OCEAN DATA DISSEMINATION: NEW CHALLENGES FOR DATA INTEGRATION

J.D. Blower[(1)], S.C. Hankin[(2)], R. Keeley[(3)], S. Pouliquen[(4)], J. de la Beaujardière[(5)], E. Vanden Berghe[(6)], G. Reed[(7)], F. Blanc[(8)], M.C. Gregg[(9)], J. Fredericks[(10)], D. Snowden[(11)]

[(1)] Environmental Systems Science Centre, University of Reading, RG6 6AL, United Kingdom, Email: j.d.blower@reading.ac.uk
[(2)] NOAA/PMEL, 7600 Sand Point Way NE, Seattle, WA, USA, Email: Steven.C.Hankin@noaa.gov
[(3)] Integrated Science Data Management, Dept. of Fisheries and Oceans, Ottawa, Canada, K1A 0E6, Email: Robert.Keeley@dfo-mpo.gc.ca
[(4)] IFREMER, BP 70, 29280 Plouzané, France: Email: sylvie.pouliquen@ifremer.fr
[(5)] National Oceanic and Atmospheric Administration, 1100 Wayne Ave #1225, Silver Spring MD 20910 USA, Email: Jeff.deLaBeaujardiere@noaa.gov
[(6)] Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, New Jersey 08901-8521, USA, Email: evberghe@iobis.org
[(7)] Australian Ocean Data Centre Joint Facility, Level 2, Building 89, Garden Island, Potts Point, New South Wales 2011, Email: greg@metoc.gov.au
[(8)] Space Oceanography Division, CLS, Ramonville-St-Agne, France, Email: fblanc@cls.fr
[(9)] U.S. National Oceanographic Data Center, 1315 East-West Highway, # 4820, Silver Spring, MD 20910, Email: margarita.gregg@noaa.gov
[(10)] Woods Hole Oceanographic Institution, Woods Hole, MA 02540, Email: jfredericks@whoi.edu
[(11)] NOAA/Climate Program Office, Silver Spring MD, USA, Email: Derrick.Snowden@noaa.gov

ABSTRACT

In the decade since OceanObs`99, great advances have been made in the field of ocean data dissemination. The use of Internet technologies has transformed the landscape: users can now find, evaluate and access data rapidly and securely using only a web browser. This paper describes the current state of the art in dissemination methods for ocean data, focussing particularly on ocean observations from *in situ* and remote sensing platforms. We discuss current efforts being made to improve the consistency of delivered data and to increase the potential for automated integration of diverse datasets. An important recent development is the adoption of open standards from the Geographic Information Systems community; we discuss the current impact of these new technologies and their future potential. We conclude that new approaches will indeed be necessary to exchange data more effectively and forge links between communities, but these approaches must be evaluated critically through practical tests, and existing ocean data exchange technologies must be used to their best advantage. Investment in key technology components, cross-community pilot projects and the enhancement of end-user software tools will be required in order to assess and demonstrate the value of any new technology.

1 INTRODUCTION

Ocean observation data are important for meeting many societal needs, including the understanding and mitigation of climate change, improving the safety and efficiency of maritime operations, reducing public health risks, managing and protecting natural resources [1]. It is increasingly recognized that individual datasets are rarely used alone: scientists typically wish to intercompare, assimilate and cross-validate many different types of data; furthermore, ocean data are becoming increasingly valuable to policymakers and decision-makers in a number of fields, including climate change, marine safety, and the management of the environment and its resources [2]. Therefore, it is a very important challenge to develop the procedures and technology to disseminate ocean data in a consistent and accessible fashion, and in a manner that supports the integration of ocean data with other data sources. There are also large social challenges, in particular to encourage and fund scientists and data providers to support the effort required to prepare and release data. Such needs are increasingly recognized by legislation, such as INSPIRE (Infrastructure for Spatial Information in the European Community) directive, discussed further in Sect. 4.2 below.

Data are currently disseminated in a variety of ways. The Global Telecommunications System (GTS) broadcasts data in real time primarily to major facilities engaged in numerical ocean and atmosphere modelling and data assembly. Nearly all of the marine observations of physical parameters – in over one million locations each month – made by international operational programs are found on the GTS. There are however major gaps with respect to non-physical parameters, observations from research programmes and coastal observations made at regional levels. A variety of Earth Observation data, including meteorological and ocean datasets, are broadcast on GEONETCast (Group on Earth Observations - global network of satellite based data dissemination systems) [3], a satellite-based

dissemination system. Although paper records are not commonly used today, there is still a rich legacy of historical data that exist only in paper form and must be converted to electronic form before they can be disseminated. There are a number of efforts to digitize such data, including the Global Oceanographic Data Archaeology and Rescue (GODAR) project of the Intergovernmental Oceanographic Commission (IOC), which has been very successful in this regard [4].

The continuous growth in capacity and availability of the Internet has led to a similar increase in its use for disseminating ocean data. Through the use of Internet and Web technologies, ocean data can be made available to a wide variety of users in a highly flexible manner. Internet-based dissemination systems, from which users "pull" data, provide extra capabilities above those provided by broadcast or "push" systems, such as the ability to monitor usage patterns and customize data feeds on-the-fly for particular users. These technologies will be the focus of this paper, although we acknowledge that other methods such as satellite distribution remain valuable, particularly for applications in which high-bandwidth Internet access is not readily available, and where high reliability, high data throughput and timeliness are important.

Historically, each project or observing platform has maintained its own data management and dissemination system. This has led to a proliferation of online data sources, meaning that users frequently experience difficulties in finding the data they require, or in finding the authoritative copy of a dataset that appears on the Internet many times. Recent trends have focused on global data assembly centres such as Argo (Array for Real-time Geostrophic Oceanography), drifters, and OceanSITES (OCEAN Sustained Interdisciplinary Time series Environment observation System) and on assembling data from various platforms into consolidated collections in support of specific goals, such as the World Ocean Circulation Experiment (WOCE) Data Assembly Centres, the Coriolis database of *in situ* observations [5], the GHRSST (Group for High-Resolution Sea Surface Temperature) project for sea surface temperature [6], the AVISO (Archiving, Validation and Interpretation of Satellite Oceanographic data) [7] project for altimetry data and the U.S. Observing System Monitoring Center [8,9,10] Work still needs to be done in helping the many users of ocean data to discover, evaluate and access the data they need [11].

It is widely agreed that users of ocean data require information to be presented in a consistent manner, irrespective of the source of the data; Sect. 2 of this paper discusses current efforts towards increasing the consistency of data across projects and platforms. Recently, much attention has been given to the adoption of open geospatial standards for the discovery, encoding, visualization and dissemination of data; Sect. 3 examines their strengths and weaknesses for the ocean community. Sect. 4 describes some recent efforts that employ new technologies to create large "virtual databases" of observations of the ocean and other elements of the earth system. Finally, the paper concludes (Sect. 5) by drawing out the main current challenges in data dissemination and making recommendations for future activities.

This paper is the third of a series of four related review papers from OceanObs '09 on the topic of data and information management: the first paper [9] describes the history and motivation behind the development of modern ocean data systems; the second paper [10] describes the current state of the art in data management, covering all processes from data acquisition to archival; this paper focuses on the most recent advances in data dissemination systems, covering some of the new operational and experimental systems; and the fourth paper [12] looks forward to developments in data systems over the next decade. We focus here on ocean observation data from *in situ* instruments and remote sensing platforms; a discussion of related technologies in ocean modelling can be found in [13].

2 INCREASING THE CONSISTENCY OF DIGITAL OCEAN DATA

A key goal for any data dissemination is to deliver data in a consistent and timely manner. It is not sustainable for each platform or project to continue to deliver data in a bespoke manner; instead, world-wide agreement on data standards is necessary. This section will discuss the main issues in increasing the consistency of digital ocean data, summarizing the progress that has been made so far and highlighting the known outstanding concerns.

2.1 File formats

A wide variety of file formats are currently in use for exchanging ocean data. On the GTS, ocean observation data are transmitted in a number of formats (including TESAC, BATHY, TRACKOB, WAVEOB, BUOY, SHIP and BUFR) and a still-wider variety of file formats is used in Internet communications. The ocean modelling community has largely settled on NetCDF (Network Common Data Form) [13] for gridded data sets. Earth Observation data, particularly orbital data, are typically encoded in HDF (Hierarchical Data Format) [14]. Historically, biological data collected concerning species observations have been highly heterogeneous, with each cruise recording data in a different way, commonly in Excel spreadsheets or Access databases. This situation is now improving rapidly, as we shall discuss later in this section.

The wide variety of data types (grids, point observations, swaths, biological species and distributions), together with the variety of tools in current use by different communities, makes it unlikely that the ocean community will be able to settle upon a single digital file format in the foreseeable future. ASCII formats, whilst easy to interpret by humans, suffer from disadvantages of inefficiency for large data volumes and a difficulty in enforcing formatting rules. Binary formats (e.g. NetCDF) can record more complex relations between data elements but are not human-readable without special tools. In practice, the capabilities of end-user tools are key drivers behind the choice of file format, an issue we shall return to later in this paper. In recognition of this, the SeaDataNet project [15] disseminates *in situ* observation data primarily in an ASCII format (compatible with the Ocean Data View tool [16]), and satellite and model output data primarily in NetCDF format (compatible with many tools that process gridded data), reflecting the typical practices of the relevant user communities.

The NetCDF format is supported by tools and libraries for many programming environments and operating systems commonly used by the oceanographic community. At heart, it is a simply a platform-independent data format for multi-dimensional arrays and their attributes. On its own, NetCDF is not sufficient to ensure interoperable data exchange: therefore, a volunteer community has developed a set of metadata conventions, called Climate and Forecast (CF, [17]). CF defines how to encode geospatial and temporal referencing information and provides a list of standard names for an ever-growing list of measured quantities (around 2000 as of November 2009, currently growing at a rate of approximately 1000 per year; Alison Pamment personal communication). The CF standard names have been mapped to other vocabularies by building ontologies [18], allowing users to more easily compare datasets that adhere to different conventions [19, 20]. The current version of CF applies only to NetCDFv3, also known as the "classic" NetCDF format. The inclusion of such metadata allows for automated interpretation of the data by tools, which brings many benefits. For example, many tools - such as the THREDDS data server (Thematic Real-time Environmental Distributed Data Services) - support *aggregation* of data files, i.e. making a set of separate data files appear as one large logical file, greatly simplifying tasks such as subsetting [21].

Although the classic CF-NetCDF format is mainly suited to array-based (gridded) data, many groups have created conventions for *in situ* observations. These include OceanSITES (for timeseries and profile data), Argo (for hydrographic profiles) and GOSUD (Global Oceanographic Surface Underway. Data) (for underway ship measurements): see [22] for more details. The recently-developed NetCDF4 contains structures that make it more suitable for encoding observational data efficiently. The CF conventions are expected eventually to adapt to these developments.

Recently, much attention has been given to the use of XML (eXtensible Markup Language) and associated technologies as a potential "lingua franca" for data exchange in many fields. XML only defines an overall structure for data: different communities must create different "dialects" or *schemas* (this situation is somewhat analogous to the relationship between NetCDF and CF). Google Earth's KML format [23] is an example of a fairly simple XML schema; the Geography Markup Language (GML) is an extremely complex one. Although XML is based upon plain text and can be written and read directly by humans, the encodings are designed for interpretation by machine, just as a web browser is needed to interpret web pages. XML plays a key role in many aspects of modern data dissemination, particularly for the encoding of structured metadata that have previously been difficult to capture in ASCII or binary file formats. Currently, there has been limited use of XML to exchange actual measurement values, for many reasons, including a lack of agreement on the most appropriate schemas to use, the lack of support in end-user tools, and, to a lesser extent, concerns over encoding efficiency. But there is a growing interdisciplinary community utilizing the features of encodings, which are specifically designed to describe sensors and associated observations and measurements, which are used by many earth observing disciplines (see Sect. 3 below).

One successful example of the use of XML to exchange ocean data can be found in the Ocean Biogeographic Information System (OBIS [24]). Recognizing that there was much valuable information on species distributions "locked up" in various formats in the archives of individual projects, the biogeographic community, led by the Global Biodiversity Information Facility (GBIF) and OBIS, defined a simple XML-based file format, known as "DarwinCore", and encouraged scientists to submit data in this common format into the OBIS database. At the present time (July 2009), OBIS holds 18.5 million records from 633 distinct datasets; on average, around 100,000 records are downloaded per day. Figure 1 shows an example of integrating this information.

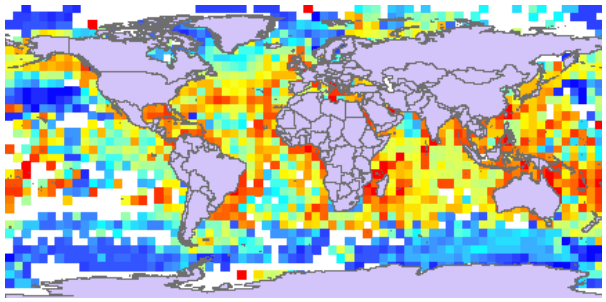

Figure 1: Map of Hurlbert's index, ES(50) – the expected number of distinct species in a random sample of 50 distribution records, calculated per squares of 5x5 degrees. Red indicates high species richness, blue low. White areas are where there are less than 50 distribution records in a square [24].

2.2 Metadata

The word "metadata" describes any supporting information that is required to interpret data correctly. The term encompasses spatio-temporal referencing, instrument descriptions, descriptions of biological or physical phenomena, quality information, contextual information (such as the contact details of the data collector) and many more [11 and 25]. The Climate and Forecast conventions encode only certain kinds of metadata, as described above.

Many kinds of metadata problems are well-documented: metadata are often incomplete or missing, or provided in a form that is not easily interpreted by machines. These factors severely limit the extent to which data can be re-used, particularly across communities. Therefore there are many current ongoing efforts in the field of metadata standardization.

QARTOD (Quality Assurance for Real-Time Ocean Data) [28] is an ongoing grass-roots initiative addressing issues of quality assurance (QA) and quality control (QC) for the growing community of ocean observing systems, specifically addressing real-time dissemination of data. As evolving technologies enable access to interdisciplinary real-time data, a common understanding of best practices in QA/QC must be developed to provide a solid foundation for global ocean observing systems. This requires international, interdisciplinary discussion and consensus on required and recommended tests and best practices in sensor preparation, calibration, and data processing.

In the biology community, it is recognized that greater consistency is needed in nomenclature of species. Some terms encompass many species (e.g. 'Turbinaria' is both a valid name of a genus of coral and of green algae) and some species have multiple names (e.g. the sperm whale, *Physeter macrocephalus*, has been known under many different names, including *Physeter andersonii*, *P. australasiensis*, *P. australis*, *Catodon colneti*, *Physeterus sulcatus* and *Tursio vulgaris*). To make matters worse, often records are documented using English or other-language 'vernacular' names, where no standardisation or nomenclatorial rules exist. Several initiatives are compiling lists of taxonomic names, to serve as tools to standardise names used in data management. For names of marine organisms, the World Register of Marine Species (WoRMS [26]) offers the most complete list. Projects such as Species 2000, the Integrated Taxonomic Information System and WoRMS are standardizing nomenclature and these look to be converging to a single solution in the Catalogue of Life [27].

2.3 Catalogues

The amount of ocean data now available through the Internet continues to grow rapidly, and new technologies are being developed to ensure that users can find the data they need. If data are available through a website then standard search engines such as Google can provide a basic means of discovering information. However, each website may be designed differently, with data files organized in very different ways. A means of creating more intelligent and focused catalogues of ocean data is clearly required.

Catalogues may be centralized or distributed. In a centralized model, data providers submit descriptions of their data holdings to a central database. In a distributed system, data providers create their own catalogues but provide standard web service interfaces (see below) to their catalogues; users can then search for data in many catalogues simultaneously. Currently the centralized model is more widely used as it places fewer demands on individual data providers; however, very large data systems may in future require a distributed model to ensure scalability. SeaDataNet [15] is a major European effort to provide a central point of access to ocean data from 35 countries and 49 European institutes and data centres. SeaDataNet adopts an approach employing a centralized catalogue but decentralized data holdings. By contrast, the MyOcean project [2] will employ a decentralized catalogue, in which each data provider exposes its own catalogue, each of which is harvested on a regular basis into a central location to enable efficient searching of all datasets simultaneously.

The overall aim of these efforts is to simplify the process of discovering data by aggregating catalogues, reducing the number of websites that a user must search for data and providing more accurate and consistent results. A key current challenge is the use of "semantic web" technologies to further help users to locate the data they need, by building computer systems that understand relationships between concepts (such as "an *Argo* is a type of *observing platform*").

2.4 Access control

Currently, ocean data are frequently not made publicly available, for reasons are often social or political rather than technical in nature: commonly the science team in question may not want to allow others to analyse their data before they have had a chance to do so themselves; this frequently leads to data never being released. Other reasons include a lack of qualified and available staff, fear of data misuse and legal reasons such as the protection of natural resources (see [29] for a discussion of these and many more issues).

The value of open access to data, permitting free and unrestricted access to data, is increasingly widely recognized by science organizations. The Intergovernmental Oceanographic Commission's *Oceanographic Data Exchange Policy* (Resolution IOC-XXII-6) recognizes that the importance of sharing data freely and in a timely fashion is essential for the effective use of ocean observations. Under this policy, IOC member states agree to provide data from IOC programmes freely, and are encouraged to do the same for non-IOC programmes. The World Meteorological Organization (WMO) has adopted a policy (WMO Resolution 40) for the international exchange of meteorological and related data and products which provides for the free and unrestricted sharing of data [30].

Open access to data is harder to implement in some communities than others. In the ocean biology community, each individual observation (e.g. an identification of a member of a species) is the result of painstaking work by an individual who must be certified to one or more international standards as being competent in species identification. It is perhaps natural that scientists are sometimes unwilling to share their data widely until they have received due credit through a publication, which may take a year or more to appear in press.

Recognizing these social barriers to data-sharing, there is a growing community investigating the possibility of publishing and citing data sources in the scholarly literature. Such opportunities would allow scientists, data managers and associated personnel to receive recognition for the hard work of data collection and management. The Scientific Committee on Oceanic Research (SCOR) and the International Oceanographic Data and Information Exchange (IODE) programme of IOC have initiated a project to promote the ability to publish datasets as unique objects and their citation by other researchers as an incentive to improve data flow [31].

When social barriers to data-sharing can be overcome, there still remain technical issues. If data are shared by individual projects and programmes through many different websites, users may be forced to maintain login details for each individual site. Therefore, many integrating efforts employ a "single sign-on" technology, which requires the user only to remember one username/password combination (in some cases, digital certificates are used instead of passwords). Some infrastructures use a central authentication database; when users log in to a data provider, they are redirected to this service, which authenticates them against a central database before redirecting them back to the data provider. SeaDataNet employs a different approach, in which each user has a single SeaDataNet login. When a user places an order for data, the system notifies each individual data provider that is relevant to the order; when every provider has "cleared" the user, the data are extracted and prepared and the user is notified that the data are ready for download.

2.5 Connecting tools directly to data services

With most existing Internet-based data dissemination systems, users navigate a website before downloading one or more data files to their local systems. They then require tools that can interpret these data files. An alternative approach that is gaining increasing use is the ability to plug end-user tools directly into the data systems. This approach has been used for many years in the ocean modelling community, in which the OPeNDAP [32] data transfer protocol is used to allow tools to download custom-selected subsets of large model datasets [13 and 21]. Many marine meteorology datasets, including NCEP reanalyses, are also disseminated using OPeNDAP. The user does not need to be concerned about file formats, metadata conventions or the way in which the data provider has organized data within the files: all of these concerns are handled automatically by the tool. The programming code required to read data from OPeNDAP servers is virtually identical to that required to read the same data from local NetCDF files, and so this method is well-supported by tools that are popular in the ocean modelling community.

In the biological community the Distributed Generic Information Retrieval (DiGIR) protocol has been used to link a series of independent databases into a single virtual collection. DiGIR tools provide a uniform interface to manage search requests to a series of "DiGIR Providers". The search requests and the response are both XML-based, as is the definition of the information available from the provider. DiGIR was originally developed at the Biodiversity Research Center of Kansas University, and is now a public open source project.

OPeNDAP and DiGIR are examples of web services[1].

[1] Here we are using a wide definition of Web Services:

Whereas a webs*ite* is a means for a computer to present information to a human, a web *service* is a means for computers to exchange information with each other. Many more types of web service are currently under experimentation, for transmitting data, providing access to remote data-processing power, for searching remote catalogues, for accessing dictionary definitions of terms (and semantically mapping between them) and many other tasks. See Sect. 3 below for discussion of a very important category of modern web services.

The use of web services for disseminating ocean observations is currently not widespread in typical operational systems, for a number of reasons. As we have discussed, ocean observation data are diverse and the process of settling upon standard file formats, frameworks and metadata conventions is ongoing; for information to be exchanged between computers, consistency is absolutely essential. There are many possible new technologies for serving observation data through web services (see Sect. 3 below), none of which has yet been extensively tested in production environments. In addition, the technical effort required to set up a robust and reliable system based upon web services is currently much greater than the effort required to distribute data through websites and FTP servers.

Much work is being done to fix these issues, and so we can confidently predict that the next decade will see a large increase in the dissemination of ocean data through web services. SeaDataNet2 plans to employ web services in addition to its current data dissemination systems. The increase in the use of CF-NetCDF for encoding observation data raises the possibility of disseminating these data through OPeNDAP, and indeed tools already exist for doing so [33]. This will allow ocean observation data to be read directly into OPeNDAP-aware tools. There are many large international initiatives for integrating data across communities (see Sect. 4 below), all of which are based around web services.

The potential gains of these developments are great. It will be much easier to automate processes in which diverse datasets are brought together for analysis. Different communities can plug "feeds" of ocean data into their own tools, reducing the need for users to learn how to use multiple interfaces to data. Custom user interfaces can be developed for specific projects or problems, reusing the back-end data feeds. A key scientific goal of this new technology is easier data integration: Figure 2 shows how ocean observations and forecast models can be brought together in a visual web environment.

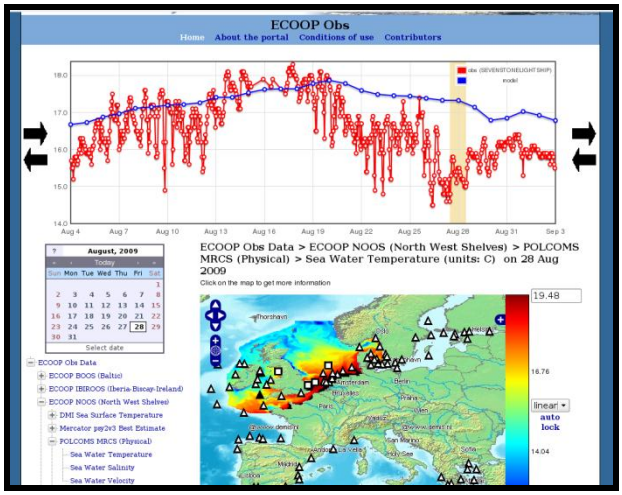

Figure 2: Intercomparison between model forecasts and in situ observations using an interactive web portal developed for the European coastal-ocean project ECOOP. Model data are overlain with observations from FerryBoxes and moorings; clicking on an observation location brings up the timeseries of measurements from that instrument, together with the model's prediction at the same location.

3 OCEAN DATA AND OPENGIS

Oceanographers rely on many sources of data from other scientific communities; conversely, other communities are users of ocean data. Climate science and earth system science are key examples of interdisciplinary science efforts in which data need to be shared between communities. Decision-makers also need access to many different data sources, for example to assess the risks posed by algal blooms [34]. There is therefore a clear need for interoperability of data between diverse users.

The common thread linking most environmentally-relevant datasets is that they are geospatial in nature, i.e. they are referenced in some way to the globe. Geographic Information Systems (GIS) have been developed over many decades to handle geospatial data and allow different sources of data to be brought together. Historically, there have been many commercial vendors of GIS tools and software, and their solutions have limited mutual compatibility. However, we are now in an era of "Open GIS", in which many of the GIS vendors, their users and other interested parties are collaborating to define open standards for data exchange, with the intention of greatly increasing interoperability between geospatial data. The Open Geospatial Consortium (OGC [35]) is the body that leads and coordinates the development of these standards. Many standards are now implemented – albeit usually incompletely – in commercial and open

---

we do not restrict our definition to SOAP (Simple Object Access Protocol) -style Web Services.

source software tools such as GeoServer, MapServer, ArcGIS and uDig.

OGC standards encompass many of the issues we have discussed above. Much effort has focused on the definition of standard Web Services for sharing data and currently there are many such standards with varying degrees of maturity. The boundaries between the capabilities of the standards are not sharp, but as a rough guide to the main standards, the Web Coverage Service (WCS) is designed for sharing raster data (such as an ocean model output), the Web Feature Service (WFS) is designed for point data (such as *in situ* observations) and the Web Map Service (WMS) defines a standard means to share accurately-geo referenced, custom-generated map imagery. The Catalogue Services for the Web (CS-W) standard defines interfaces to data catalogues. Another suite of standards (Sensor Web Enablement or SWE) defines a Sensor Observation Service (SOS) for sharing sensor data, a Sensor Planning Service for controlling or tasking sensors, Sensor Model Language (SensorML, based on XML) for describing sensors and processing (e.g. QC/QA tests [36]) surrounding the observation, and the Observations and Measurement schema for encoding the observation content and values. SWE not only encodes information about the sensor and sensor system, it also provides a framework for defining and/or describing processing components.

Sometimes different standards can be applied to the same problem. The OGC Ocean Science Interoperability Experiment (Oceans IE) – Phase 1 (2007-08) investigated the use of WFS and SOS for representing and exchanging point data records from fixed *in situ* marine platforms. Oceans IE concluded that the use of OGC SOS was better suited than WFS for this purpose and developed best practices documentation and tools for using OGC SOS [37]. Oceans IE Phase II will build on Phase I and will continue the improvement and use of OGC specifications in the marine community

The OGC approach has the potential to bring many benefits to the ocean community in addition to the possibility of increasing interoperability with other communities. Commercial and open-source software can be used "off-the-shelf" to achieve many tasks, although it is very difficult for software developers to keep pace with the variety of standards and the rapid pace of their development. OGC standards may provide solutions to fill gaps in the ocean community's current capabilities: for example, there is no widely-used method for exposing ocean catalogues through Web

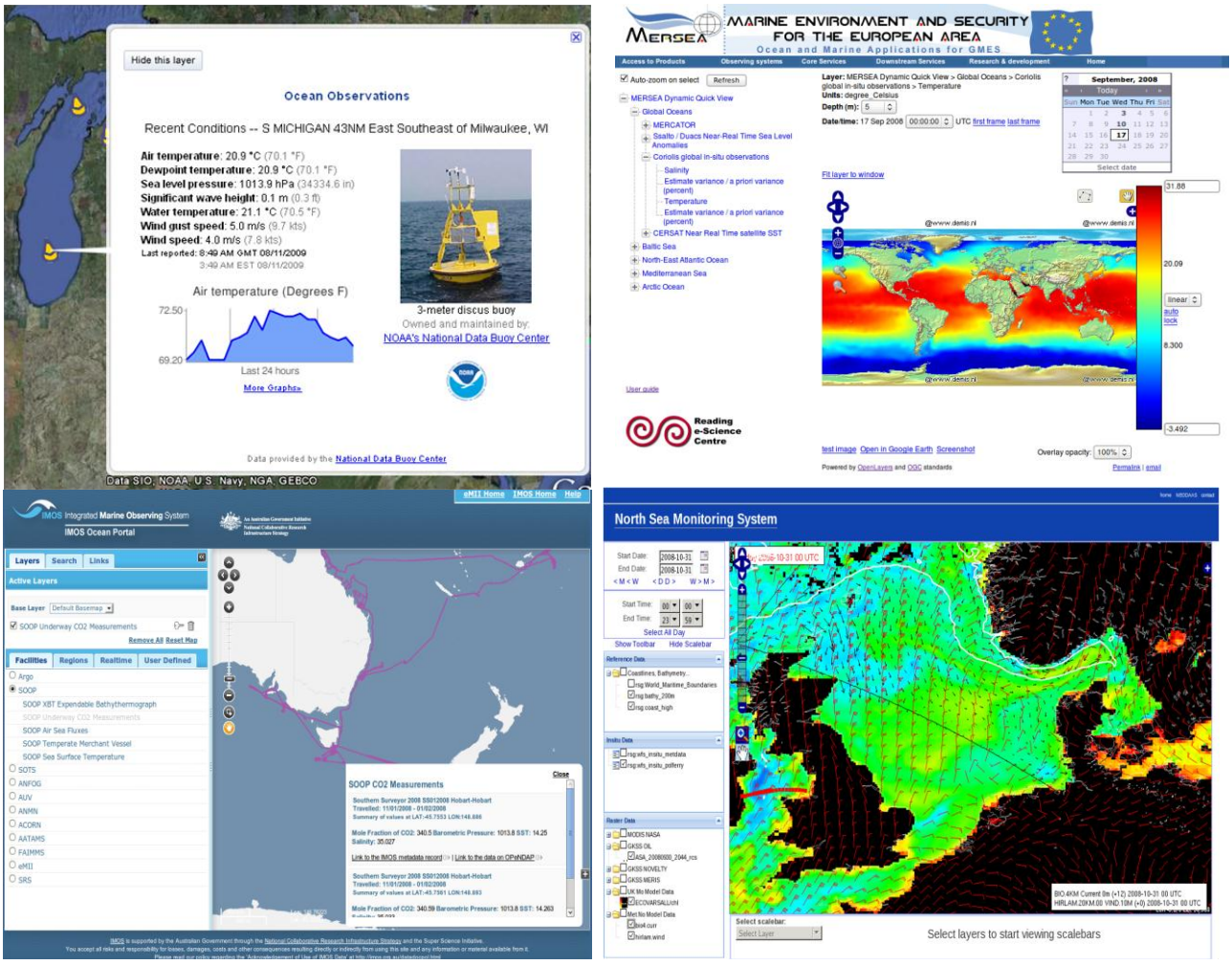

Figure 3: Examples of visualization of ocean data using GIS technologies. Top left: Buoy measurement in Google Earth (accessible through the Google Ocean layers). Top right: European MERSEA (Marine Environment and Security for the European Area) project's Dynamic Quick View system, which uses OGC Web Map Services to provide an interactive display of data, in this case sea surface temperature interpolated from observations in the Coriolis database. Bottom left: Australian integrated data portal IMOS (Integrated Marine Observing System). Bottom right: Integrating various sources of data in the InterRisk portal (Plymouth Marine Laboratory).

Services; also, there is no widely-adopted interoperable means of sharing *in situ* data. The effort of addressing commonly-shared scientific issues, such as the communication of quality information, can be shared among many communities in the OGC forums. OGC technologies have been proven to be particularly useful for visual applications in which data of various types can be plotted on maps and overlain: some examples of these are given in Figure 3.

Much experimentation is therefore currently underway to apply these standards to the sharing of ocean data. Many questions, however, still remain open, including:

- OGC standards are, by design, highly general in nature in order to support the needs of many communities and to provide the necessary flexibility to accommodate future needs. As a result, even within a given community, different groups have developed different implementations of the standards that are equally valid but are not mutually compatible. The usual approach to this is to develop "profiles" of each standard for particular communities. How can these profiles be developed so that they are useful to the ocean community, without breaking compatibility with other communities?

- What is the best way forward when OGC capabilities overlap with existing, proven capabilities? For example, OPeNDAP is a proven technology for sharing data via web services and CF-NetCDF is a proven technology for encoding data. Should the ocean community invest in adapting to new technologies, or should existing technologies be promoted in other communities? Or can the best aspects of both technology streams be combined? (Version 4 of the THREDDS Data Server [38] now includes support for the OGC WCS and WMS protocols, providing a familiar means for many providers of ocean data to serve data in the new OGC formats.)

- What is the best means of encoding data? Existing encoding formats for ocean data are not OGC standards. In fact, the OGC does not define a specific encoding format for any data types: the Geography Markup Language provides reusable XML components, but does not specify how any particular community should assemble these components to define a concrete encoding format. Each community must create a GML "application schema"; the Climate Science Modelling Language (CSML [39]) is an example that will be highly relevant for the ocean community, but is still under development and has not yet been widely tested. The fundamental importance of agreeing upon file formats (Sect. 2.1 above) means that this represents a major technological obstacle. There are ongoing efforts to standardize CF-NetCDF through the OGC process, which will promote its use in other communities; this represents an alternative way of achieving OGC compatibility without re-engineering.

- For which user communities are OGC technologies most appropriate? The answer to this is not likely to be clear-cut. We currently have very little real evidence on which to base decisions; more research into the needs and capabilities of potential user communities is required. It is likely that the needs of a decision-maker (who probably needs access to high-level data products in standard GIS formats) will be very different from those of a scientific user (who will probably require access to the original measurements and all supporting information). It is unlikely that any single technology will be able to adequately serve such a wide spectrum of users.

These questions are utterly fundamental to the problem of building interoperable IT infrastructures for ocean data, and must be addressed seriously. Doing so will require a concerted effort on the part of many stakeholders, including data providers, standards groups and the various users of ocean data. Whilst technological innovation and experimentation is to be encouraged, we strongly recommend that the costs and benefits of any new approach be assessed carefully, to ensure that future visions of interoperability can be realized. More discussion on this can be found in [12].

4 DATA INTEGRATION EFFORTS

Efforts to aggregate data from different sources to simplify user access have been going on for many years (see Sect. 1 above). Past efforts have mainly focused on transferring data to assembly centres: in this section we shall briefly describe how the new technologies described above are enabling distributed sources of data to be brought together in large "virtual" databases.

4.1 Ocean data integration

The Ocean Data Portal (ODP [30]) aims to provide seamless access to collections and inventories of *in situ* marine data from the network of National Oceanographic Data Centres of the International Oceanographic Data and Information Exchange (IODE) programme. Data can be discovered, visualized (using OGC WMS and WFS technologies) and downloaded in NetCDF, ASCII or XML formats (See Figure 4. Through a pilot project with JCOMM (Joint WMO-IOC Technical Commission on Oceanography and Marine Meteorology) and WMO, the ODP will be interoperable with the WMO Information System and hence with the Global Earth Observation System of Systems (GEOSS, see below).

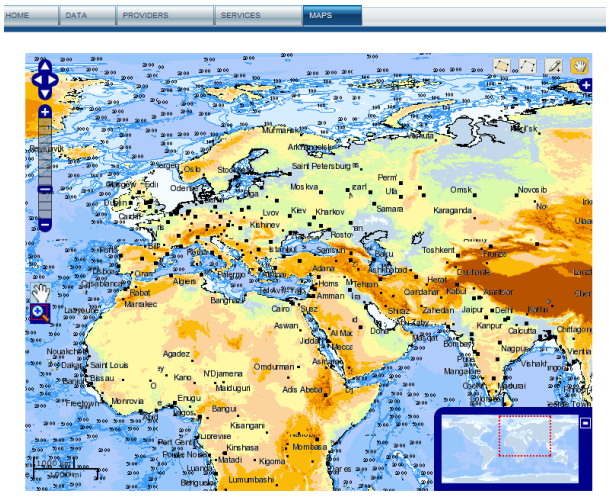

Figure 4: Locations of in situ observations accessible via the Ocean Data Portal, served through an OGC Web Map Service

The purpose of NOAA's Observing System Monitoring Center (OSMC [8], Figure 5) is to provide a tool to assist managers and scientists with monitoring the is to provide a tool to assist managers and scientists with monitoring the performance of the global *in situ* ocean observing system, identifying problems in real-time, and evaluating the adequacy of the observations in support of ocean/climate state estimation, forecasting and research. OSMC chose the ingestion of observations from the GTS as the starting point to capture the majority of observations at relatively low effort; however, its ongoing effort is to link to delayed-mode sources of data. This is necessary in order to properly monitor the formation of the climate data record. However the effort to do so is made very considerably greater by the current lack of uniform standards for data dissemination. The increasing availability of delayed-mode observations through THREDDS and OPeNDAP using CF conventions promises to greatly lower these barriers in the future.

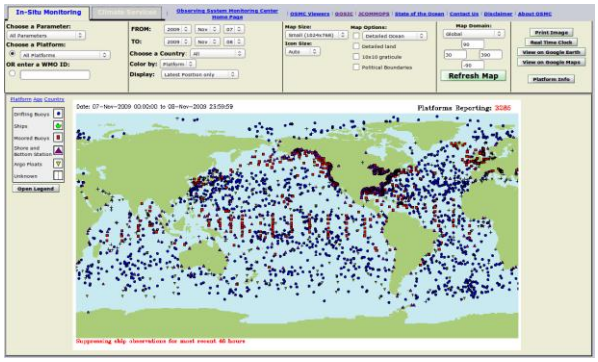

Figure 5: NOAA Observing System Monitoring Center, showing all observations from November 2007 to November 2008.

4.2 Cross-community data integration

The problem of disseminating data effectively and interoperably is common across many communities connected with earth science and earth observation. In recognition of this, many overarching bodies have been formed with the aim of coordinating data management and dissemination activities across the gamut of earth sciences.

The Group on Earth Observations (GEO [40]) is coordinating international efforts to build a Global Earth Observation System of Systems (GEOSS), which will provide access to data from a highly diverse array of instruments and monitoring systems, with the intention of supporting scientists, decision-makers, policymakers and resource managers. The GEO community has developed a number of prototype web portals to allow users to discover and visualize Earth Observation data from the GEOSS registries and to access training resources and documentation. GEOSS activities are aimed at satisfying the needs of nine societal benefit areas, all of which are in some way related to the oceans: disasters, health, energy, climate, agriculture, ecosystems, biodiversity, water and weather. The Global Ocean Observing System (GOOS [41]) is the ocean community's contribution to GEOSS and the Integrated Ocean Observing System (IOOS) and EuroGOOS are, respectively, the US and European contributions to GOOS.

The "GEOSS Common Infrastructure" includes a catalogue of datasets, a catalogue of services that provide access to those datasets, and some demonstration web portals. Phase 2 of the GEOSS Architecture Implementation Pilot, completed in September 2009, added support for custom-built web portals (to serve particular communities), data processing and new dissemination mechanisms. Data in the GEOSS can now be accessed using many of the technologies discussed in this paper, including OGC Web Services, OPeNDAP, FTP and satellite-based communications (GEONETCast [3]).

As a "system of systems", GEOSS aggregates data provided by compatible systems that exist in different communities. The WMO Information System (WIS) is an example of such a system. WIS builds upon the Global Telecommunications System, adding the capability to access data via the Internet through data and catalogue web services. Data records are aggregated in Global Information System Centres (GISCs), which allow data to be discovered from comprehensive metadata catalogues. WIS technology standards are aligned with those of GEOSS. The WIS will be the core component of GEOSS for the societal benefit areas of weather, water, climate and disasters.

In Europe, the INSPIRE directive, now in force, aims to

enable a pan-European infrastructure for sharing a wide range of environmentally-relevant data. INSPIRE defines a distributed architecture, based on web services, with a strong focus on OGC standards (e.g. CS-W for data discovery and WMS for data visualization). Services will eventually be available for data discovery, view, download and transformation (e.g. coordinate transformation), although currently most attention has focussed on the discovery and view services.

These large initiatives will be strong drivers in technology development in the field of data dissemination in the coming decade. It is very clear that architectures based upon web services will become increasingly prominent.

5 DISCUSSION

This paper has highlighted some of the latest developments in the field of ocean data dissemination, in the wider context of global-scale initiatives to increase the sharing of environmental data. The aim of these efforts is to increase the value of our investments in ocean observing systems by ensuring that the data they produce are used as effectively as possible, within the scientific community and by other users such as decision-makers.

Many scientific, technical and social challenges remain to be overcome. Based on the above discussions, we propose some high-priority activities to drive forward our goals of making ocean data more easily accessible to – and exploitable by – the various communities of interested users.

5.1 Reduce the social barriers to data sharing.

Technical solutions are powerless if scientists and data providers are unwilling to release their data in a timely manner to the wider community. Much of this reluctance stems from the current lack of means for research groups to receive due credit for data collection and preparation. We have noted (Sect. 2.4 above) that initiatives are underway to allow scientific data to be published in a way such that datasets can be cited as a publication. Another important method will be for data centres to monitor closely the usage of data. Detailed statistics on data usage patterns and other feedbacks from users would provide many benefits. Such information would act as citations to help judge the worth of a dataset; it would help data centres to provide a better service and ensure their own sustainability; and it would help technical choices to be made concerning the true efficacy of different data dissemination technologies. However, a danger of a distributed approach to data dissemination is that this usage information can be lost. This information must be preserved.

5.2 Help data providers to standardize on a minimal set of file formats.

As we have seen, the diversity of ocean data and users means that it is difficult to recommend a single standard, however it is achievable to settle on a small set of standard file formats for ocean data. Doing so would help data providers, tool developers and end users alike. The principal barrier to arriving at this short list may be another social barrier: i.e. the absence of a community structure that is sufficiently cohesive to enable agreement upon questions such as these. This issue of increasing community cohesion is discussed in more detail in [12].

5.3 Establish a forum for discussing Ocean-GIS integration issues.

Large drivers such as GEOSS and INSPIRE are motivating much engineering work connected with the use of OpenGIS standards and tools for disseminating ocean data. The results of such experiments are often subtly nuanced, leading to potential confusion: for example, two systems that are compatible with the same standard may not be interoperable in any practical sense [12]. There is a need for clear and open communication about such issues. The answers to very fundamental questions remain unclear, including whether key OGC technologies should be considered as candidates for data interchange within the met-ocean community, or between the met-ocean community and the GIS user community (see Sect. 3 above). The recently-formed Met-Ocean Domain Working Group of the OGC may be a suitable forum for these discussions. In addition, the ocean data management community could commission regular reports that summarize the current strengths and weaknesses of these technologies in a manner that is accessible to programme managers, the scientific community and other non-technical stakeholders.

5.4 Set up cross-community pilot projects.

The need to exchange data between communities is often rightly cited as a justification for pursuing radically new technological paths (see Sect. 3 above). However, it is not currently clear that the ocean data community understands in sufficient detail the technical needs of other communities such as decision-makers. What kinds of information do they really need? Which tools do they use? Such information can best be elicited through joint projects involving the relevant stakeholders. These pilots would be extremely useful in guiding a path through the technological labyrinth that currently faces us.

5.5 Invest in linking data systems with end-user tools.

Much attention is currently being placed on new software architectures for data curation and dissemination. Service-oriented architectures (see Sect. 2.5 above) are prominent, as they have the potential to

allow data to be disseminated in extremely flexible ways. However, this potential will not be realized until scientific tools are able to connect to these services reliably and in a way that frees the user from the need to understand technical details. Currently developments in tools are not systematically funded.

6 ACKNOWLEDGEMENTS

The authors are very grateful to Mike Botts, Roy Lowry, Bob Simons, Chris Little and Etienne Charpentier for helpful comments and contributions.

7 REFERENCES

1. Hankin, S. and the DMAC Steering Committee, (2005). Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems: I. Interoperable Data Discovery, Access, and Archive, Ocean.US, Arlington, VA 304 pp. http://dmac.ocean.us/dmacPlan.html

2. Bahurel, P. & Co-Authors (2010). "Ocean Monitoring and Forecasting Core Services, the European MyOcean Example" in these proceedings (Vol. 1), doi:10.5270/OceanObs09.pp.02.

3. GEONetCast website, accessed 28 February 2010 at http://www.earthobservations.org/geonetcast.shtml

4. Levitus, S., S. Sato, C. Maillard, N. Mikhailov, P. Caldwell, and H. Dooley (2005). Building Ocean Profile-Plankton Databases for Climate and Ecosystem Research, NOAA Technical Report NESDIS 117, U.S. Government Printing Office, Wash., D.C., 29 pp

5. Coriolis project website, accessed 28 February 2010 at http://www.coriolis.eu.org/

6. 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.

7. Aviso project website, accessed 28 February 2010 at http://www.aviso.oceanobs.com/

8. Observing System Monitoring Center website, accessed 28 February 2010 at http://www.osmc.noaa.gov/

9. 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

10. Keeley, R., Pazos, M. and Bradshaw, B., (2010). "Data Management System for Surface Drifters" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.47.

11. Conkright Gregg, M., Newlin, M., LeDuc, S., Keeley, R. and D'Adamo, N., (2010)."Ocean and Coastal Data Stewardship" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.18.

12. 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.

13. Blower, J.D., Blanc, F., Clancy, R., Cornillon, P., Donlon, C., Hacker, P., Haines, K., Hankin, S.C., Loubrieu, T., Pouliquen, S., Price, M., Pugh, T., Srinavasan, A. (2009). Serving GODAE data and products to the ocean community. *Oceanography* 22(3) 70-79.

14. Hierarchical Data Format project website, accessed 28 February 2010 at http://www.hdfgroup.org/

15. SeaDataNet project website, accessed 28 February 2010 at http://www.seadatanet.org/

16. Ocean Data View project website, accessed 28 February 2010 at http://odv.awi.de/

17. Climate and Forecast metadata convention project website, accessed 28 February 2010 at http://www.cfconventions.org

18. Noy, N.F. and McGuinness, D.L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

19. Marine Metadata Interoperability Semantic Framework website, accessed 28 February 2010 at http://marinemetadata.org/semanticframework

20. Latham, S.E., Cramer, R., Grant, M., Kershaw, P., Lawrence, B.N., Lowry, R., Lowe, D., O'Neill, K., Miller, P., Pascoe, S., Pritchard, M., Snaith, H., Woolf, A. (2009). The NERC DataGrid Services. *Phil. Trans. Roy. Soc. A* 367(1890), 1015-19

21. Signell, R.P., Carniel, S., Chiggiato, J., Janekovic, I., Pullen, J., Sherwood, C.R. (2008). Collaboration tools and techniques for large model datasets. *Journal of Marine Systems* 69 154 – 161

22. 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.

23. KML Documentation website, accessed 28 February 2010 at http://code.google.com/apis/kml/documentation/

24. Vanden Berghe, E., Halpin, P., Lang da Silveira, F., Stocks, K. and Grassle, F., (2010). "Integrating Biological Data into Ocean Observing Systems: The Future Role of OBIS" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.91.

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

26. World Register of Marine Species website, accessed 28 February 2010 at http://www.marinespecies.org

27. Catalogue of Life project website, accessed 28 February

2010 at http://www.catalogueoflife.org

28. Burnett, W. & Co-Authors (2010). "Quality Assurance of Real-Time Ocean Data: Evolving Infrastructure and Increasing Data Management to Monitor the World's Environment" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.12.

29. Froese, R., Lloris, D., Opitz, S. (2003). The need to make scientific data publicly available – concerns and possible solutions. *ACP-EU Fisheries Research Report (14)* 267-271

30. Reed, G., Keeley, R., Belov, S. and Mikhailov, N., (2010). "Ocean Data Portal: A Standards Approach to Data Access and Dissemination" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.72.

31. SCOR/IODE Workshop on Data Publishing, (IOC Workshop Report No. 207), accessed 28 February 2010 at http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=2457

32. Open-source Project for a Network Data Access Protocol (OPeNDAP) project website, accessed 28 February 2010 at http://opendap.org

33. DChart/Dapper project website, accessed 28 February 2010 at http://www.dchart.org/

34. Stumpf, R., Fleming-Lehtinen, V. and Granéli, E., (2010). "Integration of Data for Nowcasting of Harmful Algal Blooms" in these proceedings (Vol. 1), doi:10.5270/OceanObs09.pp.36.

35. Open Geospatial Consortium website, accessed 28 February 2010 at http://www.opengeospatial.org

36. Fredericks, J. & Co-Authors (2010). "Integrating Quality Assurance and Quality Control into Open GeoSpatial Consortion Sensor Web Enablement" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.31

37. Bermudez, L. (ed.) (2008). OGC Ocean Science Interoperability Experiment Phase 1 Report. OGC document 08-124.

38. THREDDS Data Server project website, accessed 28 February 2010 at http://www.unidata.ucar.edu/projects/THREDDS/tech/TDS.html

39. Woolf, A., Lawrence, B., Lowry, R., Kleese van Dam, K., Cramer, R., Gutierrez, M., Kondapalli, S., Latham, S.E., Lowe, D., O'Neill, K., Stephens, A. (2006). Data integration with the Climate Science Modelling Language. *Adv. Geosci.*, 8, 83–90

40. Group on Earth Observations website, accessed 28 February 2010 at http://www.earthobservations.org/

41. Global Ocean Observing System website, accessed 28 February 2010 at http://www.ioc-goos.org