DATA MANAGEMENT FOR THE OCEAN SCIENCES – PERSPECTIVES FOR THE NEXT DECADE

Steve Hankin[1], Luis Bermudez[2], Jon D. Blower[3], Benno Blumenthal[4], Kenneth S. Casey[5], Mark Fornwall[6], John Graybeal[7], Robert P Guralnick[8], Ted Habermann[9], Eoin Howlett[10], Bob Keeley[11], Roy Mendelssohn[12], Reiner Schlitzer[13], Rich Signell[14], Derrick Snowden[15], Andrew Woolf[16]

[1] NOAA/Pacific Marine Environmental Laboratory, Seattle, USA, Email: steven.c.hankin@noaa.gov
[2] Southeastern Universities Research Association, 1201 New York Avenue NW, Suite 430, USA, Email: bermudez@sura.org
[3] Environmental Systems Science Centre, University of Reading, Reading, UK, Email: j.d.blower@reading.ac.uk
[4] International Research Institute for climate and society, Lamont Campus, Palisades NY 10964, USA, Email: benno@iri.columbia.edu
[5] NOAA/National Oceanographic Data Center, Silver Spring, MD, USA, Email: Kenneth.Casey@noaa.gov
[6] USGS, National Biological Information Infrastructure, 310 Ka`ahumanu Ave, Kahului HI, 96732, USA, Email: Mark_Fornwall@USGS.gov
[7] University of California, San Diego, 9500 Gilman Drive #0446,La Jolla, CA 92093 USA, Email: jgraybeal@ucsd.edu
[8] University of Colorado Boulder, Campus Box 265, Boulder CO 80309, USA,, Email: Robert.Guralnick@colorado.edu
[9] NOAA National Geophysical Data Center, 325 Broadway, Boulder, CO 80304, USA, Email: ted.habermann@noaa.gov
[10] Applied Science Associates, 55 Village Square Drive, South Kingstown, RI 02879, USA, Email: ehowlett@asascience.com
[11] Integrated Science Data Management, Department of Fisheries and Oceans, 1202-200 Kent Street, Ottawa, Canada, K1A 0E6, Canada. Email: Robert.Keeley@dfo-mpo.gc.ca
[12] NOAA/PFEL, 1352 Lighthouse Avenue, Pacific Grove, CA, USA, Email: Roy.Mendelssohn@noaa.gov
[13] Alfred Wegener Institute, Columbusstrasse, 27568 Bremerhaven, Germany, Email: Reiner.Schlitzer@awi.de
[14] USGS, 384 Woods Hole Rd. Woods Hole, MA 02543, USA, Email: rsignell@gmail.com
[15] NOAA/CPO/COD, 1100 Wayne Avenue, Suite 1202, Silver Spring, MD, USA 20910, Email: Derrick.Snowden@noaa.gov
[16] STFC Rutherford Appleton Laboratory, STFC e-Science Centre, RAL, Chilton, Oxon, UK, Email: andrew.woolf@stfc.ac.uk

ABSTRACT

There is remarkable agreement in expectations today 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 adopt those attitudes 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". We highlight challenges and opportunities in a number of technical areas, including "Web 2.0" applications, data modeling, data discovery and metadata, real-time operational data, archival of data, biological data management and satellite data management. We discuss the importance of investments in the development of software toolkits to accelerate 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 advancements.

1. 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. We envision capabilities that will help to bring significant benefits to ocean research and to society. Sharing this vision has helped us to recognize and understand the strengths and weaknesses in the data systems that are in use today [1, 2 and 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 our attitudes to sharing data. The publication of scientific data must be handled in a manner that is as open, critical and methodical as is the current publication of scientific papers. Cultural 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]. We should understand that technological progress is always made in incremental steps, rather than "heroic leaps" [6]. We should give consideration to technology choices based upon their potential contributions to the distant vision, but we should measure them by their effectiveness at addressing today's challenges. In this paper we attempt to follow these guidelines. In the Conclusions section we recommend a few specific, near-term targets for implementation that draw upon this outlook.

