OCEAN AND COASTAL DATA STEWARDSHIP

Margarita Conkright Gregg[(1)] Michele Newlin[(1)], Sharon LeDuc[(2)], Robert Keeley[(3)], Nick D'Adamo[(4)]

[(1)] U.S. National Oceanographic Data Center, 1315 East-West Highway, Silver Spring, MD, USA, Email: Margarita.Gregg@noaa.gov, Michele.Newlin@noaa.gov
[(2)] National Climatic Data Center, 151 Patton Avenue, Asheville, NC, USA: Email: Sharon.Leduc@noaa.gov
[(3)] Integrated Science Data Management, Department of Fisheries and Oceans, 200 Kent St., Ottawa, Canada, Email: Robert.Keeley@dfo-mpo.gc.ca
[(4)] Perth Regional Programme Office of the IOC, UNESCO, 1100 Hay Street, West Perth 6005, Western Australia Email: N.D'Adamo@bom.gov.au

ABSTRACT

As the ocean community looks to the next decade of ocean observations, end-to-end data management must be considered as an integral part of the design of any ocean observing system. Ocean data stewardship ensures that observations deliver the maximum service to society. Ocean data stewardship means more than mere mechanical or electronic acts of data archiving and transfer. It consists of an integrated suite of functions to preserve and realize the full value of environmental data. Data stewards and data providers need to work toward standardization of vocabularies and formats, quality and complete metadata, and making the archival and dissemination of data a part of the initial design of an observing system. These functions must be successfully implemented to ensure optimal use of oceanographic data and information, both now and in future.

1. WHY OCEAN DATA STEWARDSHIP?

No single country can monitor the entire world ocean alone. Over the past decade, the international ocean community has launched a number of highly successful programs that have led to an understanding of how the Earth's climate system and its many components function and change over time. Data from the implementation of global ocean monitoring programs, such as the Global Ocean Observing System (GOOS) [1], Argo [2], and the Repeat Hydrography Program [3], have been used to document changes in the ocean heat and freshwater content, as well as changes in the acidity of the oceans. However, to understand how both the ocean and climate are changing, we must first quantify and understand past variability. Therefore, we must ensure the quality and preservation of the data that provide the historical record of Earth's changing marine environment so it can be used for operational applications and ocean climate research.

Observation data are unique, irreplaceable, and collected at great cost. The *acquisition, processing, preservation, quality assurance, and dissemination* of coastal and ocean data by national and international data centers are key elements of curating and archiving data – what we call "Data Stewardship". These functions are required to ensure future accessibility, readability, quality (e.g. richness, trustworthiness, etc.), and usefulness of all types of oceanographic and related ecosystem data.). National Data Centers, Data Assembly Centers, and World Data Centers (WDC) data collections and databases are a result of international data exchange and all countries and scientists benefit from this cooperation. For example, the international Global Temperature-Salinity Profile Program (GTSPP) and Global Argo Data Repository project coordinate with the US NODC/WDC World Ocean Database to archive historic, delayed-mode and real-time oceanographic profile and Argo float data and produce atlases and analysis products [4]. The importance of NODC/WDC databases, and atlases based on those databases, can be quantified: Figure 1 shows that NODC/WDC databases and products based on those databases have been cited more than 5,900 times since 1982.

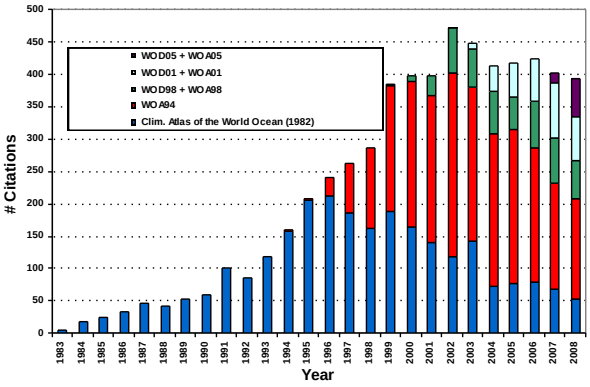

Figure 1. Citations as a function of time of NODC databases and products based on those databases

