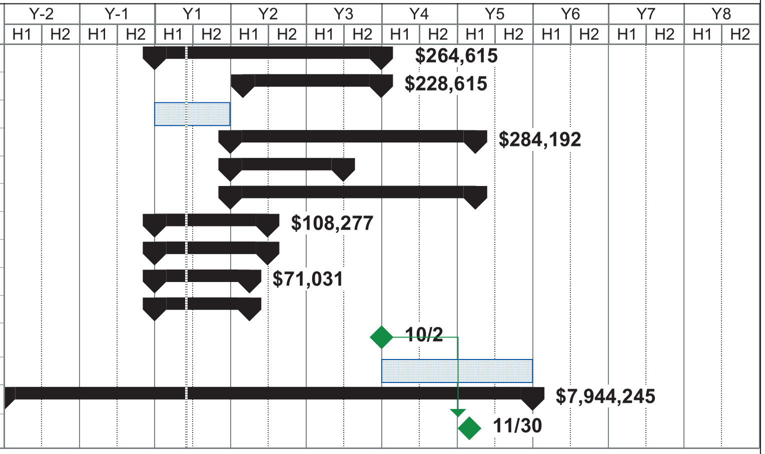
ID	Task Name	
1	DMAC OVERALL PROGRAM	$28,794,303
2	PROGRAM MANAGEMENT ACTIVITIES	$3,616,615
3	High Priority Program Management Activities	$108,415
4	Y-2 High Priority PM Activities	
5	Y-1 High Priority PM Activities	
6	METADATA AND DISCOVERY	$7,764,570
79	DATA ARCHIVE AND ACCESS	$9,360,458
80	R1: Current Archive & Access Assessment	$307,177
81	M1: Publish a current archive and access assessment report	
86	R2: Determine dataset priorities for all IOOS data disciplines	$423,000
87	M1: Develop criteria for ranking datasets and establish priority lists	
91	M2: Map unfulfilled IOOS archiving onto participating Archive System Ctrs	
93	M3: List products not available but could be developed	
97	R3: Determine IOOS dataset categorization	$650,492
98	M1: Assign dataset categorization to all IOOS datasets	
103	R4: Recruit centers for IOOS Archive System and form partnerships	$220,223
104	M1: Establish a set of guidelines for IOOS Archive System centers	
107	M2: Build international partnerships	
112	M3: Evaluate and plan for commercial overlaps w/IOOS	
115	R5: Develop archive critical metadata	$543,905
116	M1: List of archive metadata and plan for system-wide metadata std	
119	R6: Define IOOS archive and access data policy	$265,846
120	M1: Develop draft IOOS Data Policy	
122	M2: Receive community comment on draft policy	
124	M3: Create a final draft	
127	R7: Establish IOOS Data Stream Developers Guidelines	$83,077
128	M1: Publish IOOS Data Stream Developers Guidelines	
129	R8: Develop Archive System data discovery interfaces	$1,019,769
130	M1: Define human data discovery interface	
134	M2: Define machine data discovery interface	
136	R9: Receive and provide more data in real time	$1,534,292
137	M1: Pilot, etc, tasks for real-time data systems based on Data Transport	
142	R10: Establish Problem Reporting and Resolution Protocol	$1,135,315
143	M1: Establish a protocol for reporting and resolving problems	
147	M2: Determine effective ways to solve chronic problems	
149	M3: Implement a DMAC Problem Reporting and Resolution Protocol	
151	R11: Broaden base for user services	$90,554
152	M1: Accommodate New Users Services	
154	R12: Verify data security requirements for irreplaceable datasets	$351,000
155	M1: Run live tests to verify security of irreplaceable IOOS data	
158	M2: Propose regular verification procedure	

DMAC_Schedule_Sep30.mpp Page 1 2/27/04

Figure 6. Data Archive and Access Gantt Char

DMAC PHASED IMPLEMENTATION PLAN

ID	Task Name
160	R13: Establish procedures to document the Archive System metrics
161	M1: Compile a test annual report for the DMAC Archive System
163	M2: Write Archive System Metric requirement for IOOS data developer guidelines
164	R14: Improved efficiency, archive growth, and access
165	M1: Implement DMAC Transport receipt methods
167	M2: Implement DMAC Data Discovery and Transport delivery methods
169	R15: Procedure to resolve data retention issues
170	M1: Consensus resolving data retention issues
173	R16: Write Plan for ARCHIVE&ACCESS Security
174	M1: Develop Archive & Access Security Plan
177	*Archive & Access - Initial Operations*
178	R17: Initial Archive Operations
179	DATA TRANSPORT
288	*DMAC - Initial Operations*



DMAC_Schedule_Sep30.mpp Page 2 2/27/04

- **Deliverables:** Procedures to address data retention questions
- **Estimated Resources:** Funding for meetings.
- **Schedule:** During Year 1 or 2.
- **Sequencing:** Done after IOOS data archive and access policies have been established.
- **Partnerships:** Internal to IOOS, with scientific representative input.

16. Activity: Write plan for Archive and Access Security

- **Description:** The Archive System will publicly expose data and systems. To protect the data suppliers and the systems a security plan is required.
- **Milestone 1:** Develop an Archive System security plan.
- **Task 1:** Determine the level of security that is required for the DMAC and the data held for IOOS. Write a security plan.
- **Category:** Working Group activity
- **Deliverables:** Archive and access security plan
- **Sequencing:** In conjunction with the evolution and deployment of DMAC Data Transport and Metadata and Data Discovery standards in the Archive System.

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

Part III. Appendices

Appendix 1. Metadata Data Discovery

IOOS DMAC Metadata/Data Discovery Team

March 2005

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

Contents

Metadata

INTRODUCTION

Metadata is a critical component of IOOS. Metadata is information about data that captures the essential characteristics and history of a data set to ensure the data's usefulness over time. Metadata is most commonly thought of as a textual guide to understanding data. As such, metadata must describe data completely and must be written in a manner that is easy to understand. Within IOOS, metadata must be delivered along with data, and XML schema can be used as a transport "language."

The Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) defines metadata as the information required by a prospective user to determine (1) the availability of a set of geospatial data, (2) the fitness of a set of geospatial data for an intended use, (3) the means to access the data, and finally (4) the means to transfer the data successfully. In general, the role of metadata in the IOOS Data Management and Communications (DMAC) Subsystem is consistent with this standard. Specifically, metadata will provide the semantic content required to seamlessly connect all the components of the IOOS DMAC.

Data discovery, another integral facet of IOOS, will be accomplished through the use of metadata. Metadata is commonly indexed with keywords to provide a means to search for data that meets a user's needs. This use of metadata is comparable to the indexing of catalog records within libraries to help patrons locate items of interest. IOOS will develop a catalog system to help users locate data of interest. To do this will require that data providers not only write metadata that is comprehensible to a reader, but also write it to be used by software. Writing metadata for use in software requires that defined formats be followed.

Traditionally, metadata is used in data discovery to support searches through geospatial and temporal extents and parameter keywords. Metadata can also be used to provide all the information necessary to access and use the data. This kind of information can range from contact information so a user may call and order data, to a URL where a user can download a data set, or to information on how software can access and deliver subsets of data directly to a user. Metadata that contains information of this latter type can be used to develop very sophisticated and powerful systems that allow users to get direct access to data or portions of data sets that are needed. This type of metadata, frequently referred to as syntactic metadata, requires consistent use of fields and terminology.

Metadata used for data archival include versioning, lineage, and reference information. Versioning and lineage metadata are required to support modifications and corrections to data in archives. The metadata framework will also be used to maintain reference information for the archived data.

This information will include reference documentation, bibliographic references, and citation of the data. Potentially, the metadata framework should allow users of the archive to publish findings on the data.

For product generation, metadata will be used to document how the product was generated and what, if any, measured data were used as input to the process that generated the product. Metadata will also be used to enable access to data products in the same manner that metadata are used to enable access to measured data. Quality-control metadata will be important to determine the fitness of IOOS data for particular uses in generating products. For complex data sets, metadata can be used to represent the structure of the data collection, thereby enabling operations such as reformatting and sub-setting.

Metadata will be a key component of the data transport and assembly operations envisioned by the IOOS DMAC. The data transport component will support access to data from applications and enable transmission of data to assembly and archive centers.

Within IOOS, a requirement upon data providers must be to provide both semantic and syntactic metadata in a form that is useful to both readers and programmers. The IOOS data delivery system cannot work without quality metadata that provide information in a consistent and controlled manner. Although the goal of IOOS may be to provide automatic access to data, it may be necessary to implement this in a staged approach, particularly for historic data. The data delivery system would provide access to those data available on line with associated high-quality metadata. Eventually, IOOS will develop a catalog system that provides access to all data including those sets that are only available off line.

The metadata must be extensible within this system to allow for extensibility of the system as a whole. We know that for this system to work in the future and grow to a nationwide implementation, the full system, including metadata and all its capabilities, needs to be extensible. To facilitate access to distributed data sources, the metadata framework developed as part of the IOOS must comprise an extensible metadata schema reflecting the needs of the participating scientific disciplines both to provide and access science data for their particular applications. Different scientific communities participating in IOOS will undoubtedly have different requirements, but the metadata framework must support those differences to ensure it meets the needs of all participants. As such, the metadata framework should define a process by which participating science disciplines can extend the existing metadata schema to meet the needs of that community. A focus of that process must be to extend the existing schema to meet the needs of machine-to-machine interoperability with semantic meaning for that particular use.

Additionally, the metadata framework must comprise a metadata access and representation mechanism that supports programmatic access to metadata. To support machine-to-machine interoperability, distributed access to metadata must be as seamless as access to the data itself. To facilitate the use of distributed data sources, the metadata framework will provide transparent access to all the metadata fields, including those required to operate on the data in a semantically meaningful way. These include, but are not limited to, the units, a controlled set of geophysical parameters, horizontal and vertical datums, and others that allow remote applications to make use of the data. The ability to programmatically access metadata may have far-reaching implications in the evolution of observing systems such as IOOS. Coupled with a flexible, community-driven metadata framework and programmatic access, the metadata can provide the foundation to extend the capabilities of existing distributed systems in a number of unique and powerful ways.

METADATA STANDARDS

As mandated by an executive order, in the United States, each [Federal] agency shall document all new geospatial data it collects or produces, either directly or indirectly, using the standard under development by the Federal Geographic Data Committee (FGDC). The FGDC developed the Content Standard for Digital Geospatial Metadata (CSDGM) that provides a common set of names and definitions of compound and data elements used to document digital geospatial data. Also, under the CSDGM, individual data communities (Biological Data, Shoreline Data, etc.) have created supplemental standards for their various disciplines. Initially, IOOS will use the FGDC Content Standard (FGDC-STD-001-1998), and any of the applicable supplemental profiles (i.e., the Biological Data Profile, Shoreline Profile), as its standard for metadata. However, a review of the IOOS community (initially starting with the expert teams for this implementation plan and expanding to data providers and users) will be done at the earliest possible time in order to address the needs of the standard set for IOOS.

The International Organization for Standardization (ISO) has developed a standard for geospatial metadata. This standard (ISO 19115) was formally accepted in May of 2003. It is anticipated that the next version (Version 3) of the FGDC CSDGM will be a form of the international standard. Acceptance of the new version of the FGDC CSDGM is expected in 2003, and acceptance will mandate Federal Agency implementation. A gradual transition from the FGDC CSDGM version 2 to version 3 is expected, as well as a delay in conversion of existing metadata to the new standard. The greater metadata community (outside IOOS) is developing crosswalks between these metadata standards. IOOS will remain compliant with the FGDC standard and will make the current standard available to participants.

Another issue is that some users of the metadata may be libraries or other data services that use standards other than the FGDC CSDGM. MARC21, Dublin Core and DIF are a few such standards. These standards contain basic elements but some may lack adequate geospatial characteristics potentially critical to data discovery. However, as crosswalks mapping elements between the FGDC CSDGM and these other standards exist, elements from each of these standards can easily be considered in the IOOS metadata standard. Additional work in this area will be required to support use of these standards within IOOS.

A joint effort among the expert teams to determine information required for IOOS metadata records will be one of the initial tasks within the implementation plan. Included in the determination of these mandatory elements may be a phased approach that will allow data providers to incrementally add metadata as the level of interoperability of the data set increases.

This joint effort among the expert teams to determine information required for IOOS metadata records may show the need for elements not previously included in standard metadata formats. In the case of the FGDC CSDGM, these additional elements can be inserted into the standard format as "extended elements." Documentation for these extended elements must be developed and made available to all (data providers and users). The possibility also exists for the IOOS community to develop a Standard Profile under the FGDC Content Standard.

The final issue related to metadata standards is that of keywords and data dictionaries. Without the use of controlled keywords and data dictionaries, data discovery is difficult, if not impossible, and machine-to-machine interoperability with semantic meaning will not be possible.

BIOLOGICAL METADATA CONSIDERATIONS

One area where this can be prominently seen is the area of Marine Biological Data and Species data. In practice, internationally accepted species names are the keywords for information about organisms. Biological data systems require name translators that provide accurate scientific names from synonymous names and common names. With oversight from the Global Biodiversity Information Facility (GBIF), Catalogue of Life, and organizations such as the Integrated Taxonomic Information System (ITIS), Species 2000, and OBIS, the taxonomic authority for each major group of organisms maintains the accepted list of species. Fragmentary DNA or RNA sequence data on components of genomes are linked using accepted species names. Sequence information on specific enzyme molecules (such as cytochrome oxidase I) shows promise as a Bar Code of Life for unequivocal identification of species. The common usage of accurate species names will also be facilitated by expert systems for identifications using morphological characters.

Taxonomic names and descriptions are the products of individual scientists whose careers have been devoted to describing and understanding relationships among species. Increasingly, these individuals and their colleagues take advantage of DNA or RNA gene sequence data to differentiate among species and to trace their phylogeography. Species are the units that survive through evolutionary time and each species is the unique product of its evolutionary history. Specimens of each species are stored in museums for future reference and some may be maintained in culture collections. Species are classified according to their evolutionary relationships using a well-established hierarchical system of nomenclature. New species are continually being described and the hierarchical tree of evolutionary relationships among species, and the associated hierarchical nomenclature, must continually be revised to incorporate new information. For this reason, biological data systems, unlike physical data systems, require much more attention to metadata. As a minimum quality control and quality assurance measure, the taxonomic authority and the person identifying the species should be included with each record and each revised data set.

FUTURE CONSIDERATIONS

Two of the most promising methods for translation among multiple controlled vocabularies lie with the use of thesauri and ontologies through the semantic web. In well-structured thesauri with robust input capabilities, one would be able to load multiple controlled vocabularies. The users could subsequently query one thesaurus for maximum understanding of the terminology. The use of thesauri should be looked at immediately within the IOOS system.

The semantic web is "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"[1]. This is accomplished using ontologies, which are defined as: The hierarchical structuring of knowledge about things by sub-categorizing them according to their essential (or at least relevant and/or cognitive) qualities"[2]. The main purpose of an ontology is to enable communication between computer systems in a way that is independent of the individual system technologies, information architectures, and application domain. For example, the Global Change Master Directory's (GCMD) Earth science keywords are only one example of a controlled Earth science vocabulary. Other vocabularies exist, and there is a need to investigate commonality among multiple controlled vocabularies. Ongoing research and implementation of elements of the semantic web could reveal methodologies to translate among multiple ontologies and allow the user to search among multiple controlled keywords and thesauri. Further study will be required in the areas of the semantic web and ontologies.

[1] Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001

[2] http://www.dictionary.com; The Free On-line Dictionary of Computing, © 1993-2001 Denis Howe; (1997-04-09)

DEVELOPMENT AND MAINTENANCE OF METADATA

One of the more difficult tasks for IOOS is gaining acceptance and compliance with the requirement to provide and maintain quality metadata. Learning to write metadata is no different than learning any other skill. Most skills require a lot of time and effort initially but become easier and less time-consuming with practice. The job of the IOOS system will be to provide a means for the generation and maintenance of metadata that will not unduly burden the data provider, but will provide for the quality of metadata that is desired within IOOS. To accomplish this, IOOS will select or develop a master metadata management system. This system will allow data providers the flexibility to manage their metadata within a local system or through a centralized system via remote access capabilities, and will not require the data provider to duplicate existing metadata and maintain it in two or more systems. For instance, IOOS will access existing FGDC nodes (metadata servers) and harvest or point to specific data of interest to IOOS. This will also ensure that IOOS will fit into larger projects such as the National Spatial Data Infrastructure (NSDI) and international data projects.

IOOS will make available to data providers an easy means to generate, validate, and maintain their metadata. Support will be provided for parent/child metadata, the validation and approval process put in place by IOOS, and the maintenance of metadata. Data providers will come to the table with different levels of expertise in this area, and therefore the system must be flexible enough to handle what the data providers require. This may include a "common repository" for metadata and shared toolsets for those data providers that do not have the resources to manage their metadata easily, but should also allow for the data provider to manage metadata in the way they have done it in the past.

Quality metadata can only be generated by someone who understands the data that are being documented, and therefore it is required that the metadata be generated and maintained as close to the collection and/or generation of that data as possible. Training opportunities, support networks, and tools will be made available to help the beginning and advanced metadata writers. One of the first tasks within this implementation plan is the generation of a user guide for IOOS metadata. This user guide will discuss issues such as the granularity of metadata, which should be a part of the system, parent/child metadata, the validation and approval process, duplicate metadata, maintenance requirements, etc.

Although the system will be built to minimize additional work by the data provider, it cannot be stressed enough that the data provider will have to provide high-quality metadata in order for IOOS to succeed. IOOS will do its part to encourage data providers to create metadata and keep it current by providing tools, consulting services, and help desk support.

ADAPTABILITY OF METADATA

The ability to adapt these metadata so that what is delivered is not a "generic record" but is appropriate to the specific data delivered must also be considered within this system. Situations where this becomes important includes subsetting data, aggregating data, and merging data or creating products from raw data. Each of these will be considered separately. What should be consistent in any situation where metadata are "adapted" is that the new metadata record should be tagged to show that it is not the original record used to discover the data but has been modified to be appropriate for the data delivered within the system.

Subsetting Data

A single metadata record will often point to a collection of data. One example of this is when data are collected in regular intervals over time. The time information within the metadata record is shown as a beginning date/time, which is specific, and an ending date/time, which is designated as "present," showing that the data continue to be collected. When data are then delivered using the transport system, the date/time information should be modified to show the time frame of the data delivered, and not the original metadata record, which shows what data are available.

The capability to subset a large collection of data within the data transport system also makes it a requirement that the metadata be adapted to show what data are delivered. Sub-setting can be done in the spatial or temporal domains, and will also be allowed in the attribute section by allowing for the delivery of only those parameters that have been requested.

Aggregating Data

Data aggregation can be associated with a single data provider or across data providers. When the same type of data from the same provider are merged into a single data set, we can be look at these data as having the same parent metadata record (i.e., data buoys from a single source). If a single parent metadata record applies to all data that are being aggregated, then adapting a metadata record is feasible. How this metadata aggregation should be implemented will require further study.

The second option is the aggregation of data from different sources that do not or cannot share a single parent metadata record (i.e., observational data from different sources/systems). The job of aggregation in this case becomes much more difficult and should be studied as to what, if any, aggregation is appropriate. If aggregation is not appropriate, there is still the issue of how to distinguish the appropriate metadata record for each data item delivered, which will also require further study.

Products and Merging Data

Adaptation of metadata is also an issue for products and other processes that merge data. When considering metadata associated with data products, the issues include building a new and unique metadata record associated with that specific data product, and then whether the metadata associated with the data that were used as input for the product should be delivered to the user also. The metadata record associated with the data product should be generated within the same system that generated the product, and should be a unique record associated with that product. Product metadata should take into account all the considerations of any metadata record, along with the additional consideration of associating the product to the measured data (and its associated metadata) that was used to build the product. Further study in this area will be done to develop a policy on what specific metadata should and will be delivered with products that are generated from measured data.

ADDITIONAL ISSUES

Tracking Metadata Maintenance

The issue of whether metadata maintenance should be tracked is one that needs more study within the IOOS system. In many database systems that require accountability and recoverability, data are never overwritten, but a modification is added. When a query is done, the latest modification is used to generate the results. This type of system would allow IOOS to more easily track one kind of change to the metadata. Mistakes might be more easily caught and a history would be kept. If this is considered a requirement of the system, it would initially only be imposed at the centralized metadata management system.

Data Quality Metadata

The metadata associated with data quality will need to be documented carefully in order for users to understand the appropriate uses, precision, and accuracy of the data. Precision and accuracy are not only important in the measurement taken at a particular site, but also in the determination of the location of the measuring site.

In addition, the lineage of the data provides critical information on what changes have been made through time, such as measurements that have been eliminated or corrected, filtering, and correction for instrument response. Information describing factors that might affect measurements, such as atmospheric conditions and calibration history of instruments, should be included where appropriate.

Another issue associated with data quality is the ability to modify a metadata record when a quality assessment has been completed to show the information obtained within that assessment. This is an immediate requirement of the IOOS system, and further study must be done on how this will be implemented and controlled within the system.

Several of the FGDC metadata sections contain data quality information. These sections of the metadata record need to be studied further in coordination with the Applications Team to determine whether all the data-quality issues can be resolved within the existing metadata structure or whether additional elements will be required to capture all the quality information desired within the system.

Completeness of Metadata

Data providers need to look at their data with fresh eyes and try to imagine what a user might need to know. Information that is obvious to the person who collected or processed the data may not be obvious to the potential user. It is important that writers step back from their work and try to view it with different eyes. Having a colleague who is not familiar with the data may help in the metadata review. In addition, by providing a metadata management system that is easy to use, allowing parent/child metadata, providing training and consulting services, and a means for user feedback, IOOS can minimize the burden of generating quality metadata.

Maintenance of Metadata

Metadata need to be reviewed regularly to determine if updates are needed. The need for review is obvious under a number of circumstances. New processing steps or changes in the data collection methodology need to be reflected in the metadata. Information about key contact personnel may need updating as addresses, phone numbers, and email addresses change, or as people leave or join an organization. IOOS data providers must develop a review cycle for their metadata, and the metadata management system provided must easily accommodate this review process.

Archive

IOOS will encourage data providers to archive data at an approved national data archive center. A data provider may choose to archive data for a number of reasons. One is to provide a backup for data at risk at the data provider's storage site. Risks might be fire, hurricane, or lack of climate controls. Using an archive as a backup site requires the data provider to keep the metadata at the archive up to date as changes are made.

In addition, a data provider should archive data for posterity. The archive facility then takes responsibility for any metadata updates (usually due to changes in media storage, data access, or contact information).

Data Discovery

INTRODUCTION

Data discovery in IOOS will include a way for users to search for specific data sets and to browse the data holdings. It will also include the capability for automated agents to search for data. It will begin with a capability to search metadata to find the data that are desired, and, in the future, will allow for the refinement of that search to include some types of actual data searches. Since IOOS needs to include both the research and operational communities, the amount of understanding of the actual data will be very diverse within the user communities. Users of IOOS will include those who are familiar with the types of data and those who are working on interdisciplinary projects who are less familiar with the data. In addition, there will be a number of users who will not necessarily have any in-depth understanding of the data, such as programmers or decision-makers.

Many studies have shown that information retrieval systems that combine controlled vocabulary searching with free-text (or natural language) yield the best performance[3]. One example is from the Global Change Master Directory (GCMD) where the successful retrieval of documents depends on well-structured metadata and comprehensive indexing of records with keywords from the controlled vocabulary, combined with well-populated text fields to enhance free-text searching.

Controlled vocabulary and free-text searches are two independent but complementary information retrieval systems. Searches conducted using the controlled vocabulary match the chosen word in the metadata record using a direct search of the database. Results can be refined by adding another science parameter, by combining with other controlled keywords, or by adding a free-text component to the search.

Searches by free-text can be made by entering single or multiple words (for phrase searching) and simple Boolean logic (AND/OR) for words of phrases occurring anywhere in the text.

The language used in the metadata needs to be understood by interdisciplinary users. Keywords/thesauri should be carefully created to use commonly used terms and definitions, and to incorporate new terminology. Both users and programmers need to be able to understand the metadata and find information needed using consistent terminology. For IOOS to be successful, users and programmers need to be assured that the metadata they find during data discovery is up to date, consistent, and understandable.

[3] Rowley, J. 1994. The controlled versus natural indexing language debate revisited: A perspective on information retrieval practice and research. J. Information Sci., 20(2). pp. 108–119.

When searching for data, additional parameters should include geospatial search and temporal search constraints on the data and taxonomic information for biological data. Fielded searches that allow the user to specify the metadata fields that should be used in a free-text search also may be employed. An initial implementation will include these parameters, and a user feedback mechanism will be in place to allow users input on the refinement and extension of search capabilities.

CATALOG

For this document, the catalog is defined as the information held to provide for the discovery of and access to data. It was assumed at the beginning of this process that the catalog would contain the metadata that is to be searched in the discovery process. Since full text and fielded searches are required in this system, the initial implementation of the catalog will contain the full metadata record.

Single vs. Distributed Catalog

The recommendation of whether the system should use a single catalog or a distributed catalog is something that should be studied further and must be looked at in the context of the decision on governance of the overall IOOS system. The type of governance and management structure put in place for IOOS will have major impacts on the feasibility of these options and the maintenance of the system as a whole. It should be noted that a single or small number of distributed nodes that are mirrored would be more robust to network outages. A distributed system, unless every part was mirrored, could have pieces that become unavailable when potentially needed the most. This issue is especially important if the system is to be operational. As more agencies become dependent on the resource there will be a greater need to maintain near 100% availability. Also, disaster planning and preparedness (Homeland Security issues) will force a high level of redundancy for the IOOS system.

A single catalog option allows much more control over the contents of the catalog and its overall maintenance. It is easier to do administrative functions within a single catalog, including statistics on the data and metadata and upgrades to the catalog and discovery interface. There is also the consideration of performance. A single local catalog should have better performance than a distributed system that must take into account network delays.

The distributed catalog option would be more in line with a distributed governance policy in which each “organization” would maintain its own catalog and a common catalog query mechanism would be used to search these systems—preferably in parallel. An example of this type of system is the FGDC Clearinghouse nodes that use Z39.50 search protocol. Performance issues for this option need to be looked at along with the issue of updates, general maintenance and error checking, duplicate metadata records, outdated records, and extensibility of the system.

For the initial implementation, a single search catalog will be set up to demonstrate the capabilities of the system.

Maintenance/Management of Catalog

This issue of metadata maintenance is addressed in the Metadata section of this document. How this maintenance affects the catalog is the issue to be discussed here. It is assumed that the metadata review and maintenance will be done by the organization responsible for the metadata. This means that the system must provide a capability to "harvest" metadata from the data source, require that the metadata be maintained within the catalog system, which then implies a remote maintenance capability, or allow for both of these options. It is recommended that both of these options be supported so that the system can accommodate (1) the data provider who does not have the resources or chooses not to operate and maintain a metadata generation capability, (2) the data provider who already maintains metadata and wishes to continue to do so within their own system but does not want to be a part of a distributed catalog if that option is available to them, and (3) the data provider who is willing to both maintain metadata on their own system and operate a metadata catalog. The underlying requirement of the system is that a metadata record should be maintained in one place and not require a duplication of effort to update.

Access Controls

The catalog must allow for the control of access to the metadata records, not only for the modification of those records, but also for viewing and searching on those records. There are metadata records along with data that will not be available to the general public and, therefore, securing those records must be considered within this system. Implementing security within the catalog is not a difficult task, but the process for allowing access to these metadata records is affected by the governance of this system and needs to be considered in light of those alternatives. Access controls for the data are considered as a part of the data transport section and will be discussed there. A security plan is necessary to address the level of protection required (which depends on the value of what is protected) and the appropriate method to secure the data at that level. Classified information may require that the data/metadata be encrypted before transfer or even encrypted in the database.

SEARCH CAPABILITY

The IOOS system must design a method to discover data for which a user had no prior knowledge. This search capability must be extensible so the system can adapt to future requirements within the data discovery mechanism. The initial search capability that will be a search of metadata should contain spatial, temporal, and theme searching as a minimum, and should allow the user to specify whether any, some, or all conditions must be met. The system must allow for extensibility in both the metadata search capability and the area of actually searching data. Each type of search is discussed below, along with some additional capabilities that will be considered within the initial system.

Spatial Search

A geospatial area can be discovered using both the Spatial Domain and the Place and Stratum Keyword sections within FGDC records. Both of these mechanisms will be employed within the initial search capability, and to some extent should be interchangeable. For example, choosing North Carolina as a keyword should set up a search to check the spatial domain for the area (latitude/longitude bounding box) that includes the state of North Carolina, along with the Place Keyword. Another challenge for the geospatial search is defining what place keywords are "contained" within other place keywords (example: North Carolina is a part of North America).

Temporal Search

The temporal search also has multiple sections of the FGDC record to consider, but the issues here are very different. The definition of what is contained within the Time Period Information tag is defined within the Currentness Reference tag and is not necessarily the time period to which the data apply. This must be considered within a temporal search to make sure the time tag is being used appropriately.

The other issue with temporal information is that certain types of data, such as "climatology," are not easily described within an FGDC record. The standard does not address this issue, and therefore a method must be developed within the IOOS metadata guide to describe these types of data.

Thematic Search

As described above, combining controlled vocabulary searching with free-text searches yields the best performance. Controlled vocabulary and free-text searches are two independent but complementary information retrieval systems. Thematic searches can be done on the Keyword section of the metadata record using a full-text search capability on the complete metadata record, or fielded searches, which allow "full-text" searches on specified fields within the metadata record.

One of the first tasks must be to define a data dictionary (or set of dictionaries) for the controlled vocabulary portion of a thematic search. Further work would include mapping among dictionaries. A specific research area would be the use of knowledge mapping or ontologies to provide the translation capabilities among dictionaries.

Allowing for full-text searches of the metadata record will at some point be required, although this type of search is often implemented as a fielded search where specific fields within the metadata record are searched, and not the full record. There will be the option to allow the sophisticated user the capability to specifically define what sections of the metadata record will be searched in a fielded search, along with allowing single or multiple words (for phrase searching) and simple Boolean (AND/OR) for words or phrases occurring in the text. A default set of sections within the metadata record to be searched will be defined for a fielded search for the unsophisticated user.

Biological Data and Taxonomic Search

The IOOS search capability will accommodate marine biological data from a variety of sources and integrate these databases into a distributed system. One major difference between how physical oceanographers and biologists handle data is that physical oceanographers deal in files and biologists deal with data. Say a biological data set contains the name and number of all species found in a particular net haul. To be useful, the metadata documentation needs to include all the taxonomic names found, plus the geographic location. But that's pretty much all that is in the data set.

Within the *Content Standard for Digital Geospatial Metadata, Part 1: Biological Data Profile*, a section has been included that contains taxonomic information. One option is to search this section, which can include, as a minimum, items such as Common Name, Genus, and Species. The Ocean Biogeographic Information System (OBIS—see http://iobis.org) is being developed to meet observing system needs for biological data. OBIS has found that direct searching of properly structured data is easier than a metadata search, and that content standards are more time-effective than metadata standards. OBIS also provides international standards and protocols for accessing marine biological data. Integration of this type of search into the IOOS search capability is an area that needs immediate further study and will be one of the initial efforts of the data discovery team.

Parameter Search

Being able to search for specific parameters must be included early on in the system. Parameters are defined in the Attributes section of the metadata record, and filling in this section will allow not only for this search capability, but also for the ability to subset the data set based on specific attributes.

Additional Search Parameters

The search capability within IOOS must be extensible in the future to include searching on items in the metadata record such as the quality of the data, the formats data is available in, and other items that are requested by the user community.

