OCEAN AND COASTAL DATA MANAGEMENT

Jeff de La Beaujardière[1], C. J. Beegle-Krause[2], Luis Bermudez[3], Steven Hankin[4], Lisa Hazard[5], Eoin Howlett[6], Steven Le[7], Roger Proctor[8], Richard P. Signell[9], Derrick Snowden[10], Julie Thomas[5]

[1] NOAA (National Oceanic and Atmospheric Administration) Integrated Ocean Observing System (IOOS) Program Office, 1100 Wayne Ave #1225, Silver Spring MD 20910 USA, Email: jeff.deLaBeaujardiere@noaa.gov
[2] Applied Science Associates, Inc., 55 Village Square Drive, South Kingston, RI 02881, USA, Email: cjbk@Research4D.org
[3] Southeastern Univ. Research Assoc. (SURA), 1201 New York Ave. NW Suite 430, Washington DC 20005 USA, Email: bermudez@sura.org
[4] NOAA (National Oceanic and Atmospheric Administration) Pacific Marine Environment Laboratory (PMEL), 7600 Sand Point Way NE, Seattle, WA 98115 USA, Email: Steven.C.Hankin@noaa.gov
[5] Southern California Coastal Ocean Observing System (SCCOOS), Scripps Institution of Oceanography, 9500 Gilman Drive M/C 0214, La Jolla, CA 92093 USA, Email: lhazard@ucsd.edu; jot@cdip.ucsd.edu
[6] Mid-Atlantic Regional Coastal Ocean Observing System (MARCOOS), 55 Village Square Drive, South Kingstown, RI 02879 USA, Email: ehowlett@asascience.com
[7] Central and Northern California Ocean Observing System (CENCOOS), 1275 Columbus Ave, San Francisco, CA 94133 USA, Email: leho@saic.com
[8] Integrated Marine Observing System (IMOS), University of Tasmania, Private Bag 110, Hobart TAS 7001 Australia, Email: Roger.Proctor@utas.edu.au
[9] USGS (United States Geological Survey) Coastal and Marine Geology Program, 384 Woods Hole Rd., Woods Hole, MA 02543 USA, Email: rsignell@usgs.gov
[10] NOAA (National Oceanic and Atmospheric Administration) Climate Program Office, Climate Observation Division, 1100 Wayne Avenue, Suite 1202, Silver Spring, MD 20910 USA, Email: Derrick.Snowden@noaa.gov

ABSTRACT

We introduce data management concepts, including what we mean by "data" and its "management," sources of data, interoperability, and data geometry. We then discuss various components of a data management system. Finally, we summarize some existing ocean and coastal data management efforts. We make specific recommendations throughout the paper. We are generally optimistic that ocean and coastal data management is an interesting and solvable challenge that will provide great benefit to society.

1. DATA MANAGEMENT CONCEPTS

1.1. Definition of Data Management

Data management consists of the system (or network of systems) for assembly, storage, registration, dissemination, and permanent archiving of data collections, and of the enumeration and enforcement of standards and specifications regarding data quality and data handling. The operations within a robust data management system should be tested, reliable, scalable and secure. National efforts to standardize and integrate data management practices will aid in data dissemination and will ultimately advance research, decision-making, and public awareness of Earth observations. Ocean and coastal data management is a complex and evolving field. Some of the considerations are illustrated in Fig. 1.

For the purposes of this paper, we define *data* to include numerical values of physical, chemical or biological phenomena, whether directly observed or produced by simulation models or analysis algorithms, and shall also include associated metadata about the data and the processes used to obtain, derive, analyze or forecast it. Also, we consider that data management begins after observations or simulations have been performed and the results transmitted to their initial storage facility. We do not address issues of data telemetry, satellite downlinks, protocols for cabled observatories, or data transfer between components of a numerical model.

1.2. Connecting Users to Data

The overarching goal of data management is to enable users to be able to find access and utilize the data through time, including the past, the present, and forecasts of future conditions. Traditionally this data integration has been done by scientists who engaged in the labor necessary to obtain and analyze data from different sources and formats. Examples include analyzing data from buoys and satellites to study ocean temperature, or using assimilation of atmospheric and oceanographic data to improve model results. The output from the consumption and analysis of the data are derived products that can be used for decision support. Significant efforts have been underway such as the Global Ocean Observing System (GOOS), EuroGOOS, and U.S. Integrated Ocean Observing System (IOOS) to promote standardized data management practices that will reduce effort for

existing users, make data usable by a broader class of non-specialized users, and allow the automation of routine data access, analysis and transformation tasks.

