Infrastructure and tools for serving, accessing, and analyzing ocean information from the Integrated Ocean Observing System

Derrick Snowden
US IOOS® Program,
National Ocean Service, NOAA
Silver Spring, MD USA
derrick.snowden@noaa.gov

Kelly Knee
RPS/Applied Science Associates
South Kingstown, RI USA

Richard Signell
US Geological Survey
Woods Hole, MA 02543, USA

Kathleen Bailey
US IOOS® Program,
National Ocean Service, NOAA
Silver Spring, MD USA

Filipe Fernandes
SouthEast Coastal Ocean Observing Regional Association
Charleston, SC USA

Emilio Mayorga
Northwest Association of Networked Ocean Observing Systems,
University of Washington
Seattle, WA USA

Vembu Subramanian
SouthEast Coastal Ocean Observing Regional Association
Charleston, SC USA

Abstract— **According to the Integrated Coastal Ocean Observation System (ICOOS) Act of 2009 the U.S. Integrated Ocean Observing System (IOOS®) Enterprise extends across 17 federal agencies and 11 regional associations and includes numerous actors from within those organizations. One of the primary functions IOOS provides is a Data Management and Communications (DMAC) Subsystem that aims to make discoverable and accessible data and information from multiple disciplines across the aforementioned enterprise. With such diverse participation and broad mandate for the types of data included in IOOS, it is unrealistic to expect that a single data center is capable of aggregating, managing, curating, and distributing all of the ocean data of interest to the IOOS enterprise. Instead, the IOOS enterprise implements a distributed data network bound together by a few key features of a shared vision for data discovery and access. This paper will discuss progress and lessons learned from nearly ten years of experience in creating or adapting the standards, tools, and community needed to develop and maintain the distributed data network that will support IOOS efforts in science, operational decision making, and product delivery.**

The current configuration of the DMAC subsystem of IOOS is a combination of people, process, and technology that provide a service to the nation. The primary service DMAC provides is to deliver well curated and documented ocean data and information to the public using the World Wide Web as the primary platform. Within these three areas (people, process, and technology) choices are made based on resources, policy mandates, available skills, technical maturity and capability, and customer requirements. Collectively the choices within each area determine the architecture of the DMAC system and will form the organization for this paper. The "people" form the main stakeholder groups of IOOS, both the builders and the users, so understanding how these various stakeholder groups work cooperatively to grow the DMAC system is critical to progress. The "process" area determines how the people work together and the policy constraints the system is under. Finally, the "technology" includes the software and standards DMAC implements to address the system needs.

Keywords—data management; informatics; data publication; web services; information technology; ocean observations; ocean modeling

I. Introduction

The US Integrated Ocean Observing System (US-IOOS) is a partnership between 11 regional associations and 17 federal agencies [1][2]. Each region and federal agency collects a variety of observational data and runs a variety of numerical models. The goal of US IOOS DMAC is to make this data easy to discover, access and use. Considering the diversity of these organizations, the range of their capabilities and the breadth of data types collected, the system was developed by specifying a set of standardized web services for providers, that would result in a distributed data system with centralized metadata search capabilities. This leads to more effective and efficient use of ocean information by environmental managers, emergency response teams, scientific research, forecast model skill assessment, and by developers of applications for the general public [3]–[5].

While providers are free to choose whatever means they wish to provide the specified services, it has proven useful to suggest a particular stack of technologies that are known to work effectively. In other words, tested software and recipes for providers to enable the IOOS services (Table 1).

Similarly, IOOS users are free to choose the tools they wish to access IOOS services, but it has proven useful to develop and promote libraries and tools that facilitate use of these services in commonly used languages and environments. This strategy has involved the adoption and sometimes targeted enhancement of existing community libraries as well as the development of new libraries to address important gaps.

The DMAC system has been functional since 2007 and has been continually evolving and improving based on provider and user needs. In this paper we describe the DMAC technologies and practices that have been developed over the last decade, driven by real world use. We describe the basic components of the system for providers and users, then discuss applications that demonstrate the power of the system. The applications or examples are chosen from a project called the DMAC System Integration Test which was conceived to determine the effectiveness of the DMAC system currently, and the opportunities for improvement in the future[1].