Browse Option

The option to browse the catalog is also a requirement, and should be defined to allow for flexibility within the system. Defining what the user sees within the browse function, how that information is sorted, and allowing for optional sorting capabilities are all items that need to be defined in the system and must be extensible as feedback is provided to the developers on what the users of the system require.

Results Listing and Search Refinement

Another area that must be defined is the results that are returned to the user when a search is completed and how a search can be refined. An initial task in this area is defining what will be included within the results display, how to "rank" the results, and if the number of results should be limited. Initially, search refinement will allow the user to modify the defined search parameters and allow the system to then search again either within the initially returned results or within the full catalog. Future work in this area will extend the search capability beyond a metadata search and into the area of actually specifying the data to be searched for specific values.

INTERFACE TO DATA ACCESS

Initial assumptions are that the data will be available electronically, on line, and free of charge. This makes the interface to enable data access much more focused, but it is still an area that needs coordination among the working groups. Within the design phase of this system, the data transport and discovery must agree upon a means to point to the data once it has been discovered. It cannot and should not be assumed that the data and metadata will reside in the same place. Future considerations will need to include (1) data that are not available on line; (2) non-electronic data; and (3) data that are available for a fee.

PORTAL

In the World-Wide Web dictionary, a portal is defined as, "A web site that aims to be an entry point to the World-Wide-Web, typically offering a search engine and/or links to useful pages, and possibly . . . other services. . . ." (See the Glossary of Terms). It is assumed that a portal of some type is a requirement for this system to provide access to the search and discovery capabilities, but it should not be the only access mechanism. Listed below are some of the considerations that need to be addressed and recommendations on how they should be addressed.

Architecture

The issue of governance will again weigh heavily on the portal architecture. The system should be designed to support both a single and distributed portal, along with allowing remote content management of the information contained in the portal. The scope of scientific and/or reference information contained within the portal will be defined within the scope of the governance discussions.

The search capability will have both a defined user interface and a defined access protocol to allow it to be customized for different user communities. It will also allow an Application Programmer Interface (API) connection so that applications can be directly connected to the search. Recent advances in web technologies have resulted in Web Services utilizing SOAP/XML for application-to-application operations. Web Services is a standards-based system that can be easily utilized to provide the "glue" that connects a backend metadata database (relational, object, or LDAP, depending on requirements) and a portal web site or application. Web Services includes standards for advertising both the capabilities of the service and the API for utilizing the service. An implementation-language-neutral approach, like Web Services, will help provide a longer life span for the system. This is an area that will be studied further to define its applicability to the IOOS system.

Search Content and Scope

Additional functionality for supporting searches is a part of the portal and will be considered in this section. Some of the options that will be considered and supported are the ability to search anonymously, along with the option to maintain a user account. When operating the system anonymously, the user should assume that nothing is saved when the session is completed. But, if the user chooses to maintain an account, they have the option of saving search parameters and search results. They will also have the option of sharing these parameters or results with other individuals.

Subscription services will also be a part of the portal and supported within the transport section of the system. A user that maintains an account within the portal will be able to subscribe to specific data, and as those data are updated or new data arrives, the user will be notified or the data will be delivered automatically to that user.

Dictionary services will also be supported within the portal. These can include, but are not limited to, the following broad categories:

- A way to associate events with parameters. This is usually not an issue with data that are collected for specific events such as a hurricane. When data are collected in this manner, it is relatively easy to include the event in the metadata within the keyword section. But when data are continuously collected and an event occurs, it is much more difficult to go back into that metadata record and add keywords associated with specific events. In the latter case, it would be beneficial to have the event associated with specific keywords or "types" of data in the portal itself, so the search is focused on the information contained within the metadata record. An example is "El Niño". This is an event that may have "Tropical," "Southern Pacific," and "Sea Surface Temperature" as the associated keyword, location information, and parameter that is searched within the system.

- A means to provide for both Broad and Narrow search context. A user should be able to come into the system and search for specifics such as Sea Surface Temperature (Narrow search context). But, they should also be able to search a broad category such as "Harmful Algal Blooms" and the system will then define the specific narrow search parameters such as "toxic phytoplankton," "*Karenia brevis*," "red tides," and other parameters, areas, and keywords associated with these events.

The portal will incorporate basic display capabilities to allow the user to discern whether the data they have found are of interest to their specific requirements. These capabilities will include, but not necessarily be limited to, a mapping display option, a time-series display, a method to display and manipulate volumetric data, and a display capability for biological data.

The portal itself will incorporate a User Feedback capability and Help functionality to allow the user to interact with the system, solve problems that are associated with the system, and provide vital information for its maintenance. Usage tracking will also be a part of the portal, and it must be at a level that allows the management of this system to see what sections or pages within the portal are and are not being accessed, what data are available during at least routine evaluations of the system, and what data are being accessed by the user community. The amount and level of statistics that can be collected on the user community as a whole should be addressed once the governance issue is resolved.

The portal will contain links to relevant information such as tools available for metadata generation, information on the metadata required for this specific system, and information on what is required for a group to become a data provider to the IOOS. It could also contain links to the supporting organizations if that is desired, along with allowing for other types of queries such as library and/or web searches.

Other issues such as Domain, Look and Feel, Scientific Content, and Disclaimers will be addressed once the governance issue is resolved. The main requirement from the aspect of a pilot project to provide a discovery portal is that the system will be easily portable. Then, when governance is decided, the portal can be moved, if required, and easily modified to a new look appropriate to the domain. Other specific domain issues need to be addressed by the hosting domain.

Annex A: Glossary of Terms

Accuracy

Conformity to fact. (From *The American Heritage Dictionary*, third edition)

Catalog

A list or itemized display, as of titles, course offerings, or articles for exhibition or sale, usually including descriptive information or illustrations. (From *The American Heritage Dictionary*, third edition)

DAML+OIL

A semantic markup language for web resources. It builds on earlier W3C standards such as RDF and RDF Schema, and extends these languages with richer modeling primitives. (From http://www.w3.org/TR/2001/NOTE-daml+oil-reference-20011218)

FGDC

The Federal Geographic Data Committee coordinates the development of the National Spatial Data Infrastructure (NSDI). The NSDI encompasses policies, standards, and procedures for organizations to cooperatively produce and share geographic data. The Federal Geographic Data Committee approved the Content Standard for Digital Geospatial Metadata (FGDC-STD-001-1998) in June 1998. (From http://www.fgdc.gov/)

Inventory

A detailed, itemized list, report, or record of things in one's possession, especially a periodic survey of all goods and materials in stock. (From *The American Heritage Dictionary*, third edition)

Lineage

Direct descent from a particular ancestor; ancestry. Derivation. (From *The American Heritage Dictionary*, third edition)

Ontology

Ontology is the theory of objects and their ties. The unfolding of ontology provides criteria for distinguishing various types of objects (concrete and abstract, existent and non-existent, real and ideal, independent and dependent) and their ties (relations, dependences and predication). (From http://www.formalontology.it/)

OWL

A semantic markup language for publishing and sharing ontologies on the World Wide Web. OWL is derived from the DAML+OIL Web Ontology Language [DAML+OIL] and builds upon the Resource Description Framework [RDF/XML Syntax]. The OWL Web Ontology Language is being designed by the W3C Web Ontology Working Group in order to provide a language that can be used for applications that need to understand the content of information instead of just understanding the human-readable presentation of content. OWL facilitates greater machine readability of web content than XML, RDF, and RDF-S support by providing an additional vocabulary for term descriptions (from http://www.w3.org/TR/2002/WD-owl-ref-20020729/ and http://www.w3.org/TR/2002/WD-owl-features-20020729/)

Parent/Child Metadata

A relationship between metadata records where the parent record would be considered a "master" record and contain information that is common to the group of records; the child record would contain only those items specific to that particular record. An example of this is in the case of a data source with a collection of buoys. The "parent" record would contain all common information for the collection of buoys, and the "child" record would contain the location, time, and sensor information for a particular buoy platform.

Portal

"<World Wide Web> A web site that aims to be an entry point to the World-Wide Web, typically offering a search engine and/or links to useful pages, and possibly news or other services. These services are usually provided for free in the hope that users will make the site their default home page or at least visit it often. Popular examples are Yahoo and MSN. Most portals on the internet exist to generate advertising income for their owners, others may be focused on a specific group of users and may be part of an intranet or extranet. Some may just concentrate on one particular subject, say technology or medicine, and are known as a vertical portal."

Precision

The exactness with which a number is specified; the number of significant digits with which a number is expressed. (From *The American Heritage Dictionary*, third edition)

Quality MetaData

Metadata quality consists of several components, including correct information, complete information, and having the information in a standard form/vocabulary.

RDF

The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. This specification describes how to use RDF to describe RDF vocabularies. This specification also defines a basic vocabulary for this purpose, as well as conventions that can be used by Semantic Web applications to support more sophisticated RDF vocabulary description. (From http://www.w3.org/TR/2002/WD-rdf-schema-20020430/)

Semantic

Of or relating to meaning, especially meaning in language. (From *The American Heritage Dictionary*, third edition)

Semantic Web

The abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be defined. It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners. (From http://www.w3.org/2001/sw/)

SOAP/XML

SOAP is a lightweight protocol for exchange of information in a decentralized, distributed environment. It is an XML-based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined datatypes, and a convention for representing remote procedure calls and responses. (From http://www.w3.org/TR/SOAP/)

Syntactic

Of or relating to the rules of syntax. Conforming to accepted patterns of syntax. (From *The American Heritage Dictionary*, third edition)

Syntax

The rules governing construction of a machine language. A systematic, orderly arrangement. (From *The American Heritage Dictionary*, third edition)

Web Services

The World Wide Web is more and more used for application-to-application communication. The programmatic interfaces made available are referred to as Web Services. (From http://www.w3c.org/2002/ws/)

XML

The Extensible Markup Language (XML) is the universal format for structured documents and data on the web. (From http://www.w3c.org/XML/)

Annex B: Committee Membership

Susan Starke – NOAA/NCDDC (Team Leader)
Anne Ball – NOAA/CSC
Julie Bosch – NOAA/NCDDC
John Caron – UCAR/Unidata
Cheryl Demers – NOAA/NWS
Donald Denbo – NOAA/OAR
Dan Holloway – URI/OPenDAP
Lola Olson – NASA/GCMD
Karen Stocks – UCSD

Annex C: Reference

Kinzig, A. P., S.W. Pacala, and D. Tilman (eds.). 2001, *The Functional Consequences of Biodiversity, Empirical Progress and Theoretical Extensions*. 365 pp. Monographs in Population Biology, 33, Princeton University Press, Princeton, New Jersey.

I. Interoperable Data Discovery, Access, and Archive

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

Part III. Appendices

Appendix 2. Data Transport

IOOS DMAC Data Transport Team

March 2005

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

Contents

Introduction

The fundamental objective of the IOOS Data Transport System is *machine-to-machine interoperability with semantic meaning* in a highly distributed environment of heterogeneous data sets. IOOS users will be able to summon and analyze fresh and archived data using both familiar and new tools when the data transport component operates in concert with other elements of IOOS.

INTEROPERABILITY, FLEXIBILITY, AND URGENCY

Machine-to-machine interoperability with semantic meaning allows data to be exchanged between computers without human intervention. For example, consider a would-be user of the envisioned interoperable IOOS data system who is interested in the relationship between red tide occurrence and wind speeds above a given threshold. Such a user, working within a computing environment familiar to him or her, would request all red tide data for which there is a wind observation both exceeding the given threshold and located within a specified distance and time of a red tide observation; the system would return observation pairs to the user's environment without further input from the user.

For this to occur, all computers involved in such a transaction must be capable of determining both the syntax and the semantics of the exchanged data. Thus, in order to fulfill the request for red tide data, certain basic information has to be associated with each observation—variable name, units, location, and time. With this information, plus knowledge of the organization of the data and a directory of data sets, all the syntactic[1] and semantic[2] information necessary to meet the request is available.

This suite of syntactic and semantic information might not be sufficient to fulfill a more intricate request, however. For example, if a user were to ask for the averages of all red tide observations touching, say, one degree squares, then additional semantic information might be required to properly average data from different sources: it might be necessary to know the uncertainty of an observation or the size of the area over which it was made. The point here is that while the syntactic information needed to satisfy a variety of requests in a distributed system of heterogeneous data remains the same, the semantic information needed to meet different requests is quite likely to vary.

The foregoing example illustrates some of the functionality of a data management and communication system that exhibits machine-to-machine interoperability with semantic meaning.

[1] Syntactic metadata is information about the data types and structures at the computer level, the syntax of the data, for example, variable T represents a floating point array measuring 20 by 40 elements.

[2] Semantic metadata are what one normally thinks of as metadata—information about the contents of the data set.

Note that it is virtually impossible to envision the complete set of semantic information that will be needed for all future requests that might be deemed important to a given community. Therefore, extraordinary flexibility must be built into the system so that it will be able to meet the community's needs well into the future. Also, over the next 10 years changes will certainly occur in hardware and software and are likely to occur as well in data types and in the user community's interests—*the data transport component of the IOOS data system must be designed with sufficient flexibility to allow it to evolve gracefully with time.*

The system should be in place in the relatively near future (by the end of 2004, if not earlier), when the first elements of IOOS will come on-line. This urgency places constraints on the process of developing the Data Transport System. Fortunately, in anticipation of IOOS, the National Oceanographic Partnership Program funded a three-year effort beginning in spring 2000 to design and implement a data system that would provide the basis for the IOOS data system. Much of the material in this report has been drawn from that effort.

THE LAYERED (OR MODULAR) APPROACH

The Format Layer and the Syntactic Data Model—The IOOS Data Transport System must be capable of moving data from a site where they may be stored in one format to a client application that may require them in another format. There are a variety of ways in which this can be achieved. To reduce the number of translators required in the system as a whole, it makes the most sense to transform the data to an intermediate representation and then from the intermediate representation to that of the client. The intermediate representation represents the basic system data model. To retain the most flexibility in the system, the Data Transport Team recommends that this data model be discipline-neutrals, i.e., that no presumptions related to the semantics of the data be made at this point. We refer to this as a syntactic data model. In general, data values would not be altered as data pass through this layer nor would the organizational structure of the data be modified. The degree to which this is achieved would of course depend upon how comprehensive the syntactic data model is: the more comprehensive the syntactic data model, the less alteration of the data.

The Structure Layer—As noted above, the fact that data sets can be, and typically are, organized in a variety of ways imposes an additional burden on clients. To reduce this burden, the system should provide the capability of delivering data to clients in a structurally consistent form where appropriate. Modules that participate in the modification of the structural representation of data sets constitute the structure layer. The structure layer protocol would define the organization of like data objects in a data set: data at a given site meaningfully represented as an n-dimensional object would be required to be presented as such to the client. However, this need not be accomplished directly from the originating server; it might simply constitute a link in the acquisition chain. In addition, modules in this layer should provide the ability to aggregate data from multiple sites into new data sets that are structurally and semantically consistent. It is clear that some of the operations performed by modules in the structure layer will be discipline-neutral while others will either add semantic content to the data or make use of semantic information to structurally reorganize the data. For example, all gridded data that involve space and time might be required to be represented as a four-dimensional array of (Longitude, Latitude, Depth, Time); in such a case, if a given data set were initially organized as (Latitude, Longitude, Time), it would be reorganized to the preferred 4-d form by inserting a null Depth dimension. As a matter of design philosophy and to provide the most flexibility in the use and evolution of the Data Transport System, operations that can be performed in a discipline-neutral fashion will be separate from those that require a semantic understanding of the data, certainly logically and probably when implemented by enclosing them in distinct structure layer modules.

The Semantic Layer and the Semantic Data Model—To make use of data sets stored in a heterogeneous interoperable system, at least some semantic metadata must be available; moreover, they also must be consistent (or the ability to map them to a consistent form must exist), and it must also be possible to add new metadata terms as the need arises. The semantics implicit in the structural transformations that the system may provide and the semantic information transported in the data access protocol together define the semantic data model. The core of a semantic data model is the set of translational use metadata, which must be well defined and easily extensible. Interoperability within the system must be defined to depend only upon this core; that is, while the system must be able to carry additional metadata, interoperability within the system must not depend on the presence of such additional information. Note that the semantic core need not be collocated with the data that it describes, nor does it have to be collocated with the additional semantic metadata. The system infrastructure must, however, be capable of appending such metadata to the data stream on request. IOOS must be capable of providing access to metadata in a variety of forms to take advantage of the metadata developed by different communities and it must be capable of providing access to metadata from a site other than that of the data server.

Required and Optional Routes through the Layers—The most fundamental operation performed in the system outlined above is format translation from the storage format to the format expected by the client. This means that the data access protocol related to the transformation of the data must lie at the core of the system. Other operations might be performed on the data as they move from originating server to end client but they are not, in general, required.

A PROCESS FOR DESIGNING AND IMPLEMENTING THE IOOS DATA TRANSPORT SYSTEM

Design goal and assumptions

The Data Transport Team's planning process leads from interoperability requirements to system specification to system implementation as a consequence of these assumptions:

- The chief goal of the IOOS data system is interoperability over distributed data providers and users.
- The design goal (i.e., what the system is to achieve/deliver, in this case, interoperability) determines the system specifications.
- System specifications (protocols, interfaces, standards) should be addressed separately from and prior to system implementation.

Specifications

The following Plan by the Data Transport Team of DMAC-SC has been designed to provide the function and flexibility required in the examples above. The basic concept is to build a modular system that consists of a number of interoperable layers and which allows layer elements to be added or substituted as needed. This is the same design concept that underlies TCP/IP. Indeed, it could be advantageous for the IOOS data transport system's design to follow closely the directions in which various components of the Web appear to be evolving.

At the base of all services in the system is the data transport protocol. This protocol must provide each modular element required in the system. It must be flexible, straightforward to implement, and comprehensive at the lowest levels. The base layer, in this approach, would almost inevitably use TCP/IP as its low-level protocol; although there are other options, they are certainly not the direction in which the Web is evolving at present.

Requisite Data Transport System Capability #1: The IOOS Data Transport System must be capable of accessing data in a variety of formats.

At the higher layers, however, more considered choices must be made. As noted above, the primary objective of the data transport component of the IOOS data system is machine-to-machine interoperability with semantic meaning. Now, it is theoretically possible to achieve this objective by using ftp, very thick client(s), and a great deal of metadata to describe how the data are organized and what they mean. The level of effort required to develop clients for such a system as well as to serve the data, however, makes such a design quite impractical:

- The client would have to handle both format conversions and structural reorganization of the data, as well as transformation of the data to a consistent set of units.
- The data server would have to develop an extensive amount of metadata that describe the details of the organization of the data in addition to the data semantics.

At the opposite extreme is the requirement that all data be stored using the same data model. This is even less practical, because:

- Archives of historic data will be vital to IOOS and these are not stored in a consistent format with consistent metadata descriptions.
- IOOS does not control all real-time data streams that are of potential interest to the IOOS community.

Requisite Data Transport System Capability #2: It must be capable of providing access to data via a variety of client programs, and it must translate from the format in which the data are stored to the format required by the client program.

It is quite clear that at least for the foreseeable future, users of the system will access data from the system via a variety of applications and interfaces; the data must be made available to the user's client application in the format desired by the client.

Requisite Data Transport System Capability #3: It must be capable of delivering data of a given data type in a structurally (syntactically) consistent form across all data sets in the system.

It is important to stress that format consistency does not imply consistency in the structural organization of the data delivered. For example, a sea surface temperature archive at one site might consist of a number of two-dimensional (longitude, latitude) files, one per time step, while another site might present similar data as one three-dimensional (longitude, latitude, time) file. The fact that the data are delivered directly to the application package in a consistent format substantially reduces the complexity of client-side applications as well as the metadata needed to describe the data, but the lack of structural uniformity is still a substantial burden on the clients. For example, in the multi-file 2-d case cited above, the client would have to deal with an inventory system of some form, while in the 3-d case this is not necessary. In addition to the added complexity that would have to be built into the client, additional metadata would be needed to describe the inventory and associated 2-d data objects in the first case and the 3-d data object in the second.

Requisite Data Transport System Capability #4: It must provide the metadata needed to transform the data to a consistent semantic form, or it must be capable of delivering the data in a consistent semantic form, or both.

Syntactic consistency of data sets does not imply their semantic consistency:

- variables might be the same, but units of measurement might differ, for example, one data set might be in cgs units while another might be in mks units, or
- the naming conventions used to describe variables might differ among data sets, for example, sea surface temperature might be notated as "sst," "SST," "sea surface temperature," "ts," "Ts," "surface temperature," and so on.

Semantic meaning can be communicated through translational use metadata, descriptive use metadata, and search metadata. For the level of interoperability desired, all are required.

It ranges from difficult to impossible, however, to envision all the descriptive use metadata that may be required of the IOOS system in the future. Different user communities will likely require different collections of descriptive use metadata, and they may require them in different forms.

The Data Transport Team suggests that semantic metadata be kept logically separate from syntactic metadata. Special attention should be given in the design of the system to providing a flexible structure for the semantic use metadata. The resulting system is likely to produce a variety of use metadata models all sharing a common base.

Implementation as an Extensible Suite of Modules

The Data Transport Team proposes implementing IOOS data transport as a suite of modules based on a set of transport and semantic protocols addressing the requirements identified above (format translation, structural reorganization, and metadata consistency). These modules would be layered on top of modules that implement TCP/IP protocols and would constitute, in effect, an application layer over TCP/IP.

The modular approach will allow the IOOS Data Transport System to be implemented in well-defined, manageable pieces, while at the same time providing desirable flexibility in the functionality and use of the system. It also provides a clean mechanism for extending functionality by layering modules on top of those already defined, for example, adding data reprojection.

In summary, IOOS must:

- Support format translation from the format in which the data are stored to that of the client program,
- Support delivery of data of a given data type in a structurally consistent form,
- Provide the metadata needed to transform the data to a consistent semantic form or deliver the data in a consistent semantic form,
- Provide the capability to access metadata from a location different from the data server, and
- Provide a capability to access metadata in a variety of forms, i.e., IOOS must not be confined to using a single semantic data model.

Charge to the Working Group/ Bounds of the Problem to be Addressed

GENERAL CHARGE TO THE GROUP

BASIC ASSUMPTIONS

This section lists the basic assumptions that we are making in the design of the system. Such assumptions are often hidden, but in the end drive system design. Here are some that we are making:

- Data will be heterogeneous in type and storage format.
- Data storage will be distributed.
- Data will often, but not always, reside with the data collector.
- The system to be developed will be a client-server system.

Distribution assumptions

The DMAC-SC DTT is proceeding initially under the assumption that data will be provided free of charge in the IOOS system. It remains to be determined how IOOS should handle the case of wishing to make available data for which there is an associated fee, but the system must be extensible to accommodate such a case.

Initially, also, the DTT is assuming that there are no limits to be set upon the amount of data allowed to be transmitted in response to a request. However, there ought to be a way for the user to find out the size of the data set to be received beforehand, and it may be found desirable to query the user if the response involves an inconveniently large amount of data.

The primary mode of interaction will be to receive the requested data automatically and immediately. The DTT should also consider the desirability of establishing separate, optional download routes, for example, via ftp with email notification.

The DTT should also consider whether to make it possible for users to receive data in non-electronic form: CD, for instance, or on paper.

Bounds on what this group will address

These are some issues that must be addressed somewhere in the system, but apparently not by the DTT under its present charge:

- Other groups will deal with search metadata,
- Other groups will deal with data archival,
- Other groups will generate products.

General Description of IOOS Data Transport System Needs and Implementation Considerations

In this chapter we address basic issues related to the design and implementation of the IOOS Data Transport System. We begin with a discussion of basic requirements. This is followed with a section on implementation considerations based on the direction in which we believe network computing is pointing. The final section of this chapter identifies basic functional requirements of the system.

BASIC REQUIREMENTS

An adequate data model

First and foremost, the IOOS Data Transport System shall ensure that numeric oceanographic data can be interchanged without corruption or loss of precision between arbitrary data repositories and users, as the need to facilitate the exchange of numeric data is pressing.

The task of exchanging other kinds of data (images, video, audio, etc.) can be set aside for now, as middleware specifically designed for the interchange of these formats is available. This is not to say that there should be no ability to exchange graphics and images within the system we are developing, just that transfer of numeric data is the prime consideration.

The IOOS Data Transport System must be able to express the structure of the numeric data it will encounter in oceanographic data repositories, that is, it must be able to transmit syntactic metadata. It also must be able to transmit all relevant semantic metadata, that is, translational use, descriptive use, and search metadata.

It may be desirable to poll oceanographic data repositories to ensure that the Data Transport data model contains appropriate base types. As a start, the following simple and compound types will probably be required:

- Simple types: integers (signed 16, 32, 64-bit; unsigned 8, 16, 32, 64-bit); floating point (32, 64-bit); strings; etc.
- Compound: structures; arrays; lists/nets/graphs; hash keyed "dictionaries."

This list is not meant to be exhaustive or even adequate for a successful implementation. It is fully expected that other datatypes will be found useful and necessary. This list does not specify syntax, either; the ability of one type to contain another, for example, of arrays to contain structures, is extremely important to the expressiveness possible with a given data model.

Extensibility

There are several possible ways to consider "extensibility." One addresses the ease of expanding the installed base of clients and servers by bringing on line more sites of an already supported platform type. Another addresses the ease of extending the range of supported platforms, both servers and clients. Yet another addresses the extension of portions of the system; one might consider adding a new, basic datatype to the protocol at some point. One might also want to extend the system into other disciplines: meteorology, geography, economics. It would be important that fundamental parts of the IOOS Data Transport System be discipline-neutral to enable this coalescence of terrestrial data systems.

Coexistence with existing transport systems

"Coexistence," too, may be considered in a couple of senses. First, there is "coexistence" in the sense of "do no harm." It would be unfriendly and unwise for the IOOS Data Transport System to impede or restrict any site's use of other transport systems, whether in the field of oceanography or not. For example, if a repository already uses a system which depends upon a particular data storage format, that site should not be forced to abandon their system in order to adopt IOOS.

Second, there is "coexistence" in the sense of "actively associate with." IOOS should exert itself to make it as easy as possible for its data transport system to interact with other systems. If this ability is not present universally through the IOOS system, then thought should be given to operating portals between significant systems as part of the system's base configuration.

Non-dependence on proprietary software

We suggest that it is most important that the IOOS Data Transport system not depend critically on proprietary software at any point. For it to do so could exclude one of the most powerful engines for software development: the freely given efforts of programmers who simply like to see a system work better. Also, at least for the research sector, it is of utmost importance to be able to examine source code to determine what a process actually does, rather than what a person or documentation says it does.

No need for provider to "re-engineer" existing storage capability

Sometimes it seems that if a way of storing data exists, some archive has tried it, no matter how unlikely the configuration might seem. However, IOOS data transport is unlikely to succeed in being adopted widely if it attempts to coerce all equivalent data to be stored identically everywhere. Therefore, the scheme it adopts for generating syntactic and semantic data models must be so flexible and extensible that any IOOS server can find a way to express its storage format in an IOOS data model.

IMPLEMENTATION CONSIDERATIONS

GRID Computing

An effort is being devoted to the development of GRID computing technologies. It is clear that such technologies will play an important role in distributed data systems in the near future. The IOOS must therefore track development in this area with the expectation that the underlying data transport protocol will likely have to be GRID-aware.

XML Encoding

There is at present broad community acceptance of the idea of incorporating XML into a data transport system for several reasons, among them the following. As a subset of SGML, it is a fairly comprehensive, fairly accessible means of creating specialized markup "languages" that can be tailored closely to the needs of various projects and disciplines. It is finding, at the moment, growing favor as a vehicle for transferring information in and out of databases; relational databases have acquired XML-conversant front ends, and native XML databases have arisen. Being like HTML an offspring of SGML, it should fit well into the Web; it is "friendly" towards HTTP. It is at present one of the directions towards which the Web is moving and deserves our attention as we develop the IOOS data transport system. Possible areas of use are in adopting XML as the basis for persistent forms of syntactic and semantic data models.

Even though XML is presently in favor and is supported by an increasing number of tools, we consider it to be a part of the implementational aspect of IOOS data transport, not of the intrinsic functionality. XML may be supplanted by some other information framework in time, and IOOS should be ready to reevaluate its adoption of XML every five years or so.

XML is not without its awkward aspects, though. For example, it is still not entirely settled how one should include data that the XML parser should not inspect within an XML document. One solution seems to be not to include it at all, but to include a URL that points to the data, which could be fetched by a non-XML mechanism. Another approach is to use a multi-part MIME document as an additional envelope around the XML document and to include as well the binary data as another occupant of the envelope. The IOOS DMAC-SC DTT will have to keep an eye on the developing standards relevant to this matter, and a planning activity needs to be set in motion that will track the development of non-parsed XML content. When (if) a standard emerges, then the planning activity will make a recommendation on whether the data access protocol should be modified to use the standard.

Data Discovery

A critical component of the overall data system is the ability to locate data within it. Data discovery is being addressed by another subgroup of the DMAC-SC, but there are critical areas of overlap between these two groups. The perceived difficulty associated with populating data discovery services and the resistance by the data collection community to documentation of data are issues that need to be addressed not only to provide for data discovery, but to utilize data appropriately in the future. Deploying a system that allows as much automation as possible in the area of documentation is a requirement, along with educating the data-collection community and providing consulting services that will ease the burden of documentation. To address these issues, we believe that the data discovery portion of the system must be intimately coupled with the data access portion and that this coupling must be as automated as possible.

FUNCTIONAL REQUIREMENTS

This section discusses "the core," that is, the functionalities that the IOOS Data Transport System's design and implementation guarantees to make available to any user of the system. Many of these functionalities must be available to every client/server pair. Others, however, may be so implemented that they require the participation of an intermediary, such as a specially outfitted server which, nevertheless, would be available to every system user.

The motivation behind communication protocols tends to be the desire to provide the most functionality and convenience to the greatest number of users for the lowest overhead imposed upon the parties individually and in total. The design goal of bringing data transfer functions to all IOOS clients and yet not requiring them or all the IOOS servers to become complex ("thick") may be achieved in part by giving the responsibility for fulfilling certain kinds of client requests to just a few servers. The organization of this section reflects this dichotomy:

- Section 3.3.1 discusses services which contribute to the "thickness" of every IOOS Data Transport client and server,
- Section 3.3.2 considers services which could be implemented at selected servers.

Universal minimal requests and facilities

By "universal minimal requests and facilities" we mean those requests that every IOOS Data Transport user can make in every appropriate request to any IOOS Data Transport server with the full expectation that the request will be satisfied by that server without participation of an intermediate IOOS Data Transport server.

On-line acquisition of data into legacy application packages from a variety of data sources

Requiring "on-line acquisition" is taken to mean having a Web-mediated transaction with brief, hopefully imperceptible, waiting time between requests and responses. Provision may be made for a "batch mode" where very large or very many data sets can be transmitted at times more convenient to client or server. This is a topic for further examination.

Specifying "legacy applications" means modifying, wherever feasible, existing analysis and visualization software packages to participate as clients in the IOOS DT system. These packages are in many instances expensive, familiar, capable, and customized: the investment they represent is considerable and "legacy" does not imply obsolescence. Fortunately, many such packages do provide mechanisms for adding interfaces to new data sources, as will be required here. MATLAB is a prime example of a legacy application.

The "variety of data servers," of course, means not simply that data are to be fetched from many sites. It also means that sites may store data as they see fit and yet will be able to fill requests for portions of that data submitted by any IOOS DT client site without prior arrangement.

"Acquisition of data" is shorthand for what will be, for the IOOS DT system, at times a particularly intricate process, principally because files at the servers will not be inviolate, atomic units of data transfer. Instead, IOOS DT clients will have to be able to ask a server to return just part of a file and, more than that, not simply a sequence of bytes between two positions in the file but projections and selections of the data. Specifics of such a request will have to be available to be communicated by every client, and these same specifics will have to be intelligible to every IOOS DT server, moreover. Benefits of the ability to tailor requests and responses so closely to the client's needs means that at the price of some activity by the server to extract a subset of data, there is (1) a smaller data set to transmit, which reduces demands on system resources, and (2) a properly-sized demand upon the client's resources to accommodate and begin to use the data set.