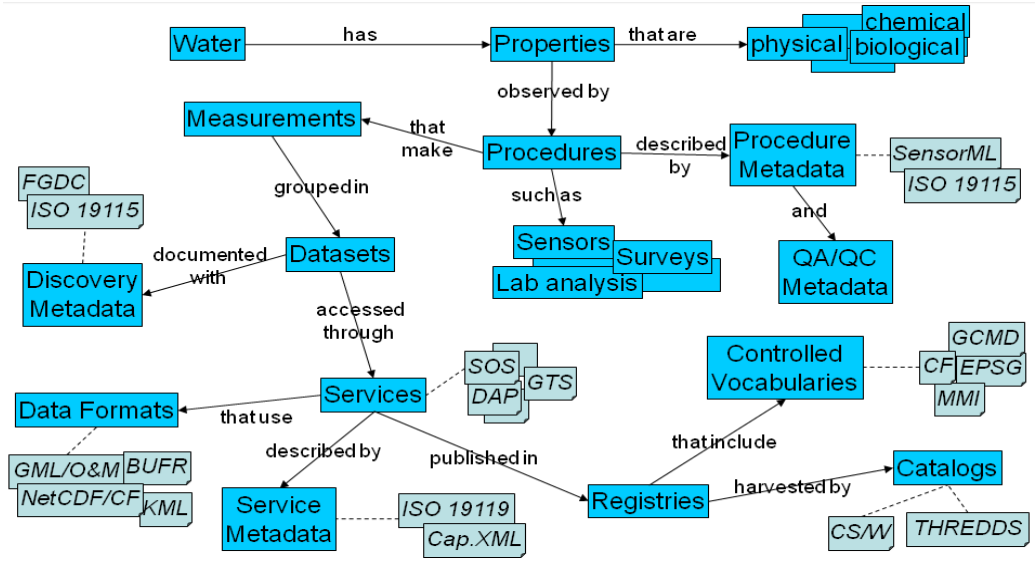

Figure 1: This information diagram suggests the complexity of the ocean and coastal data and metadata management problem.

Allowing user-determined temporal coverage is a key challenge in data access. Observations of the present or the recent past are often provided by one source, while quality-controlled archived data are often located with another source. Information access requiring aggregation of fields from the past through to the present and even into future times should be transparent to the user. Seamless access across time and disparate sources provides value to more users, particularly non-specialists. Metadata about lineage (provenance and subsequent processing) needs to be preserved.

Users include scientists, decision-makers and their advisors from policy and event driven (e.g. emergency) levels, the general public, and the operators who would monitor the integrated data management system. Access is typically mediated by software applications including analysis tools, geographic information systems (GIS), decision-support tools, and popular Internet browsers and applications.

As the need for ocean data increases among non-specialized users, data dissemination should be simplified. User-accessibility issues deserve attention, particularly in lowering any barriers preventing a user from acquiring ocean data. Access via analysis software, web browsers, and mobile devices should be supported. Any enhancement that enables data discovery and access across the integrated network of ocean observing systems, and facilitates the transformation of data to information, helps bring understanding of the ocean to people.

1.3. Open-Ocean vs. Coastal Data

Are data management and distribution problems in coastal ocean observing fundamentally different than in global climate-oriented efforts? We do not believe so. Both global and coastal ocean data management must serve a range of requirements, including: sustained measurements of high quality that can form the basis for detecting changes in climate and in ecosystems; regionally unique, one-off sets of observations made in response to events (e.g. oil spills or hurricanes); and ongoing observations suitable for short-term forecasts and interpolated estimations of state. There are needs for archiving and for breadth and speed of dissemination that are distinct for each of these classes of usage. Given the state of data management solutions that are available today, there is no proven single solution that can address this full range of requirements. Both the open ocean and coastal realms need effective community processes that can leverage the strengths of existing systems, incrementally grow those solutions based upon their strengths, and foster exploration, testing and evaluation that lead to incorporation of newer and more powerful solutions.

The coastal community – home to most of the world's population – requires a regionally-sensitive capacity-building process that exceeds what the World Meteorological Organization (WMO) and the WMO Global Telecommunications System (GTS) can alone provide. We recommend that data management practices be coordinated at the national level for each

country's waters and that a forum for international coordination and interoperability be established to ensure that regional efforts remain well integrated into the global solutions. In the US, recent legislation has directed the establishment of an Interagency Ocean Observation Committee to provide national coordination. In the EU, the Infrastructure for Spatial Information in Europe (INSPIRE) Directive includes both terrestrial and marine data in its scope. A decade ago, Australia's Marine Science and Technology Plan [1] had already recognized the need for pan-Australian coordinated marine data management.

1.4. Models as information sources

Numerical models are driven by observations. As stated earlier, for the purposes of this paper we use the term "data" broadly to include those outputs. We recommend that data management practices share standards and infrastructure to the extent feasible for both model outputs and measured values. It is important to recognize that modern models are no longer confined to rectangular grids and there is a need to support unstructured grids that use non-orthogonal cells, such as triangles and quadrilateral shapes. Advanced three-dimensional unstructured adaptive mesh circulation models require rethinking of the term "grid".

1.5. Interoperability

Interoperability is a key tenet of a successful data management system. We define *interoperability* as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [2]. This implies standardization at many different points in the system, including:

- the services which provide access to data;
- the formats and encoding conventions for those data;
- metadata about observations, observing systems and models;
- metadata about data lineage (provenance and processing) and quality control;
- controlled vocabularies for key metadata values such as physical quantities, units, and coordinate reference systems.

Interoperability includes both *syntactic* interoperability (agreements regarding data formats and request messages, for example) and *semantic* interoperability (such as agreements on vocabularies and identifiers). The advancement of technology has improved the network of ocean systems and provided better human/machine communication; however, machine/machine communication is still an ongoing challenge. Unlike humans, machines cannot comprehend data unless meaning has been embedded into that data. The "Semantic Web" promises to allow computing systems to interact and carry out specific tasks intelligently in the absence of human intervention [3]. With enhanced tools for defining data relationships and improved inference engines, the semantic web has evolved significantly since its inception over a decade ago, and its application can be found today in many industrial sectors including bioinformatics, pharmaceutical, military, seismic exploration and an increasing number of Internet applications. This offers the potential for solving a number of data integration challenges. For example, ocean observing terminology (physical parameters, sensor types, units of measure, etc.) is defined inconsistently and differently from one data provider to another. The semantic web, by using ontologies to define concepts and their relationships, can allow data providers to map their local vocabularies to shared community vocabularies.