Data management professionals can contribute only a part of the solution. We believe that progress in integrated data management cannot occur without active participation on the part of scientists and program managers. Thus, we attempt to present material in this paper in language that all stake-holder groups will find informative.

2. 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 efforts that follow these concepts within [GEOSS (Global Earth Observation System of Systems) [7], the US IOOS DMAC (Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems) Plan [8], NOAA's GEO-IDE (Global Earth Observation Integrated Data Environment) plan [9], the EU's SeaDataNet (Development of Marine Data Management Infrastructures in Europe) [10], and Australia's IMOS (Integrated Marine Observing System) [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 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.

3. 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.

Ambiguities in the meaning of the word "standard" complicate these considerations. When we achieve goals of interoperability through broadly shared practices we call those practices our "standards". The formal term for this concept is *de facto* standard. (In Latin *de facto* means "concerning fact".) Technical

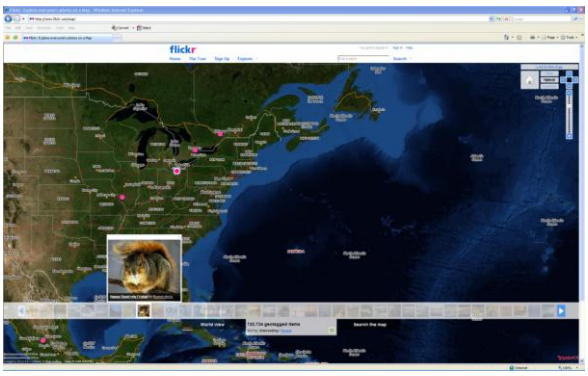

Figure 1. Example of Flickr web site that allows users to "geotag" and publish their photographs

documents that have been approved through processes with agreed-upon rules are also, though, referred to as "standards" -- formally *de jure* standards. (In Latin *de jure* means "concerning law".) Thus we have two quite distinct meanings.

De facto standards that have gone through a *de jure* process of development are rightly perceived to have higher value through being "open", since the 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. Too often they succumb to "designed by committee" failings of inconsistency, excessive complexity, and inadequate testing [12]. *De jure* standards -- even those from the most prestigious organizations -- should be evaluated on their merits in the spirit of skepticism and pragmatism advocated above. Many *de jure* standards deserve to fail at becoming *de facto* standards.

A common misunderstanding about standards is the assumption that their use will lead inexorably towards high levels of 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 effectiveness. In general, the more complex the standard is, the narrower will be its scope of applicability.

When evaluating approaches to interoperability that require major redesign of existing systems, we need to recognize that as the number and scope of innovations increases, so does the length of time that will generally be needed to implement them. The pace of change within the field of information technology is sufficiently rapid that the bold innovations one embarks upon 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 and practices should be understood as an approach that minimizes risk.