II. DMAC System Capabilities, Strengths, and Weaknesses

A *system* is a construct or collection of different elements that together produce results not obtainable by the individual elements alone[2]. A common, generic listing of these "components" typically identifies elements of technology, people, and processes or governance that somehow combine to perform a desired function. Within US IOOS, there are technology, people, and process elements that combine into the Data Management and Communications system, which aims to provide the capabilities needed to collect and deliver ocean data and information to the public. In the sections that follow we describe the composition of DMAC by identifying the people or organizations actively involved, listing the technology elements that comprise the data network, and comment on the development process by which DMAC is built and evolved. Through the description we use an example analysis to illustrate the concepts. The example chosen provides summary statistics associated with extreme events and in particular focuses on the return period of high water level events. Other examples of using DMAC services to perform calculations include model skill assessments via comparison with observational data [6].

A. A simple model of the DMAC functions

The primary function of the DMAC system is to provide a standardized, well-described mechanism for users to analyze ocean information and derive value from these analyses [3], [4]. The steps involved in this analysis workflow can be generalized to three activities. First, a user must be able to discover data of interest. Second, a user must be able to immediately access that information from tools of their choice. Last, a user must be able to apply that data to a problem. Typically this last step is labeled the "use" step and implies that the information that is accessed in step two is easily and readily understandable by the user. This "discover/access/use" paradigm provides the outline for discussing DMAC in general and the specific example. Within this paradigm, also known as the publish-find-bind architectural model, we assume that primary mode of interaction will be through machine to machine communications (Fig. 1) [7]. It is through standardized requests and responses delivered via web services that machines will discover, access, and use ocean information.

B. An example: Calculating Statistics Associated with Extreme Events

Extreme events, such as coastal inundation or severe storms, are typically characterized by comparing a particular signal with expected or average conditions. For example, a storm surge event may be characterized in terms of the increase in water level relative the predicted tidal elevation. The basic analysis requires the ability to plot or summarize the current condition and compare it to an exceedance probability curve calculated from long time series of similar data. Casting that scenario into the "discovery/access/use" paradigm discussed above we can identify four questions to be answered by the analysis:

1. Is it possible to discover long enough records of water level data to successfully perform extreme values analyses and define the return period of extreme conditions for a given location?
2. Is it possible to discover and compare modeled and observed water level data for a given geographic location or within a region?
3. What is the extent of relevant meteorological and oceanographic data available in near-real time that could be used to support rapid deployment of a common operational picture to support response?
4. Is it possible to discover and use information about the consequences of an event (e.g. determining whether extreme water levels cause overland inundation and to what extent)?

Throughout the remainder of the document we will refer to this example as the system-test notebook. This notebook is an IPython notebook utilizing client libraries written in Python.

---

[1] https://www.github.com/ioos/system-test

[2] http://www.incose.org/AboutSE/WhatIsSE

The IPython notebook code is available online[3] as are all notebooks created as part of the system-test project. While there are certainly client software libraries written in other languages the examples presented here will all be Python clients.

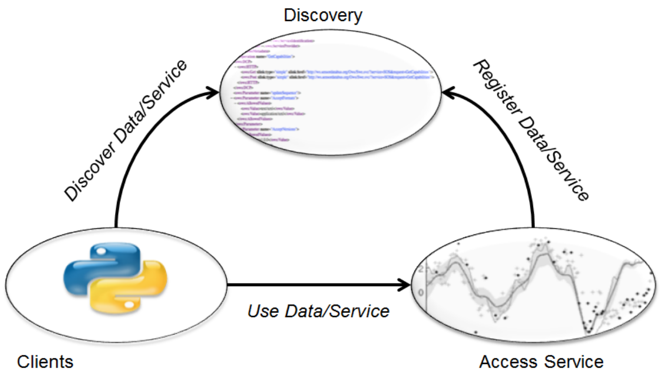

Fig. 1. Illustration of the publish/find/bind architecture pattern that provides a high level characterization of the major functions of DMAC

C. Technical Components of DMAC

The technical components of DMAC are chosen or built to help implement the discover/access/use paradigm for all types of ocean information. These components are implemented across a large and distributed network of data users and data providers. Therefore, to maximize interoperability, these components are standardized to the extent possible. Standardization means that the underlying behavior of the software, or format of the data files are governed by a set of rules documented as a specification. These rules are documented so that software developers can understand the principles and write software that adheres to the standard. However, standards alone are insufficient to ensure that all of the components within the system can communicate or interoperate effectively. A specification is not useful until it is implemented in software that is then integrated into real world systems [8]. In the sections that follow we describe the standards that underpin the DMAC system and the software implementations that are used in practice.