1.6. Standards Processes

A critical aspect of interoperability is standardization. All other things being equal, we recommend that existing open-standard approaches be used in preference to purpose-built or proprietary technologies. However, we recognize that some standards evolve through broad public adoption (e.g. Google KML (Keyhole Markup Language)), and need to be considered in addition to standards designed for specific science-based data.

Though a challenging problem, progress towards interoperability can be made independently and sequentially on various fronts. Also, we stress that interoperability does not require the abandonment of all legacy approaches: standardized practices can be adopted alongside pre-existing *ad hoc* practices.

Open standards have often been developed for purposes other than oceanographic data handling, and must then be adapted to that need. This may mean defining a profile of a broad standard wherein optional elements are made mandatory or prohibited, or defining extensions to a narrow standard. Organizations such as the Global Earth Observation System of Systems (GEOSS) and the Infrastructure for Spatial Information in Europe (INSPIRE) have standards-adoption processes to assess the suitability of standards or profiles thereof for their needs. We recommend close communication and information exchange between such groups to ensure that common standards are adopted wherever possible and duplicative or conflicting work is not performed.

Interoperability strategies require community consensus. Reaching community consensus requires a process, though the details of those processes that have proven to be successful have taken many shapes and forms. Organizations such as the WMO, the Internet Engineering Task Force (IETF), the International Organization for Standardization (ISO), the Open Geospatial Consortium (OGC) and other industry

consortia all represent variations of formal *de jure* processes for reaching consensus. Levels of openness (that is, of freedom to participate in the process and influence the standards) vary among these organizations. Grass-roots organizations such as the CF (Climate and Forecast) conventions on-line forum (http://cf-pcmdi.llnl.gov/) may spring into existence in response to a community IT needs and agree upon their own process. *De facto* standards (such as the KML format for Google Earth) may become formally approved by a standards body (such as OGC in this case).

A key lesson that should be taken from the previous decades of IT history is that it is vital that a technology not be mandated as a standard until it has demonstrated suitability for its intended purposes through testing in systems of realistic complexity and the creation of reference implementations. The so-called "fluid earth science" domain of ocean/atmosphere/climate sciences requires data management solutions to perform functions and meet thresholds of mathematical sophistication that are not commonly found in other disciplines. Hankin *et al.* [4] therefore recommend a "pragmatic and skeptical approach" to standards adoption.

1.7. Data Geometry

Too often, data management approaches are customized for individual programs or specific observed quantities. We recommend instead that the *geometry* of the dataset be the primary driver for any needed differences in methodology. For example, two-dimensional data on a regular latitude/longitude grid can be handled by the same types of encoding formats and services for data access, visualization, and subsetting and coordination transformation, regardless of whether the source was a numerical model or a Level 3 satellite image. Gridded data with a vertical or time component will need somewhat more sophisticated treatment, yet all can be represented by a general 4D (time, depth, latitude, longitude) data model. Similarly, collections of measurements at isolated points can employ the same data management methods regardless of source—buoys, stream gages, anemometers—but may require different treatment than gridded data. Vertical profiles, frequency spectra, and moving *in situ* sensors add complexity to the surface observation case. Unstructured grids (e.g. triangulated irregular networks) require different data management approaches than regular grids. Management of marine imagery from autonomous underwater vehicles must account for variable height above the seabed. Finally, some (not all) marine biological datasets may require a specialized data management treatment.

2. COMPONENTS OF A DATA MANAGEMENT SYSTEM

Data management provides the bridge between the data and its users. In this section, we discuss the role of the providers of data into the system, some of the desired functions of the system, and the role of the applications which consume data from the system. The data management system does not have control over all aspects of the observation or modeling process, but can promote interoperability by supporting a limited set of well-defined interfaces, formats and practices at each data source. Similarly, client applications are not under control of the data management system but should interoperate with its components. Figure 2 illustrates many of the components which are desirable in a system for ocean and coastal data management. We discuss these components below.

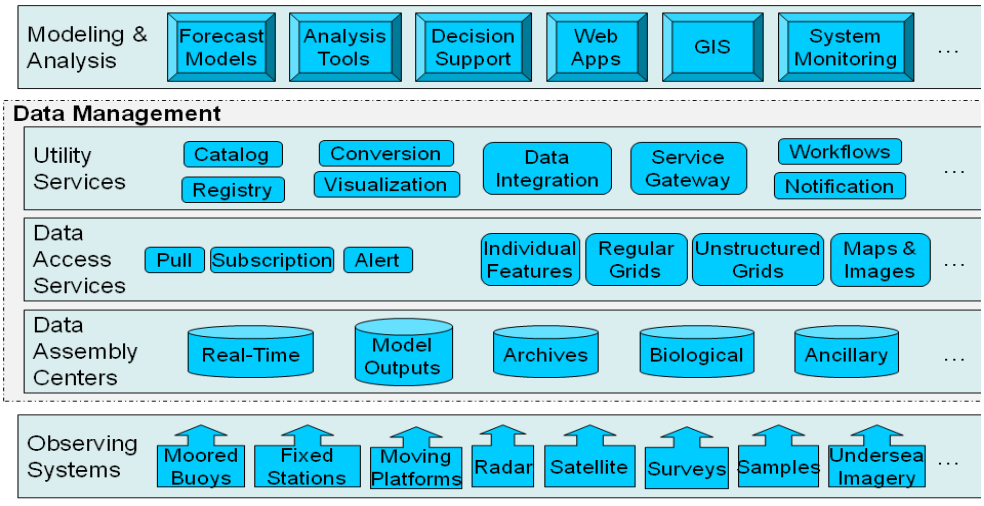

Figure 2: The data management system connects observing systems to modeling and analysis components.