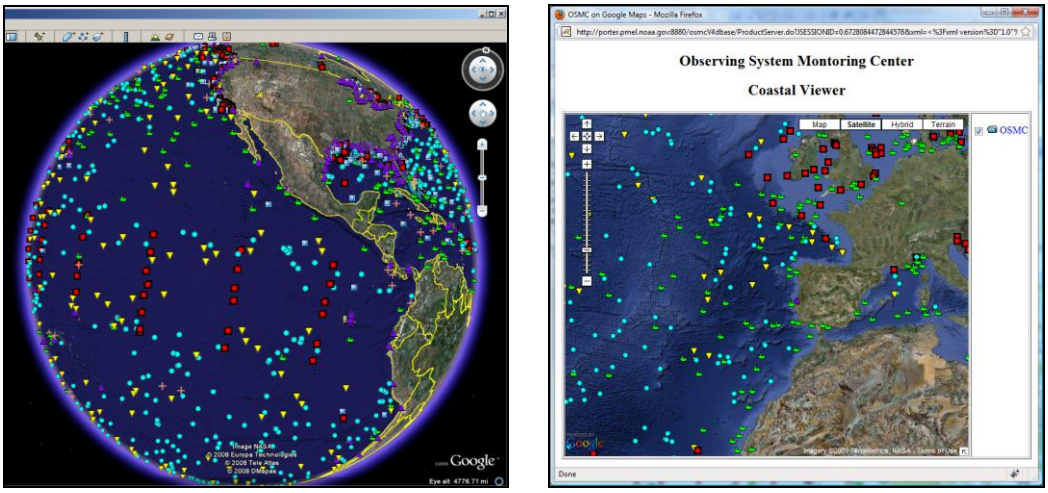

Figure 2. Many ocean data management projects have invested the relatively small efforts needed to represent their data using KML, thereby leveraging powerful applications like Google Earth® (left) and Google Maps® (right). The images we see here are just one example -- from the NOAA Observing System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

4. LEVERAGING "WEB 2.0" TECHNOLOGIES

The Web today enables millions of users to share, discover, interpret and buy information. The term "Web 2.0" refers applications that that embody principles of interactive information sharing, interoperability, user-centered design, and collaboration by means of the World Wide Web. Web 2.0 applications, such as

[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

[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.

Facebook[3] and Flickr[4], are being used by scientists to collaborate on experiments. [Fig 1.] Twitter[5] and RSS[6] (notifications delivered via Really Simple Syndication feeds) illustrate that rapidly changing data can be delivered effectively to myriads of clients 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 envision 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.

So, how do we build an effective ocean data network -- an infrastructure of data, systems, services, and tools that will allow users with divergent interests to access "live" and archived data through the tools that they prefer to use? The typical answer to these questions has been to implement "web services" – standardized interfaces through which applications can call upon heterogeneous systems across the Internet. 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[7] (WMS) for sharing maps as digital images; and Keyhole Markup Language[8] (KML) for locating information on maps and virtual globes [Fig 2.]. These web services have gained acceptance through a combination of simplicity and accessibility -- simple for data providers to make information available; and accessible through popular applications such as Google Earth™ and ArcGIS™ (Geographic Information Systems).

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. This is especially true if developments are not well funded.

[3] www.facebook.com

[4] www.flickr.com

[5] http://twitter.com/

[6] http://en.wikipedia.org/wiki/RSS

[7] http://www.opengeospatial.org/standards/wms

[8] http://code.google.com/intl/fr-CA/apis/kml/documentation/

5. STANDARDIZING OCEAN DATA THROUGH 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 ocean), and/or the origins of data (*in situ* and remote-sensed observations, model outputs, etc.). However, it is categorizing data based upon its four-dimensional structure (a.k.a. "sampling geometry") – underway ship tracks, vertical profiles, sections, grids, etc. – that brings the commonalities most useful for data management into focus. For example, a time series of temperature or current measured from an ocean mooring shares many of the concepts and tools (e.g. time axis plots, time averages, …) 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. Categorization by sampling geometry, of course, also has scientific utility, as it constrains the analytical 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 ocean data management. Many ocean data sources -- including model outputs, satellite Observations [14], OceanSites (OCEAN Sustained Interdisciplinary Time series Environment observation System) moorings [15],Argo (Array for Real-time Geostrophic Oceanography) profiles [16] and GOSUD (Global Ocean Surface Underway Data Pilot Project) underway ships [17] -- are converging on the use of the netCDF[9] [5] data model and the Climate and Forecast (CF) Conventions [18]. The netCDF data model and CF together provide the ability to share 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.

A number of 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 (Environmental Systems Research Institute) Arc Marine Data Model [20]. Two such efforts – the Common Data Model [21] from Unidata in the US and the Climate Science Modelling Language [22] (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

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

data ("GIS") concepts from the Open Geospatial Consortium (OGC). Using an agreed data model standardizes the operations that may be performed upon data – such as how to subset data or 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.

6. IMPROVING DISCOVERY, EVALUATION AND USAGE OF OCEAN DATA THROUGH METADATA

Advances in metadata are critical to many improvements in ocean data interoperability. Metadata describes data, preferably in a structured form that can be used by both machines and people. 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 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 are likely 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 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 (International Organization for Standardization) Metadata Standards (notably ISO 19115) [23] are likely to establish a worldwide practice for documenting data sets. These documents provide generalized representations of 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 and 10]. NetCDF-CF datasets are referred to as "self-describing" because metadata and data are combined into a single file and accessible together through the same programming libraries. The metadata provided by CF, which is focused primarily on the fundamentals of usability -- coordinates, units, and standardized scientific parameter names, etc. -- can be augmented with more detailed, platform- specific information as we see in such data format standards as OceanSITES moorings [15] and Argo profiles [16].