Web browser capabilities in the base system

Even a casual visitor to an IOOS facility ought to be able to browse actual data to some extent, using nothing more than a standard browser, although perhaps with a readily available plug-in.

ASCII dump of data

Users must be able to dump a data set in a human-readable form easily. As part of implementation planning, the DTT may wish to assign some members to consider whether there are formats other than plain, tab-separated ASCII (e.g., Rich Text Format, LaTeX, XML), which it would be useful to have directly available.

Security, access control, firewall penetration

Security must be considered at the outset, and each participant in the IOOS system must be assured that everything possible will be done to ensure that no harm will come to their site or their data as a result of their participation.

Access to metadata

The metadata describing data sets available through the IOOS servers must be easily retrieved in human-readable and machine-parsable form. If XML is adopted as the format for persistent metadata, then it should be possible to meet this requirement satisfactorily.

Real-time data access versus access to retrospective archives: Push-pull services

The IOOS data transport component of the data system must support real-time[3] access to data as well as access to retrospective data. The important distinction between real-time and retrospective data relates to the way in which the data are likely to be accessed. Real-time data are often desired by subscription (push[4]) while retrospective data are generally requested (pull[5]) in an as-needed mode. In addition, access to data for special events, such as hurricanes and floods, generally imposes a greater burden on real-time data sources than on retrospective data sources at the time of the event. The fully developed data access system will likely be composed of a broadcast capability such as the LDE developed by Unidata together with a pull technology such as OPeNDAP. For routine real-time data flows, the push technology is likely to dominate, although data pull will also likely play a role. This will depend on the access approach chosen by the data user. For routine access to retrospective archives, the pull technology will dominate. For access during special events, it will likely be a combination of the two forms of access that will be result. Specifically, recipients of the data pushed to the community as part of the routine data delivery system may become the servers of choice for access by many because of the relatively lighter load associated with these servers in times of heavy demand.

There is a refinement of "pull" called "informed pull," whose implementation probably lies in the future but which should be considered seriously by the IOOS data transport system. In this scheme, servers send metadata describing characteristics of newly available data, and clients decide whether or not to initiate a "pull' transaction, presumably on the basis of whether or not the new data promises to fill some previously recognized need.

Mediated facilities: a form of distributed computing

Data restructuring, aggregation, and manipulation will be absolutely essential aspects of IOOS data transport. Such facilities could be implemented at only selected server sites yet be available to the entire community of IOOS Data Transport clients. For instance, a client might have data sets organized according to one data model and wish for some reason to have them expressed according to another informationally equivalent data model instead, that is, to have its data set "restructured." This client should be able to send a request to a Restructuring Server asking that the server obtain the data set from some site, restructure it, and return the result to the client. Such mediating servers can be constructed to provide an arbitrary array of operations. COLA's GDS (http://cola8.iges.org:9090/index.html) is an example of this.

This report will recommend funding for a working group to address the issue in detail as soon as possible as part of the design implementation program.

[3] By real time we mean "shortly" after the data have been collected, not actual access to the sensor data stream.

[4] "Push" refers to data being sent from a server site to a client site. The send is initiated by the server as opposed to data being requested by the client. The client would previously have registered a subscription to the service, but, like a daily newspaper, packages of data would arrive at the client either regularly or as they became available without any further action on the client's part.

[5] "Pull" is the process we are all used to using: we request specific data, the request is filled, and the transaction ends.

Data restructuring

“Data restructuring” is any process that takes in a data set described by one data model and maps that data set into one described by another data model. The reordering of axes in a 4-d data set is a simple example of restructuring. In its purest form, data restructuring involves only mapping: while the mapping may not be complete (some input variables may not appear in the output), no data values are changed. Data restructuring may involve interesting and non-trivial design issues, particularly if it is to implement a completely general mapping from data model to data model. This potential complexity may be reason enough to implement it on selected servers only.

Data aggregation

“Data aggregation” is any process whereby a data set is generated by joining in some manner data held in more than one data set, possibly in more than one file, possibly at more than one site. It may simplify matters to specify that all the data sets input for a given transaction be described by the same data model, that is, to make a sharp distinction between “data aggregation” and “data restructuring.” Presumably if restructuring were required, the offending data sets could be piped through a restructuring server before being received by the aggregation server.

An example of a big issue in this regard is the removal of replicated data in an aggregation. Again, at first glance it would seem to be much simpler to deal with replicated data if in any given aggregation operation all the data sets were organized according to the same data model.

Data manipulation

The principal issue here is to decide which operations should be provided at every server and which should be provided by a mediating “Manipulation Server.”

There are services which will continually expand over time and are relatively complex. Thus, they probably should be implemented in a “Manipulation Server” to reduce the overhead on the basic clients and servers:

- Reprojection—e.g., Platte-Care to Mercator
- Regridding—e.g., same projection, different resolution
- Geophysical plotting

Another group, on the other hand, includes services that might fruitfully reside in the minimal server, and if they were encoded in something like a plug-in module, they could be updated readily. Some of these services could reduce the amount of data to be transmitted:

- averaging
- summing

and others might be part of an effort to establish a certain degree of semantic consistency in the system's traffic:

- Scaling of values such that they are delivered in a consistent system of units; e.g., mks or cgs
- Conversion of time from varied representations
- Conversion of Earth coordinate systems
- Conversion of measurement units
- Conversion of missing values

(This latter group of services should probably be transparent to the user requesting the data.)

Implementing the IOOS Data Transport System

It is important to stress that much of what is being attempted here breaks new ground. Certainly the integration of the various components into the envisioned data transport system is novel, and no matter how hard we try to envision all the issues that will arise in so doing, we are likely to find as we assemble the various components that significant new issues must be addressed to bring the system to the level of functionality and operation that we desire.

We recommend a modular approach to the design of the system. As noted in the Introduction, a modular architecture will allow straightforward replacement of components as new technologies render older modules obsolete. It also will allow the overall design and implementation of the system to be undertaken in stages by several groups working in parallel. Furthermore, we recommend that the system be modularized along the lines presented in the Introduction, beginning with a discipline-neutral lower layer and working up from this layer. A significant advantage to making the layers discipline neutral where possible is that it broadens the user base of support for these layers.

To provide a base on which to begin building the complete transport layer, we recommend specifying an initial configuration for several of the lower system layers. This configuration will likely change with time, but adopting a base at this time will allow more rapid implementation of the system as a whole, i.e., it will allow us to undertake a suite of pilot projects built on this base that can move forward in parallel.

In this chapter we detail the specific approach that we recommend for the design and implementation of the IOOS Data Transport System.

THE DATA ACCESS PROTOCOL

We recommend adopting TCP/IP as an operational component for the IOOS Data Transport System at the lowest level. TCP/IP is in widespread use and is likely to remain the dominant low level transmission/internet protocol for the foreseeable future.

The next layer up in the Data Transport System must be associated with a data access protocol. We recommend adopting the OPeNDAP data access protocol as a pre-operational component for the data access protocol on which the transport layer will be built. OPeNDAP is described in some detail in Annex C and in Annex B is compared to other systems of which the Data Transport Team is aware. Given the current status of OPeNDAP and the reasons outlined in the introduction to this Chapter and in Annexes B and C, we believe that selecting OPeNDAP as a pre-operational component will gain IOOS one to two years of advanced development on the Data Transport System.

In addition to OPeNDAP[6], we recommend including the netCDF client and server, the HDF 4 server, the GrADS-DODS server, the Aggregation Server, the ASCII output capability, and the user support infrastructure as pre-operational components of the system.

We also believe that it is important to keep in mind that adopting OPeNDAP as a pre-operational component does not mean that OPeNDAP is a static piece of software or protocol. In fact, we believe that all elements of the system, regardless of their designation, will and must evolve in time. Hence, even operational elements will have pilot efforts associated with them that will address increased functionality, enhanced performance, etc.

All told, maintenance and evolution of the DAP is envisioned as a 7 FTE per year effort, hence should be budgeted at approximately $1,000k per year. This includes maintenance of the code, nightly builds of system components, low level user support, code documentation, and administration.

DEVELOPMENT EFFORTS AND PILOT PROJECTS

The pilot projects outlined in this section will address what we believe to be the most critical issue faced in the full-scale implementation of the transport component of the IOOS data system. What is learned from these pilots will feed directly into the initial implementation of the system. Early implementation will allow us to learn through system pilots (outlined in Section 4.2) focusing on some of the more problematic areas what the issues in those areas really are and the way forward in addressing them.

Choice of a data access protocol addresses only a small part of the overall data transport problem. To begin with, OPeNDAP mandates a rigid syntactic description of the data to be exchanged within the system, but it does not impose any semantic requirements on these data, although it does provide a mechanism for providing access to whatever semantic information is available for the data. It is, thus, a discipline-neutral exchange mechanism. In addition, the DAP does not impose any requirements on the structural organization of the data. The DAP operates in the Format Layer only (see the Introduction for a description of layers). In the following, we propose a number of pilot efforts that must be undertaken in the near future if the transport component of the IOOS data system is to be fully operational in the three- to five-year spin-up time associated with the IOOS effort. Although these are recommended as separate pilots, they can (and many probably should) be housed within the same organization, but each effort is sufficiently different from the others that

[6] By OPeNDAP we mean the protocol itself and the core infrastructure that implements this protocol. We do not mean the clients and servers based on the protocol.

independent groups (or individuals) addressing each is important. In addition to these individual groups there also needs to be a coordinating activity that addresses the integration of the components developed as part of these pilots into the system as a whole. Again, we note that integration of a component into the system as a whole does not mean that that component is being endorsed for use. It simply becomes a candidate for future adoption. At the same time, we firmly believe that it is only through actual implementation in the overall system that we will learn the most valuable lessons associated with a given component.

We expect that to coordinate the activities outlined below will require between one and two people per year. IOOS should budget on the order of $200,000 for this activity. A cautionary note here. This coordination activity is in addition to that associated with the maintenance and evolution of the DAP, with the general administration of the data transport effort and with broader user services, and with documentation needs of the transport component as a whole.

The following pilots are presented in rank order. In addition, for each pilot we assign a priority from 1 to 10 with 10 being the highest, an absolute must that needs to be undertaken immediately, and 5 being a task that needs to be done but is not critical at the outset, to 1 for a task that would be nice to have addressed, but…

Pilot Project #1: The Semantic Data Model – Priority 10

We recommend that a pilot project be funded to examine existing semantic data models and, based on this evaluation, either to choose one or to design one for the IOOS transport component. We see this as the highest priority development task that must be undertaken at present.

A semantic data model facilitates use of the data. Immediate work must begin on identifying and layering such a model on the DAP. Several such models either exist or are in development, but the Data Transport Team was not able to obtain sufficient documentation on them to make a firm recommendation for any one at this point. The results of this pilot, a preliminary design or the decision to adopt an existing system, should be available within one year of the group being formed. The group undertaking this pilot should work closely with OPeNDAP so that the model may readily be incorporated into the DAP.

The data model should also be constructed so that data sets that are not compliant with the data model at present can be made compliant in the future without modifying the data.

Paramount to the success of the IOOS data system is that this data model deal with physical data just as well as it does with biological and chemical data; hence, representation from both communities is a requirement for this pilot.

We envision this as a two- to three-person-year effort. This pilot should be funded in the $300k range.

Pilot Project #2: DAP-OBIS Integration – Priority 10

OBIS[7] is a globally distributed network of systematic, ecological, and environmental information systems. Data held in associated archives should be seamlessly integrated with those accessible via the DAP. This means that either a gateway be established between the DAP and OBIS or that the DAP replace the current OBIS data access protocol. We recommend a pilot to address the integration of OBIS and the DAP. We envision this task to be at the same level as that associated with developing a semantic data model. Indeed, a critical component of the semantic data model will be its ability to handle biological data.

This is seen as a .5 person year effort that will be spread over a year. IOOS should budget on the order of $75k for this task.

Pilot Project #3: The Structure layer – Priority 10

Immediately above the format layer is the structure layer, which may be divided into a purely syntactic part and a semantic part. Despite the variety of ways in which gridded data are organized, syntactic restructuring of gridded data is fairly straightforward. Furthermore, gridded data sets accompanied by a semantic description consistent from data set to data set may be reorganized into structures that have semantic meaning without too much difficulty. An aggregation server that restructures many gridded data types is currently in use with DAP-accessible data sets. This aggregation server does not handle all of the cases for gridded data that have been encountered, but it is easily extensible to many of those not covered. The point is that restructuring of gridded data does not point to any serious problems. The same is not true of non-gridded data, referred to as unstructured data, sequence data, or profile data (we use “sequence” in the following). The fashion in which sequence data are organized shows a great deal more variability from site to site than do gridded data. Hence, a general restructuring algorithm is much more difficult to design and implement. This task is made more difficult by the pronounced interest in aggregating sequence data between sites, an aim not so often encountered for structured data. Aggregating sequence data

[7] See Annex B for a brief overview of OBIS the Ocean Biographic Information System.

tends to be more in demand because it generally does not require the modification of data values and hence is less threatening to the data provider and to the data user. Aggregating data on different grids, on the other hand, is less often requested, because it requires not only reorganizing the data but also modifying data values.

The restructuring of sequence data is an area that is absolutely critical to IOOS because of the large number of sequence data sets that will be collected as part of the IOOS effort. Hence, we recommend a pilot focusing on this problem. This pilot should involve a group composed of those with experience archiving and using sequence data. This group should design the procedure that should be used in IOOS to restructure and aggregate sequence and, if appropriate, the group should design the actual restructuring server that may be part of the solution. As with the data model, the group needs to include those from all oceanographic communities—physics, biology, chemistry, and geology. The pilot should also address the aggregation of data from RDMSs. This group should also work closely with the data model group, because the data model is likely to be directly relevant to the solution. This is the second highest data transport task that must be undertaken. This group should have a beta version of sequence aggregation available within one year of appointment. This version will need to be vetted by the community. We envision this as a two-person year effort that will include significant voluntary contributions from the community. IOOS should budget on the order of $350k for the sequence aggregation server prototype with the expectation that part of the effort will involve a technical workshop focusing on this problem.

Pilot Project #4: OPeNDAP server for unsupported formats– Priority 10

Servers exist for many common formats. A significant volume of data do not, however, exist in commonly used formats. OPeNDAP provides two servers for dealing with such data, the JGOFS server and the FreeForm server. Unfortunately, it has become clear that neither of the servers is optimal. A pilot project needs to be undertaken to develop a more flexible server that combines the best features of both of the existing servers. The project will require 2 FTEs to complete and should be completed in one year. It should be budgeted at the $350k level.

Pilot Project #5: GIS-SIS Interoperability – Priority 8

Many of the end users of IOOS data will be GIS users, but most of the data being collected are collected and organized from what is often referred to as the Scientific Information Systems (SIS) perspective. Accessing data between SISs and GISs is difficult at best. There is a rudimentary effort in this direction associated with the DAP, but it is clear that more work is required in this area. We therefore recommend a pilot that will address GIS access to DAP-enabled servers and DAP-enabled client access to data stored in standard GIS formats. As a precursor to full GIS access, consideration should be given in the pilot to GIS access via an intermediate file capability. The ability to do this at what is thought to be a fairly low level of effort is why this task is not ranked at a higher priority. There is a simple, inexpensive solution that will enable GIS users to access data from DAP-enabled servers, although not at the full functional level of the system—direct access to DAP-servers from within the GIS. Developing servers for some of the more widely used GIS formats will be straightforward. GIS access to data from DAP-enabled servers will likely have to be undertaken on a GIS-by-GIS basis, as has been done for such scientific information systems as Matlab and IDL.

We recommend that this pilot be undertaken in two stages. In the first stage, the pilot needs to cleanly delineate what the issues are, which GISs should be targeted, the level of support (access) that is appropriate, and the cost of building the interface. In the second stage, the pilot needs to move forward with implementation. We anticipate that the first stage is a .25 to .5 person year effort and should be budgeted at $50k. NVODS' experience with ESRI suggests that if ESRI is to develop the GIS interface to the DAP for their ArcMap products it will cost on the order of $250,000 or more. We have no experience with what it will cost to provide the similar interfaces to other GISs. It may be, however, that an alternative solution to the GIS-by-GIS DAP client will be identified in the first phase of the pilot.

Pilot Project #6: Metrics – Priority 7

Metrics on the use of the system must be collected in order to evaluate system performance. Although some lower level metrics are obvious (e.g., number of requests, volume of data moved), experience with NVODS suggests that thought and attention beyond the obvious must be given to metrics. For example, it is important to know not only the basic numbers but also how requests are being made. This is especially true as the transport layer becomes more and more hidden. An inappropriately configured client could easily make inefficient requests that loaded the system down. These need to be discovered. To do so, the information gathered about data requests needs to be carefully considered and an analysis capability needs to be built that will ferret out potential problems. We recommend a two-year pilot to address this problem. We believe that this pilot should analyze the http logs from some of the more active sites currently serving data via the DAP to learn what the existing issues are. This would be followed by the development of a suite of metrics that should be collected and the design and implementation of a module that would work with the DAP to do this. The reason for a two-year effort here is that we believe that it will take on the order of one year to perform the preliminary analysis, to design and build a metrics gathering module, and to move this module out to a significant number of DAP server sites. The second year should be devoted to the collection, analysis, and possible refinement of the metrics-gathering module. The total effort is approximately a 1.5 person year effort. IOOS should budget $200k for this pilot.

Pilot Project #7: Data Discovery – Priority 6

As noted in the section on General Description of IOOS Data Transport System (p. 142), we believe that the data discovery portion of the system must be intimately coupled with the data access portion and that this coupling must be automated. We recommend a pilot effort to identify the basic issues related to the automated population of the system's overall data discovery services to be coupled with the data transport portion of the system and to propose a solution to this problem. This has a relatively high priority for the system. This capability will likely have to be built into the system's data servers and hence needs to be in place as system population begins. We have found in the NVODS effort that data providers are generally reluctant to install new servers shortly after they are available—the "Let someone else find the problems" reaction. We believe, however, that it is a fairly straightforward task requiring approximately six person months to complete—design, implement, and test—and should be budgeted at the $75K level. The individual(s) undertaking this task need to work closely with both the data transport group responsible for the originating data servers and with the data discovery group.

Pilot Project #8: Push versus Pull and Near-Real-Time Access to IOOS Data Streams – Priority 4

The need for near-real time access to IOOS data streams is central to the goals of IOOS. It has become clear in the various IOOS-related meetings that have taken place to date that both push and pull access to the IOOS data streams are desired. The DAP speaks to the pull capability, but does not support push. There appear to be two heavily used models for push, the GTE and the IDE (developed by Unidata). In that the IOOS data system must, at least at the outset, include a push capability, we recommend a pilot study that will assess existing push technologies and that will begin experimenting with the technology that results from this assessment as the most appropriate for IOOS. An important consideration in the selection of a given technology is that it be easily integrated with the system's pull component. There are two reasons for this. First, it is very likely that some users of the system will want to pull as well as to receive pushed data. It should be straightforward for these users to use both data streams, which means that the data should look syntactically and semantically similar. Second, to address stress on the system related to high-interest events, the sites to which data are pushed should also be considered as potential pull sites. This has medium priority for the system as a whole. This function involves integrating two low level system components. Hence, the sooner it is undertaken the better. Straightforward access to data within the system (in fact, access with nearly full functionality) is possible, however, using pull technology alone, a consideration which tends to decrease the priority placed on this element.

This is envisioned as a one person year effort based on the assumption that a technology already exists and simply needs to be selected and integrated with the pull side. IOOS should budget on the order of $150k for this pilot.

Annex A: The Intersection of Data Transport and Metadata

Metadata is information about data. Generally, when metadata are discussed, one is referring to information about the contents of the data, for example, the variable T refers to sea surface temperature and its units are degrees Centigrade or the data set covers the period 8 January 1982 through 29 May 1990. We prefer to take a broader view of metadata, dividing it into two basic groups: syntactic and semantic metadata.

Syntactic metadata is information about the data types and structures at the computer level, the syntax of the data, for example, variable T represents a floating point array measuring 20 by 40 elements. This is information that is required as part of the transport protocol for the data in a network based data system.

Semantic metadata is what one normally thinks of as metadata, information about the contents and context of the data set.

THE THREE-TIERED DATA SYSTEM

Earth science data systems are generally viewed as consisting of three primary levels:

- **The Directory Level** provides a list of data sets along with the parameters available and the approximate temporal and spatial coverage for each data set. An example of an entry in such a system: a hydrographic data set at the National Oceanographic Data Center (NODC).
- **The Inventory Level** provides a detailed listing of the data granules within a data set. For the NODC hydrographic data set, this might consist of a listing of each cast along with the location (latitude and longitude) and time of the cast.
- **The Data Level** consists of the actual data objects.

Directories have generally been maintained separately from inventories and from the actual data. Inventories, when they exist, are often found collocated with the data. Inventory access functions and data access functions, however, have been kept separate.

In the past, practical considerations related to limited networking and storage capacities encouraged this hierarchical view. These constraints are relaxing rapidly now, however, and so other structures are becoming practical. For example, in a totally distributed system, in which directory, inventory, and possibly data-acquisition functions are combined, directory and inventory information could be combined and maintained at the same site(s) as the data. There would be no distinction among levels in such a system. For the purposes of this discussion, however, we shall maintain the historical three-layer view (directory—inventory—data) simply as a device to help understand the issues involved.

Whatever choices are made in implementation of these three logical levels, system-wide interoperability remains exquisitely dependent on metadata.

SYNTACTIC METADATA AND THE DATA/METADATA TRANSPORT PROTOCOL

It is virtually impossible to make use of a data stream—a large collection of bytes—without a rigorous syntactic description of the data that are being moved from one place to another.

A data transport protocol requires a data model, an organizational description of the data as they are moved between client and server. The data model generally consists of data types (e.g., byte, integer, string) and groupings of these data types (e.g., arrays, lists). HTML is effectively a data model, albeit a very simple one, consisting of string data, metadata in the form of mark-up tags, and metadata indicating inclusion of external (opaque) content, for example, GIF images. The Hierarchical Data Format (HDF) is a much more sophisticated data model designed primarily for array data, although it has evolved to include sequences and complicated data structures. The OPeNDAP data model achieves its generality by encompassing a range of such underlying models through its extensibility: complex structures may be assembled from more basic structures. The OPeNDAP data model consists of data types (Byte, Integer, Short Integer, Float, String, and URL), and groupings of these data types (Array, Structure, Lists, Sequences, and Grids). It is used only as a transport protocol, not as a storage format as is HDF.

A data model may also be considered to include operations that may be performed on the data such as subsetting and projection. In HDF, these functions are part of the Application Program Interface (API). In OPeNDAP/NVODS, they are among the operations permitted by the servers in the system, for example, NVODS.

In general, the complexity of the data model increases as one moves from the directory level to the data level. At the directory level, the data model need not be more complicated than that used for HTML, while at the data level, HTML will clearly be inadequate. This means that if the data model adopted is rich enough to accommodate the actual data, it will probably also be able to accommodate information at the inventory and directory levels.

SEMANTIC METADATA

It is useful to distinguish among use metadata, metadata required to use a data set and typically transmitted with data, and search metadata, required to find data of potential interest and typically associated with the directory level.

This distinction is quite important because most metadata discussions center around search metadata requirements and do not use metadata requirements, for example, the Directory Interchange Format (DIF) of the Global Change Master Directory (GCMD). Use metadata and search metadata overlap, but one is not a subset of the other, for example, missing value flags are not required when searching for a data set, while to use the data such information is crucial. Similarly, the ranges of the variables in a data set are not required to use the data, but they form the basis for many data set searches.

Use metadata may be further subdivided into translational and descriptive use metadata. The former refers to operations that are performed on the data values, be they the names of the variables or the digital numbers associated with them, that are required for the user to understand their meaning. For example, the variable T maps to sea surface temperature or d x 0.125 maps to °C, where d is the number stored in the data set. Descriptive use metadata, on the other hand, tells about the data—how the instrument was calibrated or what sensor was used.

Search metadata may also be further subdivided, in this case into parameter, range, and descriptive search metadata. Parameter search metadata contain the list of parameters or variables in the data set. This could be further subdivided into dependent and independent variables. Range search metadata contain the ranges of variables within the data set. In most existing directory systems, only the ranges for time and space are included. Descriptive search metadata contain other information associated with the data set, such as a generic description of the sensor used. There may be overlap between this descriptive information and those contained in descriptive use metadata, although this need not be the case. For example, a description of the sensor may be relevant to both groups, but there is no reason to include detailed information about sensor calibration as descriptive search metadata.

Figure 1 shows the different metadata types schematically. Although three levels are shown in this figure, the inventory and data levels have been treated together. This is the approach that has been taken in OPeNDAP (i.e., inventory information is treated in the same way as data).

Interoperability at the data level with inventory access requires the squares with green and magenta backgrounds. The green square requires a rigid metadata description, while the magenta squares need not be as rigid. Descriptive use metadata are not required for interoperability at the data level.

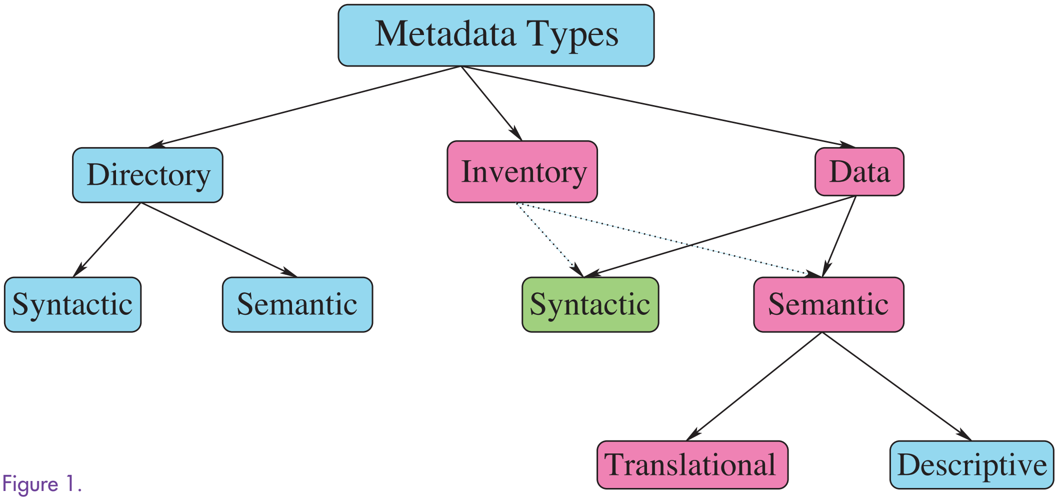


Figure 1.

Annex B: Overview of Existing Systems

Members of the Data Transport Team are aware of several data systems that address one or more of the issues raised in the Introduction for Earth science data. A brief overview of these systems is presented in this section in order to identify pieces of the systems that might be appropriate for adoption into the data transport component of IOOS.

OPENDAP/NVODS

The data access protocol of the Open source Project for a Network Data Access Protocol (OPeNDAP) forms the core of the National Virtual Ocean Data System (NVODS). NVODS was funded by the National Oceanographic Partnership Program as a step "toward an integrated ocean observing and prediction system (IOO&PS)" with the stated objective of "… develop[ing] concepts that maximize flexibility and utility of a hub-node system for the future."

The design of the OPeNDAP data access protocol (OPeNDAP, originally referred to as the DODS data access protocol) was based on two fundamental criteria: (1) servers must be easy to install, and (2) the system must interface to existing application software. A system that satisfies this basic philosophy will allow individual data collectors as well as national archives to be data providers, and it will allow researchers, operational modelers, interested hobbyists—everybody—to use familiar and appropriate software.

The OPeNDAP approach is to use the standardized interfaces defined by multiple file APIs (e.g., NETwork Common Data Format (NetCDF), HDF) as the point at which to insert the distributed data infrastructure. In this approach, existing applications—both commercial applications and those built within the science community—are "relinked" with new libraries that masquerade as the original file I/O library. The applications are unaware that they have been extended to perform network access. The data from remote files are made available through servers that invert the process—using the standard file or database APIs to read the files and then provide the data over the Internet in a format-neutral representation. The virtues of this approach are adaptability, leveraging, and invisibility:

1. The investment that each scientific project has made in its software tools is protected. Users continue to use the software tools with which they are already familiar—now extended to perform remote data access.

2. The approach leverages hundreds of already-existing, low-level file manipulation utilities. For example, utilities such as the NetCDF Operators ("NCO"), which subset, reorder, and append data from files immediately, become network tools for performing the same operations on widely distributed data sets.
3. Format independence is achievable through this approach. Applications communicate with files through standardized interfaces (APIs), without knowledge of what occurs behind those interfaces. Format translation may occur without the application being aware of it.

The first version of the OPeNDAP core software was released in 1995. There are currently in excess of 40 OPeNDAP sites in NVODS serving more than 300 data sets. A list of data sets accessible is available at: http://www.unidata.ucar.edu/cgi-bin/dods/data sets/data sets.cgi?xmlfilename=data sets.xml. In addition to these sites, the NVODS project has forged partnerships with a number of national and international oceanographic projects (e.g., WOCE/CLIVAR, AOMIP, US GOOS, GODAE, and NOMADS) as requested in the Broad Agency Announcement (BAA) on which the NOPP funding was based.

We believe that this recommendation is justified for the following reasons:

- The DAP is discipline-neutral. It is based on a small number of data types and organizations of these types. This has allowed servers based on the DAP to be built for all data types encountered thus far in the implementation of the system.
- The DAP is in wide use in both research and operational settings.
  - There are in excess of 40 DAP server installations in the United States, serving well over 300 oceanographic and meteorological data sets.
  - Most of the major ocean data archives are either experimenting with the DAP or use it routinely for their data distribution.
  - In addition to U.S. server installations, there are on the order of 10 installations overseas.
  - New installations are being added in the United States and oversees at the rate of one or more every three weeks.
  - The DAP has been adopted by the following national and international efforts as the base for their data transport needs:
    - GODAE
    - NOMADS
    - AOMIP
    - ESG II
- There is significant experience with the DAP.
  - The DAP has existed in its current form for approximately seven years.

  - There are currently on the order of 1,000,000 DAP accesses per month system-wide from more than 400 users.
- There is a rapidly growing community of system developers adding features to the system.
  - The DAP is being adapted for GRID computing by the HAO group at NCAR.
  - The group at COLA has developed a processing server referred to as the GDS. In addition to allowing in-line processing of the data, this server also provides access to GRIB data and will soon provide access to BUFR data.
  - Unidata has incorporated the DAP in their Java implementation of netCDF.
- There exists a wide range of client applications that are currently DAP-enabled (Annex C, p. 185).
- There exist servers for a significant number of formats that are in heavy use for oceanographic data (Annex C, p. 184).
- The DAP is supported on a broad range of platforms (Annex C, p. 184), and it has been implemented in both C++ and Java.
- The DAP has been vetted within the oceanographic community as part of the NOPP funded NVODS effort.

OBIS (OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM)

OBIS (Ocean Biogeographic Information System) is a globally distributed network of systematic, ecological, and environmental information systems, which have received funding from the Alfred P. Sloan Foundation, Office of Naval Research, National Oceanographic Partnership Program, and the National Science Foundation. OBIS is also the information component of the Census of Marine Life (CoML), a major international research program to assess and explain the diversity, distribution, and abundance of marine organisms throughout the world's oceans.