Data stewardship means more than mere mechanical or electronic acts of data archival and transfer. Data must be archived with all their associated metadata so that it can be reliably used and interpreted over time. Data subjected to quality control analyses makes it possible to identify and address anomalies and limitations in the data or metadata. In the best scenario, this documentation and review is done before principal investigators move on to other projects or retire, thus avoiding a possible loss of information critical to the long-term use of their data. Acquiring analog historical data and converting them to electronic form before they are lost to use by the scientific community is another important stewardship activity. The Global

Oceanographic Data Archaeology and Rescue (GODAR) project of the Intergovernmental Oceanographic Commission (IOC) has been very successful in this regard [5].

Data stewardship plays a vital role with not only the preservation and use of oceanographic data, but also with the creation of key products derived from archived data. End-to-end management of data facilitates the ability to integrate data from multiple sources and sensors (i.e. satellites, *in situ*, and model data). An example of one of these key products is the Global Ocean Heat Content (Fig. 2). This product integrates different observing systems and programs with very different quality control techniques into one format for a more enhanced climate record. Another example is the G Measurements of

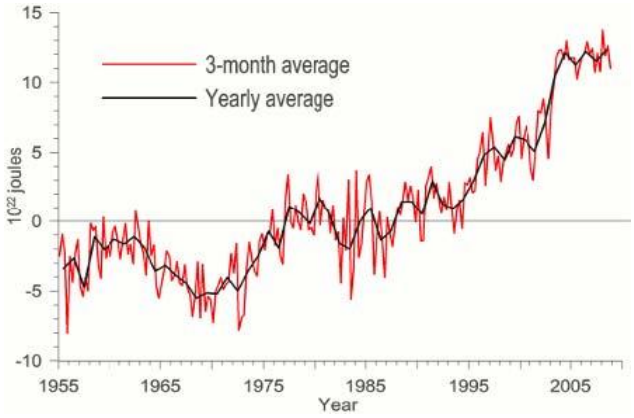

Figure 2. Global Ocean Heat Content (1955-present) based on integrating historical and modern (Levitus et al. 2009).

environmental parameters taken at specific locations and times are unique. Preserving detailed information about observed variables, along with observation instruments used, techniques, calibrations, and other metadata are critical components of data stewardship efforts. Comprehensive quality metadata about observed values enables these data to be used as a baseline for future comparisons or integrated with other parameters to be used for a wide variety of research, educational or commercial applications. With sufficient metadata, data taken from random ocean sampling can be assembled together to produce frequently used climatologies, such as the World Ocean Atlas [6]. Data may also have purposes and benefits beyond those originally intended, such as the use of surface drifter data from the Gulf of Panama to investigate the life stages of sea snakes. Ultimately, the full value of data cannot be realized without sustained, long term, and rigorous ocean data stewardship.

2. HOW DO WE ENSURE PROPER STEWARDSHIP OF OBSERVATIONS?

To achieve this goal, a truly end-to-end design and management system is needed along with adoption of the latest standards and technologies. These concepts are not new, but as we deal with increasingly heterogeneous and complex data, it is imperative that we incorporate these mechanisms into the design of any observing system.

The end-to-end design of an observing system must incorporate:

2.1 Standardized data collection and sensor calibration

Customers for real-time ocean data are also customers of weather services. Currently, the infrastructure for ocean observing systems and ocean data is not consistent with the existing infrastructure of World Meteorological Organization (WMO) observing systems. As operational oceanography transitions into its full capabilities we need to develop data services standards which are compatible with the existing WMO system. This can be achieved by close coordination during the development of the standards that facilitate interoperability. Some of these activities are already being addressed by the Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology (JCOMM). The Pilot Project for Marine Observations under the WMO Integrated Global Observing Systems (WIGOS) framework is an example of the development of standardized data transmission across the meteorological and oceanographic communities [7]. WIGOS will make appropriate datasets available in real-time and delayed mode to WMO and IOC applications through interoperability arrangements with the WMO Information System (WIS). The pilot project will integrate *in-situ* and "remotely sensed" ocean observing systems. These activities will benefit the ocean community since the addition of a standard data acquisition service, that manages data along side collection-level metadata, provides an opportunity to automate and manage elements of the process at strategic points.