A third collection of relevant standards is the Open Geospatial Consortium's Sensor Web Enablement (SWE) suite [24]. SWE defines conceptual models, web services and XML (Extensible Markup Language) encoding frameworks that can be used as a toolkit in formulating community-agreed description of ocean sensors, platforms, 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. A single concept may be known by different names across metadata standards; a single term may have different meanings. Technologies and tools to address these problems – to achieve "semantic interoperability" -- will be needed. During the next ten years we expect these technologies to advance to the point that they will routinely translate terminology, codes, conceptual models and relationship across standards.

7. INTEGRATING OPERATIONAL DATA AND METADATA

Traditionally the World Weather Watch, Global Telecommunications System (GTS) of the World Meteorological Organization (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. The next ten years will see today's operational data distribution systems evolve to embrace far greater use of the Internet. We must ensure that data and metadata can never be dissociated. For example, a BUFR (Binary Universal Form for the Representation of meteorological data) file containing an ocean observation that is distributed on the GTS in real time must have an iron-clad linkage to the Internet-based 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, WMO has mandated [25] that by 2012 messaging on the GTS will switch from the old ASCII (American Standard Code for Information Interchange) 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.

8. ARCHIVING OCEAN DATA

Ocean archives are tasked with long-term preservation of ocean observations. At the end of the 20th 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 of users for instantaneous, online access to information.

During the first decade of the 21st century, digital archives around the world began to share experiences and challenges. They discovered that these communications were hampered 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 are embracing the responsibilities with increasing enthusiasm. The adoption of OAIS-RM can help ocean data archives to improve their internal 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-systems integration should blur the divisions between management of real-time, delayed mode and archived data, so that users need only have minimal awareness of which particular level of the system is providing services.

9. 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 and 31] can help to address both scientific and management questions.

The OBIS projects represent a successful start on a much larger effort. Community data standards are needed to 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.

10. 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.

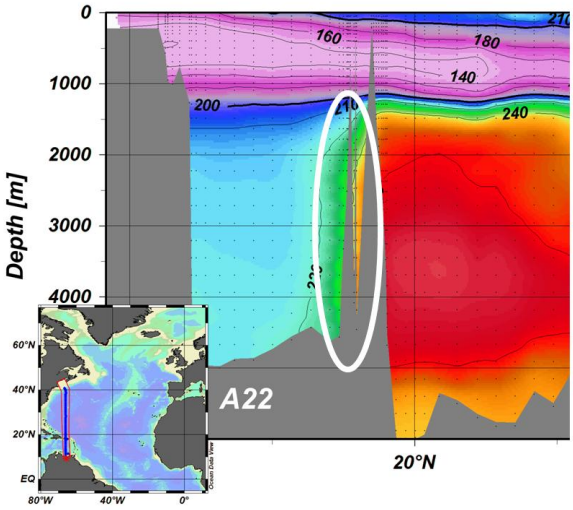

Figure 3: Oxygen distribution along the WOCE A22 section showing elevated values in the Caribbean deep water. This structure is an artifact of the weighted-averaging gridding algorithm and caused by influence of high-oxygen waters on the northern side of the ridge.

11. ACCELERATING DEVELOPMENT THROUGH SOFTWARE TOOLKITS

Most funding for ocean data system development is directed to support scientific programs such as WOCE (WorId Ocean Circulation Experiment) or JGOFS (Joint Global Ocean Flux Study); institution-specific research and development infrastructures; platform-specific data management systems (DACs and GDACs (Global Data Assembly Centers)); 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 has been re-developed (incompatibly) 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 software developers with the capability 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[10] (IDV), ncWMS [33], Panoply[11], Live Access Server[12], and ERDDAP (Environmental Research Division's Data Access Program)[13]. Thus toolkits have allowed scientists and software developers to concentrate their energies on the unique contribution that they wish to provide. Further advancement of software libraries in programming languages such as C and Python will similarly accelerate progress in data system tool development.

