# Data Management for the Ocean Sciences – the Next Decade

Steve Hankin, NOAA/PMEL, 7600 Sand Point Way NE, Seattle WA, USA, Email: Steven.C.Hankin@noaa.gov
Luis Bermudez, , SURA, 1201 New York Avenue NW, Suite 430, USA, Email: bermudez@sura.org
Jon Blower, Environmental Systems Science Centre, University of Reading, RG6 6AL,
United Kingdom, SURA, 1201 New York Avenue NW, Suite 430, USA, Email:
j.d.blower@reading.ac.ukbermudez@sura.org
Benno Blumenthal, International Research Institute for climate and society, Lamont Campus, Palisades NY 10964, USA, Email: benno@iri.columbia.edu
Kenneth S.Casey – NOAA National Oceanographic Data Center, 1315 East-West Highway, Silver Spring MD 20910 USA, Kenneth.casey@noaa.gov
Mark Fornwall, USGS, National Biological Information Infrastructure, 310 Ka`ahumanu Ave, Kahului HI, 96732, USA, Email:Mark_Fornwall@USGS.gov
John Graybeal, University of California, San Diego, 9500 Gilman Drive #0446,La Jolla, CA 92093 USA, Email:jgraybeal@ucsd.edu
Robert P Guralnick, University of Colorado Boulder, Campus Box 265, Boulder CO 80309, Email: Robert.Guralnick@colorado.edu
Ted Habermann, NOAA National Geophysical Data Center, 325 Broadway, Boulder, CO 80304, ted.habermann@noaa.gov
Eoin Howlett, Applied Science Associates, 55 Village Square Drive, South Kingstown, RI 02879, USA, Email: ehowlett@asascience.com
Bob Keeley, Integrated Science Data Management, Department of Fisheries and Oceans, 1202-200 Kent Street, Ottawa, Canada, K1A 0E6, Canada. Robert.Keeley@dfo-mpo.gc.ca
Roy Mendelssohn, NOAA/PFEL, 1352 Lighthouse Avenue, Pacific Grove, CA, USA, Email: Roy.Mendelssohn@noaa.gov
Rainer Reiner Schlitzer, Alfred Wegener Institute, Columbusstrasse, 27568 Bremerhaven, Germany, Email: Reiner.Schlitzer@awi.de
Rich Signell, USGS, 384 Woods Hole Rd. Woods Hole, MA 02543, Email: rsignell@gmail.com
Derrick Snowden, NOAA/CPO/COD, 1100 Wayne Avenue, Suite 1202, Silver Spring, MD, USA 20910, Email: Derrick.Snowden@noaa.gov
Andrew Woolf, STFC Rutherford Appleton Laboratory, STFC e-Science Centre, RAL, Chilton, Oxon, UK, Email: andrew.woolf@stfc.ac.uk

## Abstract:

Today we find remarkable agreement on expectations for vastly improved ocean data management a decade from now -- capabilities that will help to bring significant benefits to ocean research and to society. Advancing data management to such a degree, however, will require cultural and policy changes that are slow to effect. The technological foundations upon which data management systems are built are certain to continue advancing rapidly in parallel. These considerations argue for adopting attitudes of pragmatism and realism when planning data management strategies.

In this paper we follow those guidelines as we outline opportunities for progress in ocean data management. We begin with a synopsis of expectations for integrated ocean data management a decade from now. We discuss factors that should be considered by those evaluating candidate "standards". Then we highlight challenges and opportunities in a number of technical areas, including "Web 2.0" technologies, data modeling, data discovery and metadata, real-time operational data, archival of data, biological data management and satellite data management. We discuss the value of investments in the development of software toolkits to accelerating progress. We conclude the paper by recommending a few specific, short term targets for implementation, that we believe to be both significant and achievable, and calling for action by community leadership to effect these advances.

## Introduction

The Internet has altered our expectations for scientific data management, much as it has altered expectations for many other elements of society – personal communications, commerce, journalism, etc. Today we find remarkable agreement in expectations for vastly improved ocean data management a decade from now. We envision capabilities that will help to bring significant benefits to ocean research and to society. Sharing this vision has helped up to better understanding the strengths and weaknesses in the data systems that are in use today [1,2,3]

Advancing data management, however, is not merely a question of improving the use of technology. The organizational traditions that control lines of planning, funding and influence today, still largely reflect pre-Internet priorities. Our expectations for data management will not be realized until cultural and policy changes have occurred in areas such as free sharing of data. Organizational traditions are generally slow to change and often inhibit the adoption of new technologies [4]. While time is passing the technological foundations upon which data management systems are built are certain to continue advancing rapidly.

These considerations argue for adopting attitudes of pragmatism and realism when planning data management strategies [5]. In this paper we attempt to follow those guidelines. We understand that technological progress is always made in incremental steps, rather than "heroic leaps" [6]. We examine technology choices based upon their potential contributions to the distant vision, but we measure them by their effectiveness at addressing today's challenges. We conclude this paper by recommending a few specific, near-term targets for implementation that we believe both to be achievable and to address pressing problems.