2.2 Common Vocabularies

One of the major challenges facing data stewards is how to facilitate the discovery of available, relevant data, especially when one is seeking data from multiple data sources. Controlled vocabularies and standardized metadata discovery tools help make it easier to reliably discover data in distributed repositories [8]. Libraries have used common subject headings to relate items with similar content characteristics for decades, but the ocean data, metadata and stewardship communities are still working towards a similar capability.

Although a single controlled vocabulary used by all national and international data managers is unlikely especially in the near term, data managers and stewards are working together to identify best practices for some controlled vocabularies (e.g. using Global Change

Master Directory construction guidance for scientific variables or institutional names) or to share a common database of information through web-based services (e.g. observation platform names managed by the International Council for Exploration of the Seas (ICES) used by ICES and SeaDataNet in the European Union and NODC in the US). Developing basic crosswalks and richer ontologies among terms in different controlled vocabularies is an area of active development within the metadata community. Tools that enable matching terms in one vocabulary to one or more similar terms in other vocabularies and registries for vocabularies and ontologies are made freely available (e.g., http://mmisw.org/or) to the community with a goal to enhance the semantic understanding of the data and metadata.

The application of controlled vocabularies that include well-defined terms and relationships between terms helps to disambiguate results of searches and to improve our understanding of data. Building, maintaining, and managing vocabularies requires a significant effort for data stewardship, but ultimately pays for itself when unambiguous terms are widely used to produce accurate and reliable results for searches across multiple data discovery environments and for aggregations and transformations of data from multiple disciplines.

2.3 Standard Data Formats

To more efficiently ensure proper data stewardship, the selection and adoption of a small number of standardized data formats is essential. While it is believed that no one format can satisfy all needs, the use of just a few formats in all but the most extreme cases would dramatically enhance the ability of data stewards to preserve information over the long term and to make it readily accessible in a wide variety of formats today. The combination of NetCDF (Network Common Data Form) with Climate and Forecast (CF) attributes is a strong contender for the position of primary data format and attribute convention since powerful data server technologies such as Open-source Project for a Network Data Access Protocol (OPeNDAP), World Coverage Service (WCS), and Web Map Service (WMS) exist today, and can provide data which meets those standards to users in forms that they are accustomed to working with. The future of NetCDF is NetCDF-4, which uses an underlying HDF-5 file format but with a simpler application programming interface.

2.4 Quality Assurance and Quality Control (QA/QC)

There are even further considerations to be given with standards for QA/QC for indicating results, and documenting QC descriptions. The preservation of original values, even if they appear wrong, needs to be considered, and preserving the history of records could be important for possible processing questions down the road. Data should also be properly or uniquely tagged, so that as it is ingested by various centers and users, there are no questions about its origins or versions. This could help with the controlling of duplicates, and with identifying operations with added value. Another example is implementing QA/QC standards for *in situ* ocean sensors using Open Geospatial Consortium (OGC)-Sensor Web Enablement.

2.5 Data Archive

The functions required for long term preservation and stewardship of digital information are defined in the Open Archival Information System Reference Model (OAIS-RM), the ISO standard (ISO 14721) for digital archives (Fig. 3). These services span the Functional Entity areas of Ingest, Archival Storage, Data Management, Access, Preservation Planning, and Administration, plus Common Services and include a comprehensive collection of activities that should be performed by an OAIS archive.

To conform to the OAIS-RM, data stewards must accept a set of mandatory responsibilities. The data stewards must:

- Negotiate for and accept data and information from Producers of that information;
- Obtain sufficient control of the information to ensure its long-term preservation;
- Determine and understand its Designated Communities, for whom the data are being preserved;
- Ensure the information is independently understandable to the designated community, which means that members of that community should not need the special assistance of experts to make proper use of the data;
- Follow documented policies and procedures; and
- Provide to the Designated Community access to the preserved information.