An important example of the need for toolkits is found in the common problem 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 wide range 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. None of the utilities that have attempted to address this problem for scientists, for example the Data Interpolating Variational Analysis package (DIVA[14]), have yet been designed in the form of software libraries that can be readily integrated into other applications. It is only with difficulty that stand-alone gridding utilities can be into integrated general purpose software packages. (An example of this approach may be seen in the Ocean Data View[15].)

12. CONCLUSIONS -- CONCRETE COMMUNITY TARGETS FOR IMPROVED DATA INTEGRATION

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 wish to suggest what a few of those steps should be. The list provided here is by no means comprehensive. It does, however, provide some directions for which there is 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 actions to pursue these concrete steps and identify others.

12.1. Ocean Observations Made 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 relatively straight forward applications of the same techniques. Many of the techniques are applicable to biological data such as continuous plankton recorder observations. This trio of practices has been accepted as a standard for gridded data by US IOOS (United States Integrated Ocean Observing Systems) [34] and is working its way through the NASA ESDSWG (National Aeronautics and Space Administration /Earth Science Data Systems Working Groups) standards processes [35]. Efforts to achieve standardization within OGC have begun[16].

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. Since the examples enumerated in the previous paragraph already address most platform types the technical barriers to standardizing the remaining observations are in most cases significantly smaller than the barriers that have already been surmounted.

Broad convergence on the use of netCDF-CF-OPeNDAP is but a milestone along a path to the 10 year

[10] www.unidata.ucar.edu/software/idv/

[11] www.giss.nasa.gov/tools/panoply/

[12] http://ferret.pmel.noaa.gov/LAS/

[13] http://coastwatch.pfeg.noaa.gov/erddap/

[14] http://modb.oce.ulg.ac.be/projects/1/diva

[15] http://odv.awi.de

[16] 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

vision that we outlined at the introduction of this paper. Indeed these technologies have acknowledged imperfections that must be addressed. A strong motivation to achieve convergence despite known imperfection is the multiplier effect that convergence brings. Each new tool that is developed and every technical improvement that is made thereafter will have wider applicability; each will yield on average greater benefits to the community.

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

12.3. 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 (Joint Technical Commission for Oceanography and Marine Meteorology) 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

12.4. Addressing organizational, cultural and policy issues

The strategy for making progress in the face of an evolving technology base is to implement changes in an incremental fashion guided by a shared “heroic” vision of future capabilities. The paper suggests the following guidelines for community actions:

- build active participation on the part of scientists, program managers and data management professionals into all activities;
- encourage data sharing policies that are as open as possible in order to deliver data as quickly as possible to all;
- wherever feasible take advantage of work done by others (software, data modelling, standards) that has demonstrated its effectiveness in realistic setting, rather than building capabilities from scratch.

13. REFERENCES

1. Pouliquen, S. & Co-Authors (2010). "The Development of the Data System and Growth in Data Sharing" in these proceedings (Vol. 1), doi:10.5270/OceanObs09.pp.30

2. Blower, J. & Co-Authors (2010). "Ocean Data Dissemination: New Challenges for Data Integration" in these proceedings (Vol. 1), doi:10.5270/OceanObs09.pp.05.

3. Keeley, R., Woodruff, S., Pouliquen, S., Conkright-Gregg, M. and Reed, G., (2010). "Ocean Data: Collectors to Archives" in these proceedings (Vol. 1), doi:10.5270/OceanObs09.pp.25.

4. Clayton, M. Christensen (2003), *The Innovator's Dilemma: The Revolutionary Book that Will Change the Way You Do Business*, Harper-Collins Publishers, Inc. New York, 320 pp.

5. 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.

6. Diamond, J., (2005), *Guns, Germs, and Steel: The Fates of Human Societies*, W.W. Norton & Co., 512 pp.

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

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

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

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

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