We believe that progress in integrated data management cannot occur without active participation on the part of scientists and program managers as well as data management professionals. We attempt to present material in this paper using language that all of these stake-holder groups will find intelligible.

## The Vision of Interoperable Ocean Data Management

How do we envision ocean data management a decade from today? We see a future in which ocean data systems are managed by many independent organizations, yet they behave like a unified "system of systems". (See planning for these concepts with [GEOSS [7], the US IOOS DMAC Plan [8], NOAA's GEO-IDE plan [9], the UK's SeaDataNet [10], and Australia's IMOS [11].) We see volumes of data flowing that would overwhelm today's capabilities. We see a future in which ocean data are broadly shared, and users can locate it system reliably and quickly. We see rich descriptive information (metadata) available for all data and products. We see all sorts of users -- scientists, educators, industrialists, planners and recreationists -- accessing the data and information that is derived from it with little effort. We see these users doing their work with client software that addresses their particular needs, including sophisticated decision-support tools that incorporate both real time and historical ocean data. We see

planners utilizing such tools to make better-informed decisions that provide clear societal benefits

In this future we see providers of ocean data sharing data freely. We see careful tracking of provenance through the life-cycle of data usage. We see observing platforms that are able to alter sampling behaviors under sensor-automated, model-driven, animal-directed and human control. And we see all data that are of lasting value securely archived inside the context of this system-of-systems.

**Understanding Data Standards and Interoperability**

Most data management experts agree that adopting and using effective standards that define the interfaces between systems is a sound strategy for building a system of systems. However, viewpoints diverge over which standards and practices are "best"; what our highest priorities are; and what are the appropriate metrics for evaluating the quality of standards. The data management community finds itself divided into "camps". Some see critical weaknesses in standards that are currently delivering satisfactory levels of service to their intended customers, but were developed to address visions that were more limited than today's. Others have reservations about reliance on emerging technologies that have not yet demonstrated their effectiveness in settings of realistic complexity.

The meaning of the word "standard", unfortunately, is context-dependent. When we achieve our goal of interoperability through broadly shared practices we will call those practices our standards. The formal term for this concept is *de facto* standard. (In Latin *de facto* means "concerning fact".) In other contexts the word standard often refers to technical documents that have been approved through processes with agreed-upon rules: *de jure* standards. (In Latin *de jure* means "concerning law".) *De facto* standards that have gone through a *de jure* process are generally perceived to have higher value through being "open" – i.e. design documents are available for scrutiny and the future evolution of the standard is controlled by the *de jure* process. However, *de jure* processes do not reliably produce high quality standards, succumbing often to "designed by committee" failings of inconsistency, needless complexity, and inadequate testing [12]. Many *de jure* standards, including those from the most prestigious *de jure* organizations, deserve to fail at becoming *de facto* standards.

A common misunderstanding about standards is the assumption that their use will lead inexorably towards interoperability. As a counter-example, simply consider the Roman alphabet. While clearly a *de facto* standard that is vital to interoperability, the Roman alphabet is the foundation for written Italian, German and English -- languages that are manifestly non-interoperable. Information technology standards may have similar (sometimes unintended) consequences of dividing communities into non-interoperable sub-groups.[1] Matching a standard to its appropriate scope is critical to its acceptance.

[1] Subtle scalability considerations underlie community interoperability considerations. Data management requirements expand as the technical diversity of the "community" that we define expands. There will always be a level of diversity at which it becomes impractical to address detailed requirements with uniform standards

A second reason to be cautious when considering major redesigns is that as the number and scope of innovation increases, so does the length of time that will generally be needed to implement them. The pace of technological change for the IT industry as a whole is rapid. If too much time passes the bold innovations one may embark on today may be rendered obsolete before they can be realized in operational systems.[2] Adopting standards is fundamentally about managing risk [13]. Building interoperability through incrementally enhancing established standards should be understood as an approach that minimizes risk.