2.1. Data Assembly Centers

The concept of a large-scale distributed data network where data providers can push their data into the network by simply using the appropriate standards is appealing. Data can come from a variety of sources such as radar, satellite, in-situ measurements, drifters and gliders, models, and synoptic analysis from domain scientists. The challenge is that many data providers or research scientists do not have the IT capacity or infrastructure to maintain data servers and manage the auxiliary requirements such as registration, metadata, and quality assurance. Smaller providers can therefore arrange for a larger entity to provide data management services, as described below.

2.1.1 Real-Time Data Assembly Centers

A Data Assembly Center (DAC) is defined as a facility that obtains data from multiple observing systems or platforms, aggregates the information into local databases or file structures, performs quality control tests, and adds (or organizes) appropriate metadata. The DAC is the cornerstone of a data management system, providing the initial stewardship and dissemination services for the data. A DAC may enable access to the data by end-users, by forecast models, and by archives for permanent storage.

A DAC may be one of the WMO operational meteorological service centers. A DAC may receive all or part of its data from the GTS, and may send all or part of its data out via the GTS. A coastal DAC should provide the services needed to ensure effective bidirectional integration of the coastal and global data streams. Specific funding may be required to ensure the sustained, operational functioning of DACs.

A real-time DAC focuses on current and recent observations. A DAC may send older observations to an archive and delete the local copy. Algorithms must quickly perform automated quality control, and bandwidth must be able to support surges in demand based on emergency conditions. Users must be able to pull data on request and to subscribe to data streams.

Observations may include *in situ* measurements from platforms such as moored buoys, fixed stations, drifters, volunteer observing ships, gliders, as well as remotely sensed data from satellites or coastal high-frequency radar installations. Different observation methods may require different access services, formats and metadata, but we recommend that similarities be exploited as much as possible rather than inventing *ad hoc* approaches for each observing project. We recommend that observations be made available in their native coordinate reference systems, and transformed to other space and time axes only as needed for derived products and analyses.

2.1.2 Archives

The role of an archive is to preserve data indefinitely and reliably in a manner that allows retrieval in the future. Ocean data stewardship is discussed at length in [5]. Traditionally, some archives have adopted a policy to keep copies of submitted data exactly as they were submitted. This allows for improvements in quality control or interpretation in the future. However, the risk exists that legacy formats will not always be readable, or that the cost of conversion of many different formats will be prohibitive. Therefore, we recommend that data also – or instead – be archived in a modest number of well-defined and preferably self-describing formats that allow for the possibility of automated translation to other formats in the future with sufficient metadata to describe their origination.

An archive focuses on historic data. Aggregation across the temporal boundary between the more recent data at the DAC, and the archived data, perhaps physically located elsewhere, should be as transparent to the user as possible, so an archive should support data request services and formats in common with those used at the DAC.

2.1.3 Model Outputs

Models produce a large variety of different data, including retrospective analyses, short term met/ocean forecasts, and predictions of climate change. Many models generate output files on their native grid system with a native output format. Essential to maintaining the maximum scientific content from these models and allowing for the greatest range of potential derived products, numerical model data should be maintained and delivered by the data management system on the native grid, but delivered to clients in an interoperable fashion [4]. Models with triangular or unstructured grids pose some additional difficulty, as data models and methods for aggregating, subsetting or transforming those grids are currently less standardized.

2.1.4 Biological and Other Environmental Information

Some observations, especially in the realms of marine biology and water quality, are not sensor-based at all but instead depend on trawl surveys, laboratory analysis or other techniques. Such data should also be made available on-line using, to the greatest extent possible, data management practices that are interoperable with those used for physical real-time data. Geospatial information on migratory birds is an interesting data question, as most spend time *over* land, *over* water and even foraging *in* water.

Anthropogenic chemicals of interest for water quality frequently undergo chemical reactions within the water. Data streams for reactive chemicals need to consider precursors chemicals, degradation products, and suite

measurement characteristics (what was or was not measured, what was measured but not found, etc.).

Systems such as Ocean Biogeographic Information System (OBIS, http://www.iobis.org/) allow spatial exploration of locations of marine animals and plants, including tools for creating tables and predicting distributions using environmental information.

2.1.5 Ancillary Information

Observations are generally more useful in a human context. Ancillary information refers to geographic framework information that is independent of the measurement data and model forecasts, such as political boundaries, shorelines, and marine or terrestrial features. Because viewing data in the context of ancillary information is often necessary, we recommend that interoperability be enhanced among sources of ancillary information and between that information and the actual data.

2.2. Data Access Services

Data access services enable a human user or software application to obtain data stored in one location and to transfer it to a different location for actual use. In this paper we focus on internet-accessible services as opposed to, say, replication between master and slave databases.

As noted above, different data geometries – *in situ* features, gridded coverages, unstructured grids, etc. – may require different access services.

The most basic service type allows the user to "pull" data by explicitly requesting it. We recommend that pull services allow the user to constrain the geographic location and the time covered by the information and receive an aggregated dataset if possible.

Subscription services "push" data to registered users. The subscription may be for all observations, or may be alert-based and only send data when some threshold value has been reached. The WMO GTS is an example of an existing subscription service that serves a core set of national operational meteorological service centers. Other service types may be necessary to enable ad hoc, unofficial or short-term subscriptions by data users who are not qualified to serve as WMO centers.

2.3. Utility Services

Utility services provide functionality beyond data access. Utility services include transformation, aggregation, integration and discovery.

2.3.1 Transformation and Integration

Transformation services include visualization, format conversion and coordinate transformation. These are functions that can be applied to the data by network-accessible services. Data management systems often do not provide such functions, leaving client applications to do this work after data retrieval. We recommend that standalone services be made available. This allows light-weight clients (web browsers, cell phones) to access data and enables the creation of service chains. (An example of a service chain would be a "script" that fetches data from one service, feeds it to another service for transformation into the desired coordinate system, and then to a third service for visualization before handing off the transformed data to the client for display.)

Data integration services are very useful. Users often need a unified presentation of all measurements of some quantity regardless of source. That can be accomplished visually (by having each source an independent layer in a display) or numerically (through suitable concatenation, interpolation, or assimilation into a model). Preservation of data lineage and metadata is important, and should always be available to the user when lossy or algorithm-dependent integration is performed.

Information integration – the ability for the user to assemble and maintain a heterogeneous set of data, metadata and annotations relevant to a topic or phenomenon of interest – is likewise desirable. An analogy with the commercial web might be a "shopping cart" or "wish list" that a user can create at an on-line store and retrieve later by logging into the same web site.

2.3.2 Catalogs and Registries

The ability to find data in a distributed system is essential. Users should be able to find information based upon geography, time and observed property, without regard for the source of the data. However, the source should be indicated, and ideally qualified by maturity level. A Catalog Service should be based upon open standards, and queryable both via a human user interface and a software query language.

A Registry is closely related to the Catalog. For example, a Registry might include the list of all known data access services of a particular type. This list changes infrequently, as new services are added or removed from the network. The Catalog can query the Registry to get the list, and then regularly harvest the table of contents of each service to determine what data holdings are currently available. A Registry may also support the semantic web approach by holding controlled vocabularies, coordinate reference system identifiers, and other metadata about classes of objects.

Various efforts exist to standardize registry and catalog query interfaces. We applaud those efforts, but we also recommend further research in the topic of making ocean data and metadata discoverable by commercial web search engines in a semantically rich way. For

example, instead of a search for "temperature" merely returning URLs of web pages that include that word, including advertisements for thermometers and aspirin, we would like the ability to search for temperature as an observed property of the ocean on a given date, in a given range of latitude and longitude, and available from a particular service types, and to be shown a list of URLs that return actual observation values from data access services. Physical data are more easily amenable to these types of services, but other environmental data (chemical, biological, etc,) should be added as much as possible.

2.4. Crosscutting Considerations

Crosscutting considerations are those elements of a data management system that apply to the entire system or to all of its components individually. These considerations include metadata, data quality, and operational reliability.

2.4.1 Metadata

In the simplest case, metadata is information about data, such as a high-level description of a dataset including the source, coverage area and time, and so on. Ocean and coastal data management requires more rich metadata about not only datasets but about the sensors that make measurements, stations that contain multiple sensors for different phenomena, networks of stations, numerical models, and data lineage information algorithms for deriving physical values from observed quantities, quality control processes, and derived products. Metadata can be managed as actual files on computer disks, but is more useful when the "documents" are metadata instances generated upon request from a database.

We recommend the adoption of standardized metadata formats and profiles. We also recommend that metadata be treated as a linked set of "documents" that each contains a different subset of the metadata. Figure 3 illustrates this concept, where separate resources describe a type of sensor, a specific sensor of that type, the station on which that sensor resides, the type of station, etc.

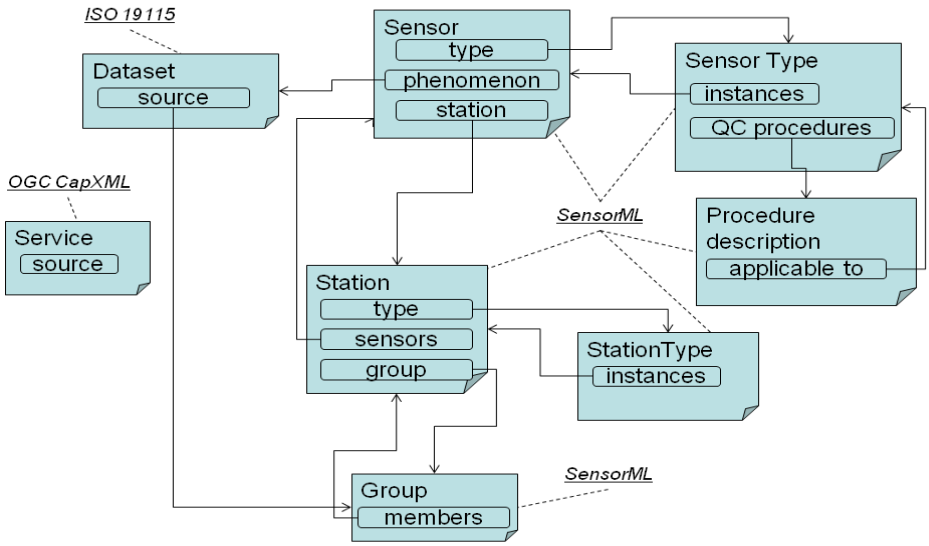

Figure 3: Conceptual diagram of a linked set of metadata resources that describe a particular sensor, the station on which it is mounted, the network to which the station belongs, and the general types of each instance.

2.4.2 Data Quality

Quality Assurance and Quality Control (QA/QC) are integral to addressing data management. No matter how efficient and successful the process for data handling, the end goal is to have high quality and accurate data. Quality assurance ensures that the instrument is calibrated to the highest standard, and quality control addresses the data stream, ensuring that the best QC methods and metadata are employed. Data QA/QC should be a continual process of the data production to ensure data always meet specified data standards. Ideally, all DACs would conform to defined standard QA and QC methods and analysis. In practice, there will be levels of maturity regarding QA/QC.

Differing QC methods and procedures are being used in various ocean observing communities for other ocean data variable types. For consistency, the same

set of algorithms should be uniformly applied to a specific data type to ensure that all data meet a known level of quality. At minimum, descriptions of the QC procedures and their results should be expressed using the same metadata standards even if the methods themselves cannot be harmonized. Users may need to select a variety of levels of data quality when constructing long time series. For example, ocean water characteristic data collected 50 years ago will not have the same accuracy as observations collected today, but are still desirable for use in analyses of long term trends. So long as the researcher understands the different quality in individual time series segments, useful insight can emerge.

Within a distributed data network where data are initially collected by data providers and then aggregated at a regional or national level, the QC process can be applied at different stages. Ideally data should be quality controlled immediately after the initial data collection; however, many data providers at this level may not have the adequate infrastructure or resources. QC methods may therefore be applied when data get aggregated at the regional level. Once a dataset is quality controlled by the community's sanctioned QC methods and algorithms at a collection or an aggregation point, and documented with appropriate quality metadata, the dataset should be trusted with a high level of confidence and may not need to be quality controlled again downstream.

2.4.3 Operational Reliability

Not all components in a comprehensive data management system will have the same level of performance, reliability, and sophistication. For example, a data center with redundant hardware and power, staffed by personnel 24 hours a day even during a hurricane, and with dedicated resources for quality control, is clearly of a different class than a university effort performing observations with graduate students. Similarly, an industrial data customer is of a different class than a casual visitor. Nevertheless, the volunteer data provider and the occasional data user both have relevant contributions and requirements. We recommend that levels of capability maturity [6] be defined for various roles in the data management system, and that the maturity of data and service providers be indicated to users when searching for data.

2.5. Customer and client applications

Customer and client applications are as varied as the users of the data. Scientists may download raw data for further analysis within custom applications and models. Third-party providers may produce value-added products such as descriptions, summaries, and visualizations and then provide packaged information via subscription. The public may browse freely-available data from national or regional organizations. The maintainers of these data management systems themselves may need to monitor the system and data flows. Disaster response teams may need to receive alerts when critical thresholds are exceeded. We recommend that robust, well-documented data formats be supported to enable conversion as needed to simpler representations for end users. We see that a key factor that will define the success of an integrated ocean observing system is the ability for different users to access data with different software clients.

2.5.1 Integration with GIS

Geographic Information Systems (GIS) are an important class of client application that have not traditionally been used by ocean scientists but are in broad use elsewhere. GIS allows users to "layer" a variety of disparate geographic information typically in two main classes: features (e.g., points, lines, and polygons) and coverages (gridded data, typically on uniform rectangular grids), and to perform complex spatial analysis on these data. (Note the correspondence of these classes to data geometry discussed in Section 1.7.) Once a tool only for professionals, the advent of popular tools such as Google Earth and other GIS-like web-based mapping applications have greatly broadened the use and understanding of GIS by students, government officials, the general public, and scientists from a broad range of disciplines. These new users now expect to view science data in a map-based environment.

The challenge in meeting this demand is that GIS specialists do not always share a common foundation of concepts with fluid earth scientists. In fluid-earth–science, the atmosphere and the oceans are regarded as 3-dimensional, time-dependent and continuous. Seemingly simple GIS concepts like a "feature" become ambiguous when it is realized that a 1-dimensional sequence of points in the vertical (a profile) is actually a discrete sampling of a continuous field, which meets the definition of a "coverage". Many of the most common features in ocean science, for example the location of an eddy or meandering current, are time dependent and do not readily fit the traditional GIS concepts of a 'feature'.

Thus, while some problems of ocean GIS integration are simple, and effective interoperability bridges are rapidly going into production (e.g. surface ocean conditions at a point in time as GIS map layers), a deeper integration of the fluid-earth-sciences with GIS concepts is likely to require many years. As a practical matter, much progress can be made in the near term simply by working with ocean data providers to ensure that the fullest possible geo-referencing information is included in datasets where often such considerations are commonly ignored today – e.g. a circulation model set on a spherical earth, rectangularly gridded coastline

and missing small islands. Such georeferencing can help users integrate these outputs with fine-scale biological data (e.g. beach nesting areas of an endangered species) in a GIS framework.

2.5.2 Modeling and Analysis

Numerical models are essential for scientific understanding, for weather and climate forecasting, and for tracking and predicting oil spills, algal blooms, and providing the optimal search pattern for persons lost at sea. Models create continuous data fields and predictions for further examination. We have previously discussed models as a source of data to be managed. Models are also consumers of data, in the form of initial and boundary conditions, assimilation fields during model runs, and assessment of model performance. Data management infrastructure must support the needs of modelers.