The OAIS-RM highlights some key areas that are necessary to ensure proper stewardship but are not often considered part of an end-to-end design:

- Feedback between Archives and Data Providers, as well as feedback between Archives and the Data Consumers;
- Planning for the future by working closely with the Designated Communities (the primary constituencies for whom the data stewardship is being provided) to understand the way they access and use data; and

The concept of “Active Archives”, which utilize the data themselves in various applications, integrate the data into “value-added” products, and other uses. The process of actually using the archived data is a key

component of proper stewardship since it ensures the data are accessible, readable, and understandable at the time the archive receives it.

In order to conform to the OAIS-RM and therefore enable the long-term preservation and dissemination of data, data producers and data archives need to work together to generate the information needed to be able to understand and re-use data. Well-defined file naming conventions and format descriptions, data with all related descriptive metadata, plus information added by data centers or assembly centers to manage data and facilitate data dissemination are critical for long-term accessibility and data use.

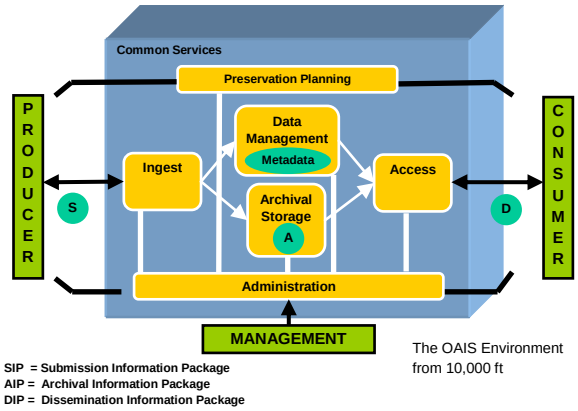

Figure 3. The OAIS framework

A growing international community is converging on the adoption of standards as the solution for ingesting highly complex and heterogeneous data into models, multidisciplinary studies, and resource management. The ocean community needs to agree on these standards and implement them as a routine part of the design for the next decade of observing systems.

Other papers in this volume address these issues. For instance, de La Beaujardière and authors examine integration, access, and standards [8]; Hankin and authors [9] also address standards with a focus on interoperability; Snowden and authors [10] suggest improvements on how we manage the metadata describing ocean observations. These and other similar papers illustrate the key role end to end data management plays in ensuring a "successful" observing system.

3. WHAT ARE THE CHALLENGES FOR TODAY AND THE FUTURE?

Scientific inquiry demands new data and the re-casting of historical data to help reduce uncertainty in understanding. Forecasting sea level rise, understanding the impact of climate on coastal ecosystems, forecasting the "weather of the ocean", loss of sea ice, and changes in ocean circulation, all require a data integration/assimilation framework that accommodates multiple data types, formats, and user needs. For example, Integrated Ecosystem Assessments (IEA) will benefit from a data management architecture that synthesizes and registers all metadata and data relevant to physical, chemical, ecological, and human processes in relation to specified IEA management objectives (Fig. 4). Without the creation of an integrated and interoperable data management architecture NOAA and the nation risk developing an architecture that is "stovepipe" in nature, i.e., not connected, lacking standards, protocols, and the organization to promote data discovery, access, and archiving across diverse and distributed data networks.

Data stewards face many challenges in today's world. Understanding issues such as the effect of ocean acidification on plankton and corals or hypoxia-triggered harmful algal blooms along our coasts highlight the need for integrating oceanographic data, particularly chemical, biological, and fisheries data. To support this effort, complex data collections are obtained from increasingly complicated systems,

Figure 4. Model of an Integrated Ecosystem Assessments (IEA) for the Gulf of Mexico

which are more difficult to annotate, describe, organize and disseminate. Data volume necessarily increases as the number of observation types expands to address new information requirements. Designing the technological infrastructure that balances scalability and retention requirements with the needs of archivists and end users is particularly difficult.