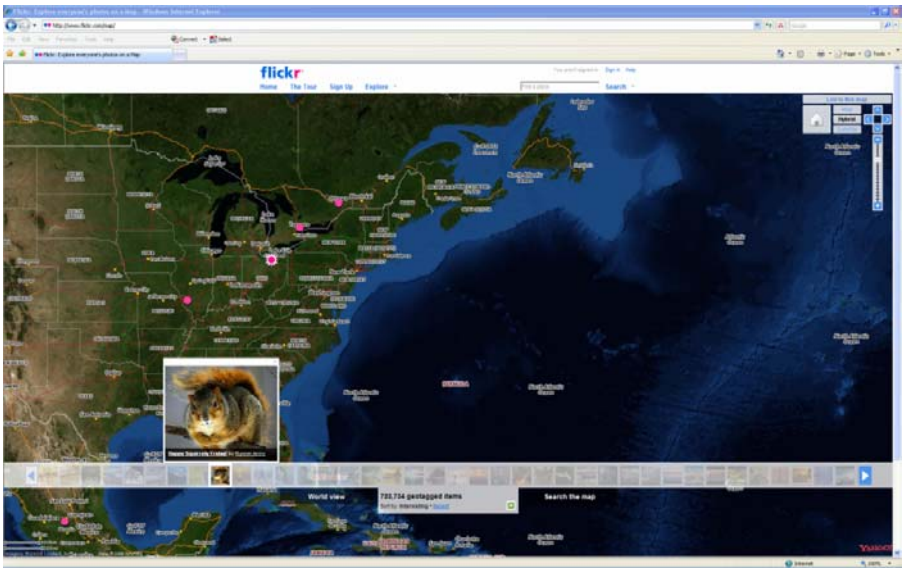


Figure 1. Example of Flickr web site that allows users to "geotag" and publish their photographs

**Leveraging "Web 2.0" technologies**

The Web today enables millions of users to share, discover, interpret and buy information. Technologies such as Facebook[3] and Flickr[4], that were developed to promote social interactions, are being used by scientists to collaborate on experiments. [Fig 1.] Twitter and RSS feeds show that rapidly changing data can be delivered effectively to myriads of devices including mobile devices. No-cost commercial search engines such as Google™ help us to locate vast amounts of information; no-cost tools such as Google Earth™ provide remarkable visualizations of geospatial data. To build an ocean data network in isolation from these transformative technologies would be foolish. These trends shape end-user expectations and provide low cost (or even no cost) solutions.

[2] With sufficient investment -- assuming that resources are well-managed and coordinated among the appropriate stake holders -- the length of time to implement an innovative technology can be reduced. Thus we come to the (intuitively obvious) conclusion that there is a direct relationship between the level of investment available and the boldness of the innovations that should be considered.

[3] www.facebook.com

[4] www.flickr.com

So, how do we build an effective ocean data network -- an infrastructure of data, systems, services, and tools that will allow users of divergent interests to access "live" and archived data through the tools that they prefer to use? The standard answer to these questions has been "web services". Indeed it is clear that web services can provide a useful bridge between successful data management solutions that are in use today and emerging Web 2.0 tools. As examples, consider two of the Open Geospatial Consortium (OGC) standards for sharing geospatially-referenced data: the Web Mapping Service[5] (WMS) for sharing maps as digital images; and Keyhole Markup Language[6] (KML) for locating information on maps and virtual globes [Fig 2.]. These web services have gained acceptance through a combination of simplicity and utility -- simple for data providers to make information available; and popular applications (e.g. Google Earth™ and ArcGIS™) are available to manipulate or display the information.

These themes – 1) simplicity for software developers and 2) highly functional tools for users to access information -- must drive the planning and implementation of ocean data management if it is to progress rapidly – especially in a low budget environment.

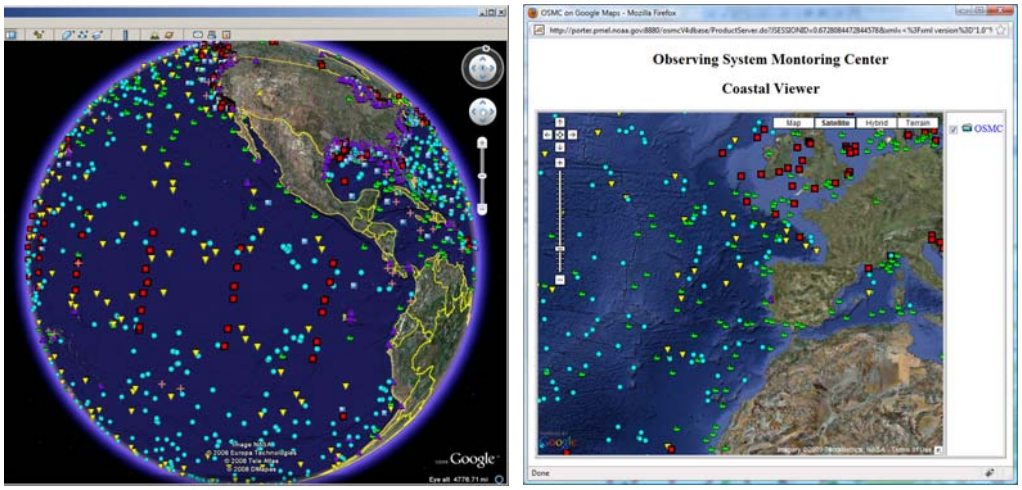


Figure 2. Many ocean data management projects have invested the relatively small efforts needed to represent their data using KML, thereby leveraging powerful applications like Google Earth® (left) and Google Maps® (right). The images we see here are just one example -- from the NOAA Observing System Monitoring Center[7].