3. SOME EXISTING OCEAN AND COASTAL DATA MANAGEMENT EFFORTS

(Note: The efforts described here are representative, not exhaustive.)

3.1. WMO Information System

The WMO Information system (WIS) is the pillar of the WMO strategy for managing and moving weather, water and climate information in the 21st century. WIS will provide an integrated approach suitable for all WMO Programs to meet the requirements for routine collection and automated dissemination of observed data and products, as well as data discovery, access and retrieval services for all weather, climate, water and related data produced by centers and Member countries in the framework of any WMO Program. WIS is being designed to dramatically extend WMO Members' ability to collect and disseminate data and products. It will be the core information system utilized by WMO Members, providing linkages for all WMO and supported programs associated with weather, climate, water, and related natural disasters. It is being built upon the Global Telecommunication System of WMO's World Weather Watch, using standard elements and at a pace feasible for all Members.

We recommend coordination between WIS and the other efforts described in this section on standards adoption and technology development.

We note that the pace of change at a global coordination level is typically much less rapid than the rate of change of technology. We therefore recommend that the WMO establish clearly-separated roles and responsibilities for, on the one hand, high-level policy, guidelines and functional requirements, and on the other hand, the technical implementation details implementation. The latter should be able to respond nimbly to technical changes in ways that are transparent to users, allow differing practices behind standardized interfaces, and do not violate the high-level policy, guidelines and requirements.

3.2. Integrated Ocean Observing System

The Integrated Ocean Observing System (IOOS; http://ioos.gov/) is the US coastal component of the Global Ocean Observing System (GOOS), which is the marine component of the Global Earth Observing System of Systems (GEOSS; see below). IOOS includes both US Federal agencies and Regional partners; the National Oceanic and Atmospheric Administration (NOAA) is the lead agency. IOOS plans explicitly call for a Data Management and Communications (DMAC) subsystem to link observations to models, analysis tools and users. The NOAA IOOS Data Integration Framework (DIF) project [7] is establishing DMAC capability on a small scale to assist specific customers and to assess the viability of particular technical approaches. The customer groups include models or decision support tools relevant to coastal inundation, hurricane intensity forecasting, harmful algal blooms, and ecosystem assessment.

The IOOS DIF project has worked with several DACs to establish standardized data access services including Open Geospatial Consortium (OGC) Sensor Observation Services [8] for *in situ* data, OGC Web Coverage Service [9] and OPeNDAP/CF/NetCDF (Open-source Project for a Network Data Access Protocol/Climate and Forecast/ Network Common Data Form) [10] subset service for gridded satellite data and model output, and OGC Web Map Service [11] for images of data. IOOS is also developing metadata profiles for observing systems using Sensor Model Language [12].

Current NOAA DACs include the National Data Buoy Center (NDBC), the Center for Operational Oceanographic Products and Services (CO-OPS), and CoastWatch. IOOS also supports data assembly and quality control at DACs such as NDBC. Data from NDBC and CO-OPS are also disseminated to official subscribers via the WMO's pre-existing Global Telecommunications System (GTS). IOOS is considering establishing a service gateway that would broaden the subscription and alert capability using open-source standards and additional formats.

In collaboration with the US National Science Foundation, the DIF project is also testing the use of "cloud computing" (virtual server capacity hosted by commercial providers) to provide a scalable format-conversion service. In addition, IOOS is elaborating metadata profiles for discovery, sensor descriptions and QA/QC information, and is planning to use or establish Registry and Catalog components.

Besides the DIF effort, IOOS funds observing and data management capacity at regional coastal ocean observing system nodes in the US. IOOS arranges for regional observations to be fed onto the GTS via NDBC. IOOS is committed to ensuring that US coastal observations will be included in the global ocean data framework to the greatest extent feasible, and will work with the global community to expand the opportunities for integration of new parameters, such as biological and chemical observations.

3.3. Australian Integrated Marine Observing System

Marine data and information are the main products of the Integrated Marine Observing System (IMOS, http://www.imos.org.au/) and data management is therefore a central element to the project's success. The eMarine Information Infrastructure (eMII) facility of IMOS provides a single integrative framework for data and information management that will allow discovery and access of the data by scientists, managers and the public. The initial strategy has focused on defining specific data streams and developing end-to-end protocols, standards and systems to join the related observing systems into a unified data storage and access framework.

IMOS data streams can be categorized in four ways: gridded data from satellites and HF radar systems; time series data from moorings, Argo floats, gliders and ships of opportunity; image data from Autonomous Underwater Vehicles; biological data from continuous plankton recorders and acoustic tagging. The first two provide real-time and delayed-mode data sets whereas the latter are delayed-mode delivery only.

The IMOS data management infrastructure employs OGC standards wherever possible. The main components of the system are: the Australian Research Collaboration Service (http://www.arcs.org.au/) Data Fabric 'cloud storage' incorporating OPeNDAP/THREDDS (Thematic Real-time Environmental Distributed Data Services) servers hosting CF-compliant NetCDF, HDF (Hierarchical Data Format) or GeoTIFF (Geospatial Tagged Image File Format) data; the open-source GeoNetwork (http://geonetwork-opensource.org/) Metadata Entry and Search Tool (MEST) for metadata cataloguing; SensorML, which provides standard models and an XML (Extensible Markup Language) encoding for describing sensors and measurement processes; the open-source DataTurbine (http://www.dataturbine.org/), data streaming middleware providing the foundation for reliable data acquisition and instrument management services; a web portal (http://imos.aodn.org.au/) using the open-source ZK Ajax framework (www.zkoss.org) and the OpenLayers geospatial framework (http://openlayers.org/) incorporating access to Web Services.

A distributed network of OPeNDAP/THREDDS servers around Australia forms the primary data storage. This complements the regional nodal structure of IMOS and allows rapid access to data by the local research community. Each local server also supports the GeoNetwork catalog with, wherever possible, automatic harvesting of metadata from the OPeNDAP/THREDDS system. An IMOS NetCDF standard ensures that all necessary metadata complying with ISO 19115 can be automatically extracted from the NetCDF files. Automation of metadata creation from non-NetCDF datasets is also being investigated. A master GeoNetwork catalog at the University of Tasmania (http://imosmest.aodn.org.au) routinely harvests new metadata records from the regional catalogs to maintain a central registry.

Data storage and retrieval in IMOS is designed to be interoperable with other national and international programs. Thus, it will be possible to integrate data from sources outside IMOS into IMOS data products, and IMOS data will also be exported to international programs such as Argo and Oceansites. Also, most of the real-time data of physical parameters will be exported to the GTS.

3.4. GEOSS

The Global Earth Observation System of Systems (GEOSS) is an international infrastructure that is connecting users, producers and integrators of environmental information. One of the GEOSS goals is to make environmental information publicly available to a broad set of users.

The core components of GEOSS are the "Components Registry" and the "Standards and Interoperability Registry". The Components Registry's main purpose is to provide a centralized place to register and access GEOSS components (e.g., organizations, web services, software, models). The Standards and Interoperability Registry's main purpose is to provide a centralized place to register and access standards and "special agreements" among communities.

Ideally, multiple Registries will exist and will communicate with each other via standardized protocols and interfaces. An organization that makes available ocean observations could register those services in an existing GEOSS registry or could create a community registry that will connect to other GEOSS registries. In practice, however, this registry infrastructure is not yet fully developed.

3.5. Other Projects Relevant to Data Management

The Marine Metadata Interoperability (MMI; http://marinemetadata.org/) project is providing

registry services, guides and workshops to facilitate creation of vocabularies and mappings that could work with Semantic Web tools.

Quality Assurance of Real-Time Data (QARTOD; http://qartod.org/) is a NOAA-funded effort addressing data QA/QC. At this time, three data standards have been submitted to the IOOS DMAC standards process by QARTOD:

- Real-Time Quality Control Tests for In Situ Ocean Surface Waves
- High Frequency Radar Surface Currents
- Quality Control Standards for Real-Time, In-Situ Currents Measured by Teledyne RD Instruments

4. CONCLUSION

We have come to expect instantly available data and information. Stewardship of our planet requires interdisciplinary information integrated into a variety of decision support frameworks. Access to different types of data, stored at different locations and crossing a variety of temporal boundaries (past, present, future) and length scales (local, global) should be as seamless to the user as possible. From the computational world of numerical models to the real world observations and locations of natural resources, we need to be able to find, access and use disparate data, and, from it, to derive information, knowledge and understanding.

However, our ability to take observations and make predictions has outpaced our data management capabilities. This paper includes a number of specific recommendations for enhancing those capabilities in order to support our global need for ocean and coastal data.

5. REFERENCES

1. Jensen, R. (1999). *Australia's Marine Science and Technology Plan*, (Canberra: Dept. of Industry, Science and Resources).
2. Institute of Electrical and Electronics Engineers (1990). *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries* (New York, IEEE).
3. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). "The Semantic Web," *Scientific American*, May 2001.
4. Hankin, S. & Co-Authors (2010). "NetCDF-CF-OPeNDAP: Standards for Ocean Data Interoperability and Object Lessons for Community Data Standards Processes" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.41.
5. Conkright-Gregg, M., Newlin, M., LeDuc, S., Keeley, R. and D'Adamo, N., (2010). "Ocean and Coastal Data Stewardship" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.18.
6. Paulk, M. C., Weber, C. V., Curtis, B. & Chrissis, M. B. (1995). *The Capability Maturity Model: Guidelines for Improving the Software Process* (Boston: Addison Wesley).
7. de La Beaujardière, J. (2008). "The NOAA IOOS Data Integration Framework: Initial Implementation Report," in *Proc. MTS/IEEE Oceans'08*.
8. Na, A. & Priest, M., eds. (2007). *Sensor Observation Service, version 1.0* (Wayland, MA: Open Geospatial Consortium).
9. Whiteside, A. & Evans, J., eds. (2008). *Web Coverage Service (WCS) Implementation Standard, version 1.1.2*, (Wayland, MA: Open Geospatial Consortium).
10. Cornillon, P., Gallagher, J., & Skouros, T., "OPeNDAP: Accessing Data in a Distributed, Heterogeneous Environment," *Data Science Journal*, **2**, 5 Nov 2003, p. 164.
11. de La Beaujardière, J., ed. (2004). *Web Map Service Interface, version 1.3.0* (Wayland, MA: Open Geospatial Consortium).
12. Botts, M. & Robin, A., eds. (2007). *Sensor Model Language (SensorML) Implementation Specification, version 1.0* (Wayland, MA: Open Geospatial Consortium).