A Federation of Autonomous Members

OBIS is managed as a federation of database sources that agree on the means of achieving interoperability while respecting and preserving the autonomy of the member sources, each of which has complete freedom to choose the data format, data management systems, and semantic data model for its own site. Within OBIS, members typically associate themselves with one discipline-oriented interest group. This interest group develops and publishes standards to enable interoperability within that discipline. OBIS members are free to participate in other interoperation efforts and are encouraged to link to other data systems and programs.

Interoperability, Extensibility, and Flexibility

OBIS recognizes that interoperability at the semantic level must be supported, that semantics is discipline-dependent, and that a semantic federation core must be flexible and extensible. Interoperability is achieved through protocols and standards agreed upon by the members. OBIS' members have jointly defined a call interface and semantic data exchange format which serves as a federation core.

To become a provider, each federation member devises a schema that expresses the structure of the data and metadata within its database and also implements a method of mapping from this schema to the federation core and vice versa; the result of this work is to give each provider's database a uniform appearance from the outside. The providers are capable of understanding and satisfying requests couched in XML and transmitted via HTTP. This traffic typically happens between the OBIS provider and another, more central kind of site called an OBIS portal.

Portal functions may be implemented by any OBIS member who desires to do so. A likely configuration may be for each discipline represented in the federation to implement a portal in addition to the federation's main portal. Portals contain two distinct components: (1) a presentation layer and (2) PortalServices. The former is envisioned as an application server/Web server, and the latter is to handle all external network activity and decides upon, schedules, and issues requests to providers. Users, whether within or outside the federation, wishing to obtain information from OBIS direct their requests via HTTP (i.e., the Web) to a portal by filling out and submitting one or more query forms issued by the portal. The portal communicates with providers as needed to complete the query and returns the requested information to the user.

OBIS places considerable emphasis on the design of its semantics. On the one hand, multiple disciplines are represented by OBIS members, and each discipline has its own terminology and perspective, which OBIS has decided should be reflected in the semantic model. On the other hand, interoperability among disciplines requires that the federation have a semantic core capable of unifying the disciplines. OBIS takes the approach that each discipline should agree upon (1) a semantic data model for use within the discipline and (2) a method of mapping between the discipline semantic data model and the semantic core. OBIS assists this process by providing mapping tools for data model translation. This mapping is typically performed by the original data provider.

Disciplines are represented by interest groups within OBIS, and so OBIS can logically be seen as a federation of interest groups. The interest group for fishes, for example, works with OBIS to develop a standard module which contains a data exchange format and the necessary input/output software and which enables fishery data to be exchanged interoperably. Other interest groups have their own standard modules. All the modules, however, share certain concepts: geospatial, taxonomic, physical, chemical.

Standards for interoperability among modules are promoted by OBIS. Indeed, OBIS actively incorporates standards developed by such national and international standard bodies as the Open GIS Consortium (OGC) and the Taxonomy Database Working Group (TDWG).

Communication within OBIS and between OBIS and the rest of the world occurs via HTTP, and XML is heavily used.

The portal/provider software will be upgraded in March 2003 to a product of the open source software project Distributed Generic Information Retrieval (DiGIR), and the federation core schemas are available to all.

OGC/OGIS

The following was taken from the OGC website.

The OpenGIS Consortium operates to develop interoperable technologies involving geospatial information. The organization comprises two primary programs, the Specification Program and the Interoperability Program. The Specification Program operates through Technical (TC) and Planning (PC) committees to identify the standards required to foster interoperability among various groups (SIGs) within the geospatial information community. The Interoperability Program (IP) identifies and manages testbed and pilot projects that implement the proposed standards originating in the Specification Program, feeding the results of those activities back into revisions of the underlying standards. Combined, these two programs function to develop publicly accessible implementation specifications for the development of standards-based commercial off-the-shelf software (SCOTS) for the geospatial information community.

There are two important sets of OpenGIS Implementation Specifications grouped by service category:

OpenGIS Web Mapping Services: This is a family of specifications that enable servers to dynamically query, access, process, and combine different types of spatial information over the web with OpenGIS Specification conformant servers developed by other companies and organizations. To date, OGC has developed three Web Mapping Service specifications: OpenGIS Web Map Server Specification (Approved), OpenGIS Web Feature Server Specification (Candidate), and an OpenGIS Web Coverage Server Specification (Candidate).

[Note added by the Data Transport Team: The Web Map Server delivers pictures over the Web; it is not a data server. The specifications for the feature and coverage servers which are the OGIS data servers have not been finalized; at present, only candidate specifications exist.]

OpenGIS Geospatial Fusion Services: Non-map information—text, video, audio, digital photographs, mpeg movies, sensor data, word processing documents, etc.—often refers to place. It would be useful for many people in many situations to be able to "fuse" information such as addresses, place names, coordinates, pinpoints on photographs, and descriptive directions into one information management framework that would support search, discovery, and sharing of spatial information stored in non-map formats. This is the goal of OGC's "Geospatial Fusion Services (GFS)," which currently include: OpenGIS Gazetteer Service Interface (GAZ) Specification (Candidate), OpenGIS Geocoder Service Geocoder (GeoC) (Specification (Candidate), OpenGIS GeoParser Service (Geoparser, or GeoP) Specification (Candidate), and OpenGIS Location Organizer Folder (LOF) Specification (Candidate).

DATA EXCHANGE INFRASTRUCTURE (DEI) AND THE FIELD SPATIAL DATA MODEL (FSDM)

DEI began as a project within the Naval Research Laboratory. Physitron (Huntsville, AL), which is also a member of OpenGIS, accomplished much of the design work and helped develop the OpenGIS standards to which DEI adheres. To provide standardized data representation, DEI adopted the Environmental Protection Agency's Field Spatial Data Model (FSDM) developed by Todd Plessel. Largely due to fluctuations in funding support, DEI and the FSDM are in regular use only at the National Coastal Data Development Center (NCDDC), where a C++/Java implementation is the middleware for transporting data.

DEI is a set of CORBA-based interfaces used to locate and transfer geospatial data. It is a client/server architecture in which the server, or gateway, is responsible for (1) translating a request for data into a data source's access method to acquire the data and (2) putting the resulting data into a representation that is then ingested into the FSDM. That representation is called the Field Data Markup Language (FDML) Streamer, which is transported (via CORBA's IIOP transport protocol) to the client, which translates it into FSDM internal objects. Note that the data, unless they are relatively small, are not transported at this time. The data are transported from the gateway as needed by the client.

The FSDM consists of approximately 140 classes of objects that describe, contain, and/or manipulate geospatial data. One of the key features of this model is the separation of data storage from its geometry and topology. The geometry and topology information are contained in an object called a Mesh and the data values are contained in a Data object; a Field object contains the relationships that bind the Mesh and Data objects together. Where the Mesh object contains simply a sequence of locations related one-to-one to an array of values in the Data object, there is some similarity to the Grid object of OPeNDAP, which contains exactly that: a sequence of locations and then the associated values. The FSDM, however, elaborates upon the basic pattern in several ways.

First, metadata are contained within (or closely associated with) both the Mesh object and the Data object, so the Field object is an entry point to more information about the data than just their values and points of origin. The Field object, along with the objects it binds, seems capable of providing the information needed for "interoperability with semantic meaning."

Second, the Mesh object contains "cells," which are spatial objects such as Points, Lines, and Hexahedrons. These cells include implementations of geometric "behavior" including facilities for forward and reverse transformations through projections (e.g., Lambert Conformal-Conic) and definitions of the coordinate systems (e.g., Polar).

Third, there are issues currently under study of one-to-many and many-to-one relationships between Data and Mesh components and of arbitrary, user-defined data types for Data values.

A desired effect of the Field object having an abstract interface which hides internal data representation and storage is that there can be a consistent appearance to all geospatial data, regardless of local choices about physical data storage.

Further study of this system, especially of FSDM, seems warranted, as it seems to embody a genuine attempt to conjoin geospatial data and metadata in (or beneath) one object and to present a consistent view of geospatial data, such as will be required in the mature data transport system of IOOS. The long-term utility of FSDM would likely be increased if it were to prove amenable to extension beyond solely the geospatial context to, for example, biologic, taxonomic, spectral, and finite element modeling contexts.

SRB

The Storage Resource Broker is a generic data management infrastructure that supports digital entities. It treats each digital entity as an atom from the viewpoint of a logical name space. However, the SRB does support all of the traditional latency management functions required for wide area networks, including:

- partial file read/write
- streaming
- caching
- replication
- staging
- aggregation of digital entities into containers
- aggregation of metadata into XML files
- aggregation of I/O commands into remote proxies

The aggregation of I/O commands allows the user to specify a series of operations that require knowledge of the encoding format/data model. The operations are performed at the remote storage system to minimize the number of I/O commands that are sent over the network.

Similarly, the aggregation of metadata into XML files is done through the application of templates at the remote storage system. The templates specify how a digital entity can be parsed to identify relevant metadata attributes, which can then be shipped in bulk over the network for ingestion into a database.

Since the SRB deals with digital entities, it is possible to register files, directories, databases, URLs, SQL command strings, etc. into the logical name space. One can manipulate the organization of the logical name space independently of the physical names used on the actual storage systems. This freedom makes it possible to assemble a collection that spans multiple administration domains/sites/storage systems.

The SRB can be used as generic collection management software, independent of the structure of the digital entities. The challenge comes in when latency management functions are invoked that require knowledge of the structure of the digital entity. Hence, the desire to encapsulate the knowledge that is required to manipulate a digital entity as a remote proxy or template that can be moved to the remote storage system for application.

SRB also is used to interact with databases. The results of the application of an SQL query can be formatted as an XML file or HTML file for presentation. The system has been used to extract metadata records from remote databases.

SRB containers also have been used to aggregate digital entities before storage into archives. This makes it possible to store large numbers of small files in an archive without overloading the archive name space. Simultaneously, this can result in minimizing the number of tapes onto which the digital entities are stored, and minimizing the latency for the retrieval of a large collection.

Thus, the issue of structure applies not only to the internal structure of a digital entity for latency management, but also to the external structures that are used to aggregate data and information.

CONTRIBUTIONS OF EXISTING SYSTEMS TO THE MATURE IOOS DATA TRANSPORT SYSTEM

It is probably fair to say that the goal of providing networked oceanographic data transport and retrieval through "machine-to-machine interoperability with semantic meaning" has not been achieved by any existing system discussed above. Indeed, it may not be possible to point to any discipline for which such a system is in place. The target that this Data Transport Team has set, therefore, will be a novel creation. Fortunately, the desire for a data and information interchange facility of this power is not utterly unfamiliar: OPeNDAP/NVODS, OBIS, OGC/OGIS/BMS, DEI/.FSDM, and SRB are each the product of very considerable thought and effort to develop services that represent some aspect of the mature IOOS data transport system.

At the end of Section I, five characteristics of IOOS's mature system were listed:

1. format translation (e.g., from server archive to client application),
2. consistent structures for datatypes, in fact or potentially,
3. consistent semantic form, in fact or potentially,
4. freedom to distribute storage of related information (e.g., data vs. metadata, over multiple sites),
5. freedom to employ different semantic models for equivalent information.

The following describes how the five existing systems we have examined may contribute to these five aspects.

Format translation

In theory, format translation could be the province of an all-knowing, mediating site capable of accepting data in the native format of any registered archive and transforming it into any registered application's format. Such a design is both delicate and complex. It would be subject to overloading, if it broke the data transport system would stop, and each added input or output format would increase algorithmic complexity exponentially (number of translation methods equals number of archive formats times number of application formats).

An alternative approach is to suppose the existence of a mediating format capable of representing all input and output formats and to distribute the responsibility of being able to translate from native to mediating format, and vice versa, to every server and client site respectively. This is the tack taken by OPeNDAP and used with success in the NVODS community, and by envisioning "machine-to-machine interoperability" rather than "machine-to-mediating-server-to-machine interoperability," this IOOS team endorses this second format translation topology.

The idea underlying OPeNDAP's treatment of numerical data formatting is not that a single mediating format can express all server and client formats, but that a single grammar can be devised that describes and stipulates a universe of mediating formats, at least one of which can be translated to and from the formats used by any server-client combination. This grammar is modeled on that used to declare programmatic datatypes and variables in such modern computer languages as C++. As it happens to be true that a depth-first tree traversal of any instance of this grammar produces an unambiguous, unique interpretation, then fidelity of transmission can be attained. It can be argued that the grammar is both terser and more familiar to those accustomed to looking at program source code than other grammars such as XML, although an instance of OPeNDAP's grammar could be contained within (or referenced by) an XML-encoded document.

Presently, OPeNDAP does not enforce any uniformity upon transmitted data other than using its predefined basic datatypes to represent values. How the values are organized is left open, and can be any well-formed instance of the grammar, conveyed as a Data Description Structure (DDS). As precisely equivalent information can be stored in different formats, and commonly is, OPeNDAP will transmit these equivalent data sets somewhat differently, too. A time series of observations of two variables collected at a fixed location would be transmitted as a sequence of pairs of variable values or as two sequences of one variable value each, depending upon how the data were archived. Work remains to be done on whether, where, and how to specify that such data sets be transmitted identically. The principal motive for such an effort would presumably be to minimize the complexity and workloads of clients while avoiding as much as possible the need for a mediating, format-translating server. In general, servers would assume the responsibility for translating archived data into agreed-upon configurations.

OGC/OGIS/BMS has concentrated on designing and implementing specialized mediating servers, and the datatypes it has used to date tend to be restrictive, compared to the needs of IOOS.

DEI's FSDM provides an interesting contrast to OPeNDAP. The separation of location information (under the Mesh object) from variable values (under the Data object) is: (1) unlike OPeNDAP's Grid datatype, which contains a sequence of locations followed by the variable values, but (2) reminiscent of OPeNDAP's separation of data values (the DataDDS), grammar instance (the DDS), and grammar-instance-structured metadata (the DAS). Perhaps the thinking that has gone into the design of FSDM can inform an effort to integrate metadata more gracefully into IOOS's system than OPeNDAP does at present, particularly metadata applicable to a data set as a whole. Conversely, the OPeNDAP grammar may suggest ways to extend FSDM's capabilities well beyond its current restrictive geospatial (location-keyed) data.

Syntactically consistent delivery of equivalent data

In OPeNDAP, one can imagine syntactic uniformity being imposed upon equivalent information by a software module inserted at each server between the module that extracts archival data to its closest OPeNDAP DDS-grammar representation and the module that transmits the DDS/DataDDS/DAS family. This interposed module would be capable of examining a DDS and recognizing whether it was an instance that could be re-expressed in a conventionally agreed upon, informationally equivalent form, and then performing the transformation. The problem of designing such a module could be made easier by first ascertaining whether, as seems likely, most requests are for such simple data structures as arrays involving one or two variables in four-dimensional space-time and then concentrating on recognizing such cases.

DEI's FSDM should be examined to see whether it has dealt with the issue of reorganizing the syntactic structure of data.

Potential or actual remapping of metadata to a consistent semantic form

OBIS is organized as a federation-about-a-core, and the concentration on metadata has resulted in effort being spent on devising mappings between metadata structures adopted by federation interest groups and that of the core. Apparently there is an ongoing effort to devise a general solution to this remapping problem; the Data Transport Team should stay apprised of progress in this endeavor.

OGC intends to support sharing of spatial information with its Geospatial Fusion Service (GFS); presumably attention will be given to a common format, and the Data Transport Team should become well-informed about decisions and developments in this area. Even if an OGC format is not adopted, IOOS will have to develop a gateway at least for sharing information with OGC.

It is not clear whether DEI/FSDM faces the problem of remapping syntactic and/or semantic structures, as it seems that its specification may impose uniformity upon participants from the outset. However, this effort has worked out serviceable semantic structures in some detail, and the data Transport Team should be certain to mine its specifications for good ideas and features.

Distributed storage of data and metadata of a data set and its coordinated retrieval

In OpeNDAP, perhaps fortuitously, metadata and data values are held and transmitted in two separate entities, the DAS and the DataDDS, with the underlying structure of both held and transmitted in a third, the DDS. Current practice has these three identified by URLs that differ only in their filename suffix, signifying that all three, in the absence of some redirection at the target site, are stored on the same system. However, there seems to be no compelling reason why each could not reside at a distinct site or why there could not be multiple files at multiple locations to make up the DAS entity, for example. It seems likely that OPeNDAP can be extended to continue to serve in the foundation of the mature IOOS system because of this inbuilt separation of value from syntax from semantics, an example of a good initial design decision paying unforeseen dividends.

If OGC coordinates metadata from multiple sites, it seems likely to do so by using a mediating server.

The similarity between the underlying tripartite nature of OPeNDAP and FSDM suggests opportunities for each to help sharpen our vision of the mature IOOS system and to reuse thoughts, algorithms, and perhaps even code from both. For the sake of argument, erase any memory of the earlier comparison of OPeNDAP's Grid datatype to FSDM's Field object, with the Grid's location sequence being analogous to the Mech object and the variable values to the Data object. Instead, compare the Field object to the DDS, the Data object to the DataDDS, and the Mesh object to the DAS. FSDM may suggest ways in which the OPeNDAP portion of the early IOOS system can be brought to the level of expressiveness required in the mature IOOS system. Perhaps the DataDDS will acquire translational use metadata, as the Data object apparently does. Perhaps the DDS or a root-DAS will become a directory to metadata distributed among multiple sites and capable of being updated and added to over time by users of the data, as the Mesh object can contain more varied information than does DAS as presently employed.

SRB comes into its own in this aspect of the mature IOOS system: it embodies much thought and effort at making intentional, coordinated use of a capability provided by the World Wide Web, that is, that a collection of identifiers of things available over the Web (designated by URIs and URLs) can be constructed and manipulated, modulo synchronization issues, as though they were the things themselves, provided you have sufficient information about them, that is, metadata. Perhaps the SRB collection is analogous to the FSDM Field object or the OPeNDAP DDS. Perhaps the collection groups all the metadata associated with a data set before or after aggregation and/or subselection. Perhaps the operations envisioned for OPeNDAP/NVODS mediating servers, for example, such as the Aggregation Server or the Restructuring Server, can be viewed as manipulations of SRB collections and their components.

Free choice of semantic models for metadata storage and retrieval

The fifth requirement, that IOOS servers and clients be free to use different semantic models for the same information, makes a point which may at first seem to contradict some implication of the third requirement, which says that information which is potentially or in fact equivalent either could be or is transmitted identically. To the extent that two design goals are sought (1: identical transmission of equivalent data, and 2: preference for direct, unmediated server-client transactions), the third goal implies that each server will be able to determine whether requested information could be expressed in a conventionally agreed upon form and transmit it accordingly. The fifth goal, however, speaks to the question of the structures imposed upon metadata and means that users ought to be able to ask for metadata, that is, information about a data set, in more than just one way and get their answers as well assembled in various patterns. The implications of this desideratum are not entirely clear; all that is evident is that the Data Transport Team would like it to happen.

OPeNDAP's contribution to this issue may be its emphasis upon tailoring software libraries to the expectations of client applications. Up to now, most effort has been directed toward moving numerical data that originates as subsets of archived data sets from a remote server into a local application such as MATLAB^{TM}. There is no apparent reason, however, why metadata could not be downloaded even from multiple remote sites into a local database, for instance, in a similar fashion, there to be queried in any way the user likes. Presumably software would trace query results and maintain links to the data, not yet downloaded, at remote sites, perhaps not even the same sites as served the metadata.

NVODS already contains gateways to distributed archived information; as these increase in number and sophistication, users should be able to choose one or more of these mediating sites that suit their purposes and preferences for query and response formats.

OBIS explicitly concentrates on facilitating storage of and access to distributed metadata.

OGC's requirements for metadata storage and query are not known at present; DEI/FSDM's are unknown, too.

SRB's experience may well be relevant here; each installation of SRB requires access to a database. It may be that functions already exist to download and query a collection's metadata; this Team will become better informed about SRB's relevant capabilities.

Annex C: OPeNDAP/NVODS

For several years, a data exchange network, the Distributed Oceanographic Data System (DODS), has connected scores of oceanographic data servers and an even greater number of clients. Because the data access protocol underlying this network is discipline-neutral and has begun to be adopted by groups in other disciplines, a nonprofit corporation called the Open source Project for a Network Data Access Protocol (OPeNDAP) has been formed to maintain and evolve the data access protocol. The overlying features that are specific to the exchange of oceanographic data have been encapsulated in the National Virtual Oceanographic Data System (NVODS). Together, OPeNDAP/NVODS is for all practical purposes equivalent to DODS.

In the following, "OPeNDAP" refers to OPeNDAP's data access protocol itself.

THE OPENDAP DATA MODEL

OPeNDAP's task is to ensure that all data in numeric form can be interchanged between arbitrary data repositories and users. It does not assume the task of exchanging, for example, multimedia (images, video, audio, etc.). OPeNDAP was designed based upon faith that all numeric data storage formats could be replicated and all client application data needs could be met by using a three-part information transfer from server to client:

- the DDS, which would express the structure of the numeric data, that is, syntactic metadata,
- the DAS, which would contain all relevant semantic metadata, configured according to the structure expressed in the DDS, and
- the DataDDS, which would contain the numeric data itself in the linear form generated by a depth-first traversal of the structure expressed in the DDS.

Thus, the success of OPeNDAP depended upon the adequacy of the DDS to express any arbitrary data storage scheme.

Acting on the supposition that it was unlikely to encounter a data storage format that could not be expressed in a modern programming language, and since C, FORTRAN, and Lisp data declarations should be able to express all the likely possibilities for data storage format components, the OPeNDAP DDS provides 13 data structures (eight simple types and five compound types) analogous to the ones in these languages:

- SIMPLE TYPES
- 8-bit unsigned integers; characters
- 16-bit signed integers
- 16-bit unsigned integers
- 32-bit signed integers
- 32-bit unsigned integers

- 32-bit floating point numbers
- 64-bit floating point numbers
- strings
- COMPOUND TYPES
- structures
- arrays
- lists
- sequences
- grids

plus two datatypes relevant to the Web environment; these are considered to be strings and thus to be simple datatypes:

- URLs
- pointers to URLs

(There is also the "data set" datatype, which is used to wrap the entire data declaration; one DDS contains one "data set.")

For more details on these data types see section 6.3 of the DAP users' guide: http://www.unidata.ucar.edu/packages/dods/user/guide-html/guide_58.html.

The potential complexity of the data structures expressible in a DDS is suggested by the following:

- Structures can contain simple, array, list, sequence, and grid datatypes as well as other structures;
- Arrays can contain simple, structure, sequence, grid datatypes;
- Sequences can contain simple, array, list, structure, sequence, and grid datatypes;
- Grids contain arrays.

To date, the DDS has been able to express all formats used by actual servers and clients; the bounds of its expressiveness have hardly been challenged.

QUERIES TO A SERVER

OPeNDAP is not a mechanism for transferring files as such. Instead, it a mechanism for transferring the information contained in files by using the DDS, the DAS, and the DataDDS, and this requires that the server open and parse the datafile containing the information to be transmitted. Since the server will be processing the contents of the datafile, OPeNDAP was also designed to permit the client to instruct the server to return only data which fit the conditions in a query from client to server.

Two basic processes are triggered by a query: projection and selection.

Projection specifies which variables are to be returned to the client.

Selection specifies which conditions the values of a returned variable must meet.

Projection

Projection is quite simple. The names of the desired variables are written in the query as a comma-separated list.

Selection

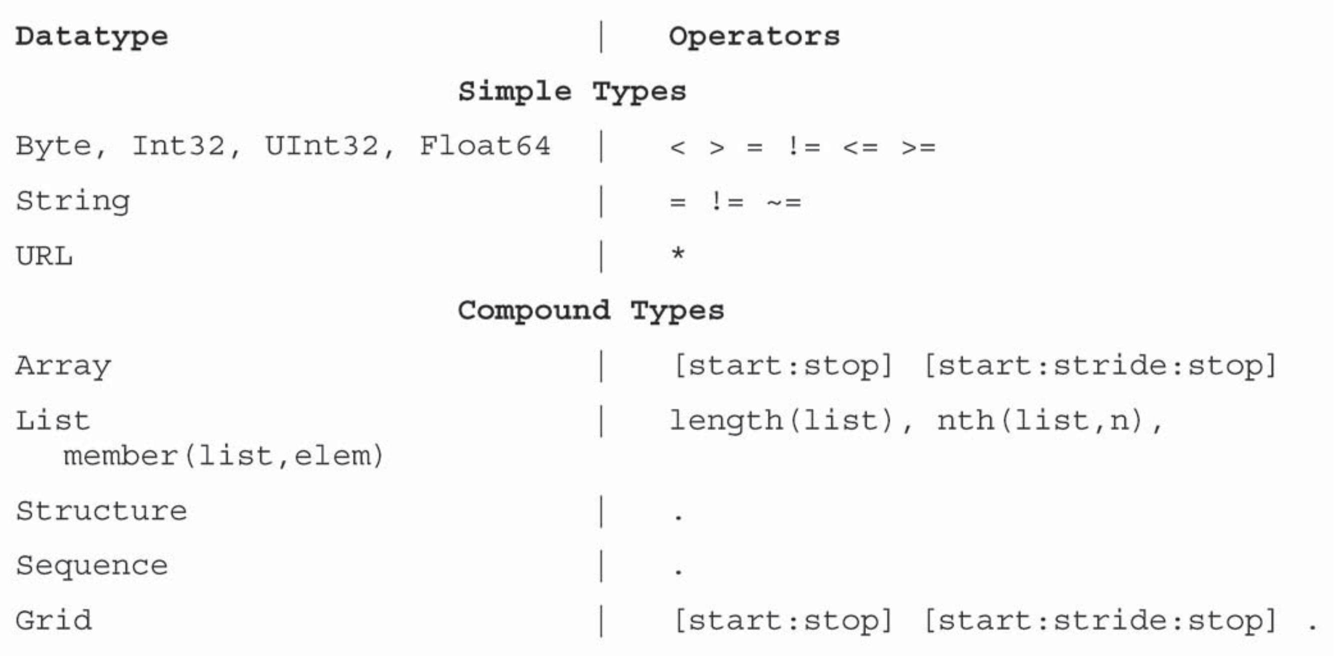
Selection can become quite complex. Ranges can be selected from within arrays, or they can be sampled at regular intervals. Values can be accepted or excluded on the basis of comparison tests. Servers can implement functions and advertise their availability. Then, clients can use these functions in their selections. For example, if a server implements statistical functions, a client could request the moments of a range of data. The full power of OPeNDAP queries has probably not been regularly utilized.

Since every OPeNDAP request takes the form of a qualified URL, every OPeNDAP-compliant server must be prepared to interpret certain features of these request URLs. The first characteristic that must be interpreted is the filename extension, the token that follows the last period in the filename. A client requesting the DDS associated with the stored data set specified by the filename adds ".dds" to the actual filename when sending out the URL. When requesting the DAS, ".das" is added. When requesting the data set itself, ".dods" is appended; if the URL has no extension, then ".dods" is assumed as the default. There are special-purpose request extensions as well.

HTTP URLs may contain query data which are appended to the pathname following a question mark and passed to the server. In OPeNDAP, the question mark and the query data immediately follow the request extension just described. Currently, such queries only make sense following a DataDDS request, that is, the ".dods" extension (or no extension). These queries can direct the server to look inside the datafile and return a subset of the variables (projection) and values (selection) contained there.

Details of the constraint expression which is used to convey the projection desired from the client to the server may be found in section 4.1 of the Users' Guide: http://www.unidata.ucar.edu/packages/dods/user/guide-html/guide_32.html.

The user may select subsets of the data; basic access to variables can be modified using operators. Each type of variable has its own set of selection and projection operators which can be used to modify the result of accessing a variable of that type. The permissible operations associated with each data type are listed below:



```
Datatype                       |   Operators
                         Simple Types
Byte, Int32, UInt32, Float64   |   < > = != <= >=
String                         |   = != ~=
URL                            |   *
                         Compound Types
Array                          |   [start:stop] [start:stride:stop]
List                           |   length(list), nth(list,n),
  member(list,elem)
Structure                      |   .
Sequence                       |   .
Grid                           |   [start:stop] [start:stride:stop] .
```

For more details on the data type see section 6.3 of the DAP users' guide: http://www.unidata.ucar.edu/packages/dods/user/guide-html/guide_58.html.

The operators listed above have the meaning defined by ANSI C except as follows:

- the array hyperslab operators are as defined by netCDF,
- the string operators are as defined by AWK, and
- the list operators are as defined by Common Lisp.

Two of the operators deserve special note. Individual fields of type constructors may be accessed using the dot (.) operator or the virtual file system syntax. If a structure "s" has two fields, time and temperature, then those fields may be accessed using s.time and s.temperature or as s/time and s/temperature. Also, a special dereferencing "*" operator is defined for a URL. This is roughly analogous to the pointer-dereference operator of ANSI C. That is, if the variable my-url is defined as a URL data type, then my-url indicates the string spelling out the URL, and *my-url indicates the actual data indicated by the URL.

FUNCTIONALITIES OF OPeNDAP/NVODS

The motivation behind communication protocols tends to be the desire to provide the most functionality and convenience to the greatest number of users for the lowest overhead imposed upon the parties individually and in total. Certainly, OPeNDAP's design and implementation were guided by this desire. The design goal of bringing data transfer functions to all OPeNDAP/NVODS clients and yet not requiring them or all the system's servers to become complex ("thick") is being achieved by giving the responsibility for fulfilling certain kinds of client requests to just a few servers. The organization of this section reflects this dichotomy:

- Section 1 discusses services which contribute to the "thickness" of every OPeNDAP/NVODS client and server;
- Section 2 considers services which are presently implemented at selected servers.

The next three sections list platforms, formats, and applications currently in use and supported by OPeNDAP/NVODS:

- Section 3: Platforms and operating systems
- Section 4: Server-side data storage formats
- Section 5: Clients (applications)

The remaining section examines one aspect of OPeNDAP/NVODS which is in flux:

- Section 6: XML

Universal minimal requests and facilities

By "universal minimal requests and facilities" we mean those requests that every OPeNDAP-compliant client can make in every appropriate request to any OPeNDAP-compliant server with the full expectation that the request will be satisfied by that server without participation of an intermediate OPeNDAP server.

Since all communication from client to server takes the form of a qualified URL, these universal services depend on the expressiveness of that URL. As noted above, the client, by using various filename extensions, is able to ask for the DDS, the DAS, the DataDDS, an information page about the server, and help documentation from the server. A query (constraint expression) added to the DataDDS request can trigger projection and selection by the server. Since the constraint expression can invoke functions implemented on the server and made available to clients, the constraint expression is actually a powerful tool.

Mediated facilities: a form of distributed computing

Data aggregation: The "Aggregation Server"

Recall that requests from a client to a server are by qualified URL and that in its usual form, a URL refers to a single file. When the data that a client wants reside in more than one file, we use "data aggregation," which is any process whereby a data set is generated by joining in some manner data held in more than one data set, probably in more than one file and possibly at more than one site.

Aggregation is presently supported by the OPeNDAP/THREDDS Aggregation server. This makes it possible to combine multi-file grids and arrays into single file data sets. For more details on the types of aggregations supported by the existing Aggregation Server see: http://www.unidata.ucar.edu/projects/THREDDS/tech/AggServerConfig.html#aggTypes

The existing Aggregation Server does not support the reordering of axes, e.g., x(i,j,k) ==> x(j,i,k), and by the reasoning laid out in this Plan, it should not. Such a mapping would be the province of a Restructuring Server.

Platforms/operating systems currently supported by OPeNDAP/NVODS

The following are already supported:

- Intel x86: Linux 2.2 and 2.4 (Red Hat 6.2, 7.1, 7.2, 7.3, plus other Linux distributions that use the 2.2 or 2.4 kernels; i.e., virtually all of them).
- SGI/IRIX 6.5
- Sun/Solaris 2.6 (but that runs on 2.7 also)
- Dec Alpha: OSF 4.0f
- Intel x86: Windows2000, NT, XP

Server-side data storage formats currently supported by OPeNDAP/NVODS

Given that the IOOS data access protocol will be based on open source software, nothing restricts development in supporting data storage formats. Nevertheless, certain data storage formats are widely enough used in oceanography that they warrant explicit support in the IOOS effort. The first items in the list are currently supported by NVODS (or by NVODS affiliates); following these are items which IOOS should probably support.