1) Data Discovery

Any new interaction with ocean information begins with a discovery step. Typically, the user is confronted with a problem and some sort of data is required to solve the problem. Strategies to discover data online range from using commercial search engines, browsing through known Data Center web pages, or querying a federated system of data catalogs using query languages that are designed to address spatio-temporal queries. Likely, all three strategies will be employed in many situations but here we focus on geospatial catalog queries.

The Open Geopspatial Consortium Catalog Service for the Web (CSW[4]) is the primary interface standard in use across IOOS/DMAC particpants. The CSW interface provides the capability to query metadata catalogs, which are populated with standardized dataset descriptions such as the ISO 1911* family of geospatial metadata standards. These ISO documents describe various elements of a data set including geospatial and temporal extent, data content, keywords, and web service location (i.e. a URL). Common software implementations of CSW data catalogs include the Esri Geoportal Server, pycsw, and CKAN. IOOS/DMAC relies primarily on the ISO 19115-2 metadata standard for geospatial data. An IOOS/DMAC service provider can stand up its own CS/W server to advertise its web service endpoints and metadata, or register its service endpoints and metadata collections with the IOOS/DMAC Registry that also incorporates mappings from standardized services to ISO 19115-2 metadata. This centralized Registry and a related Esri Geoportal CS/W server are hosted by the NOAA National Centers for Environmental Information (NCEI) for the network of providers associated with IOOS.

From the extreme water level example the following steps outline the parameters of the query that will be made of select CS/W servers, such as the NCEI CS/W server described above.

1. Define spatial bounds of interest
2. Define temporal bounds of interest based on minimum desired record length
3. Define parameters of interest using standard variable names
4. Select the discovery services to be queried and issue the query
5. Process the results, which should contain the URLs for data access services

These query parameters are encoded as Python statements in an IPython notebook (Fig. 2 screenshot of the notebook with the folium map). In this case the IPython notebook is the client and the NCEI CSW[5] is the discovery service (Fig. 1). The Client issues the query to the Discovery service and processes the results.

2) Data Access

DMAC data providers offer several standard data access services. These services provide application programming interfaces (APIs) through which client software can perform sophisticated subsetting operations. Using APIs to access the data of interest is more efficient than the alternative, which is downloading a possibly large and unstructured data set and performing all of the preprocessing and subsetting on a user's personal computer. The particular standards for data access endorsed by IOOS include the OGC Sensor Observation Service (SOS) and OPeNDAP (Table 1). SOS services are associated with observational data, while the OPeNDAP

[3] https://github.com/ioos/system-test/blob/master/Theme_2_Extreme_Events/Scenario_2A/Extremes_Inundation/Extreme_Water_Level.ipynb

1 [4] http://www.opengeospatial.org/standards/cat
[5] http://www.ngdc.noaa.gov/geoportal

services can be associated with both observational or model data [5].