Large scale partnerships are required to address these and other large scale issues. Through partnerships in the ocean and coastal communities, we can ensure stewardship of all new data, develop the capability to provide increasingly complex data in a form that is usable for multiple purposes, and engage the user community to generate products that meet their needs. Leveraging our knowledge, time, and experience, as demonstrated by projects such as GEOSS, JCOMM, SeaDataNet in the European Community, IOOS, Argo,

GTSPP, GHRSST (Global High Resolution Sea Surface Temperature), GODAR, etc., has proven to be a successful mechanism for improving our ability to understand the world's oceans. However, there is still much to do.

For instance, we need to create distributed data systems where it is easy not only to request the latest versions from the appropriate national/international body, but also ones that have them available. These systems also need to be available to users from all nations. Data managers and data collectors must collaborate more closely to ensure that observation data is well documented and organized so that it is more easily assimilated into information products and archival collections. Blower and authors [11] describe current efforts and challenges in ensuring data are used as effectively as possible but more importantly, they propose high-priority activities for making this a reality such as how to promote data sharing, understand the user community, and standardization and integration of data,.

These activities all require resources. We can optimize the available resources by using agreed-upon standards and building a strong relationship between data producers and data stewards from the outset of a project or program. By including data managers in the planning and execution phases of new data collecting programs (such as ocean acidification [12], the proposed global high resolution SST integrated observing system [13], and deep ocean measurements [14]), the likelihood of having data collections that are ready for assimilation into other products and archival collections is significantly improved.

4. CONCLUSIONS

We have made great progress in observing and understanding the ocean. We now have an excellent opportunity to realize the full return on the investment we have made in observing the oceans by investing in managing those observations for future generations. Data stewards and data providers need to work toward standardization of vocabularies and formats, quality and complete metadata, and making the archival and dissemination of data a part of the initial design of an observing system. These functions must be successfully implemented to ensure optimal use of oceanographic data and information, both now and in future.

5. ACKNOWLEDGEMENTS

We wish to acknowledge the additional contributions from Kenneth Casey, Charles Sun, Tim Boyer, Eric Roby, Julie Bosch, Donald Collins, Sydney Levitus, Russ Beard, Rost Parsons, and Andy Allegra (from NOAA's National Oceanographic Data Center), Sharon LeDuc (NOAA's National Climatic Data Center), and Ming Ji (NOAA National Weather Service).

6. REFERENCES

1. Koblisnky, C. and Co-Authors (2010). Plenary Talk From the in situ perspective. Full paper not submitted in these proceedings.

2. Hood, M. & Co-Authors (2010). "Ship-Based Repeat Hydrography: A Strategy for a Sustained Global Program." in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.44.

3. Pouliquen, S., Schmid, C., Wong, A., Guinehut, S. and Belbeoch, M., (2010). "Argo Data Management" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.70.

4. Sun, C. & Co-Authors (2010). "The Data Management System for the Global Temperature and Salinity Profile Programme" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.86.

5. Boyer, T. P., J. I. Antonov, H. Garcia, D. R. Johnson, R. A. Locarnini, A. V. Mishonov, M. T. Pitcher, O. K. Baranova, and I. Smolyar, 2006: World Ocean Database 2005, Chapter 1: Introduction, NOAA Atlas NESDIS 60, Ed. S. Levitus, U.S. Gov. Printing Office, Wash., D.C., 182 pp., DVD.

6. Locarnini, R. A., A. V. Mishonov, J. I. Antonov, T. P. Boyer, and H. E. Garcia, 2006. World Ocean Atlas 2005, Volume 1: Temperature. S. Levitus, Ed. NOAA Atlas NESDIS 61, U.S. Gov. Printing Office, Washington, D.C., 182 pp.

7. 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.

8. de La Beaujardière, J. & Co-Authors (2010). "Ocean and Coastal Data Management" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.22.

9. 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.

10. Snowden, D. & Co-Authors (2010). "Metadata Management in Global Distributed Ocean Observation Networks" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.84.

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

12. Feely, R. & Co-Authors (2010). "An International Observational Network for Ocean Acidification" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.29.

13. Donlon, C. & Co-Authors (2010). "Successes and Challenges for the Modern Sea Surface Temperature Observing System" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.24.

14. Garzoli, S. & Co-Authors (2010). "Progressing Towards Global Sustained Deep Ocean Observations" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.34.