* Supported by groups not funded by NVODS. Included here for completeness.

- netCDF
- HDF 4
- HDF 5
- HDF-EOS
- SQL
- JGOFS - Data system developed for U.S. JGOFS
- DSP - U. Miami satellite analysis package
- FreeForm - Developed at National Geophysical Data Center
- Matlab
- GrIB *
- BuFR *
- CDF *

Clients currently supported by OPeNDAP/NVODS

OPeNDAP/NVODS currently is implemented on these clients.

- ArcView
- EASy
- netCDF-supported apps
- Ferret
- GrADS
- Matlab
- IDL
- others?

XML

A development effort is presently underway to use XML to preserve the DAS and DDS of data sets; in fact, the physical distinction between the DAS and the DDS is being eliminated in the new XML version of the metadata containers. For some time, the DAS and DDS will be available in their present forms and will coexist with the XML versions.

The benefits also can move to the XML community. OPeNDAP adds a strategy for encoding binary data within XML. OPeNDAP provides a rich existing infrastructure and concepts under development—standard APIs, THREDDS style catalogs, applications already linked, LAS, ...—OPeNDAP adds lots of value to the basic XML starting point.

Annex D: OpenGIS Basic Service Model Review

INTRODUCTION

The following review of specifications targets the OpenGIS Basic Service Model (BSM) as the basis for comparison with OPeNDAP/NVODS. The BSM defines a suite of services providing equivalent functionality for the Geographic Information System (GIS) problem domain as might be required in an IOOS data system.

The target for the BSM are GIS users and applications. Input data is abstracted into Layers and Features consistent with the expectations of GIS users and applications. All data must be geo-referenced, with the implicit assumption that the various BSM servers are capable of performing spatial reference system (SRS) transformations from the data's native SRS to the client's requested SRS. In response to a client's request, each BSM server returns its output response formatted in a well-known binary (WKB) representation, though the set of WKBs differs between BSM server types.

For the sake of brevity, this review will focus on the basic interoperability approaches taken by the OGC BSM and OPeNDAP/NVODS. Documentation for the various OGC specifications described in this review are available at the OpenGIS website, http://www.opengis.org/.

OPENGIS ORGANIZATIONAL STRUCTURE

The OpenGIS Consortium operates to develop interoperable technologies involving geospatial information. The organization comprises two primary programs, the Specification Program and the Interoperability Program. The Specification Program operates through Technical (TC) and Planning (PC) committees to identify the standards required to foster interoperability between various groups (SIGs) within the geospatial information community. The Interoperability Program (IP) identifies and manages testbed and pilot projects that implement the proposed standards originating in the Specification Program, feeding the results of those activities back into revisions of the underlying standards. Combined, these two programs function to develop publicly accessible implementation specifications for the development of standards-based commercial off the shelf software (SCOTS) for the geospatial information community.

OGC BASIC SERVICES MODEL (0.0.8)

This review will focus on the Basic Services Model (OGC Document 01-022r1). The OGC Basic Service Model comprises four servers, Map, Coverage, Feature, and Registry. The primary distributed computing platform (DCP) for all servers is HTTP, though the Feature server provides DCP specifications for CORBA and OLE/COM. Currently underway in the OGC/IP is a Web Services testbed (OWS), which is working to develop the technologies required to evolve the existing BSM servers into web services using SOAP, WSDL, and UDDI. However, the underlying BSM approach to data interoperability should not substantively change upon completion of the web services initiative.

Following is a brief description of each server, taken directly from the BSM Draft Candidate Specification document.

The Web Map Server (WMS) generates "pictures" of georeferenced data, independent of whether the underlying data are simple features (such as points, lines, and polygons) or coverages (such as gridded fields). The WMS produces an image of the data that can be directly viewed in a graphical web browser or other suitable picture viewing software.

The Web Feature Server (WFS) offers access to the geographic features (points, lines, polygons) in a data store. A "basic" or "read-only" WFS implements operations to describe and retrieve features, while a "transaction" WFS also implements operations to lock and modify (create, update, delete) features.

The Web Coverage Server (WCS) offers access to the actual numeric values of gridded georeferenced data or imagery.

The Web Registry Server (WRS) is a catalog of OGC Web Services. It is a "stateless" catalog in that it relies on the single request/single response mechanisms of the HTTP DCP. The WRS supports registration, metadata harvesting and descriptor ingest, push and pull updates of descriptors, and discovery of OGC Web Service types and instances.

Currently public Implementation Specifications exist only for WMS and WFS; the WCS and WRS activities remain in draft candidate status.

INTEROPERABILITY APPROACH

The basic interoperability approach within the BSM comprises four elements.

1. Each service type within the BSM provides a well-defined request interface that provides a consistent, abstract representation for the underlying data or service. All data is abstracted into named layers and provide a set of sample dimensions for subsetting operations. The minimal set of sample dimensions is the layer's latitude/longitude bounding box.

2. All data are explicitly geo-referenced. Each server (WMS, WCS, WFS) must advertise the supported spatial reference systems that the service instance is capable of returning the data in. All servers must support the standard geographic reference system (WGS84).

3. Each service type within the BSM produces its response formatted in predefined, well-known binary formats (WKB). The WKBs differ between server types but are defined to meet the requirements for the particular use.

4. Each service type within the BSM has an integral advertising capability. In response to a GetCapabilities request, each server will return an XML-encoded response document listing the layers available, and operations the service instance is capable of performing on them.

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

Vision

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.

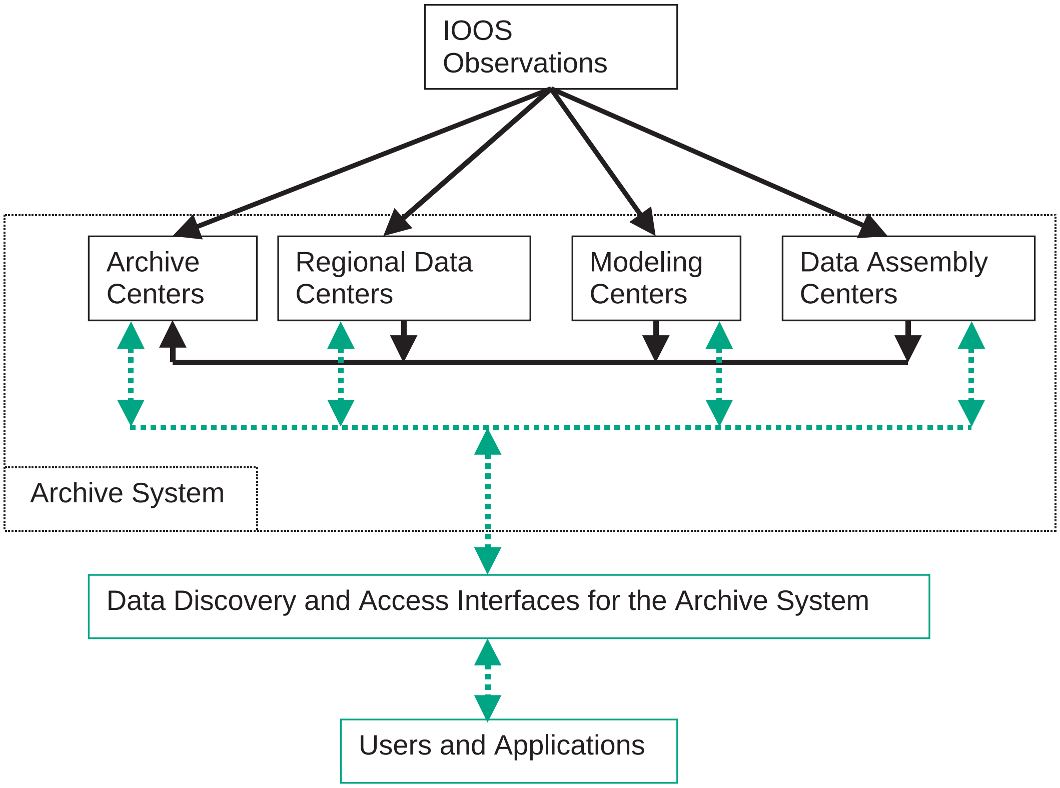


Figure 1. Primary archival (solid lines) and access (dashed lines) data flow within the DMAC Archive System of IOOS. Not shown are the secondary bi-directional archival data flow between all the centers and IOOS observations flowing directly to users and applications external to the Archive System.

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 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.

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,

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	• Raw, ancillary satellite observations • Instrumental measurements • Biological samples • Model reanalyses • Complex merged data analyses	Two
Replaceable	Derived from irreplaceable data, can be regenerated through systematic processing	• Calibrated satellite radiances • Simple composites or analyzed data	One
Perishable	Real or near-real-time data; typically replaced by higher-quality data	• Direct broadcast satellite data • Operational analyses • Quick-look analyses based on uncalibrated or incomplete data	One
Virtual	Data provided through on-demand processing	• Subsets from GUI • Analyses from a Live Access Server	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.

- 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 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.

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				

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.

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 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.

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: (Data set Adaptation Cost) × (1 – Structure Factor),
- Annual Maintenance: 15% × (Hardware Cost + Staff Implementation),
- IOOS Start Up: All first year costs for IOOS data sets,
- IOOS Maintenance Per Year: (Annual Maintenance) × (1 – 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

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.

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).

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).

I. Interoperable Data Discovery, Access, and Archive

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

Part III. Appendices

Appendix 4. User Outreach

IOOS DMAC User Outreach Team

March 2005

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

Contents

User Outreach Team Members

Philip Bogden, Chief Executive Officer, Gulf of Maine Ocean Observing System, Portland, ME

Carol Dorsey, Microbiologist, Alabama Department of Public Health, Mobile Division Laboratory, Mobile, AL

David L. Eslinger, Oceanographer, NOAA Coastal Services Center, National Ocean Service, Charleston, SC

Larry Honeybourne, Water Quality Program Chief, County of Orange Health Care Agency, Environmental Health Division, Santa Ana, CA

Mark Luther, Director, Ocean Modeling and Prediction Lab, University of South Florida, College of Marine Science, St. Petersburg, FL

Michael McCann, Monterey Bay Aquarium Research Institute, Monterey, CA

Roy Mendelssohn, Operations Research Analyst, NMFS/Pacific Fisheries Environmental Laboratory, Pacific Grove, CA

Phillip R. Mundy*, Science Director, Gulf of Alaska Ecosystem Monitoring and Research Program, Exxon Valdez Oil Spill Trustee Council, Anchorage, AK

Malcolm Spaulding, Professor of Ocean Engineering, University of Rhode Island, Narragansett, RI

Margaret Srinivasan, Altimeter Applications Lead, Ocean Science Element, NASA/Jet Propulsion Laboratory, Pasadena, CA

Joseph J. Tamul, Jr., Cooperative Program Director, Oceanography Department, Naval Oceanographic Office, Stennis Space Center, MS

Suzanne Van Cooten, National Data Buoy Center, NOAA, Stennis Space Center, MS

* Team Leader and Report Editor

Introduction

The User Outreach Team was formed in June 2002 to serve the U.S. Integrated Ocean Observing System (IOOS) Data Management and Communications Steering Committee (DMAC-SC). The DMAC-SC was formed in March 2002 at Airlie House, Warrenton, VA during a national meeting convened by Ocean.US, the national office for integrated and sustained ocean observations. The Steering Committee is working to implement the DMAC vision for the data and communications component of the IOOS. The Airlie House workshop defined an ambitious process for developing a detailed, phased implementation plan to make this vision a reality. The first step in this process was to establish a DMAC Steering Committee, whose responsibilities include oversight of the planning process, oversight of initial pilot projects, guiding the Expert and Outreach Teams, and writing the plan.

PURPOSE OF THE TEAM

The User Outreach Team serves as support to the four DMAC expert teams, to help them define and refine their products in terms of user-defined issues that ultimately will become data system requirements. The makeup of the outreach team, listed above, was balanced with respect to the subject areas defined by the seven Airlie objectives outlined below. The User Outreach Team members are "scientists whose disciplines are data intensive (i.e., modelers) or who interface with other users (some scientists, some not) who need products based on the IOOS "data stream," such as oil spill trajectories and maps of natural hazards."

The User Outreach Team fulfilled two primary immediate roles: (1) produced the Community Issues Lists and (2) in the process of compiling the lists, served as a communications clearinghouse for the other teams on identifying user groups, getting feedback from user groups, and identifying their issues and related requirements for the system. It is essential for the DMAC to be in touch with the definition of "community" and "issues" from the outset of its work. The entry-level definitions of community and issues are directly derived from the seven objectives for the IOOS identified in the Airlie Conference:

1. improve the safety and efficiency of marine operations (marine operations),
2. more effectively mitigate the effects of natural hazards (natural hazards),
3. improve predictions of climate change and its effects on coastal populations (climate change),
4. improve national security (national security),
5. reduce public health risks (public health),
6. more effectively protect and restore healthy coastal marine ecosystems (Coastal Marine Ecosystems), and
7. enable the sustained use of marine resources (sustained use).

The entry level issues with respect to the DMAC concern data management and information transfer aspects of the top 30 or so key variables (e.g., ocean temperature, ocean salinity) identified by the Airlie process. Other issues will be provided by User Outreach Team members, under the responsibilities outlined below.

DUTIES OF THE TEAM

The duties of the User Outreach Team are summarized as follows:

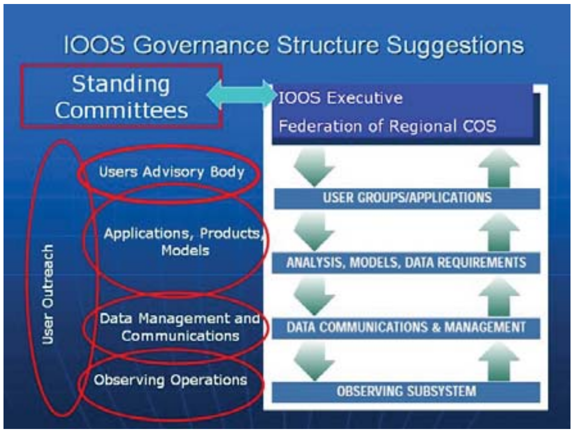
1. Community Issues Lists—Primary responsibility is to serve as a point of contact to solicit inputs to the DMAC Plan for the designated community, as captured in the Community Issues Lists. Together with white papers prepared by the other non-outreach teams, the Community Issues Lists will form the first draft of the user outreach section of the DMAC report. Issues are to be focused on the top 30 variables identified at the Airlie conference, as a starting point. Each team member is to edit, refine, and prioritize her/his list, as a reflection of the inputs received from the community.

2. Recommend to the DMAC-SC a list of requirements of users, represented by the seven communities of the seven objectives established at the Airlie conference.

3. Keep Team Leader informed of progress on a regular basis.

4. Make recommendations to the DMAC-SC on a structure that ensures ongoing communications between IOOS DMAC managers and user groups—identifying new user needs and providing feedback on any identified inadequacies within the evolving system.

What is the definition of users for the purposes of DMAC? IOOS is, by definition, user-driven (see Figure 1), and user groups were identified, to an extent, during the process of putting the Airlie meeting together, and by the choice of the seven goals, above. Based on these beginnings and the experience of operational programs like Rutgers' Long-term Ecosystem Observatory's (LEO) and Coastal Ocean Observation Laboratory public service web site known as the "COOLroom" (www.thecoolroom.org), there is a dynamic pool of end users, fishers, recreationalists, and private companies that support commercial marine transport and other marine industries that is beyond the reach of the DMAC Plan writing process. This pool is out of reach because it changes depending on circumstances, such as natural marine disasters, and changing threats to national and seasonal activities of users. It is also out of reach at the moment because of the limited time available to this phase of the DMAC process.



IOOS Governance Structure Suggestions

Standing Committees

IOOS Executive
Federation of Regional COS

Users Advisory Body

Applications, Products, Models

Data Management and Communications

Observing Operations

User Outreach

USER GROUPS/APPLICATIONS

ANALYSIS, MODELS, DATA REQUIREMENTS

DATA COMMUNICATIONS & MANAGEMENT

OBSERVING SUBSYSTEM

Figure 1. Recommended committees and advisory bodies (inside ovals) necessary for Ocean.US to implement the end-to-end, user-driven ocean observing system.

The users for the DMAC purposes are found at the Data Communications and Management and Analysis Models and Data Requirements levels (Figure 1). People at these levels are either (1) data management (IT) professionals trained primarily in computer science who are concerned with Data Communications and Data Center (and Product) Management, or (2) other scientists whose disciplines are data intensive (i.e., modelers) or who interface with other users (some scientists, some not) who need products based on the IOOS "data stream," such as oil spill trajectories and maps of natural hazards. One task of the Facilities Team is to work with outreach to the first category (IT professionals), whereas the User Outreach Team is primarily to be focused on the second category (seven scientific and technical communities).

METHODS

Members of the User Outreach team were selected for expertise in one of the seven communities defined by the IOOS goals, as identified in each of the seven Community Issues sections. The starting point for the definition of user needs was the 2002 Airlie House workshop, where approximately 100 national experts in the seven communities of IOOS met to define needs and priorities for the U.S. observing system. Within each community issue, members were encouraged to consult with as many other members of the community as possible, taking into account the short time frame available for this portion of the planning process. Different methods were used by different community groups, so no standardized approach was applied to develop the community issues lists. Overall, due to the extensive use of the Internet, it is estimated that 1,500 to 2,000 concerned individuals were contacted, informed of the IOOS needs, and given the opportunity to comment, across all seven communities. Nonetheless, the approach to development of user-based requirements for the data system is an iterative process, where broad sections of the concerned communities will have future opportunities to review and comment, and the lists will be broadened and deepened as a consequence of this ongoing review process.

Community Issues

MARINE OPERATIONS

Team Members: Mark Luther, Phil Bogden

Introduction

One goal of the Integrated Ocean Observing System is to improve the safety and efficiency of marine operations. The composition of this community of consumers includes users involved in near-shore port operations, as well as mariners operating in open-ocean conditions near the coast. Harbor pilots who are responsible for maneuvering large ships and tankers in dangerous waters have been especially strong supporters of widely available real-time observations. Their needs are mirrored by other users operating at sea, such as commercial fishermen, recreational boaters, commercial shippers, the U.S. Coast Guard, and others. These groups are interested in forecasts, but they are much more interested in immediate access to buoy data. They seem less trusting and consequently less interested in model forecasts. The Coast Guard is apparently a big user of the real-time data for planning their own sea-going activities, as demonstrated by web-site hits, but we do not believe they are using information about real-time ocean currents (either from models or from HF radar) for Search and Rescue.

Issues

Port Operations

Users: Harbor Pilots, Ship Masters, Port Authorities, Shipping Agents, Shipping Companies, Ship Yards, Tow boat operators, Dredging contractors, USCG Marine Safety Office

1. **Timing of slack water for safe maneuvering of ships in harbors**—Many vessel maneuvers cannot be made except near slack water (currents less than 0.1 m/s or 0.2 kts.). Having real-time current measurements available to users (primarily pilots and masters), rather than relying on tidal predictions, widens slack water window in which maneuvers can be made safely. In the United States the vanguard program for delivering real-time oceanographic data to mariners is the Physical Oceanographic Real-Time System (PORTS), which dates to 1991. PORTS is a public information acquisition and dissemination technology developed by the National Ocean Service (NOS) in cooperation with the Greater Tampa Bay Marine Advisory Council. In Tampa Bay, the pilots state that having real-time current information available from PORTS aboard ship widens the slack water window from 1-2 hours to 3-4 hours for approaches into Port Manatee and Old Port Tampa, both of which have channel entry paths that perpendicular to the main tidal flow.

2. **Real-time water-level and density data to estimate under keel clearance**—Large bulk carriers often are loaded to the minimum under-keel clearance. Availability of real-time water-level data allows for more efficient use of vessel draft. In areas of highly variable fresh water inflow and salinity, real-time data on temperature and salinity also are useful in computing vessel draft. Published estimates are that one foot of additional draft for a bulk carrier is worth $66,000 to $250,000 (depending on cargo) in additional revenue per transit. In Tampa Bay, during the five years prior to the installation of PORTS, there were 35 ship groundings. In the five years after PORTS became operational, there were 14 ship groundings. The Tampa Pilots Association states that the majority of this decrease in groundings was attributable to the availability of real-time water level, wind, and current data. A single grounding can cost hundreds of thousands of dollars in lost revenue, ship operation costs, tug boat fees, hull damage, and environmental damage. Costs can be much higher if the hull is breached and hazardous cargo is spilled.

3. **Meteorological and oceanic conditions (waves and currents) for collision avoidance**—Availability of real-time current, wind, and water-level data aid in collision avoidance by giving pilots and masters better estimates of vessel set and drift and better estimates of maneuvering room in passing or overtaking situations.

Greater availability of more accurate predictions and observations of current, water level, winds, temperature, and salinity will aid in all of the above.

Coastal Operations

Users: Mariners of all types

1. **Search and Rescue (SAR)**—USCG needs accurate trajectory simulations (hindcast and forecast), with some probability distribution, for persons in the water or vessels in distress. Present estimates of trajectory (based on available winds and tidal models) can be misleading in places like the Gulf of Maine where poorly estimated low-frequency currents can dominate tides and wind drift. Chances of survival in the waters of the Gulf of Maine are almost negligible after 2 hours in the water, so a fractional improvement in SAR effectiveness could save many lives per year. A conservative NOAA cost/benefit analysis put the potential savings at six lives/year and $24M/year for marginal improvement in the Gulf of Maine alone[1].

[1] See: Economics of a US Integrated Ocean Observing System, Prepared by Hauke Kite-Powell, Charles Colgan, Rodney Weiher. Airlie House, 2002. http://www.ocean.us/documents/docs/BAKDOC9_Economics.doc. See also: An Economic Case For An Integrated Ocean Observing System, NOAA Magazine, 2002. http://www.noaanews.noaa.gov/magazine/stories/mag71.htm

2. **Recreational Boating**—Huge user group in Florida and in the Gulf of Maine, as examples—Recreational boaters are interested in general availability of current information, water level, winds, temperature, and salinity for a variety of different reasons related to port operations. They use all of the data they can get their hands on. Competitive sailors use wind and current information to determine tactics during races. Fishermen use wind, current, water level, and wave information to determine the best fishing spots or even whether to go fishing. In terms of number of hits or calls to the Tampa Bay PORTS, this is the largest user group.

3. **Real-time open-ocean meteorological and oceanic buoy data coverage for safe operations**—Coastal waters in the Gulf of Maine require pilots to maneuver tankers and merchant ships over tracks of hundreds of kilometers of open-ocean conditions. Pilots need real-time buoy data because ocean and weather conditions vary rapidly on scales that remain unmeasured by the NDBC network of buoys and C-MAN stations. They use the data for trip planning and performing safe operations at sea. Existing NDBC buoys are too distantly separated, and C-MAN stations don't provide waves. These users have grown dependent on the enhanced spatial and temporal coverage provided by GoMOOS buoys.

4. **Commercial fishing and trip planning**—Anecdotal information from fishermen of various types in the Gulf of Maine (e.g., scallopers, ground fishermen, lobstermen) indicates that many mariners either don't believe or don't trust weather forecasts, and they use the last 12 hours of real-time data (whenever available) to determine their ability to go to sea. User testimony indicates that the enhanced coverage of GoMOOS buoys allows fishermen to determine the location of weather fronts and whether it makes economic sense to go to sea on any particular day.

5. **Real-time measurements of fog for trip planning for large and small vessels**—GoMOOS provides visibility measurements from its buoys. Mariners of various types have reported that the visibility measurements are accurate and helpful for planning a variety of sea-going activities. USCG representatives have indicated that fog data can influence SAR response (e.g., aircraft and sea-going vessel needs), but we're not aware that USCG in the Gulf of Maine is actually using the visibility data in this manner right now.

6. **Commercial and Recreational Fishing and Sea-Surface Temperature**—Tuna fishermen, for example, are knowledgeable about the relationships among ocean temperature, productivity, and fish location. They use AVHRR data and estimates of front location to plan trips, and desire access to more and higher-resolution data. Such data products underlie the business model for some private companies, and there has even been objection and legal action to prevent federally funded groups from providing this information for free.

7. **Hazardous Material Spills (HAZMAT)**—HAZMAT activities needs accurate trajectory and dispersal simulations/predictions for most efficient deployment of containment/clean up resources. Accurate map-based data on locations of sensitive/endangered natural resources are also needed.

8. **Forensics for law enforcement**—Forensic experts need accurate trajectory hindcasts to determine probable point of origin of bodies found in the water (the authors of this report section have been contacted by law enforcement officials two or three times in the past regarding cases like this).

9. **Trip planning and forensics for ship operations**—Ocean Routes, Inc. has based its business model on meeting identified needs for weather information and oceanographic conditions (e.g., waves) along planned (or past) ship tracks. The company's focus has been on open-ocean conditions, but applications of this kind of service in near-shore regions remains untapped. Pilots in the Gulf of Maine are using the real-time data for planning, but might also make use of data products and services provided by companies such as Ocean Routes.

10. **Data Availability at Sea**—Mariners emphatically want "dial-a-buoy." GoMOOS and NDBC have partnered so that data from both the Florida and Gulf of Maine buoy systems is now presented on NOAA's dial-a-buoy service where mariners can use cellular telephones to access current sea conditions.

NATURAL HAZARDS

Team Members: Malcolm Spaulding and Suzanne Van Cooten

Introduction

Information on the communities concerned with Goal 2 of IOOS, i.e., more effectively mitigating the effects of natural hazards, was gathered by means of an email message (Annex A) sent to two email server lists, coastal_list@udel.edu and mem@appsci.com. The first list has approximately 800 subscribers and reaches most of the coastal engineering community, while the second has 400 subscribers and is targeted to the marine environmental modeling community. Both lists are international in scope, but the majority of the subscribers are from the United States.

Respondents to the natural hazards solicitation were requested, at a minimum, to provide the following: (1) brief explanation of natural hazards that would benefit from an IOOS, (2) the principal community of interest and their characteristics, (3) principal data variables that are required, (4) issues of concern or attributes that are critical to the application (e.g., timely access to data, ease of access, accuracy). This section summarizes the responses received from this email survey on each natural hazard that would potentially benefit from IOOS. A record of individual responses is presented in Annex B. The vast majority of the responses received concern storm impacts on coastal resources. All other natural hazards, including tsunamis, received substantially less input, and so responses are summarized for two groups, storm impacts and tsunami hazards. Storm impacts are generated by the phenomena of storm surge, wind, and wave, which act on coastal areas and structures (buildings, infrastructure—roads, power lines, sewer and water distribution systems—beaches/shorelines, drainage systems, groins, breakwaters, piers, and bulkheads). Other natural hazards are functions of tsunami generation, propagation, and run-up.

The communities of interest for storm impacts include property owners, residents, or users of the impacted coastal area. In addition, those people with a financial or personal interest in structures and infrastructure subject to damage, such as roads, bridges, marinas, ports, and harbors, are clearly concerned about mitigation of coastal hazards. Included in this group are beach users, boat owners, and businesses dependent upon the impacted area.

Also concerned with storm impacts are governmental entities with responsibilities for insuring, maintaining, regulating, or protecting the groups outlined above. These include the federal (Federal Emergency Management Agency [FEMA], Army Corps of Engineers, U.S. Coast Guard, NOAA), state, and local (police, fire, sanitation, permitting, health, etc.) governments. Coastal states often have a variety of agencies with responsibilities paralleling their federal counterparts.

Communities concerned with mitigation of storm impacts need to receive timely information on variables of interest for storm impacts (all with time), such as water elevations (near shelf break and in selected coastal areas where impacts are expected to be significant), and directional deep (near shelf break) and shallow water wave heights and periods. Associated with the provision of wave height data is an additional need to provide the capability to validate wave models and account for local wind and wave transformation effects on the shelf and nearshore area. Storm impact variables of interest also include wind and atmospheric pressure measurements at selected offshore stations, mapping of pre- and post-storm impacts on shorelines and near shore areas (beaches, dunes, cliffs, coastal wetlands).

The principal communities of interest for tsunami hazards are similar to those for storm impacts, but more limited in geographic scope. Other natural hazards communities include property owners, residents, or users of the coastal areas prone to tsunami-generated waves. Those people with a financial or personal interest in structures and infrastructure subject to tsunami damage, such as roads, bridges, harbors/ports, and marinas, are clearly concerned about mitigation. Included in this group are beach users, boat owners, and businesses dependent on the facilities or beaches. In addition, governmental entities with responsibilities for insuring, maintaining, regulating, or protecting the property owners, residents, and users are concerned with tsunami impacts. These vary from the federal (FEMA, Army Corps of Engineers, U.S. Coast Guard, NOAA), state, and local (fire, police, sanitation, permitting, health, etc.) government.

Although variables of interest for storm impact mitigation are certainly relevant to mitigation of tsunami hazards, additional variables are required for tsunami hazard mitigation, such as water elevations (in network surrounding tsunami generation areas, in selected coastal areas where coastal impacts are expected to be significant), and horizontal water displacements measured from moored buoys with GPS. In addition, information on directional shallow water wave heights and periods are required, which need to provide the capability to validate wave run-up models. Run-up models are critical in areas where impacts are likely and important (e.g., harbors). Other required information needs for mitigating tsunami hazards can be derived from a network of hydrophones deployed in areas of potential tsunami generation to provide early warning of events and identification of the tsunami source. Finally, mapping of the impact of tsunamis on shorelines and near shore areas (river deltas, beaches, dunes, cliffs, coastal wetlands, marinas, harbors) is essential.

Issues

Storm Impacts Issues

- Measurements must be converted to useful products (easily understood data, maps, and graphs) and distributed by communications channels (the internet (web pages), radio, TV, warning systems) that reach those at risk (home and business owners) and response personnel promptly and often.
- System must provide accurate information, be highly reliable, and provide real-time access to observations and to forecasts (every hour for the next day).
- Data must be archived for use in future hindcast studies and research on fundamental coastal processes.
- Need to develop multiple self-contained wave and water level gauges in coastal states with more numerous pre-established mounts that would be deployed in the event of a storm to measure wave and water-level conditions that will likely impact coastal structures.
- Data standards and appropriate conversions need to be developed for key variables (water levels, waves) to ensure that the data are consistent and comparable. This is particularly critical for wave data where different sensing systems can provide substantially different results.

Tsunami Hazards Issues

- Measurements must be converted to useful products (easily understood data, maps, and graphs) distributed by communications channels (the Internet (web pages), radio, TV, warning systems) that reach those at risk (home and business owners) and response personnel promptly and often.
- System must provide accurate information, be highly reliable, and allow real-time access to observations and to forecasts every few minutes for the next 12 hours. Note that most tsunamis impact shorelines very quickly after they have been detected on the continental shelf.
- Critical to have accurate information on shelf and nearshore bathymetry to ensure accurate run-up forecasts.
- Critical to have as much information as possible about the source (i.e., landslide, volcano) and its location as tsunami wave conditions (wavelength, amplitude, and directionality) are strongly dependent on the source characteristics.
- Data must be archived for use in future hindcast studies and research on fundamental tsunami generation and propagation processes.

CLIMATE CHANGE

Team Members: Mike McCann (MBARI) and Margaret Srinivasan (JPL)

Introduction

Another stated goal of the Integrated Ocean Observing System is to improve predictions of climate change and its effects on coastal populations. The composition of this community of data users includes research scientists, modelers, climatologists, GIS data system users, and policy makers, particularly in coastal communities. This last group of users includes city governments, harbor districts, port authorities, county and state governments, planning commissions, and consultants. Global and long-term climate considerations are of particular interest to this user community. As an example, coastal development proposals in the Monterey Bay, California area must take into account expected sea level rise due to global warning estimates. Better data and improved access to data can improve these estimates.