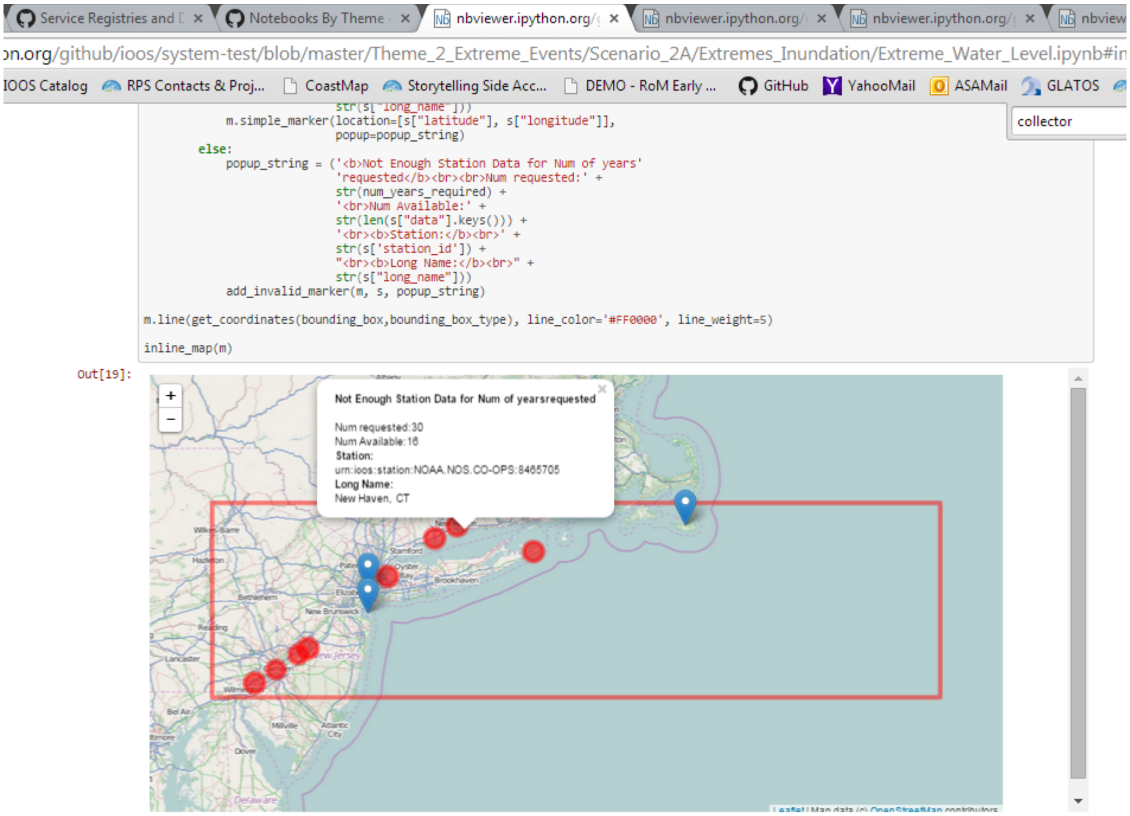

Fig. 2. IPython notebook showing the geographical locations of select stations matching the CSW query issued as part of the system-test water level analysis.

TABLE I. IOOS RECOMMENDED DATA ACCESS SERVICES

Interface Standard	Recommended Software
OGC SOS	IOOS fork of the 52north.org server[a]
OGC SOS	ncSOS plugin for THREDDS[b]
OPeNDAP	THREDDS[c]
OPeNDAP	ERDDAP[d]
OGC WMS	Geoserver[e]
OGC WMS	ncWMS plugin for THREDDS[f]
OGC WMS	Sci-WMS[g]

a. https://github.com/ioos/i52n-sos
b. https://github.com/asascience-open/ncSOS
c. http://www.unidata.ucar.edu/software/thredds/current/tds/
d. https://github.com/BobSimons/erddap
e. http://boundlessgeo.com/solutions/solutions-software/geoserver/
f. https://github.com/Unidata/ncWMS
g. https://github.com/sci-wms/sci-wms

Once the records of interest have been identified via the catalog CSW search, the IOOS supported data service endpoints are extracted from the metadata. To be able to retrieve geospatial data in a common way using OPeNDAP, the Climate and Forecast (CF) conventions must be used in the underlying data. For unstructured grid and staggered grid model data, the CF conventions must be supplemented with UGRID[6] and SGRID[7] convention metadata.

Often the user does not require the actual data set to be accessible and has simpler needs for visualizing data in a preview mode. This type of access is typically accomplished via OGC Web Map Services. The records returned from the catalog CSW search can also be mined for OGC Web Map Service (WMS) endpoints. These endpoints produce georeferenced and projected images of data to be returned, allowing quick preview of data that can range from simple image returned from a specified URL to sophisticated web portals that allow interactive exploration. Several implementations of WMS compliant software exist and are designed to complement the netCDF based storage and OPeNDAP access via THREDDS [9], [10]. The IOOS Coastal Ocean Modeling Testbed has invested significantly in the sci-wms software package, a python-based web map service that delivers products from both UGRID and SGRID convention model output.