**Standardizing Ocean Data through "Sampling Geometry" (Data Modeling)**

To gain understanding from a complex collection of information it is reasonable to divide the collection into categories. Ironically, the categories that are best suited to gaining scientific understanding of ocean data are sometimes poor choices for simplifying the challenges of data management. Ocean scientists generally categorize data based upon disciplines (e.g. chemical versus physical parameters), regions (coastal versus open

[5] http://www.opengeospatial.org/standards/wms
[6] code.google.com/apis/kml/documentation/
[7] www.osmc.noaa.gov/

ocean), and/or the origins of data (*in situ* and remote-sensed observations, model outputs, etc.). However, categorizing data based upon its structure (a.k.a. "sampling geometry") – time series, profiles, grids, etc. – brings the commonalities most useful for data management into focus. For example, a time series of temperature measured from an ocean mooring shares many of the concepts and tools with a time series of sea level measured by a tide gauge or a sequence of repeated phytoplankton counts made at the same location over time. Such classifications also have significant scientific utility, as sampling geometry constrains the scientific purposes to which data may be put.

The development of conceptual data models that capture ocean/atmosphere data structures is a relatively recent effort that has already accelerated progress in data management. Many ocean data sources -- including model outputs, satellite observations [14], OceanSites moorings [15], Argo profiles [16] and GOSUD underway ships [17]-- are converging on the use of the netCDF[8] [5] data model and the Climate and Forecast (CF) Conventions [18]. The netCDF data model and CF together provides the ability to share the data transparently across the Internet using the OPeNDAP protocol [19]. Through OPeNDAP users of the data are often unaware whether the data reside locally or remotely.

Two significant efforts – the Common Data Model [20] from Unidata in the US and the Climate Science Modelling Language [21] (CSML) from the Natural Environment Research Council (NERC) in the UK – are collaborating to develop an over-arching data model that unites the netCDF data model with spatial data ("GIS") concepts from the Open Geospatial Consortium (OGC). Using a data model standardizes the operations that may be performed upon data – how to subset data, to change its resolution, to "regrid" it from one coordinate system to another in a manner that conserves mass and energy – which enables the development of sharable software toolkits that greatly accelerate the development of the system. A number of other independent data management initiatives in the ocean and atmospheric sciences have adopted classifications around sampling geometries, for example the NOAA GEO-IDE Concept of Operations [9], the ESRI Arc Marine Data Model [22].

**Improving Discovery and Documentation of Ocean Data**

Metadata describes data, preferably in a structured form that can be used by both machines and people. Advances in metadata are critical to many improvements in ocean data interoperability. To appreciate the potential role of metadata consider the recent advances in handling the data represented by audio files, particularly in consumer applications like iTunes™. Structured metadata that describes performers, albums, genres, ratings, etc., supplied by both data suppliers (vendors) and consumers, has enabled effective strategies for us to locate, organize, understand and better utilize the data.

The effective use of metadata for scientific applications lags behind analogous commercial applications. Scientists, like other data users, tend to continue using what has worked in the past until sufficiently attractive alternatives become available. In the

[8] www.unidata.ucar.edu/software/netcdf/

next few years, an explosion of alternative tools and techniques for discovering data is likely to occur, followed by a consolidation of the most successful ones. Science users can expect to see commercial search engines such as Google™ advance to meet many of their needs, but should recognize the vital role that improvements in standardized metadata must have in enabling these advances.

As the data discovery challenges are increasingly met the focus of metadata is likely to return to "documentation" -- information needed for a more complete understanding of the data. Several standards are advancing to address these needs. The ISO Metadata Standards (notably ISO 19115) [23] are likely to establish a worldwide practice for documenting data sets. These standards provide structures (abstractions) that address data quality; processing algorithms; spatial and temporal extents; linkages to descriptions of sensors; collections of and subsets of datasets; annotations by users (a social networking concept); and many other documentation needs.

A second set of specifications likely to continue gaining traction is netCDF (comprising a data model, software libraries and file format) together with the CF metadata conventions [4, 10]. A third collection of relevant standards is the Open Geospatial Consortium's Sensor Web Enablement (SWE) suite [24], which defines conceptual models (and web services) to describe sensors and sensor data streams. Increasingly manufacturers are supplying metadata in various forms at the sensor level.

It is inevitable that scientific datasets will be described with multiple, independent but overlapping metadata standards. Similar concepts may be known by different names across metadata standards; the same term may have meanings that differ. Technologies and tools to achieve semantic interoperability will be needed to address these problems. During the next ten years we expect knowledge engineering (ontology) technologies to advance to the point that they will routinely translate terminology, codes, conceptual models and relationship across standards.

**Integrating Operational Data and Metadata**

The next ten years will see today's operational data distribution systems evolve to embrace far greater use of the Internet. Traditionally the World Weather Watch, Global Telecommunications System (GTS) of the WMO has provided data dissemination services for operational meteorology and oceanography. While the ocean observing system has derived immense value through this association, significant problems have become apparent. We must ensure that data and metadata can never be dissociated. For example, a BUFR file containing an ocean observation that is distributed on the GTS in real time must have an iron-clad linkage to the metadata that contains the manufacturer, model, and calibration history of the sensor and platform that generated the observation.