For climate change research, the top priority is having high-accuracy data that are consistent for long-time-series data sets. Understanding the implications for long-term archiving of data as the technology advances is also a key element in successful data management and usage. Both data and metadata must be updated to reflect the current best solution to most effectively manage the delicate balance among new technology, existing hardware resources, and personnel resources, not only for current research but for archiving and storage as well. Solutions should endeavor to be as system-independent as possible, while realizing that other constraints may exist. Evaluating the status of data holdings should be a continuous process .

The issue of metadata is also an important element, primarily to data managers, but also to data users. The science that comes from the data is the ultimate resource for the research effort, but the ease in data management affects every phase of climate change research from acquisition to scientific results.

Issues

1. **Data Accuracy**: Future mission requirements should incorporate improved data accuracy. Sometimes there is a trade-off between data accuracy vs. data latency. Sufficient resources need to include research and re-processing efforts that would improve the quality and accuracy of data measurements.

2. **Consistency**: Data products need to be as consistent as possible for follow-on missions. As data accuracy improves, so do the geophysical algorithms. Therefore, resources need to be available for "data engineering" and re-processing efforts to provide data that will have identical geophysical models across missions. An excellent example of this is the SSM/I Pathfinder data available from Remote Sensing Systems. One of the problems with this particular data set is that the processing algorithm is not in the public domain, but is a closely held model "secret."

3. **Quality control**: Assessing the quality of a data set is difficult. It is often not known what sort of quality control a data set has been subjected to. Some methods of quality control often require knowledgeable personnel and significant amounts of human intervention.

4. **Modeling**: the application of models in (near) real time, like in AOSN II, will provide an interesting forcing function on improving some of these capabilities, since automation and access are so critical to correcting models in (near) real time. Problems with linking models and data are ones of time and space scales. Models are generally coarse while data can have very fine granularity.

5. **Data archiving**: Long-term data archiving, for future missions as well as heritage missions, should be a high priority. Both data products and expertise should be maintained, so that as algorithms improve the data can be re-processed 10 to 20 years later. Archive the original source observations (level 1b data) to enable data users to return to this level to fix problems.

6. **Format**: Sharing and blending of data sets from all the entities that collect and archive oceanographic data is difficult because these entities use different data formats and standards. In addition, improved subsetting engines will allow users to quickly access the data in different formats and regions. Without this capability, full use of the data sets will be impeded.

7. **Error Statistics**: In order to fully use the many Earth science data sets, attention must be paid to the necessity of providing error statistics and/or quality of information with the data sets. This involves careful thought to the quality information provided with each data set. Such strategies will need to be developed in conjunction with any metadata models. This is a major point in the SST effort of merging data sets from different satellites.

8. **Data management**: For data collection systems, there are no standardized systems or processes for up-front collection and management of important sensor, instrument, and platform metadata. Better automatic observatory data management requires this kind of metadata data management.

9. **Data cataloging**: Knowing what data sets exist, in what format (raw data, near-real-time data), and their appropriateness for a specific use is difficult because most oceanographic data sets are not cataloged. Being aware of data resources such as those in the Global Change Master Directory (http://gcmd.nasa.gov).

10. **Metadata**: The use of metadata, or rather the lack of it, impacts operational use, processing, analysis, archiving, QC, visualization, access/integration/reuse, and subsetting of data, and the automation of all the above. Better metadata definitions will improve all of these functions, in many cases by orders of magnitude.

11. **Data Discovery**: Not all climate data sets are registered with the Global Change Master Directory (GCMD). Data providers do not always realize the benefits of taking a few moments to do this. More generally, they don't always see beyond the needs of their immediate user community.

12. **GIS Access**: Typical GIS are not well suited to climate studies. They do not handle time or elevation properly and don't generally give direct access to the numerical values. The major GIS vendors need to be made aware of these needs.

NATIONAL SECURITY

Team Member: Jack Tamul

Introduction

Another goal of the Integrated Ocean Observing System is to provide information to support improved national security. National security may be broadly defined to encompass not only the protection of U.S. persons and interests, but also the promotion of the economic and social interests of the U.S. government and its citizens. Using this broader definition, many of the other themes of the IOOS have aspects that can be considered as contributing to national security. However, the broader national security will not be explicitly handled here. Instead, the scope of the National Security theme will be limited to the military's missions of war-fighting, peacekeeping, and humanitarian assistance. It includes maritime national security interests around the world, in every ocean, as well as maritime homeland security.

The oceans profoundly affect those whose job it is to ensure national security in the maritime environment (e.g., the Navy, Marine Corps, and Coast Guard). Knowledge of the ocean makes for better decision making and employment of people, platforms, and systems, increasing their effectiveness, and decreasing risks to those resources. This knowledge is used both operationally in the planning and execution of military missions, and by researchers supporting the development of new national security capabilities. "Operational" refers to those data and products for which availability is assured for time frames needed to support practical decision-making.

It is anticipated that a number of the elements of the IOOS will be useful in addressing a variety of national security issues. For example, a network of coastal radars would not only support the prediction of waterborne contaminant movement, but could also be used for port security and tracking ship traffic. Additionally, a robust U.S. coastal component of IOOS will enable the U.S. Navy to use the U.S. littorals as "surrogates" for denied areas in order to assess its coastal prediction and forecasting capabilities through data deprivation and forecasting experiments and exercises.

The variables and products required from IOOS to further national security interests are grouped below by issue.

Issues

National Security Issue 1

Improve the effectiveness of maritime homeland security and war-fighting effectiveness abroad, especially in the areas of mine warfare, port security, amphibious warfare, special operations, and antisubmarine warfare.

- Product NS-1.1: Estimates/predictions of near-surface currents on hourly to seasonal (i.e., climatological) time scales.
- Product NS-1.2: Estimates/predictions of near-bottom currents on hourly to seasonal time scales.
- Product NS-1.3: Estimates/predictions of tidal-period sea level/water level and velocity fluctuations.
- Product NS-1.4: Estimates/predictions of near water clarity on hourly to seasonal time scales.
- Product NS-1.5: Estimates/predictions of sediment transport on hourly to seasonal time scales.
- Product NS-1.6: Estimates/predictions of acoustic performance, especially on the continental shelf on daily to seasonal time scales.
- Variables required for National Security Issue 1 include:
  - 3-D Vector Currents
  - 3-D Water Temperature
  - 3-D Salinity
  - 3-D Suspended Sediment (for density)
  - Flux estimates of momentum, heat, moisture/freshwater, and radiation. Usually these are provided by NWP models. There is a need for verification by observations, such as:
    - Wind Vectors
    - Water temperature
    - Air temperature
    - Humidity
    - Long-wave radiation
    - Solar radiation
    - Precipitation amount
    - River discharge
  - Wind Vectors
  - Water Temperature
  - Air Temperature
  - Humidity
  - Long-Wave Radiation

- Solar Radiation
- Precipitation Amount
- River Discharge
- Bathymetry
- Sea Level/Ocean-Sea Surface Height
- Bottom Characteristics (type, vegetation, sediment composition and thickness, acoustic stratigraphy)
- Ambient Noise
- Nutrients
- Bioluminescence
- Optical Properties
- Ocean Color
- Surface Roughness

National Security Issue 2

Improve safety and efficiency of operations at sea.

- Product NS-2.1: Improved wave forecasts at the 3–7 day range, especially for storms and tropical cyclones.
- Product NS-2.2: High-resolution (to include variability at scales of meters) shallow-water wave and surf forecasts, especially in denied areas.
- Product NS-2.3: Real-time near-surface velocity estimates and forecasts for search and rescue.
- Product NS-2.4: Improved navigational products.
- Variables required for National Security Issue 2 include:
  - Directional Wave Spectra
  - Bathymetry
  - Wind Vectors
  - 3-D Vector Currents
  - Ice Concentration
  - Ice Thickness
  - Atmospheric Visibility

National Security Issue 3

Establish the capability to detect airborne and waterborne contaminants in ports, harbors, and littoral regions at home and abroad, and to predict the dispersion of those contaminants for planning, mitigation, and remediation.

- Product NS-3.1: Background levels of nuclear, biological, and chemical (NBC) contaminants.
- Product NS-3.2: Analyses and predictions of NBC concentrations on scales from the sub-hourly to weekly.
- Variables required for National Security Issue 3 include:
  - 3-D Vector Currents
  - Wind Vectors
  - Water Contaminant Observations (both initial conditions and real-time updates)
  - Bottom Characteristics (sediments composition)

National Security Issue 4:

Support environmental stewardship

- Product NS-4.1: Physiological descriptions of sensitivity to and utilization of acoustic signals by classes of marine mammals
- Product NS-4.2: Real-time and climatological marine mammal/protected species distributions.
- Product NS-4.3: Real-time velocity fields in locations of hazardous material spills or potential spills.
- Variables required for National Security Issue 4 include:
  - Marine Mammal Abundance
  - All variables listed for Issues 1 and 3.

National Security Issue 5:

Improve at-sea system performance through more accurate characterizations and prediction of the marine boundary layer.

- Product NS-5.1: Improved prediction of electromagnetic/electro-optic propagation through the marine boundary layer in support of strike warfare, antiaircraft warfare, and antisubmarine warfare.
- Product NS-5.2: Improved prediction of near-surface visibility
- Variables required for National Security Issue 5 include:
  - Water Temperature (especially sea surface temperature)
  - Humidity
  - Marine Boundary Layer Height
  - Directional Wave Spectra (especially, wave height)
  - Aerosols
  - Atmospheric Visibility

PUBLIC HEALTH

Team Members: Carol Dorsey and Larry Honeybourne

Introduction

Public health stakeholder issues of concern for the coastal component of IOOS include exposure to pathogens during body-contact recreation, chemical, and microbial contamination of seafood and anomalous weather, marine organisms, and/or surf events. Stakeholders and product consumers' use of data related to public health issues may include, for example, regulators, commercial shellfish harvesters, researchers evaluating raw water quality data to assess harmful algal blooms, or a Midwestern tourist checking the quality of coastal marine waters for swimming or fishing activities. Though the stakeholders and consumers are varied and have differing degrees of technical expertise, they are united in a need to access relevant data for decision making. The diversity of the public health group is reflected in the responses of individuals to requests for information for this report (Annex D).

Some public health data collection activities are rooted in regulatory decision making such as swimming advisories for recreational waters. According to EPA's *National Beach Guidance and Required Performance Criteria for Grants*, June 2002, "'Good' quality data are those that enable the user to make the decision at hand with an acceptable risk of error within the required time frame." Regulatory actions in the interest of public health require reliable, accurate data based on good science and delivered in a timely manner. The process of continual quality assurance helps ensure that the data meet specified standards and is legally defensible.

For example, regional bacterial water-quality observing systems for body-contact recreation purposes have been extensively implemented along the Southern California coast for many years. Coastal water surf-zone monitoring is conducted by local health departments and publicly owned wastewater treatment works (POTWs) to meet statutory and NPDES requirements, respectively. Data are compiled from both sources by local health departments to determine compliance with the State of California, public-health-based, body-contact recreation standards. The development of software for data transfer, assimilation, analysis, and compliance determination has recently been successfully completed by the Southern California Coastal Water Research Project in conjunction with several Southern California county health departments and POTWs. This regional observing system includes data acquisition, management and analysis. Regional products include Internet-accessible public health beach reports and metadata. This cooperative, operational pilot project could be utilized as a model for the data management portion of the recently enacted federal BEACHES bill as part of IOOS.

Another example of data sets for regulatory purposes is the water quality for shellfish-growing waters, which exist as required components of the National Shellfish Sanitation Program. These data may be in paper files, digital data sets in assorted forms, and with varying availability and accessibility. Though the program does not stipulate how long data are retained, many states archive decades of microbiological, chemical, and physical data related to the classification of shellfish growing waters. NOAA and NOS are developing a demonstration project of these data in the Shellfish Information Management System (SIMS). Coastal state agencies, FDA, and EPA also participated in workshops to prepare a single source of shellfish growing water information with GIS functionality. The regional project is considered platform independent and may be tailored to the state's need for data manipulation. The data are generated by FDA-evaluated laboratories, which are held to a high degree of accountability. However, the present restrictions on access prevent use of the site except by permission from the users. The data are considered proprietary and are not available to consumers outside of the project. Such issues of accessibility, security, and availability must be addressed in an integrated ocean observing system.

Current buoy and satellite-based technologies have limited value in most public health applications. Satellite imagery is used successfully to identify and monitor HAB in offshore, Gulf of Mexico waters, but the resolution and specificity render its use inappropriate in the coastal-zone areas. Bacteriological water quality, harmful algal cell densities, mercury in finfish, and shellfish toxin concentrations continue to be lab-based analyses. Promising new technologies could eventually be employed in buoy modules, but there must be an IOOS commitment to pursue the development and quality of these products.

Numerous stakeholder and consumer groups have an interest in assessing the need for immediate and long-term databases. These data sets include a variety of subjects (biological, chemical, oceanographic, epidemiological, atmospheric, model output, demographic data) and varying degrees of technicality. The proposed national backbone could assist in correcting problems with existing data communications and management by standardizing the way data are edited across applications, languages, and platforms. Planning protocols now will assure that new data sets can be appropriately formatted and assimilated into the national platforms. Regionally developed observing systems and databases will provide the functional, standardized products for the federally funded backbone. This network will provide critical information to users of ocean and coastal information and service.

Some data users for the public health issues of interest:

- EPA
- FDA
- CDC
- State and local health departments
- POTWs
- Commercial shell fishers
- Educators
- Recreational water users
- Marine safety organizations
- Coastal counties and cities
- Researchers
- Regulators
- Health professionals
- Environmental groups and non-governmental organizations
- Hospitality/Tourism Industry

Issues

The author of this report section polled several public health professionals concerning their data management issues with respect to a national ocean observing system. Individuals' responses to requests for feedback are documented in Annex D. The individual responses are summarized by the following points, which serve as an introduction to the issues:

- Multi-source integration,
- Geographic layering,
- At least two layers of technical depth-general consumer and technical user,
- Security,
- Standardized protocols and platforms,
- Increased fishery data of the appropriate type (fisher-dependent, such as onboard boats and interviews),
- Communication in place prior to information dissemination so that there is adequate alert for situations with a minimal false alarm element,
- Physical and chemical telemetered data could be used in modeling efforts with public health applications.

The major data management issues with supporting detail from the public health perspective are as follows;

- Assess current and future public health needs and goals so that data sets and the integration of data will best serve the system.
    - Many states use a multiple-agency approach to managing the coastal zone. In such an approach, there may be overlap or gaps in coverage. Sanitary surveys, bacterial source tracking and water chemistry may be measurements taken by one or several agencies on the same areas. This makes data integration important. Issues of data availability, accessibility, distribution, and integration should be addressed to improve use.

- Identify on a nation-wide basis, existing databases related to public health issues.
    - Extensive monitoring of coastal waters has been recorded for the purpose of classifying shellfish growing waters, recreational quality, and illness related to the consumption of shellfish and finfish. These data sets often cover decades of data, generated using standardized methods of analysis.
    - Bacterial water-quality monitoring databases for recreational waters are available from Local or State Health Departments, POTW's, and Water Quality Regulatory Agencies. Data set access, quality, and formats are highly variable.

- Develop QA/QC standards to evaluate existing and yet-to-be-developed data products.
    - Data products must be assessed for accuracy, precision, reproducibility, etc. by technical experts and data managers. Written standards for procedures such as those employed in the National Shellfish Sanitation Program and EPA certified laboratories, which are used to generate the data, are critical to the quality and reliability of the measurements. For data products, programs such as the National Coastal Data Development Center offer guidance as they develop and maintain a catalog of available coastal data, ensuring the quality of these data and associated metadata, populating and maintaining databases. Quality assurance is integral throughout the process of data production. Public health regulatory action must be supported by "good," legally defensible science delivered in a timely manner.

- Evaluate the data sets for availability and accessibility to consumers.
    - Surveys of coastal and ocean areas generate data used in the determination of water quality in recreational areas and seafood harvest. These data sets vary in their levels of accessibility.
    - Levels of accessibility
        - General consumer
        - Pre-arranged approval
        - Proprietary, with time limitations
        - Proprietary, not available outside of project or network

- Ensure relevancy of observations to public health users by identifying the update intervals of the data sets and adequacy of the frequency of measurement.
  - Timely measurements and posting of data are important to use of and incorporation into a public health response to a situation. An example of this might be the issuance of swimming advisories when water samples exceed standards or response to seafood- related illness outbreak. Nearshore sensing stations producing real-time data streams of swells, tides, and wave heights could be useful in the public health and safety issues for swimmers, surfers, and fishermen. Other data sets may yield sufficient coverage monthly, seasonally, or annually.

- Evaluate data sets for level of processing (raw data points vs. analyzed with interpretation). Regulatory compliance requirements at the federal, state, and local levels and the subsequent usable product will require processing and interpretation; however, raw data could be available to specified user groups, i.e., researchers.
  - Data sets range from local paper files to national digitized databases. Within these instruments users may find raw data, for example, telemetry from buoys such as tide levels that may be seen at http://www.co-op.noaa.gov/. Some databases contain numeric data points with some interpretation as is published on BEACH water monitoring in Alabama http://www.adem.state.al.us/FieldOps/Monitoring/monitoring or data which have been interpreted according to state standards as seen on the Florida Marine Research Institute's Red Tide Status or satellite images, http://floridamarine.org/features/category_main.asp?id=1884. The degree of processing influences the extent to which the data may be used and by whom. For example, research may find numeric data points (CFU) more useful than a Red, Yellow, Green warning system, but for swimmers the color code will suffice.

- Determine in what formats data are stored and how should new data elements or objects be designed and delivered.
  - Paper,
  - Digital with substantial manipulation of format to meet platform specifications,
  - Digital with easy conversion and assimilation into specified platform.

- Evaluate and/or develop new technologies for the detection of human pathogens, indicators of pollution, or hazardous conditions using remote sensing or permanent monitoring stations for the timely communication of information used in public health decisions. The development of new technologies should be integrated into the enhanced platforms envisioned for the IOOS system (moored buoys, satellite sensors, remote sensing, etc.)

COASTAL MARINE ECOSYSTEMS

Team Member: Dave Eslinger

Introduction

The Integrated Ocean Observing System is intended to enable efforts to more effectively protect and restore healthy coastal marine ecosystems. The community of data users for these ecosystems includes those who derive economic benefit from healthy ecosystems (e.g., the commercial fishing, sports-fishing, and eco-tourism industries), those who derive recreational benefit from these ecosystems (e.g., beach-goers, sports-fishers, divers, boaters, surfers), those who derive aesthetic benefit from a healthy coast (e.g., coastal residents, tourists), and those whose job it is to understand, manage, and protect these environments (e.g., state and local departments of environmental protection, fish and game, and health; academics). These coastal ecosystem stakeholders share a number of concerns about the data they need. These can be summarized as needing: (1) operational and (2) archival data, (3) collected at appropriate times, with (4) high spatial and (5) temporal resolution, and delivered in a (6) user-friendly format.

Issues

1. **Operational**: Operational data are consistent, timely data that are available on a regular schedule.
   a. Consistency: Although collection instruments wear out, get upgraded, and change through time, data streams delivered to the end user need to remain constant in terms of accuracy and format. This will require a data-delivery system capable of delivering data that may require reformatting, conversion of units, and other operations, in a manner that is transparent to the end user.
   b. Timeliness: Coastal ecosystems are physically and biologically dynamic. IOOS data must be delivered to users quickly enough to be of use in understanding ongoing processes. In many cases, this means within 1 hour to 1 day, at a maximum. For many management applications (e.g., harmful algal blooms, pollution events), fast information may be more valuable than absolutely accurate information. Therefore, the data system should be capable of rapid delivery and of reprocessing data to a high level of accuracy and quality.
   c. Regular delivery: Data that cannot be consistently counted upon may be interesting, but not useful. The maximum utility in the IOOS observations will come when the data streams can relied upon to be there—same time, every time.

2. **Archival data**: Archival data are older data sets that are available for comparison with current measurements.
   a. Older data sets: For the data management system, this issue implies data mining, data rescue, and keeping an ongoing archive as operational data become archival data.
   b. Comparison: This could require data mangers to find and understand older metadata, translate older data sets into appropriate units, and reformat older data for consistency.

3. **Collected at appropriate times**: Data are most useful when they can be easily integrated with other data sets for analysis. In the coastal ecosystem, that means that data from different sensors must be collected at almost the same time. The data management system must be capable of keeping the data streams organized and of delivering the needed section of multiple data sets.

4. **High spatial resolution**: Coastal processes occur over relatively small spatial scales. IOOS data must be collected at high spatial resolution to observe and monitor these processes. This high resolution could come from large numbers of *in situ* sensors or from high-resolution, remote-sensing systems. For the data management and delivery system, storing and delivering these data sets to users will require massive storage capacity, excellent cataloging/relational data base capability, and a high-volume delivery system:
   a. Massive storage capacity: It takes 900 times the data volume of currently available IKONOS imagery (approximately 1 m resolution) to cover the same area as one pixel of "old" Landsat (30 m resolution) imagery.
   b. Excellent catalog/relational data base capability: Multiple data sets must be able to be searched, sub-set, and selected areas extracted to be useful.
   c. High-volume delivery system: Users need these large amounts of data delivered in a useful period of time. This will require efficient compression technologies and fast, reliable delivery mechanisms.

5. **High temporal resolution**: Coastal ecosystems have processes that users need to monitor occurring over time scales of storms and tides to El Niño and on to sea level rise. To address this variety of issues will require high temporal resolution data collected over long periods of time. This will add to the requirements for massive storage capacity, excellent cataloging/relational data base capability, and high-volume delivery systems.

6. **User-friendly format**: IOOS data will be of no use if they cannot be found, related and used.
    a. Found: Users must be able to easily locate the data they need. This will require a data management system with an understandable interface for conducting searches of the data by type, location, time, and other parameters. The system must work with a variety of different computer types.
    b. Related: Users must be able to conduct queries to get different types of data that they may wish to relate. For example "all wind and wave observations within 50 miles from lighthouse X and 3 months prior to…"
    c. Used: Data must be delivered in a format that end users can readily use. It should be understood and imported into a variety of readily available software packages.

SUSTAINABLE USE OF MARINE RESOURCES

Team Member: Roy Mendelssohn

Introduction

The Integrated Ocean Observing System is intended to enable the sustained use of marine resources. This sustained use of marine resources is a cross-cutting issue, as it depends on healthy coastal ecosystems, natural hazards and marine operations, and for proper long-term management on the effects of climate change. The composition of this community of consumers includes, on the research and management side, a variety of interests that are dominated by federal, state, and tribal government scientists and policy makers working in the management of fisheries. The fishing industry itself, from fishermen through processors, includes both potential users of the data from the system, as well as groups that may be affected by the data needs of the system. Harbormasters, recreationalists, and educators also have interests in the sustained use of marine resources and are potential consumers. Other highly visible potential IOOS user communities are composed of the scientists and agents of user groups representing commercial, recreational, subsistence, and non-consumptive interests.

Issues

1. **Data Formats**. At present, OpeNDAP on the server side supports relatively few formats and supports relatively few programs on the client side. The OBIS format is mainly used in museums and universities. For this to become the standard "middle-layer" of the communication system, much more work would have to be done to be consistent with formats and programs used in both state and federal government agencies concerned with sustained use. Many groups are now applying GIS-based systems, so easy ingestion of IOOS data into such systems would appear to be a necessity.

2. **Data Entry Timing, and Quality Control**. The currently applied model for data collection appears to be based largely on systems of sensors, etc., where the data are readily available in some electronic form immediately after collection. Much fisheries data are collected on paper forms, and there is often long lag times before the data are entered into an electronic format and subjected to data quality assessment procedures. The design of the IOOS system will to take into account some amount of time delay before certain types of data would become available. Quality control in general is more difficult with biological data—for example, sea surface temperature data exhibit a certain consistency in time, seasons, and space that allows possible outliers to be flagged. Such "neighbor consistency" does not often for biological data, which make the prescription of quality-control indicators for biological data more difficult to define.

3. **Data Confidentiality**. Unlike measurements of sea surface temperature and winds, much biological data are provided by individual businesses, so there may exist a legal obligation to maintain confidentiality of the raw data. If the raw data are to be put into the IOOS system, how will confidentiality of the data be protected? Is there necessary information available to create adequate algorithms to safeguard the confidentiality of these data? If only some form of aggregate data are to be put into the IOOS system, what are the guidelines for maintaining privacy and confidentiality while still providing useful data to the system?

4. **Data Archiving**. Where, how, and in what format (e.g., aggregated, raw) biological data should be archived has not been fully addressed so far in the IOOS DMAC planning document. Given the confidential nature of much of the raw biological data, this remains a non-trivial issue.

5. **Right to publish**. Data collected by government scientists, even when obtained in a format that allows for immediate availability (e.g., pop up tags, satellite), generally are not available for sharing until technical papers have been published in the open literature. What mechanisms does the IOOS system plan to provide to protect a researcher's right to "first publication"?

Variations on these five issues may be found in the literature. For example, see Boehlert and Schumacher (1997).

Team Conclusions

REQUIREMENTS

The lists of community issues are the ultimate source of the user's requirements for the data system being designed by DMAC. As IOOS evolves, the development of institutional infrastructure, such as committees and regional representatives, should make it possible to capture and use more detailed information on user requirements in the design and implementation of the data system. For the interim, the following are highlights of data management and communications issues that appear to be common to all types of consumers of oceanographic data, as entry-level system requirements. Note that end users make a sharp distinction between data (as raw observations) and information (as data organized in ways that make it easy to use). These requirements are more extensively defined and discussed in the references of the bibliography.

1. Data integrity must be assured. The origin, chain of custody, accuracy, precision, and other vital characteristics of the data must be known and verifiable.

2. End users want useful information products. For the majority of users of IOOS data, raw data need to be converted to descriptive statistics, other types of mathematical summaries, such as models, and visualizations, such as graphs and maps (GIS).

3. Data need to be available in a timely fashion. The length of time between observation and dissemination needs to be minimized, recognizing that certain types of data, such as biological, will require different lengths of time to complete the three-step cycle of observation, QA/QC, and dissemination. Timeliness is especially critical to users in government agencies with regulatory responsibilities.

4. Easy access to data through commonly available hardware and software should be provided. Users expect to be able to get the information and data needed.

5. Open access is provided to all data collected with public funds. Access to data collected with government funds needs to be open to all, save for considerations of national security, scientific professional courtesy, QA/QC being performed.

6. Data need to be preserved indefinitely. Although some information products are ephemeral, i.e., the flight conditions now prevailing, or the wave heights at the surfing beach today, the data that goes into those products should be stored for future retrieval.

7. Continuity of time-series observations needs to be preserved. The establishment and maintenance of long time series is of vital interest to a variety of different types of users, but especially to physical and biological modelers working on systems with "long memories."

8. The size of the 4-D cube of interest is defined by user groups. There are user-specific requirements for data to be packaged into 4-D cubes, where 4-D cubes are observations grouped by time, and by space in three dimensions. Cooperation among users, data managers, and those managing the observing system is essential if IOOS is to meet the needs of the maximum possible number of users with the available sampling budget.

RECOMMENDATIONS

User Outreach needs to fulfill roles within IOOS that transcend the needs of the Data Management and Communications Subsystem. IOOS will need to establish an infrastructure of standing committees in order to function. As illustrated by Figure 1, the functions of the standing committees correspond to the subsystems associated with the end-to-end user driven system, as originally envisioned (Nowlin and Malone, 1999). In the future, it is recommended that DMAC maintain a strong and active connection to users, through formally structured interactions with users, i.e., via a User Outreach Committee that seeks out and understands the needs of current and potential users. The User Outreach Committee would work with a Users Advisory Body, and Applications, Products and Models Committee, a Data Management and Communications Committee, and an Observing Operations Committee. In this way the needs of users could be represented in all of the key subsystems of the end-to-end user-driven observing system.

User Outreach should function at the level of an IOOS Standing Committee (SC) and it should provide liaison to the following standing committees;

- Users Advisory Body
- Applications/Products
- DMAC
- Observing Operations

User needs should be a common currency that is used to some extent in all of the operations of IOOS.

Bibliography

Boehlert, G. W., and J. D. Schumacher. 1997. Changing Oceans and Changing Fisheries: Environmental Data for Fisheries Research and Management. Proceedings of a workshop held 16–18 July, 1996, Pacific Grove, California. NOAA-TM-NMFS-SWFSC-239.

Nowlin, W., and T. Malone. 1999. Toward a U.S. Plan for an Integrated, Sustained Ocean Observing System. National Ocean Research Leadership Council, Washington, D.C.

Opishinski, T., and M. L. Spaulding. 2002. Application of an integrated monitoring and modeling system to Narragansett Bay and adjacent waters incorporating Internet based technology, Proceedings of 7th International Conference on Estuarine and Coastal Modeling (ECM 7), November 5–7, 2001, St Petersburg, Florida.

Oceans US. 2002a. An integrated and sustained ocean observing system (IOOS) for the US: Design and Implementation, prepared by Oceans US, Arlington, VA., 21 pp.

Ocean US. 2002b. A multi-year phased implementation plan for an integrated ocean observing system for the US, prepared by Oceans US, Arlington, VA. (Draft)

Ward, M., and M. L. Spaulding. 2001. A nowcast/forecast system of circulation dynamics for Narragansett Bay, Proceedings of 7th International Conference on Estuarine and Coastal Modeling (ECM 7), November 5–7, 2001, St Petersburg, Florida.

Annex A: Solicitation to Natural Hazards Communities

Dear Colleagues,

Ocean.US, the national office for integrated and sustained ocean observation system (ISOOS), convened a workshop in March 2002, which resulted in a report entitled: "An integrated and sustained ocean observing system for the US, Design and Implementation" (May 2002). This report is available at the Oceans.US web site for those interested. The goals of ISOOS are to:

1. improve the safety and efficiency of marine operations (marine ops),
2. more effectively mitigate the effects of natural hazards (natural hazards),
3. improve predictions of climate change and its effects on coastal populations (climate change),
4. improve national security (national security),
5. reduce public health risks (public health),
6. more effectively protect and restore healthy coastal marine ecosystems (CM Ecosystems), and
7. enable the sustained use of marine resources (sustained use).

During this workshop a Data and Communications (DAC) Working Group (DACWG) was formed to develop a plan for the data and communications component of the ISOOS. The DACWG was divided into Expert and Outreach Teams. The User Outreach Team is to play two roles, (1) produce the community issues lists, and (2) serve as a communications clearinghouse for the other teams on identifying user groups, getting feedback from user groups and identifying their issues and related requirements for the system. I have been selected to serve on the User Outreach Team and assigned the responsibility to develop an outline of user community-specific issues that a national data and communications subsystem of the ISOOS would have to address in its development and operations. I have been specifically assigned primary responsibility in the natural hazards (storm surge and coastal flooding, coastal waves, tsunami, coastal erosion) area, with secondary responsibility in marine operations.

In the interest of obtaining input from the coastal engineering and the marine environmental modeling community I solicit your thoughts on user specific community issues. As a minimum I need the following: (1) brief explanation of natural hazards that would benefit from a ISOOS , (2) the principal community of interest and their characteristics, (3) principal variables that are required, (4) issues of concern or attributes that are critical to the application (i.e. timely access to data, ease of access, accuracy, etc.).

I must provide a summary of my input to the team not later than August 25, 2002 and hence would appreciate any input you might have. You can contact me at 401-874-6666 if you would like to discuss your input in more detail.