6 https://github.com/ugrid-conventions
7 https://github.com/sgrid

The system-test query described previously resulted in 417 data services that include data in the region around Long Island advertised through the NCEI Geoportal Server. These records were further filtered to reveal 79 SOS services and 106 OPeNDAP services, which were available for subsequent analysis.

3) Data Use via Client Software

There are many ways to access and download data hosted online. Data available via OGC Web Services (OWS) standards can be easily queried and accessed via the low-level Python library OWSLib[8]. Pyoos[9] and Ulmo[10] are two Python packages designed to address the need for higher-level, more user friendly access and parsing for observational, mainly *in situ* data. Ulmo focuses on access to hydrology and climatology data sources using a variety of access mechanisms and data encodings, while Pyoos, which leverages OWSLib, provides access to met/ocean data via SOS and other access mechanisms and different data encodings. The development of Pyoos was supported by the IOOS PO specifically to facilitate access to data sources relevant to the IOOS community.

The Iris[11] and xray[12] are high level libraries that create data abstraction layers which isolate analysis and visualization code from data format specifics. Both can read a variety of formats (PP, netCDF3/4, HDF4/5, Grib, etc) and access data via OPeNDAP. Iris is more restrictive than xray because it enforces the CF Data Model, while xray extends the widely used Pandas [13] DataFrame data structure and API into multidimensional arrays typical of netCDF files. CDMs remove the burden of writing data format specific code and let the user perform more fundamental data manipulation operations, such as arithmetic, interpolation, and statistics; and a range of integrated plotting options.

In the system-test example, data from the stations in Fig.2 were selected for further analysis. These data, served via THREDDS and SOS services, were accessed using PyOOS. Stations with more than 30 years of data were identified and the time series were accessed for further comparison with model simulations.

D. Preliminary Conclusions from the System Test

The system-test example described here is just one of over twenty notebooks written to exercise the DMAC system. When embarking on this project, the stated goal was to identify areas in need of improvement; however, testing the ability of the infrastructure to answer specific scientific questions was an extremely worthwhile exercise.

Some challenges encountered fell into categories of data missing from the system, problems with dataset granularity, service issues, metadata accuracy and server stability and configuration. With many system tests, we knew that certain data products existed but were not yet connected to the data infrastructure. For others, the data were accessible but had not yet been aggregated appropriately, so that hundreds of records were found instead of only one.

The ESTOFS data analyzed as part of the system-test notebook (not shown) are a good example of many of these issues. The data existed as a collection of daily forecast files on a THREDDS server, but a procedure needed to be developed so that the data could be accessed as an aggregation. Once the aggregation was assembled, the aggregation catalog simply needed to be added to the registry to become available to the system. However, it was then discovered that ncISO was not properly calculating the geospatial extent for unstructured grids, and we worked with the developers to release a new version of ncISO. This dataset also revealed a longer-term development need: to have the THREDDS Data Server Forecast Model Run Collection Aggregations work with unstructured grids. We instead needed to invoke a custom script to generate non-overlapping sequences of data that could be joined as a supported aggregation.

The power of working within a standardized framework is that when we fix problems to enable a specific workflow to function, we fix a whole class of problems that strengthens the entire infrastructure. In this way, we continually build a stronger and more capable infrastructure, while embracing Gray's Law of "go from working to working"[11].

III. BEYOND TECHNOLOGY: WHO'S INVOLVED AND HOW DO THEY WORK TOGETHER?

A distributed data network like DMAC is much more than a collection of technology. It involves people and organziations with missions addressing broad national or even global concerns down to very specific local concerns. This network requires coordination, shared vision, strategy and funding to be effective today and to evolve.