Currently the GTS messaging formats do not include detailed sensor metadata. However, JCOMM has mandated [25] that by 2012 messaging on the GTS will switch from the old ASCII driven codes to Table Driven Code (TDC) formats such as BUFR, and its ASCII cousin, CREX. These TDC formats will support enhanced metadata content by referencing external tables describing the data. A BUFR message might contain a code,

or descriptor, that references an entry in a table that defines the name, size, and units for the upcoming data packet. This design makes BUFR flexible, but also potentially complicated. There is a similar need to define the templates that encode particular data types. For example, an operational template for XBTs, defines the subset of descriptors from the tables that will be used to describe all XBT observations.

Populating the code tables and defining templates is the role of the WMO with input from national and international programs. The templates for ocean observation data types will be designed by the JCOMM Cross-cutting Task Team on Table Driven Codes. The observing system operators must play a critical role for this approach to be a successful. They must communicate detailed metadata requirements for their platform type to the Task Team. After the transition to the new BUFR formats has occurred the operators must also ensure that the templates are fully populated as data are disseminated on the GTS.

**Archiving Ocean Data**

Ocean archives are tasked with long-term preservation of ocean observations. At the end of the 20$^{th}$ century the archives were struggling with the rapid flux of technology, escalating data volumes, and dramatically more complex and varied data types. At the same time archive budgets were often flat or decreasing in real terms, and user demands were increasing due to the exploding use of the World Wide Web and the associated expectations from users of instantaneous, online access to information.

During the first decade of the 21$^{st}$ century, digital archives around the world began to share experiences and challenges. They discovered that these communications were made challenging by a lack of a common vocabulary and understanding of "archive" functions. The community of archivists tackled the issue through the establishment of the Open Archival Information System Reference Model (OAIS-RM), the ISO standard for digital archives (ISO 14721). The OAIS-RM defines common terminology and a suite of responsibilities that must be accepted by an OAIS archive. Ocean data archives around the world have embraced the responsibilities with increasing enthusiasm. They are seeing how OAIS-RM can help them improve their internal archive operations as well as their interchange functions with other archives, data producers, and data consumers.

Looking to the next decade as the demands placed upon ocean archives will continue to grow the OAIS-RM will provide a foundation that positions them to improve efficiency, and to better meet the needs of their users. Funding agencies should require that data-generating ocean projects work with archives to preserve the observations. Ocean archives must be prepared to support them in doing so. Journals must begin to incorporate data set citations as they do for journal citations. Ultimately system-of-system integration should blur the divisions between management of real-time, delayed mode and archived data, so that users need have minimal awareness of which particular level of the system is providing services.

**Integrating Ocean Biological Data**

Data management systems must increasingly mobilize available marine biological data and ensure their interoperability with physical and chemical data in order to advance our

understanding of the complex ocean ecosystems. Marine biodiversity data is often difficult to find or not available for anything but well-studied, economically important taxa. Although the necessary observations exist for many regions of the oceans, inadequate data integration leaves us unable to answer fundamental biodiversity questions such as "what biodiversity has been found in region X?" and "has previous sampling been sufficient to support confidence in biodiversity estimates?"

Roughly 3 billion records of biological diversity [26] collections are housed in repositories world-wide. Only a small proportion (~ 5-10%) [27]of these are digitized. These data are the best possible resource with which to construct baselines to measure changes in biodiversity over time [28]. Multiple agencies, notably the Global Biodiversity Information Facility (GBIF) and International Ocean Biogeographic Information Systems (iOBIS), have developed a worldwide information infrastructure into which natural history collections can be published. This distributed global network of databases [29,30,31] can help to address both scientific and management questions.

The OBIS project represents a successful start on a much larger effort to develop community data standards support the sharing of population, community ecological, genetic and tracking datasets. Future biological data systems must track the flow from initial biological observations to the application of this information to address scientific and social problems. These systems must be extended to achieve interoperability the with non-biological data management systems in order to be able to assess the connections between changes to biological systems and the surrounding physical and chemical systems.

**Integrating Satellite Data**

Over the last decade best practices and standards have begun to emerge for satellite-based ocean observations, addressing key areas such as file format, metadata, data quality, and data access. Projects like the Group for High Resolution SST (GHRSST, Donlon et al., 2007), which began shortly after the OceanObs '99 convention, provided clear demonstration of the benefits achievable when the community self-organizes around common principles. As a result of GHRSST, de facto standards were established for the global satellite SST community. Many groups that were not part of the original consortium joined the collaboration, and the principles developed by GHRSST are now being applied in other areas. Thus a major goal for the next decade is to ensure that international coordination programs are developed, implemented, and sustained for all ocean observations collected from space-borne platforms.

The GHRSST program has demonstrated the need for feedback loops between scientific activities, data management, and production activities. For example the requirement that GHRSST collaborators provide interoperable SST observations with associated uncertainty estimates facilitated intercomparisons, which in turn revealed the need for better understanding of the diurnal cycle and led to improved error estimates. The data management strategies provided feedback of these scientific improvements into the data productions systems. The concept of "crossing the valley of death" as a one-way street from research to operations in the management of earth observing satellites is evolving into a concept of an iterative feedback between research and operations. The second key

goal for the coming decade is thus to ensure that scientific activities and operational data management and production activities support one another in an iterative feedback loop.

Achieving this level of data interoperability requires agreements in several areas. Data content standards that ensure consistent representation of variables must be agreed upon by the science communities. File formats must be used uniformly, with netCDF-4/HDF5 emerging as the format of choice. ISO19115 and its XML representation, ISO19139, are emerging as standards for collection-level metadata. For "use metadata" in the file, the Climate and Forecast (CF) conventions have become widespread, and are supported by numerous data clients. Best practices are under development for pixel-by-pixel quantified error estimates. As standards for interoperable access OPeNDAP's Data Access Protocol (DAP), and the OGC's Web Coverage Service, and Web Mapping Service (for images) have emerged. Finally, data access policies must be in place to support widespread access to the observations. Thus the third challenge for the coming decade is to implement the set of international standards and policies for file format, content, and metadata; data quality information; and data transfer and access.

**Accelerating Development through Software Toolkits**

Most funding for ocean data system development is directed to support scientific programs such as WOCE or JGOFS; institution-specific research and development infrastructures; platform-specific data management systems (DACs and GDACs); or tools for end-users. The target users for the capabilities that are developed are scientists and managers. Remarkably little investment identifies the software development community, itself, as the target audience.

A consequence of this (unplanned) community investment strategy is significant inefficiencies in the development of new software. The layer of software development that builds earth-fluid-specific concepts from industry-wide software frameworks is duplicated time after time. A notable counter-example to this is the modest investments by the US National Science Foundation in Unidata [32]. With stable funding for just a handful of software developers (an average of 12 programmers over the past decade), Unidata has created the software libraries and Web server tools that form the foundation for many of the capabilities we rely upon in oceanography.

The lesson to be learned from NSF's investments in Unidata is that progress in data management can be greatly accelerated through investment in toolkits that support software development. The NetCDF-Java library, for example, provides routines able to extract geospatially referenced data from time series, profiles, track lines, simple grids, and complex ocean model grids (e.g. curvilinear horizontal grids and stretched vertical coordinates) that follow the CF conventions. The availability of this toolkit has greatly increased the utilization of data through applications such as Unidata's Integrated Data Viewer[9] (IDV), ncWMS [33], Panoply[10], Live Access Server[11], and ERDDAP[12]. These

[9] www.unidata.ucar.edu/software/idv/
[10] www.giss.nasa.gov/tools/panoply/
[11] http://ferret.pmel.noaa.gov/LAS/

the toolkits allow scientists and software developers to concentrate on the unique contribution that they wish to provide. Development of toolkits for commonly used languages like C, Python, IDL, and R will similarly accelerate progress in data system tool development.

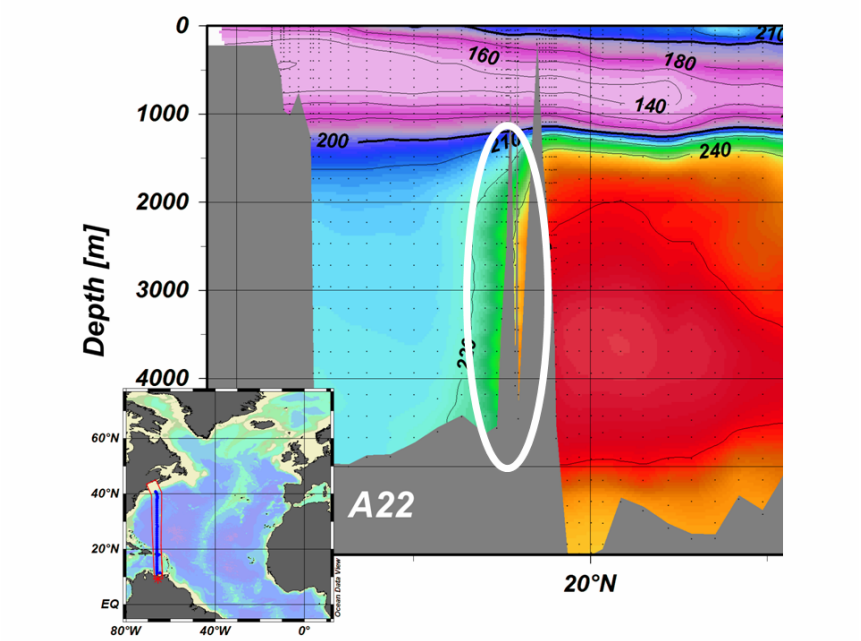


Figure 3: Oxygen distribution along the WOCE A22 section showing elevated values in the Caribbean deep water. This structure is an artifact of the weighted-averaging gridding algorithm and caused by influence of high-oxygen waters on the northern side of the ridge.

A second example of the need for toolkits is found in the common practice of visualizing 2D tracer fields as maps, sections or time-evolution plots. This process involves the mapping of the heterogeneously distributed observed data values onto a set of grid nodes (gridding). While a vast variety of gridding algorithms exists, presently a user only has two broad choices: (1) to use simplistic algorithms, which often create significant artifacts in the distributions (see Fig 3); or (2) to use advanced gridding algorithms that utilize computationally demanding objective analysis methods. To use the advanced techniques, however, requires expert knowledge and considerable effort, because toolkits do not exist in a form easily-used by software developers. Some utilities have been developed for use by end-user scientists, for example the Data Interpolating Variational Analysis package (DIVA[13]), however, it is only with difficulty that stand-alone utilities can be into integrated software packages. (An example of this approach may be seen in the Ocean Data View[14].

**Conclusion -- Concrete Community Targets for Improved Data Integration**

[12] coastwatch.pfeg.noaa.gov/erddap/
[13] http://modb.oce.ulg.ac.be/projects/1/diva
[14] http://odv.awi.de

In the preceding sections of this paper we have outlined an ambitious vision for ocean data integration a decade from today. We have discussed a number of approaches and technologies that may help us to achieve those goals. We have pointed out, however, that the technological foundation, upon which ocean data integration must be built, is in a period of rapid evolution. The approaches used to build ocean-specific capabilities must be agile to adjust to rapidly changing circumstances.

The development of an integrated system-of-systems must proceed in concrete, incremental steps. In concluding this paper the authors suggest what a few of those steps should be. The list of suggestions provided here is by no means comprehensive; but it represents an informal consensus within the ocean data management community. The authors of this paper encourage leaders in both the ocean science and data management communities to call for meetings to plan these concrete steps and identify others.

1. **Ocean Observations Universally Accessible through NetCDF-CF-OPeNDAP**
   The past decade has seen a striking convergence on the use of netCDF-CF-OPeNDAP for delayed mode ocean data. Above we discussed the use of these standards for model outputs, satellite products, OceanSites and Argo. Solutions are in development for underway ship observations and surface drifters. Solutions for XBTs, tide gauges, gliders, etc. are simple applications of the same techniques. Many of the techniques are applicable to biological data such as continuous plankton recorder records. This trio of practices has been accepted as a standard for gridded data by US IOOS [34] and is working its way through the NASA ESDSWG standards processes [35]. Efforts to achieve standardization within OGC have begun[15].

   The trend represented by this convergence should be sustained and strengthened until all ocean observations and models are on-line and available through netCDF-CF-DAP. The effort to do so can be modest if the organizations that remain to make a transition leverage the efforts of other organizations that have preceded them.

   Convergence on "files" is by no means the desired ultimate foundation for interoperability. It falls well short of the 10 year vision for a service-oriented architecture that we imagine today. But the process of convergence – achieving *de facto* standardization -- would be a very significant milestone. Technical progress will greatly accelerate thereafter, as new tools that are developed by any one institution become applicable across the community.

2. **Develop a Common Data Model and associated software toolkits**
   The efforts shared by Unidata, the CF community and other community members to develop a Common Data Model should be supported and accelerated. The resulting model should be implemented in software toolkits that are able to store, retrieve and perform operations in a uniform manner on the widest feasible range of ocean-

[15] The formal process to advance these technologies through the OGC standards process was initiated at the OGC Technical Meeting held in Mountain View, California in December 2009

relevant data structures. These toolkits should be advertised and made available to software developers.

**3. Completing the transition to BUFR to improve WWW support for ocean observations**

As discussed above, a plan has been agreed upon within WMO and JCOMM to ensure that all ocean observations on the GTS have improved metadata contents by 2012. To achieve this, it will be necessary that:

- observing platform communities complete the task of defining the metadata that are required at the time of data collection;
- templates be designed that encode this information;
- unambiguous linkages between real time messages and enhanced metadata content on shore be developed.

[1] Pouliquen, S., Hankin, S.C., Keeley, J.R., Blower, J.D., Donlon, C., Kozyr, A. Guralnick, R. (2010). The development of the data system and growth in data sharing. In these proceedings (Vol. 1).

[2] Blower, J.D., Hankin, S.C., Keeley, R., Pouliquen,S., de la Beaujardière, j., Vanden Berghe, E., Reed, G., Blanc, F., Conkright Gregg,F., Fredericks,J., Snowden,D. (2010). Ocean Data Dissemination: New Challenges for Data Integration, In these proceedings (Vol. 1).

[3] Keeley, J.R., Woodruff, S., Pouliquen, S., Conkright Gregg, M., Reed G. (2010). Data Assembly Infrastructure: from acquisition to archives. In these proceedings (Vol. 1).

[4] Clayton M. Christensen (2003), The Innovator's Dilemma

[5] Hankin, S.C., and coauthors (2010). NetCDF-CF-OPeNDAP: Standards for Ocean Data Interoperability and Object Lessons for Community Data Standards Processes. In these proceedings (Vol. 2).