Malcolm Spaulding, Professor of Ocean Engineering, University of RI

Annex B: Natural Hazards Correspondence

X-Sender: mooers@mail.rsmas.miami.edu
X-Mailer: QUALCOMM Windows Eudora Version 5.1
Date: Mon, 19 Aug 2002 13:07:09 -0400
To: "Malcolm L. Spaulding" <spaulding@oce.uri.edu>
From: "Christopher N.K. Mooers" <cmooers@rsmas.miami.edu>
Subject: Re: User Input on ISOOS

Malcolm - I am glad to know you are working on this challenging task. Perhaps I could best help by responding quickly to a strawman since I will be in town for the next few weeks. - Chris

At 11:05 AM 8/19/02 -0400, you wrote:

****

Date: Thu, 22 Aug 2002 10:12:09 -0400
From: Spencer Rogers <rogerssp@uncwil.edu>
Subject: Re: Coastal_List: User Input on ISOOS
X-Sender: rogerssp@pop.uncwil.edu
To: "Malcolm L. Spaulding" <spaulding@oce.uri.edu>
Cc: houston@soest.hawaii.edu
X-Mailer: QUALCOMM Windows Eudora Version 5.1

Mr. Spaulding,

In response to your coastal-list email, US programs measuring waves and coastal engineers in general are unaware that the most common design use of wave predictions in the US is for building design not beaches or other marine structures. The FEMA-prepared flood maps publish minimum floor elevation requirements for most coastal communities. They assume depth limited waves on a numerically modeled storm surge that is calibrated using historical storm surge data. Though crude, the depth limited wave is probably not unreasonable. "Hurricane Storm Surge and Wave Conditions: Research Needs" by Sam Houston, then with the NOAA Hurricane Research Division and I, was published in the conference proceedings for Ocean Wave Measurement and Analysis (1997, ASCE, v. 2, p. 1414). We compared traditionally collected post-storm still water elevations with nearby evidence of the lower limit of wave induced damage or gaged storm surge elevations and concluded that there are major inaccuracies in the reported storm surge elevations. Localized setup and wave runup appear to routinely cause unimpeachable still water level elevations to exceed even the wave damage elevations nearby. In short, we are taking good measurements of water levels but we do not know what we are measuring.

The issue is interesting science but becomes a significant national problem when the erroneous water marks are eventually used to calibrate the next round storm surge model studies. The water level errors are further amplified when depth-limited waves are added.

The paper concludes that the only way to make sense of the measured water marks is to install wave gages where we intend to apply the data, around oceanfront buildings when a hurricane threatens. Multiple self-contained wave gages in multiple states with more numerous pre-established mounts are necessary to have a reasonable chance of catching a direct hit of a design level storm. I encourage you to include the issue in your summary of coastal hazard data collection needs. I can fax a copy of the paper if you do not have access to the proceedings.

Waiting for a hurricane may seem far fetched to some, but a similar effort to measure hurricane wind pressures on coastal buildings is already underway. Twenty houses have been pre-installed for multiple pressure transducers in Florida and ten are planned in South Carolina, both projects through Clemson University. One building in North Carolina has been instrumented since the late 1990s.

Thank you for the opportunity to make suggestions.

Spencer Rogers

Date: Thu, 22 Aug 2002 11:37:58 -0400
From: "C-S. Wu" <Chung-Sheng.Wu@NOAA.GOV>
Organization: DOC/NOAA/NWS - National Weather Service
X-Sender: "C-S. Wu" <cs.wu@hqmail.nws.noaa.gov>
X-Mailer: Mozilla 4.77 [en]C-CCK-MCD (Win98; U)
X-Accept-Language: en,zh-TW,pdf
CC: "Malcolm L. Spaulding" <spaulding@oce.uri.edu>
Subject: Re: Coastal Hazard Input on ISOOS

Malcolm,
>

> I'd suggest that you contact FEMA (HAZUS program), which seems to focus on the areas interested. A good contact is Claire Drury at FEMA in Washington (202-646-2884).

c-s Wu

****

From: "David McGehee" <bigwave@emeraldoe.com>
To: "Malcolm L. Spaulding" <spaulding@oce.uri.edu>
Subject: Re: Coastal_List: User Input on ISOOS
Date: Thu, 22 Aug 2002 10:44:25 -0500
X-MSMail-Priority: Normal
X-Mailer: Microsoft Outlook Express 5.50.4522.1200
X-MimeOLE: Produced By Microsoft MimeOLE V5.50.4522.1200

Dear Malcolm:

I'm afraid I'm going to ignore your specific request and bring up a single issue that I feel strongly about, and I which I believe should be of overriding interest - standards for ocean data, most especially wave data. The free exchange and wide utilization of wave data is severely hampered by the babble of formats, definitions, and methods of computation in use by the various agencies and organizations that collect and/or distribute wave data. I managed the US Army Corps of Engineers Field Wave Gaging Program for 12 years, which funded the operation of more wave observation stations than any other program (yes, including NDBC).

One of my most vexing problems was simply intercomparing data from two stations operated by two different organizations, or even the same organization using different instruments. In most cases, it required several hours of a technicians time to simply lay the two data sets on the same plot. Validating model output with data taken within its domain was even more trying. If you make any attempts to integrate the available wave observations into a single distribution system, you'll soon run into the same problem. While the average scientist familiar with writing code or tools such as Matlab may not find this a difficult problem, it is a recurring one, and a major impediment to the utilization of wave data by managers and the general public.

I began a solution by development of standards. The first was a standardized method for converting time series into spectra and wave parameters (Earle, M., McGehee, D., Tubman, M. 1995. "Wave Data Analysis Standard" Technical Report FWGP 95-1, US Army Engineer Waterways Experiment Station, Vicksburg, MS, April 1995.) I was in the process of developing wave data file format specifications (including model output), and had held a series of workshops on wave climate statistics standards, but I went into private practice before either issue was completed. One manifestation of the effort was the "Harvest Project", a gage intercomparison experiment with the synchronous goal of developing a gage intercomparison engine http://cdip.ucsd.edu/harvest_experiment/).

One of the ultimate objectives of this plan was to provide a one-stop "clearinghouse" for access to ALL measured and modeled (hindcast, nowcast, and forecast) wave data. In addition, a standard suite of intercomparison tools (plots and statistics, such as are shown on the Harvest page) would allow anyone to compare any two data sets - or evaluate any model. There was some resistance by the modeling community to this concept.

In any event, I still feel strongly that the full benefits of wave information to the engineering and research community and to the general public will never be realized until these issues are addressed. Please contact me if you wish to discuss this further.

David D. McGehee, P.E., M.Oc.E.
Emerald Ocean Engineering
www.emeraldoe.com
850.932.9111

X-Sender: rjs@splash.ucsd.edu
X-Mailer: QUALCOMM Windows Eudora Pro Version 4.1
Date: Thu, 22 Aug 2002 10:16:13 -0700
To: "Malcolm L. Spaulding" <spaulding@oce.uri.edu>
From: Dick Seymour <rseymour@ucsd.edu>
Subject: Re: Coastal_List: User Input on ISOOS
Cc: rtg@coast.ucsd.edu

Malcolm:

1) The most damaging natural hazard on the coast is, of course, the impact of wind waves. Unlike earthquakes, damaging wind waves can be forecast to some degree (we are publishing reliable 3 day forecasts on the web at this time.) Wave runup is roughly proportional to the deep water wave height such that large waves, especially if they coexist with high tides, can cause flood damage and increased structural impact damage. In areas with broad shelves (most of the coasts) strong onshore winds can cause significant storm surge, greatly increasing the flooding and structural damage. This makes local wind forecasting extremely important. Wave driven flooding, in addition to direct inundation, often causes overwash of beach sand that clogs storm drains and prevents the runoff of rainfall, adding to the flooding hazards. The loss of beach or dune sand during storms can create a hazard in that the normal protection from wave damage has been compromised or lost. Therefore, a knowledge of the state of the beach is important to decision making on both a short term (temporary protection of vulnerable areas) and a long term basis (planning for beach nourishment programs.)

Tsunamis are rare events and damage from distantly-generated tsunamis is limited to particular coastal geometries that amplify the runup. Because of the lack of reliable warning, especially for locally generated tsunamis, loss of life as well as significant property loss is often suffered.

Deep water directional wave measurements coupled with effective models can provide nowcasts of nearshore wave conditions sufficiently in advance to allow for local evacuation or protection measures. Large scale events, such as hurricanes, must depend on meteorological forecasts to provide sufficient warning time.

2) The community of interest can be readily divided into two groups. The first are the property owners, residents or users of the coastal area. Those people with a financial or personal interest in structures and infrastructure subject to damage such as roads, bridges and marinas are clearly concerned about mitigation of coastal hazards. Included in this group are beach users, boat owners and businesses dependent upon beach tourism. The second group are governmental entities with responsibilities for insuring, maintaining, regulating or protecting the first group. These vary from the federal (FEMA, Corps of Engineers, Coast Guard, NOAA, etc.) to the local (police, sanitation, permitting, health, etc.) Coastal states have a variety of agencies, often paralleling their federal counterparts.

3) The parameters that require measurement by a system like ISOOS include:

a. Deep water directional wave gaging
b. Sufficient nearshore directional wave measurement capability to validate propagation models, and to account for modification by wind over the shelf
c. Wind measurement on the shelf
d. Long wave and sealevel variation measurement in deep water (GPS makes this easy, now)
e. Rapid response broad area assessment of beach and dune health following severe hazards.
f. Seasonal broad area 3D mapping of beaches, dunes and cliffs.

4) The measurements must lead to useful products. Although they will have archival value, they must provide pre-hazard warnings or post-hazard data in a timely and useful form. Therefore, models which convert measurements into easily understood web pages, radio announcements or even siren blasts are a necessary part of the system. The overall reliability of the system must be very high because users will rely on it. Forecasts must be updated rapidly, often and be sufficiently accurate that they will be taken seriously by those in harms way. For coastal hazard information, hourly updates would seem to be sufficient.

Richard J. (Dick) Seymour, Ph.D., P.E.
Head, Ocean Engineering Research Group
Scripps Institution of Oceanography
University of California, San Diego
La Jolla, CA 92093-0214
Voice (858) 534-2561 FAX (858) 455-5575
[Room 100 Isaacs Hall, 8855 Biological Grade, for Deliveries Only]
http://rseymour.ucsd.edu

Subject: RE: Coastal_List: User Input on ISOOS
Date: Tue, 27 Aug 2002 11:46:52 +0200
X-MimeOLE: Produced By Microsoft Exchange V6.0.4417.0
X-MS-Has-Attach:
X-MS-TNEF-Correlator:
Thread-Topic: Coastal_List: User Input on ISOOS
Thread-Index: AcJJ8oEk+2xGGeeOReG7GeAqLf+PSgDt1/zg
From: “Gavin Hough” <ghough@intervid.com>
To: “Malcolm L. Spaulding” <spaulding@oce.uri.edu>
Cc: <dphelp@csir.co.za>

Dear Prof Spaulding

After a couple of trips to the Antarctic, recording auroral displays with low light level cameras, we have applied various space-time imaging techniques to digital video & radar based sea state surveillance. I would be very interested in contributing to a program like ISOOS, so I've included a few video time stack (or keogram) wave scans for your interest. These and related digital image processing techniques have been used to measure the following:

[1] Wave height, Period, direction & celerity
[2] Plumes during surfacing events near marine outfalls
[3] All weather 24/7 wave period, direction & celerity
[4] Moored ship motion (6 DOF)
[5] Remote (line of site) buoy tracking
[6] Monitoring breakwater damage (breakage & displacement of armor units)

Looking forward to getting a better understanding of the services which ISOOS can offer.

Regards

Gavin.

Dr Gavin Hough.
Development Director, InterVid Ltd.
Chairman, KZN Innovation Support Center.

Annex C: Climate Change Comments Not Yet Fully in Issues

- Internet access to data: Knowing what data are available is part of the science. The big change is the web that makes not only information about the data available (hate to call it metadata) but also makes a lot of the data available online. I think the web is the big area to push and all data systems should be minimalist systems that simply employ the web for data access and distribution. We should work to build clever interfaces to access the data but we should not invent another access path. The web is becoming part of daily life and the common man knows how to deal with it. Scientists are generally a bit more skilled with it and so it should become the pervasive element of the data system. The role for the scientist is to sort out the crap from the good stuff....but then that is what they pay us for. The big change is that data is there, is available, can be ordered, can be worked with. You don't need an antenna, or to know someone with an antenna. This is all doable with regular internet access. Again formats need to be minimal allowing the maximum number of people access to the data. All efforts to force unification should be avoided and the web community should vote with its mouse clicks.

- Metadata: These are many of the basic problems that motivated EOSDIS and what it was supposed to become. The management of metadata became an all-consuming passion that unfortunately did not succeed wildly in delivering more data to us users. The goal is great to have "automatic" metadata that makes it possible to open and work with all kinds of data. The problem is that some specific decisions have to be made and once those are made you life becomes either simpler or more complex. The problem is efforts to make universal formats that do everything for everybody all fail. HDF has been an outstanding example. While the motivation was fine to pick a standard it has become the source of more difficulty in getting data from EOS instruments than any other. Don't make metadata rule the data.

- Data management: "This is a most abused term…..data management needs to be minimal. What we want is data analysis-enablers to make it possible to work with the data. We really don't want the data to be managed since the end goals are not clear."

- The reality is that when you want to put data sets together you face the music, figure out how to work with each of the datasets and then formulate a strategy to put them together. You need to learn about the instrument, its characteristics, the data, how they are generated and what will happen when you put the data together. In the present system this largely takes people, mostly students, postdocs, etc.

- Sharing data with others is generally easy once you have mastered working with them yourself. You can help others avoid the pitfalls and you will have the answers since you had to work them out for yourself. It is always a good idea to talk to someone else who has worked with that data before you try it yourself.

- As for assessing the quality of data we get into all sorts of problems. First remotely sensed data must be calibrated first in the lab, then in space. The measurements are only as good as this calibration. Second satellite sensors drift and change their characteristics making recalibration and validation a necessity. So keeping track of accuracies is important. As you say the actual value of the quality often becomes a very subjective measure and it is difficult to get agreement on the metrics that must be used to say whether or not a measurement is of value.

- Dealing with various data formats is a headache, but it is manageable by the user. However, these points are extremely difficult for an individual researcher to handle and should be managed at a higher level.

- In addition to climate studies more and more of the SST data sets are being applied to regional and coastal problems. Additionally the merging of previous global data sets can lead to better and higher quality coastal data sets with enhanced spatial and temporal resolutions. However, as a previous response mentioned, progress in merging these data sets is not hindered by lack of merging strategies or algorithms but by standard formats in reading the data. My overall impression is that the scientific community is leaning more towards net CDF as a standard. Thought needs to be put into developing the right metadata in order to fully implement the merging of data sets from different satellites.

- Data Access: Accessing and using data typically are still challenging tasks. The push to use HDF was a misguided effort that set the field back by ten years. Data formats are well known only by those who “know them well”. All others have a major struggle. An emerging trend is “interface” standards. This is a specification of how to access the data, rather than of how the data are stored. Again, data providers need to see beyond their immediate community. Climate studies typically require combining data from multiple sources and multiple disciplines. Our standards should be consistent with those from other disciplines.

Annex D: Public Health User Feedback Quotes

1. I think the Issues document sounds great. There is already some effort underway within different agencies to gather some of this data together - hopefully it won't result in duplication of effort. It would be fantastic that water sampling agencies within one state would all agree on a method and recognize each other's data as being valid. For that matter, it would be even better if the state agencies would recognize data collected by non-state agencies and processed in certified labs - consulting company data is frequently looked at by the state with great disdain.

   Diana Sturm

2. Thanks for allowing me the opportunity to review the documents. I am glad that someone is considering management of the wealth of information that exists not only in Alabama, but also around the nation. After reviewing the documents, I envision the development of a data system that would function much like the different layers of a GIS system; in that different, but related layers of information may overlay a specific geographic region/water body.

   I would also agree that the foundation for this system should be national in origin and standardized for use by all providers/users. A "regional" or "body of water" subset of this system may also prove useful from a management/use perspective. Regardless, critical to the success of this system is a standardized platform that would support the data sets. Security of the data is critical.

   Also, I believe there exists 2 basic users of the data.....the general public and those who need the data for technical purposes. Perhaps the system could be configured to allow for basic consumer/user information in one format and a different format for the technical details available for research and other uses.

   Once again, Thanks!
   jackie

3. In general:

   Increasing the level of sampling for fishery independent data is a sound concept. However, there is already a huge overemphasis toward fishery independent data in federal fisheries management. The feds generally don't want to spend the money it takes to get fishery dependent data (aboard commercial vessels or through interviews), however, I believe that new money should have at least as much emphasis in placing observers aboard vessels or in some way gathering the data directly from the fishery. Fishery independent data are often the "best available" information because they are the only available information. However, they may not be very representative either in quality or quantity of actual commercial operations.

Also, the use of temperature and salinity in the coastal zone as indicators of elevated levels of pathogens (naturally occurring or sewage related) is intriguing and probably should be pursued. This may especially be helpful for recreational use of coastal waters to alert users of increased risk for would infections (due to Vibrios) or gastroenteritis (due to enteric viruses or bacteria).

I would hate to see, however, shellfish harvesting areas opened and closed based on a remote sensing device. I doubt that is even remotely possible anyway (pardon the pun). However, it is conceivable that remote indications of lowered salinities would trigger closer investigation of growing waters for fecal contamination.

I believe the ability to detect conditions which indicate HABs is also a laudable goal. One cautionary tale: in MS a researcher actually found a potentially toxic algal bloom offshore. The researcher took it upon herself to notify the media of a potential public health threat. The bloom never reached any commercial shellfish harvesting waters, yet the media caused a "shellfish scare" with the info. The point is that proper risk communication must be in place prior to gathering the information. This is especially important if a new technique or enhanced coverage is employed.

Hope some of this is of use. Please contact me if there are any specific questions you have regarding my thoughts on the documents.

4. As I think I've discussed with you and Monica in the past, I'm working with the Corps to develop a numerical model of coastal processes that can assist us in determining potential health risk in the coastal waters adjacent to Newport and Huntington Beaches. Our plan is still evolving but essentially we will use historic data to determine relationships among various factors like surface currents, currents at various depths, flood channel flow rates, water temperature in the water column, wave height and direction, wind speed and direction, tide stage, meteorological conditions and bacteria hits on the beach. If the model can make any sense out of the data we will maintain a real time telemetered data system that will continuously feed this type of data to the model which will analyze the data and tell us when conditions are present that historically were present when bacteria standards were violated. The model can then tell us based on real time data when the health risk is low, moderate or high. As faster bacteriological analytical

methods or other indicators get developed the model can get further refined. It is not something you would use to post or close the beach, its more of an early warning system of potential risk that could change continuously.

How this relates to what you have sent me, is that we will have to perform big data crunches and for the circulation and dispersion phases of the model look farther afield than HB&NB. The IOOS would very helpful to us in looking at some of the more bight-wide macro level processes. We are thinking that our model development may take 5 years or so. Are we in the right time frame for the IOOS? Can I share the attachments with the Corps of Engineers?

Annex E: Pilot Project Proposal: Integrated Distribution System

An Integrated Marine Environmental Monitoring, Modeling, and Information Distribution System for Regional Seas: Application to the Southern New England Bight

Prepared by:

Malcolm L. Spaulding
Ocean Engineering
223 Sheets Laboratory
Narragansett, RI 02881
Tel: 401-874-6666
Email: Spaulding@oce.uri.edu

Date:

August 28, 2002

Submitted to:

User Outreach Team
Data and Communications Steering Committee
Ocean.US
National Office for Integrated
and Sustained Ocean Observations
http://www.ocean.us.net/
(703) 588-0848 (Voice)
(703) 588-0872 (FAX)

BACKGROUND

The consensus reached by the participants of the May 2002 US Oceans workshop was that the coastal component of the Integrated Sustained Ocean Observing System (IOOS) should be structured as a national federation in which regional systems are nested in a national backbone (Oceans US, 2002a). The backbone will measure, collection, process, and distribute core variables required in the regional system (but on a sparse network of stations) and will provide national scale capabilities in nowcasting and forecasting. The regional systems will increase the resolution at which core variables are measured, measure other variables of local interest, and provide data, information products, and predictions tailored to meet the needs of the user community in the region. The ideal regional system will provide end-to-end capability (Oceans US, 2002b). The system will include: a monitoring subsystem (platforms, sensors, measurement techniques) to measure key variables on the space and time scales appropriate to the region and issues of concern, a data communications and management subsystem to collect, quality control, disseminate, and archive/store data, and model products and a data analysis, modeling, and assimilation subsystem to nowcast and forecast variables of principal concern to the regional/local user community.

Researchers at University of RI, Ocean Engineering (M. Spaulding) and Drexel University (M. Piasecki), Civil and Architectural Engineering are leading a 3 three year, National Ocean Partnership Program (NOPP) sponsored study to develop a globally re-locatable, integrated system for real time observation, modeling, and data distribution for shelf, coastal sea, and estuarine waters. The project seeks to integrate Global Ocean Data Assimilation Experiment (GODAE) data and other global and ocean basin scale data products into the system. Additional partners include the National Oceanic and Atmospheric Administration (NOAA)/ National Ocean Survey, the Naval Research Laboratory, Brown University, Applied Science Associates, Inc, Narragansett Bay Commission, RI Department of Environmental Management, and the University of Rhode Island Transportation Center. The system is being developed and applied to Narragansett Bay and Rhode Island coastal waters as a demonstration of the practical use of the system.

The core of the project is the development of COASTMAP, a marine environmental monitoring, modeling and management system, that operates on a personal computer. This approach allows the cost of the system to remain low and at the same time provides the end-to-end functionality called for in IOOS regional sub-systems (Oceans US, 2002b). A geographic information system (GIS), data processing and analysis tools, and environmental nowcasting and forecasting models form the basic components of the system. Linkages with real time environmental monitoring stations allow users to collect, manipulate, display, and archive local environmental data through embedded data management tools (e.g. time series analysis including filtering, power spectral analysis, and harmonic analysis) with the system presenting a real time status display of all data sources. Spatial representations and animations of the data, within the context of the GIS, are also provided by the system. Environmental models, linked with the system, can access the environmental data for assimilation, validation, predictions, or comparative studies.

An Internet based data collection and distribution system has been developed and incorporated within the COASTMAP framework. This system allows GODAE and other global and basin scale model nowcasts and forecasts and real time observations to be accessed via the Internet. COASTMAP also has the capability to collect data from local monitoring systems (i.e., monitoring equipment operated through direct connection such as serial, radio, cellular or modem communications). Data collected from the various online sources is subjected to quality control processes, archived alongside traditional data sets, and automatically distributed to support high resolution coastal modeling efforts.

COASTMAP's Internet based data collection and distribution system is composed of web, data, and map server applications. Presently the system is configured for operation on three separate computers making the separation of server application functionality clear. The system is scalable and hence can be operated in a variety of multi-server/platform configurations. These might include operation of all server applications (i.e., web, data and map) on one PC or simultaneous operation of multiple data and map servers (each operating on their own platform) on a networked system with operations coordinated by the web server. To communicate with each other the web, data, and map servers require only a communication path utilizing TCP/IP protocol. This arrangement allows the servers to be located in different geographic locations and even on different network domains. Multiple map and/or data server configurations offer increased flexibility and improved efficiency when downloading and accessing large volumes of information. Such scalability allows for future expansion of the existing system and application to large-scale systems without sacrificing efficiency. For example, one might expect access to environmental data from additional data sources to occur in the near future thus increasing the bandwidth and processing time required by the single data server presently in operation. Additional data servers would allow the tasks performed by the data server application to be divided amongst two (or more) data servers, thereby reducing bandwidth and processing requirements for each individual server.

COASTMAP and its associated Internet server applications (i.e., web, map and data servers) are presently operational for Narragansett Bay and adjacent Rhode Island coastal waters (Southern New England Bight). In the present application the system provides access to real time data collected by the NOAA PORTS system, the RI Road Weather Information System (RI RWIS), and a network of water quality monitoring buoys distributed throughout Narragansett Bay. In addition the system allows access to nowcasts and forecasts from the NOAA East Coast, Coastal Ocean Forecasting System (COFS) and the National Weather Service's Extra-tropical Storm Surge (ETSS) model.

Access to predictions from the Navy's global ocean models are available based on special arrangements. As part of the NOPP project NOAA/NOS personnel have implemented a high resolution meteorological model to nowcast and forecast winds, atmospheric pressure, and air temperature fields for the southern New England Bight and adjacent areas. The forecasts are available via the internet from NOAA. Finally researchers at Brown University are providing high resolution (50 m) remotely sensed sea surface temperature data derived from Landsat ETM+ for Narragansett Bay and nearby coastal waters shortly after each satellite pass.

NOPP researchers have applied a state of the art, three-dimensional, boundary fitted hydrodynamic model to Long Island Sound, Block Island Sound, Rhode Island Sound, Buzzards Bay, and Narragansett Bay study area. The model has been applied in a two- dimensional vertically averaged mode and shown an excellent ability to predict tidal circulation and elevations in the study area. Model predictions can be visualized through the model's user interface or via COASTMAP. Efforts are currently in progress to implement forecasting of tidal and wind driven circulation in the study area using the high resolution meteorological model predictions and the ETSS waters levels as forcing and boundary conditions, respectively.

The current state of COASTMAP's development and its application to Narragansett Bay are summarized in Opishinski and Spaulding (2002) and Ward and Spaulding (2002).

One of the goals of the current NOPP project is to assess the market for COASTMAP and transition the system from a research project to a commercial operational system that can be used globally. One of the ultimate measures of success of the project is the extent to which the system is adopted and used in other locations and becomes a commercially viable product. The project team has had very good initial success in this area. The Smithsonian Institution has licensed the system to provide real time monitoring for its Carrie Bow facility off the coast of Belize. The system has also been licensed by the Georgia Port Authority (GPA) as a real time monitoring and modeling system for the Savannah River. The system has been specifically configured to collect data in support of evaluating the impact of port dredging projects on marine water quality. Most recently NAVOCEANO's, Ocean Modeling Division will be acquiring a license to COASTMAP (October 2002) and will employ the system for use in monitoring and modeling activities related to homeland security at key naval installations throughout US coastal waters. The system will be configured to integrate output from the Navy's hydrodynamic models for each facility of interest. These initial successes in the market place clearly demonstrate that the COASTMAP concept and its implementation are viable and applicable to a range of different users.

It is clear from the presentation above that COASTMAP is an excellent candidate for a regional subsystem within the IOOS national backbone. It is reasonably well developed, provides end-to-end capability, is extremely flexible, low cost, includes well developed user interfaces with data and information products that meet user needs, and is commercially viable. It is therefore proposed to extend our current NOPP project and apply COASTMAP as a regional subsystem, pilot project for the southern New England Bight (Long Island Sound, Block Island Sound, Rhode Island Sound, Buzzards Bay, Narragansett Bay and adjacent southern New England Shelf.). This area forms a logical division in regional monitoring systems for the northeast; between a system for the Gulf of Maine (GoMOOS) and one for the mid Atlantic Bight (New York/ New Jersey Harbor, Delaware Bay). It is a microcosm of these larger systems with many of the same environmental and use issues.

TASKS

The following major tasks are proposed:

- Provide links in COASTMAP to allow access to national and international databases using the Open Source Project for Network Data Access Protocol (OPeNDAP) middleware system (Oceans US, 2002b). This will allow direct link to the national backbone system when it becomes available and to other regional systems. This linkage will include the ability to use the evolving mega data structure contemplated for the backbone system.
- Access to existing data collection systems is well developed for Narragansett Bay, access to similar data collection systems for the remainder of the study area needs to be implemented. This includes not only coastal ocean monitoring systems but relevant geographic information system data sets as well.
- Extend the suite of models* and associated products in the current NOPP project to include the following: Note the models selected are based on a comprehensive survey and assessment of the needs expressed by the regional user community.
- Water level, depth, and current forecasting model (navigation aid for shipping)
- Storm surge and directional wave models (hurricane and nor'esters)
- Oil and chemical spill models
- Models to predict evolution of fecal coliforms discharged from combined sewer overflow models (CSOs) during storm events
- Search and rescue model
- Dredged material disposal predictions
- Model of thermal discharges from power plants.
- Extend the NOAA/NOS's meteorological modeling program to include the entire operational area. The output of this model will be used directly and as input to many of the models above.

The models noted above will be validated for selected areas and problems within the region.

Once the system is fully operational presentations will be made to all potential major users groups (regulatory agencies, emergency responders, power industry representatives, state environmental agencies, etc.). Pilot systems will be implemented in the facilities of key user groups and an assessment made of the how well the system meets user demands. Feedback from the assessment process will be used to better target the system and its products to meet user needs.

An assessment will be made of a variety of options for ensuring sustained operation of the system and recommendations made on the most viable approach.

* Note that all the models are available for the above problem areas. They however need to be integrated into COASTMAP, linked with appropriate input data sets, and the output customized for the regional user community.

POTENTIAL PROJECT PARTICIPANTS

Potential project participants are listed below. They represent the major oceanographic institutions in the area that have been active in marine environmental monitoring and modeling of the Southern New England Bight and the key government agencies currently working on the NOPP project. Not all of the potential participants have been contacted at this time. This will be done if there is interest in pursuing the project further.

- Ocean Engineering University of Rhode Island (Spaulding, Opishinski)
- University of Connecticut, Marine Sciences (O'Donnell, Bohlen)
- University of Massachusetts, Dartmouth, Marine Sciences and Technology (Brown)
- State University of New York (SUNY), Stony Brook (Bowman, Wilson, Wang)
- Applied Science Associates, Inc. (Swanson)
- Brown University (Mustard)
- NOAA/NOS (Kelley)

Estimated Budget and Time: $2 M per year for 3 yrs.

REFERENCES

Opishinski, T. and M. L. Spaulding, 2002. Application of an integrated monitoring and modeling system to Narragansett Bay and adjacent waters incorporating Internet based technology, Proceedings of 7 the International Conference on Estuarine and Coastal Modeling (ECM 7), November 5-7, 2001, St Pete, Florida.

Oceans US, 2002a. An integrated and sustained ocean observing system (IOOS) for the US: Design and Implementation, prepared by Oceans US, Arlington, VA., 21 pp.

Ocean US, 2002b. A multi-year phased implementation plan for an integrated ocean observing system for the US, prepared by Oceans US, Arlington, VA. (Draft)

Ward, M. and M. L. Spaulding, 2001. A nowcast/forecast system of circulation dynamics for Narragansett Bay, Proceedings of 7th International Conference on Estuarine and Coastal Modeling (ECM 7), November 5 to 7, 2001, St Pete, Florida.

I. Interoperable Data Discovery, Access, and Archive

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

Part III. Appendices

Appendix 5. System Engineering Approach

Contributed by John Lever and Landry Bernard

March 2005

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

The other sections of this document describe a wide variety of requirements that represent a diverse group of stakeholders. The resultant complexity would likely render ineffective any uncoordinated approach to satisfying these requirements. Accordingly, there is strong evidence that the Data Management and Communication Subsystem of the Integrated Ocean Observing System can only be accomplished using a formalized System Engineering process. The following provides a brief description of three System Engineering process models and recommends the approach that should be used for the DMAC development and integration. The three process models discussed are the Waterfall Model, the Rapid Prototype Model, and the Spiral Model.

The Waterfall Model is the most commonly used approach for major acquisition systems over the past several decades. Under this approach there are a series of steps that will have to be achieved from system concept to system operations and all will be preformed in series, not in parallel. The transition from each step to the next is only accomplished after successful completion of a very structured review process. Table 1 shows the typical process steps and corresponding reviews for each phase in the Waterfall Model.

Table 1. Typical process steps and reviews for Waterfall Model

	Task/Step	Review
1.	Requirement Definition	System Requirement Review
2.	Analysis	Risk Assessment Review
3.	Design	Preliminary/Critical Design Reviews
4.	Coding	Walk Through Review
5.	Testing	Technical Evaluation Review/Operational Evaluation Review
6.	Operations	Initial Operational Capability

For each task and review there are many structured documents that are prepared, reviewed, and maintained for the life of the system. The highly structured nature of the waterfall method makes it quite applicable for large well-defined projects. This linear approach has been adopted by the Department of Defense for major acquisition programs (DoD Instruction 5000.2-R).

The third process model to be considered is the Spiral Model. This model accommodates the waterfall "task-oriented" highly structured approach while allowing rapid prototyping and risk-analysis to be performed at juncture points of the project. Another difference between the waterfall approach and the Spiral Model is that in the former, all requirements are known up front and are all developed throughout each step; in the spiral model selected requirements are chosen for development through requirements to operations. Then more requirements are added and the process from requirements to operations is repeated through this “spiral” until all requirements are accomplished.

A variant of the spiral model is the phased approach. In this method, the system requirements are allocated to phases where a preceding phase may have influence on the subsequent phase requirements. The phases can be executed using a waterfall-like process, i.e. with requirements specification (or update), analysis and design, system development, and verification performed for each phase. Each phase (sometimes referred to as effectivity), then, would represent a complete end-to-end execution of a subset of the requirements.

Figure 1 illustrates the tasks and sequence associated with each phased cycle in the spiral model. After steps one through seven additional requirements are specified and the cycle is repeated.

Figure 2 represents a full-blown project implementation following the Spiral Model; this implementation includes the formalism of the Waterfall Model.

Table 2 compares all three types of process models. Based on a review of the Data Management and Communication Subsystem requirements and a view of the Comparison Table it is recommended that the Spiral Model for Systems Engineering be selected.

A phased approach that would fit this purpose is shown in Figure 3.

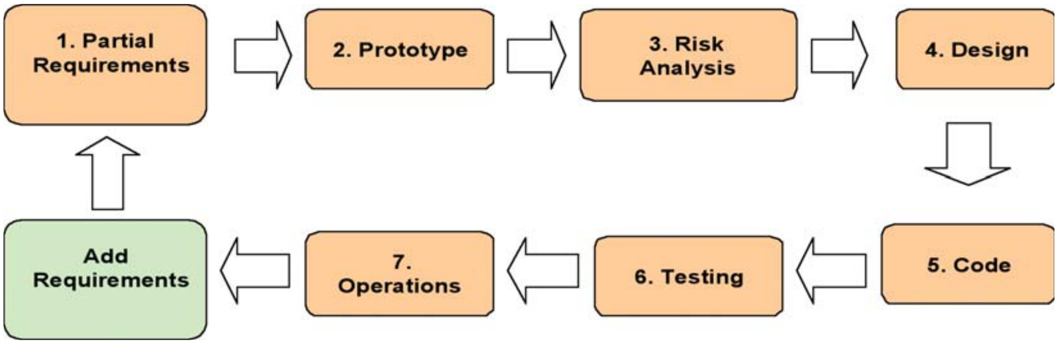


Figure 1. Tasks and sequence cycle for Spiral Model

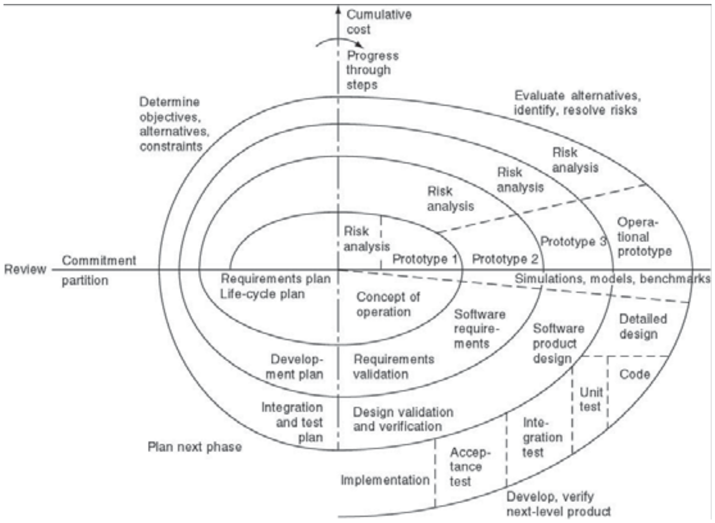


Figure 2. Full Spiral Model

Table 2. System Engineering comparison of Waterfall, Rapid Prototype, and Spiral Models

Waterfall Model	Rapid Prototype Model	Spiral Model
Pro	**Pro**	**Pro**
Documentation	Allows frequent changes	Risk analysis preceding each phase
Maintenance easier	Helps define user requirements	Allows for changing requirements
Quality product at finish	Rapid return on investment	Allows prototyping
Con	**Con**	**Con**
Specification document	Increased maintenance costs	Once risk cannot be mitigated the project is terminated
Have to get it right first time	How do you know you are finished?	Not effective for large-scale projects
Does not allow for prototype	Build-and-fix	
Time consuming		
Costly		
Hard to accommodate		
Doesn’t accommodate new requirements		

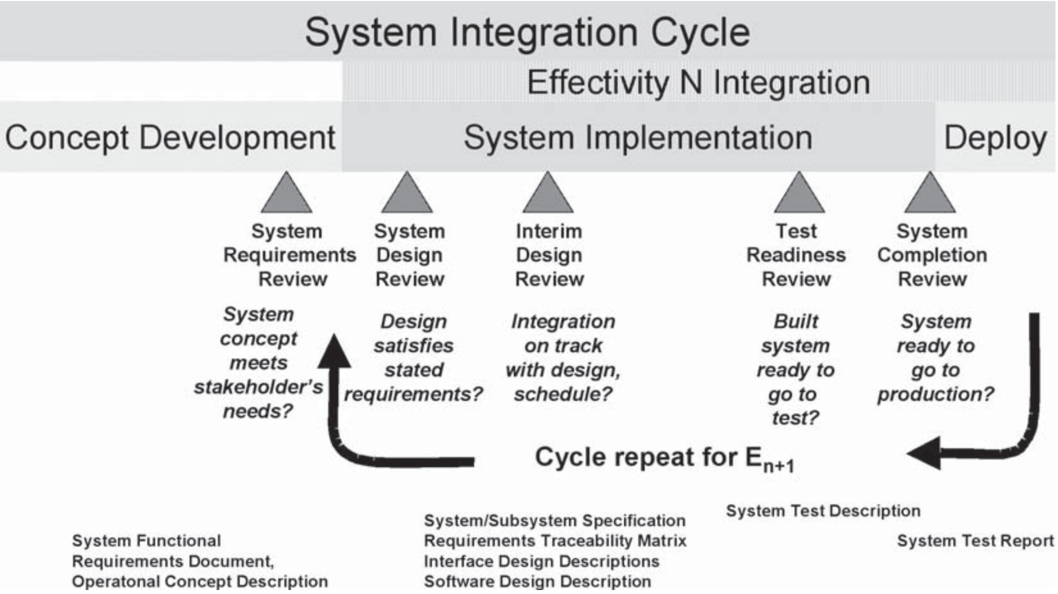

System Integration Cycle
Effectivity N Integration
Concept Development
System Implementation
Deploy
System Requirements Review
System Design Review
Interim Design Review
Test Readiness Review
System Completion Review
System concept meets stakeholder's needs?
Design satisfies stated requirements?
Integration on track with design, schedule?
Built system ready to go to test?
System ready to go to production?
Cycle repeat for E_{n+1}
System Functional Requirements Document, Operatonal Concept Description
System/Subsystem Specification
Requirements Traceability Matrix
Interface Design Descriptions
Software Design Description
System Test Description
System Test Report

Figure 3. System Integration Cycle for the Spiral Model

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

Part III. Appendices

Appendix 6. Technology Maintenance and Refreshment

Contributed by John Lever and Landry Bernard

March 2005

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

This section presents the reasoning and issues concerning the life cycle maintenance and refreshment of the technology components of the DMAC. The other chapters of this document describe systems capabilities that are all reliant on technology, whether systems, hardware, or software. To ensure that the IOOS/DMAC meets the program goals, it is critical to ensure that the technology stays current and operational; to do so requires a concrete plan for maintenance and refreshment.

A Price Systems LLC paper defines Technology Refreshment as "*the periodic replacement of commercial off-the-shelf (COTS) components; e.g., processors, displays, computer operating systems, commercially available software (CAS) within larger … systems to assure continued supportability of that system through an indefinite service life.*"[1] (We would add communications capabilities to this list.) That is, systems are being acquired that are ever larger and more complex, and that are built up out of components that are designed and built by commercial third parties. If the DMAC is viewed as a system of systems, this situation still applies; whether "component" in this context means a server or an application software suite, it still represents an item that must be replaced periodically in order for the overall system to meet its mission requirements.

For the IOOS and the DMAC to remain current, the technology must be refreshed during the system life for a number of reasons, including the following:

- The existing system component has malfunctioned and either cannot be repaired, or the repair costs exceed the replacement costs,
- The existing system component has reached its life expectancy,
- The surrounding technical infrastructure has evolved and is incompatible with the existing component under consideration,
- Newer technology has come to market that provides more capability for the same or lower Total Cost of Ownership,
- The requirements of the system have evolved to the extent that the system cannot meet the requirements with the existing technology.

Accordingly, this section recommends that the DMAC systems integrator be tasked with the development of a Technology Refreshment Plan (TRP) to incorporate the ideas introduced in this section. The above-referenced paper reflects three categories of Technology Refreshment:

- Technology Upgrades—A change that incorporates the next generation product or product upgrade to an existing technology or component which improves overall system functionality,

[1] "Technology Refreshment - A Management/Acquisition Perspective," available at http://www.pricesystems.com/downloads/pdf/technology%20refresh.pdf

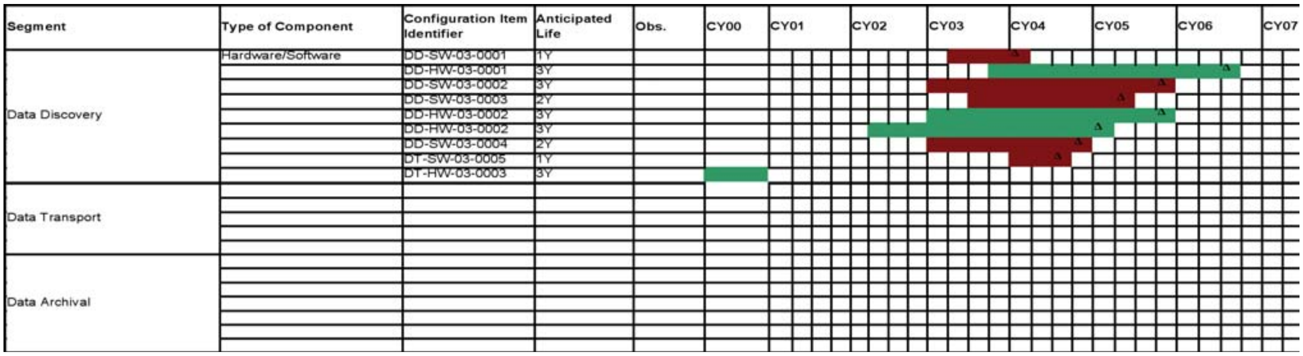
- Technology Refreshers—A change that incorporates a new product to avoid an ensuring end of life or product/COTS obsolescence, or to correct a problem identified via a customer,
- Technology Insertion—A change that incorporates a new product or function capability which is a result of industry growth or advanced development.[2]

The TRP should consider indicators that correspond to the categories of change shown above. For example:

Category	Indicators
Technology Upgrades	Vendor announcement of new technology
	Industry trends (e.g., Linux vs. proprietary operating systems)
Technology Refreshers	Reaching predefined age
	Component failure
	Repeated maintenance calls on the component
	Failure to meet the system requirement
	Mission failure
	Planned obsolescence of component resulting in vendor's inability to maintain
	The component's vendor has gone out of business or been acquired
Technology Insertion	Vendor or developer announcement of product line that meets or increases component capability
	Vendor or developer announcement of product line that decreases cost
	Industry trends (e.g., Linux vs. proprietary operating systems)
	Announcement of milestone of research and Development effort resulting in a new capability that can be applied to the DMAC

Initial phase of integration of the DMAC must consist of inventory and baseline of components in accordance with a Configuration Management Plan (CMP). The CMP will define the level of granularity at which to document the Configuration Items (CIs), for example, CIs could consist of integrated systems on down to individual computers, COTS software, etc. (Two types of Configuration Management can be conducted; an inventory and control of a baseline of inventoried items or management of specification items defining a baseline against which a repeatable manufacturing process is conducted. It is envisioned that the DMAC CMP will be concerned with primarily the first type, but both types will be addressed.)

[2] op. cit. Note 1.



Segment	Type of Component	Configuration Item Identifier	Anticipated Life	Obs.	CY00	CY01	CY02	CY03	CY04	CY05	CY06	CY07
Data Discovery	Hardware/Software	DD-SW-03-0001	1Y					🟥	🟥Δ			
		DD-HW-03-0001	3Y					🟩	🟩	🟩	🟩Δ	
		DD-SW-03-0002	3Y					🟥	🟥	🟥Δ		
		DD-SW-03-0003	2Y					🟥	🟥	🟥Δ		
		DD-HW-03-0002	3Y					🟩	🟩	🟩Δ		
		DD-HW-03-0002	3Y				🟩	🟩	🟩	🟩Δ		
		DD-SW-03-0004	2Y					🟥	🟥Δ			
		DT-SW-03-0005	1Y						🟥Δ			
		DT-HW-03-0003	3Y		🟩							
Data Transport												
Data Archival												

Figure 1: Example Refreshment Schedule against Configuration Items.

It is essential that the TRP be integrated with the Configuration Baseline; a sample template (partial) is shown in Figure 1. This represents a mapping of the Configuration Items to their planned life expectancy and replacement date. When CIs are to be replaced primarily on age, such a chart will allow the systems integrator to schedule and budget for replacements of the CIs. Note that DMAC is a system being integrated from existing components (e.g., smaller systems) that vary in age from current to legacy. The TRP must consider phasing of older components.

Issues resulting from the development of a detailed TRP must be reflected in the Systems Engineering Management Plan, as such issues must be considered in the design and integration of the DMAC. Some of the potential guidelines resulting from the TRP are as follows:

- Open Architecture—To the extent possible, systems should be acquired using components that are designed and built to open industry standards. Proprietary designs should be avoided. Open Source Software is preferred, as it provides for the highest degree of maintainability.
- Preference for COTS—Commercial off the shelf items are to be acquired preferentially over developmental items.
- Synchronization of components—Integrators should consider anticipated life spans of components and report a system life span that is the minimum of the components contained therein.

Just as a preference is indicated for COTS products rather than home grown, so too here, the DMAC will not attempt to reinvent the wheel on system life cycle maintenance. The authors recommend that in the preparation of the TRP, tailored approaches be considered, such as the Navy's Technology Assessment and Management Methodology (TeAM)[3] or the Technology Refreshment Cost Estimating and Planning Model[4].

[3] Technology Assessment and Management Methodology – An Approach to Legacy System Sustainment Dynamics, available at http: //smaplab.ri.uah.edu/dmsms98/papers/samuelson.pdf

[4] Technology Refreshment Cost Estimating and Planning Model: User s Guide, available at http://www.its.berkeley.edu/nextor/pubs/RR-00-5.pdf

I. Interoperable Data Discovery, Access, and Archive

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

Part III. Appendices

Appendix 7. Biological Data Considerations

Contributed by Dr. Fred Grassle

March 2005

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

Contents

Introduction

The Data Management and Communications (DMAC) Subsystem for ocean data from the Integrated and Sustained Ocean Observing System (IOOS) and NSF's Ocean Observatory Initiative (OOI) will accommodate marine biological data from a variety of sources and integrate these databases into a distributed system. Few studies of the marine environment provide information on accurately identified species placed in a context that is spatially and temporally resolved. The general availability of such data, along with modern navigational capabilities and real-time information on ocean processes at grid scales less than one kilometer, would bring about a revolution in management of marine resources by allowing the marine habitats on which species depend for survival to be physically and chemically defined. Present efforts to sample living resources on continental shelves or in deep-sea areas do not adequately consider habitat characteristics and related features of the underwater environment at the time of sampling, and as a consequence managers are unable to relate the abundance of natural resources to changes in the marine environment that alter the amount and quality of available habitats. Such changes in habitat are of serious concern to resource managers and the public because they can profoundly influence the availability of natural resources. Use of real-time data from IOOS will improve the information content and statistical reliability of biological sampling and reduce uncertainty associated with management decisions.

Data that can alert investigators to changes in the physical characteristics of the environment will allow targeted, adaptive sampling to measure specific physical-biological interactions. Understanding interactions among naturally occurring physical and biological factors is key to understanding how human actions may impact the animals and habitats of the oceans. Tools, such as acoustic and optical swath-mapping of the ocean bottom, provide essential information reflecting the distribution of marine organisms in relation to the physical and chemical factors that define their habitats. Each species inhabits a characteristic habitat that is defined by the environment, so that marine populations often respond sharply to ephemeral events such as climate oscillations, the passage of fronts, conditions favoring localized plankton blooms, and myriad other oceanographic phenomena at a variety of scales. Synoptic data on ocean conditions and associated plant and animal species is thus of vital importance to the managers, members of industry and the public whose livelihoods and well being depend on marine resources.

Ocean Biogeographic Information System

The Ocean Biogeographic Information System (OBIS—see http://iobis.org) is being developed to meet observing system needs. It provides international standards and protocols for accessing marine biological data, and can provide a biological subsystem for DMAC. OBIS offers a distributed Web presence for marine biological data in an up-to-date biological context. The many categories of marine biological data are typically maintained separately by appropriate authorities; many will be integrated into a distributed system accessible through a common portal to serve educators, scientists, and the general public. The OBIS portal is already providing a growing arsenal of tools for data analysis, synthesis, and visualization, as well as various access functions. Information from the heterogeneous assemblage of sources listed in the following section will be part of a distributed system accessed and integrated through functions of OBIS.

The OBIS system will allow anyone to click on a point or area of the ocean and obtain information on what lives there. Where does the blue-ringed octopus live? Where and how far does the Atlantic bluefin tuna travel in a year? What habitats have the most coral species, where are these places, and how are they changing? How do deep-sea animals find and distribute themselves with respect to the hydrothermal vents along the mid-ocean ridge? How do we compare the abundance of life from one place to another, and how do we map the myriad patterns of individual movement and behavior that enable each species to survive despite predators and competitors for resources? Answers to these questions are of vital importance to natural resource managers, and to many others who use marine information for work and recreation.

The basic information required is a species name (bioreference) and the latitudes, longitudes, and depth where it is found (georeference); time is also an essential dimension to understand variability within years or changes between years. If the information is from an ecological survey, information on abundance and sampling are necessary. The existing catalog of accessible, bioreferenced, and georeferenced information on life in the oceans is, at present, surprisingly small.

Marine Biological Data

SOURCES

- Government archives—In the United States, NODC and NASA maintain archives for ocean biological data. In Australia this activity is centralized in the newly formed Australian National Oceans Office, and in New Zealand the parallel organization is NIWA.
- Fisheries databases—Commercial fish species data are held for fisheries management areas within each country. International treaty organizations, such as the North Pacific Anadromous Fish Commission, the International Pacific Halibut Commission, and the Inter-American Tropical Tuna Commission have long time series of observations on catches of certain fish species shared among nations. The FAO has an international database that includes part of this information.
- Environmental protection agencies—For the United States, bottom assemblage data and habitat information are maintained by EPA and NOAA.
- Conservation organizations—Species information is used to define hot spots, endangered species, and harmful algal blooms.
- Museums—The major museums of the world have specimen-based databases on the species of the world. The value of these specimens is increasing as new methodologies for morphological and genetic analysis develop.
- Marine Laboratories – A number of marine laboratories, such as the Sir Alistair Hardy Foundation for Oceanographic Studies and the Scripps Institute of Oceanography, have geospatially referenced collections of plant and animal specimens, and related environmental data that span decades.
- Individual scientists—Individual taxonomic specialists and marine ecologists have extensive databases that have not been archived.
- Major oceanographic research programs—Programs such as JGOFS, GLOBEC, and RIDGE have important data sets that need to be accessible through an IOOS portal. OBIS is the data component of the Census of Marine Life (CoML) (http://coml.org) and its field programs including the Oceanic Pacific Pelagic (TOPP) program to track movements of large migratory species (e.g., tuna, swordfish, sharks, marine mammals, turtles, birds, etc.) using individual acoustic tags (http://www.toppcensus.org/). These data are being integrated with open-ocean IOOS data. Coastal listening arrays used to measure movements of individual salmon and other coastal migratory species will become an integral part of coastal observing systems (http://www.vanaqua.org/POST/).

Efforts are in progress to develop a standard classification of coastal habitats and this is one of the goals of the Ocean Biogeographic Information System.

SPECIES

The basic units for biological data are species. The names of these species and descriptions of them are the products of individual scientists whose careers have been devoted to describing and understanding evolutionary relationships among species. Increasingly, such individuals take advantage of DNA or RNA gene sequence data to differentiate among species and to trace their phylogeny. Each species is the unique product of its evolutionary history. Species are classified according to their evolutionary relationships using a well-established, internationally accepted hierarchical system of nomenclature. New species are continually being described (Figure 1 in Grassle, 2000); the hierarchical tree of evolutionary relationships among species and the associated nomenclature must continually be revised to incorporate new information. For this reason, biological data systems require more attention to metadata than do physical data systems.

The names of each species of plant or animal are the key words for information about organisms. Although ideally each species is known by a single, unique name, in practice, a species may be named more than once (creating synonyms) and the same name may be applied to more than one species (creating homonyms). Therefore, biological data systems require name translators that provide accurate scientific names from synonymous and homonymous names. In addition, translators are needed to relate common names to their scientific counterparts. With oversight from the Global Biodiversity Information Facility (GBIF), Catalogue of Life, and organizations such as the Integrated Taxonomic Information System (ITIS), Species 2000, and OBIS, the taxonomic authority for each major group of organisms maintains the accepted list of species.

Biological specimens are stored in museums and/or are maintained in culture collections, depending on the type of organism, to provide reference material for identification. Expert systems for identifications are being developed; identifications, be they based on morphology, on DNA or RNA sequence data, or on some other sort of distinctive feature, use accepted species names. As a minimum quality control and quality assurance measure, the taxonomic authority (the name of the person who originally described the species) and the name of the person identifying the specimen is typically included with each specimen record; such a practice is advisable for each data set based on such a specimen.

However, the units commonly used in biological oceanographic research are frequently not based on species, but instead are based on habitat, taxa above the species level (genus, family, order, etc.), size, chlorophyll biomass, optical or acoustic signatures, and trophic position in food chains or food webs. Such units are often used to quantify ecosystem function, and the precise identity of the component species is not considered germane to the question being studied. However, the relevance of species-level information to the inference of ecosystem function of species assemblages is an important area of research (Kinzig et al., 2001). The ambiguity or outright lack of information on the number and characteristics of species represented may be the product of a particular method of sampling (net tows, optical plankton recorder, flow cytometer, satellite data), which is incapable of making taxonomic determinations. There is also a shortage of expertise for resolving identifications to the species level.

The most accurate species-level information on organisms in the marine environment may, in general, reside in notebooks or computers of individual scientists, or in publications based on these sources. For terrestrial data, agencies and organizations responsible for environmental protection, land use policy, natural resource management, and protected areas, establish basic requirements for archiving the data and provide access to them. By contrast, for the most part, research activities in the marine environment do not have a system of governance to assign responsibilities for accurate environmental information, so marine biological data are insufficiently accessible and are accompanied by too little metadata to meet most user needs.

RELATING DATA ON PRESENCE/ABSENCE, ABUNDANCE, AND BIOMASS TO ENVIRONMENTAL PARAMETERS

Museum accessions usually record the existence of a particular species at a particular location (ideally, but not always, latitude, longitude, and depth). Data on age and/or size of individuals, and life-history stage are often included. Museum records are not designed to provide information about the overall species composition of a locality, since only the species found on a particular field trip or in a particular survey are recorded. However, if the data are arguably the result of an exhaustive comprehensive survey using a consistent sampling design, the possibility exists of inferring the probable presence of common species that may have been absent in a given collection. Thus the interpretation of counts of number of individuals, amounts of biomass, and presence/absence data require description of geospatial sampling patterns, effort expended and type of sampling gear as part of the metadata.

References

Grassle, J.F. 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional context. Oceanography, 13(3), 5–7.

Kinzig, A. P., S.W. Pacala, and D. Tilman (eds.). 2001. The Functional Consequences of Biodiversity, Empirical Progress and Theoretical Extensions. Monographs in Population Biology, 33, Princeton University Press, Princeton, New Jersey. 365 pp.

DMAC Team Members

DMAC-SC Data Archive and Access Team

Steve Worley, Team Leader, UCAR/NCAR
Landry Bernard, NOAA
Donald W. Collins, NOAA
Bob Cushman, CDIAC
Mark Fornwall, Maui Research and Tech. Center
Stephen Hale, EPA
Alex Kozyr, CDIAC
Sydney Levitus, NOAA
Chris Lynnes, NASA
Kenneth Rahn, EPA
Steven Rutz, NOAA
Kurt Schnebele, NOAA
George Sharman, NOAA
Gus Shumbera, NOAA
Shawn Smith, Center for Ocean Atmospheric Prediction Studies, Florida State University

DMAC-SC Data Discovery and Metadata Team

Susan Starke, Team Leader, NOAA
Anne Ball, NOAA
Julie Bosch, NOAA
John Caron, UCAR
Cheryl Demers, NOAA
Donald Denbo, NOAA
Dan Holloway, Univ. of Rhode Island
Lola Olsen, NASA
Karen Stocks, Univ. of California, San Diego

DMAC-SC Data Transport Team

Peter Cornillon, Team Leader, Univ. of Rhode Island
Steve Collins, NOAA
Donald Denbo, Univ. Washington
Allan Doyle, International Interfaces
James Gallagher, Univ. of Rhode Island
Dan Holloway, Univ. of Rhode Island
Tony Lavoi, NOAA
Ken McDonald, NASA
Reagan Moore, San Diego Supercomputer Center
Richard Owens, NOAA
John Ulmer, NOAA
Phoebe Zhang, Rutgers University
Eben Oldmixon, Consultant, Technical Editor

DMAC-SC Data Applications and Products Team

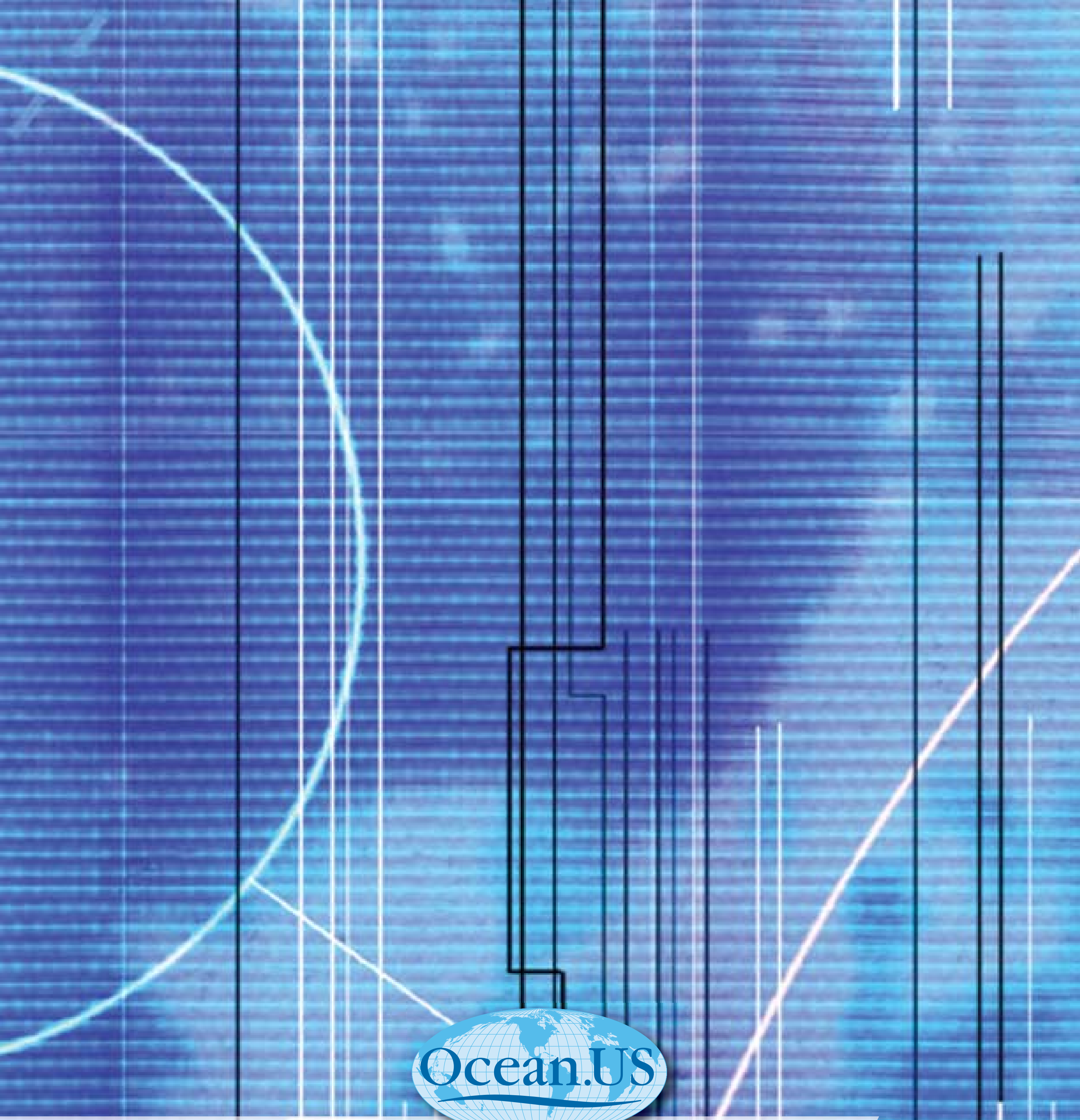
David Legler, Team Leader, US CLIVAR Office
Russ Beard, NOAA
Margarita Conkright, NOAA
James Cummings, Navy
Daphne Fautin, Univ. Kansas
Craig Kelly, Navy
Bernie Kilonsky, SOEST, Univ. of Hawaii

DMAC-SC Data Facilities Team

Landry Bernard, Team Leader, NOAA
William Birkemeier, USACE
Donald J Collins, NASA
Lee Dantzler, NOAA
Mr. John Jensen, NOAA
J. Edward Johnson, Navy
Dr. Tom Karl, NOAA
Joe Stinus, NOAA

DMAC-SC User Outreach Team

Phillip R. Mundy, Team Leader, Gulf of Alaska Ecosystem Monitoring and Research Program, Exxon Valdez Oil Spill Trustee Council
Philip Bogden, Gulf of Maine Ocean Observing System
Carol Dorsey, Alabama Department of Public Health
David L. Eslinger, NOAA
Larry Honeybourne, County of Orange Health Care Agency, Santa Ana, CA
Mark Luther, University of South Florida
Michael McCann, Monterey Bay Aquarium Research Institute
Roy Mendelssohn, NOAA
Malcolm Spaulding, Univ. of Rhode Island
Margaret Srinivasan, NASA
Joseph J. Tamul, Jr., Navy
Suzanne Van Cooten, NOAA



2300 Clarendon Boulevard, Suite 1350
Arlington, VA 22201-3667 USA

www.ocean.us

March 2005