Association to DMAC or participation in DMAC can be described in several ways (Fig. 4). The IOOS Program Office in NOAA has the responsibility to prescribe the standards and protocols that are the foundation of the DMAC system, and to fund a significant portion of the implementation through the IOOS Regional Associations [3]. Through these funding arrangements, the Regional Associations engage in three broad activities: (1) to operate coastal ocean observing systems, (2) integrate data from existing observing activities at the academic, state, local, or tribal level, and (3) to engage regional stakeholders to better understand their needs for ocean information. Data management and publication are central to all three of these activities and the regional associations therefore serve a foundational role for the IOOS enterprise.

Other Federal agencies provide significant input to DMAC, through the coordinating efforts of the Interagency Ocean Observation Committee (http://www.iooc.us/), an oversight committee for the IOOS enterprise. This input includes hosting data access services or contributing to the development of software.

[8] https://github.com/geopython/OWSLib
[9] https://github.com/ioos/pyoos
[10] https://github.com/ulmo-dev/ulmo/
[11] https://github.com/SciTools/iris
[12] https://github.com/xray/xray
[13] https://github.com/pydata/pandas

Finally, within the outer ring, are ocean observations communities with common goals and an interest in leveraging efforts rather than replicating them. Herein lies the most

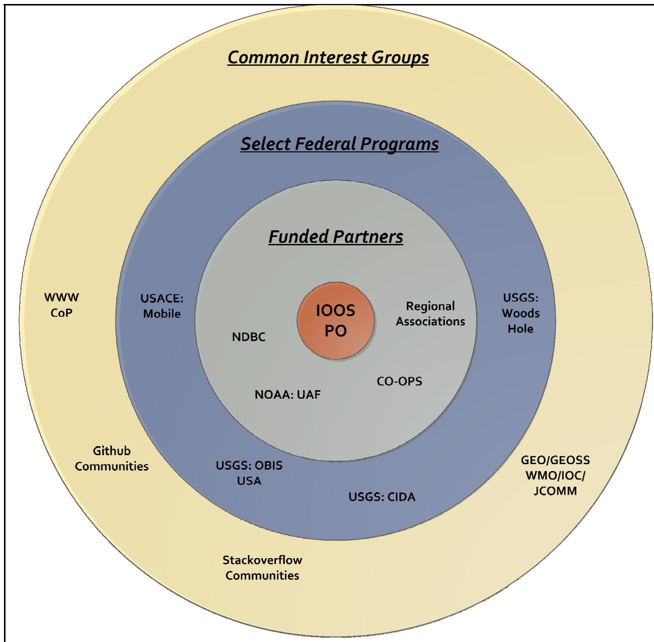

Fig. 3. Schematic depicting the level of collaboration the IOOS Program Offices engages in with other DMAC contributors.

potential for efficient development and for developing a robust system with enterprise-wide support.

With such broad participation and inadequate financial resources to address all requirements in a centralized activity, DMAC has to adopt strategies to increase coordination and cooperation and to leverage the efforts of similar programs.

Most important in creating a network of distributed data sets accessible via standard services is to ensure that the content of the network is rich, reliable, and desired. A data network must be populated by relevant data sets. The assumption of open data has not historically been part of the funding model for oceanographic research. In the past, principal investigators were afforded time (two years or longer) to work with data sets and to publish results in the scientific literature. While there were important reasons for this, we are currently in the midst of a policy shift based in the recognition that federally funded data is a public asset, whether it is collected by the governemt or through extramural grants[12], [13]. Focused and collaborative development of the technology needed to implement these policies will benefit the federal agencies who struggle to comply with these new requirements.

Another strategy for increasing collaboration is a commitment to developing software in the open. Open source software development affords the opportunity to contribute to an existing project rather than begin a new one. Furthermore, open source principles are becoming best practices in software engineering and have been shown to lead to higher quality software. The examples described here rely on over 100 software packages, all of which are available freely and openly. More important than free software is a framework to contribute to its evolution. For example, the IOOS PO has funded a small number of software projects directly (approx. 10) and more than half of those are based on pre-existing open source projects. For example, the i52N SOS service is based on the existing OGC reference implementation SOS service created by 52north.org. All IOOS supported modifications are contributed back to the 52North.org trunk so that the larger community can benefit. Other IOOS supported packages, such as a data format compliance-checker, have received contributions from ocean observing colleagues in Australia, Canada, and Europe. Open source development is a foundational principle for all IOOS software projects.