[6] Jared Diamond , (2005), Guns, Germs, and Steel: The Fates of Human Societies

[7] *The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan* (2005) accessed on 23 December 2009 at http://www.earthobservations.org/documents/10-Year%20Implementation%20Plan.pdf

[8] *Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems* accessed on 23 December 2009 at http://dmac.ocean.us/dmacPlan.html

[9] *NOAA Global Earth Observation Integrated Data Environment (GEO-IDE) Concept of Operations Version 3.3 (2006)* accessed on 23 December 2009 at www.nosc.noaa.gov/docs/products/NOAA_GEO-IDE_CONOPS-v3-3.doc

[10] *Development of Marine Data Management Infrastructures in Europe (SeaDataNet) (2009)* accessed on 23 December 2009 at http://www.seadatanet.org

[11] *Integrated Marine Observiong System (IMOS)* accessed on 23 December 2009 at http://imos.org.au/about.html

[12] Henning, M. (June 2006). The Rise and Fall of CORBA, ACM QUEUE, *http://www.zeroc.com/documents/riseAndFallOfCorba.pdf*

[13] Cargill, C. and Bolin, S. (2004). Standardization: A Failing Paradigm, http://www.chicagofed.org/news_and_conferences/conferences_and_events/files/cargill.pdf (paper presented at the Standards and Public Policy Conference, Federal Reserve Bank of Chicago, May 13-14)

[14] *The Recommended GHRSST-PP Data Processing Specification GDS* accessed on 23 December 2009 at http://www.ghrsst.org/documents.htm

[15] *OceanSITES User's Manual* accessed on 23 December 2009 at http://www.oceansites.org/docs/oceansites-user-manual.pdf

[16] *Argo Data Management Handbook* accessed on 23 December 2009 at http://www.usgodae.org/argodm/manuals/argo_data_management_handbook_v1.2.pdf

[17] *GOSUD: User's Manual* accessed on 23 December 2009 at http://www.ifremer.fr/gosud/doc/gosud-dm-user-manual-08-064.pdf

[18] *NetCDF Climate and Forecast (CF) Metadata Conventions* accessed on 23 December 2009 at http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.4/cf-conventions.pdf

[19] *OPeNDAP User Guide* accessed on 23 December 2009 at http://www.opendap.org/user/guide-html/guide.html

[20] *Unidata's Common Data Model Version 4* accessed on 23 December 2009 at http://www.unidata.ucar.edu/software/netcdf-java/CDM/

[21] *CSML User Guide* accessed on 23 December 2009 at http://proj.badc.rl.ac.uk/csml/browser/Documentation/trunk/CSMLUsersManual.pdf

[22] *Arc Marine (The ArcGIS Marine Data Model)* accessed on 23 December 2009 at http://dusk.geo.orst.edu/djl/arcgis/

[23] International Standard, Geographic information — Metadata, ISO 19115:2003, 1st ed. May 2003.

[24] *Sensor Web Enablement (SWE)* accessed on 23 December 2009 at http://www.opengeospatial.org/ogc/markets-technologies/swe

[25] *SUMMARY OF THE PLAN FOR MIGRATION TO TABLE-DRIVEN CODE FORMS (TDCF)* accessed on 23 December 2009 at http://www.wmo.int/pages/prog/www/WMOCodes/MigrationTDCF/Plan/SummaryMigraPlan_en.pdf

[26] Beaman R and B. Conn. 2003. Automated geoparsing and georeferencing of Malesian collection locality data. Telopea 10:43–52.

[27] Krishtalka, L. & Humphrey, P.S. (2000). Can natural history museums capture the future? Bioscience, 50, 611–617.

[28] Suarez, A.V. & Tsutsui, N.D. (2004). The value of museum collections for research and society. BioScience, 54, 66–74.

[29] Edwards,J.L. (2004) Research and societal benefits of the global biodiversity information facility. *BioScience*, 54, 485–486.

[30] Lane,M. (2006) Information infrastructure for global biological networks. *Microbiol. Aust.*, 27, 23–25.

[31] Guralnick,R.P. *et al*. (2007)Toward a collaborative, global infrastructure for biodiversity assessment. *Ecol. Lett.*, 10, 663–672.

[32] *Unidata 2008: Shaping the Future of Data Use in the Geosciences* accessed on 23 December 2009 at http://www.unidata.ucar.edu/staff/mohan/Unidata_2008_Final_Report.pdf, the final report for

[33] Blower J D, Haines K, Santokhee A and Liu C L 2009 GODIVA2: Interactive visualization of environmental data on the web *Philosophical Transactions of the Royal Society A* 367 1035-9

[34] Standards package for the representation and transport of gridded data: netCDF+CF+OPeNDAP+aggregation, accessed on 21 December 2009 under Recommended Standards at http://ioosdmac.fedworx.org/

[35] *Earth Science Data Systems Standards Process Group*, accessed on 21 December 2009 at http://www.esdswg.org/spg/docindexfolder