

I. Interoperable Data Discovery, Access, and Archive

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	127
Introduction	127
Metadata Standards	129
Biological Metadata Considerations	130
Future Considerations	131
Development and Maintenance of Metadata	132
Adaptability of Metadata	133
Additional Issues	134
Data Discovery	137
Introduction	137
Catalog.....	138
Search Capability	140
Interface to Data Access	143
Portal	143
Annex A: Glossary of Terms	146
Annex B: Committee Membership	149
Annex C: Reference	150

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.

Part III. Appendix 1: Metadata and Data Discovery

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.

Part III. Appendix 1: Metadata and Data Discovery

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.

Part III. Appendix 1: Metadata and Data Discovery

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.

Part III. Appendix 1: Metadata and Data Discovery

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.

¹Tim 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 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.

Part III. Appendix 1: Metadata and Data Discovery

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 performance³. 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 or 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.

³Rowley, 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.

Part III. Appendix 1: Metadata and Data Discovery

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 sys-

Part III. Appendix 1: Metadata and Data Discovery

tem 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

Part III. Appendix 1: Metadata and Data Discovery

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.

Part III. Appendix 1: Metadata and Data Discovery

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 <i>The American Heritage Dictionary</i> , 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 <i>The American Heritage Dictionary</i> , 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 <i>The American Heritage Dictionary</i> , third edition)
Lineage	Direct descent from a particular ancestor; ancestry. Derivation. (From <i>The American Heritage Dictionary</i> , 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

Part III. Appendix 1: Metadata and Data Discovery

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 <i>The American Heritage Dictionary</i> , 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/)

Part III. Appendix 1: Metadata and Data Discovery

Semantic	Of or relating to meaning, especially meaning in language. (From <i>The American Heritage Dictionary</i> , 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 <i>The American Heritage Dictionary</i> , third edition)
Syntax	The rules governing construction of a machine language. A systematic, orderly arrangement. (From <i>The American Heritage Dictionary</i> , 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.