Many users of DMAC software and data will not contribute to enterprise software development activities, but will create scripts and software aimed at analyses of some type. These efforts can also benefit from open source software principles such as the use of distributed version control tools like Github. Openly developing anaysis routines provides for scientific reproducibility and serves as an excellent educational resource for the next generation of scientists and application developers [6].

IV. Conclusions and Challenges for the Future

The IOOS DMAC system is a functioning and effective data distribution network enabling discovery, access, and use of thousands of data sets. DMAC is complementary to other networks in the US and abroad because of the use of open standards and software. It is well positioned to respond to policy changes toward open data and to technology changes due to the commitment to open development.

The implementation is not without flaws as detected through the system integration test. As we increasingly rely on networked interactions we must respond with even more standardization of tools and practices. Patterns that are immediately recognized by the human eye are not detectable by computers so interfaces need to be rigidly defined and implemented. Standardization is not something that is only implemented at the web service level, rather efforts focused on interoperability through standardization need to be implemented at the earliest stages of planning an observational campaign.

Finally, making use of the vast amount of ocean information available to us requires a new skill set for existing and future scientists and engineers. The emerging field of data science is poised to provide new insghts into how the ocean works by combining data in ways that were impossible recently. The ocean observations community must continue to evolve our data management practices and data distribution infrastructure to be responsive to new user needs.

References

[1] J. Lubchenco, "Ocean observations: Essential for good stewardship," *Mar. Technol. Soc. J.*, vol. 44, no. 6, pp. 6–9, 2010.

[2] R. Rayner, "The U.S. Integrated Ocean Observing System in a Global Context," *Mar. Technol. Soc. J.*, vol. 44, no. 6, pp. 26–31, Nov. 2010.

[3] E. Howlett, D. P. Snowden, R. P. Signell, K. R. Knee, and D. Wilson, "Data Management Update for the Integrated Ocean Observing System (IOOS®)," *Mar. Technol. Soc. J.*, Submitted 2014.

[4] US IOOS Program Office, “Guidance for Implementation of the Integrated Ocean Observing System (IOOS®) Data Management and Communications (DMAC) Subsystem.” 12-Mar-2010.

[5] R. Signell and D. Snowden, “Advances in a Distributed Approach for Ocean Model Data Interoperability,” *J. Mar. Sci. Eng.*, vol. 2, no. 1, pp. 194–208, Mar. 2014.

[6] Vembu Subramanian, Debra Hernandez, Rich Signell, and Filipe Fernandes, “Reproducible Workflows for Coastal Ocean Model Skill Assessment,” in *MTS/IEEE*, Washington, DC, 2015. (in press)

[7] The Open Geospatial Consortium, “Open Geospatial Consortium Reference Model,” Informative/Educational OGC 08-062r7, Dec. 2011.

[8] S. C. Hankin, J. D. Blower, T. Carval, K. S. Casey, C. Donlon, O. Lauret, T. Loubrieu, A. Srinivasan, J. Trinanes, and O. Godoy, “NetCDF-CF-OPeNDAP: Standards for ocean data interoperability and object lessons for community data standards processes,” in *Oceanobs 2009, Venice Convention Centre, 21-25 septembre 2009, Venise*, 2010.

[9] J. D. Blower, A. L. Gemmell, G. H. Griffiths, K. Haines, A. Santokhee, and X. Yang, “A Web Map Service implementation for the visualization of multidimensional gridded environmental data,” *Environ. Model. Softw.*, vol. 47, pp. 218–224, Sep. 2013.

[10] V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, *Machine Learning and Data Mining Approaches to Climate Science: Proceedings of the 4th International Workshop on Climate Informatics*. Springer, 2015.

[11] A. J. G. Hey, S. Tansley, and K. M. Tolle, *The Fourth Paradigm□: Data-Intensive Scientific Discovery*. Redmond, Wash.: Microsoft Research, 2009.

[12] *Open Government Directive*. https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf

[13] *Public Access to Research Results*. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf