OCEAN DATA: COLLECTORS TO ARCHIVES

Robert Keeley[(1)], S. Woodruff[(2)], S. Pouliquen[(3)], M. Conkright-Gregg[(4)], G. Reed[(5)]

(1) Integrated Science Data Management, Department of Fisheries and Oceans, 200 Kent St., Ottawa, Canada, K1A 0E6, Email: Robert.Keeley@dfo-mpo.gc.ca
(2) NOAA/ESRL (PSD3), 325 Broadway; Boulder, CO 80305 US, Email: Scott.D.Woodruff@noaa.gov
(3) Oceanography from Space and In-Situ Departement, IFREMER, BP70, 29280 Plouzané, France, Email: sylvie.pouliquen@ifremer.fr
(4) Director, U.S. National Oceanographic Data Center, 1315 East-West Highway, # 4820, Silver Spring, MD, 20910, Email: Margarita.gregg@noaa.gov
(5) Australian Ocean Data Centre Joint Facility, Wylde Street, Potts Point, NSW Australia, Email: greg@metoc.gov.au

ABSTRACT

Preserving data for future use involves the acquisition, processing, preservation, quality assurance or control (QC) or assurance, archival preservation (including accompanying platform and instrumental metadata), and dissemination by data centres serving national and international users [1]. Many organizations, both national and international have a primary objective to assemble, preserve into the future and disseminate measurements from the ocean and the overlying atmosphere. The few that are mentioned illustrate the common objectives and issues of all. This paper is a companion to other papers [2, 3 and 4] that cover the scope of ocean data management. This one deals with the processes of moving data from acquisition to archives and leaves the other aspects of data management to the other authors (other than touching briefly on issues related to timely delivery).

1. AN ORGANIZATIONAL OVERVIEW

Present data systems have been set up at different times and this has influenced their development and evolution. For example, the Voluntary Observing Ship (VOS) Scheme [5] can trace its origins back to the 1853 Brussels Maritime Conference; elements of modern physical and biological oceanography arose from the *Challenger* Expedition (1872-76); and elements of traditional alphanumeric codes used in the current World Meteorological Organization (WMO) Global Telecommunication System (GTS), can be traced to a 1913 meeting in Rome of the International Meteorological Committee (forerunner of WMO). Some of the more recently established systems have taken lessons from previous experience, together with technological advances, to improve how data are managed. Various community white papers (CWPs) of this conference have described in more or less detail some of these different systems that manage data coming from operational and scientific programmes. They have also commented on desirable attributes that the global data system should have.

At national governmental levels, many countries have established centres to manage the data generated by observing programmes within their country. It is unusual for a national data centre to actually manage all of the various kinds of data that are collected. Normally, they act as a focal point for the nation and coordinate activities in their country.

Other centres have been established in nations at academic institutions or other non-governmental organizations. These often start out as project data assembly centres, and over time some of them gain status as a primary source on the national or international scene for the kind of data they manage. These centres make an important contribution and must be recognized and included in the considerations of a future data system.

In the oceanographic arena, the international coordination of national data centre activities is managed by the International Oceanographic Data and Information Exchange (IODE) Committee of the Intergovernmental Oceanographic Commission (IOC). The IODE was established in 1960 and has grown from a few national centres to 64 as of today. This group is focused on management of all kinds of ocean data and provides a mechanism for national centres to discuss and share solutions to their data management challenges. For more information consult http://www.iode.org/

The Joint Technical Commission on Oceanography and Marine Meteorology (JCOMM) was formed in 2001 as a joint activity between IOC and WMO. It united organizations in both WMO and IOC that were focused on marine meteorological and physical oceanographic data with an emphasis on real-time data delivery (readers interested in the observing systems that are considered to be a part of JCOMM can visit http://www.jcomm.info/and click on the "Observations Programme Area"). One of the three major components of JCOMM is the Data Management Programme Area (DMPA). Though real-time data continue to dominate the focus of JCOMM, delayed mode (including

historical) data processing is part of the considerations of the DMPA. This Programme Area coordinates data management activities of JCOMM but also has developed close ties with IODE and other organizations to minimize overlaps in function.

The International Council for Science's (ICSU, http://www.icsu.org/index.php) World Data Centres (WDCs) are non-governmental partners in the global data management system. They were originally established to ensure long-term security of solar, geophysical, and environmental (including marine meteorological and oceanographic) data for future use. But since the 1960's when these were established, there have been large changes in computer capabilities, in communications facilities, and the establishment of national and international organizations, such as IODE and JCOMM, that contribute to managing data. The WDC system has recently undergone a review and is now in the midst of change.

There are other organizations that are multi-national in membership and with defined geographical areas of interest that also carry out data management activities. Two examples of these are ICES (International Council for the Exploration of the Seas, http://www.ices.dk/indexnofla.asp) and PICES (North Pacific Marine Science Organization, http://www.pices.int/). The nations that support these organizations are obligated to provide them with data and information. Sometimes this is through the national data centre, but not exclusively so. Both of these organizations have a strong emphasis on biology and fish stocks which is sometimes outside of the domain of national data centres, and is not a strong component at this time in IODE or JCOMM. However, with the recent inclusion of the Ocean Biogeographic Information System (OBIS, http://www.iobis.org/) into IODE, biology will play a more important role in that organization.

As an important additional consideration, it is also necessary to take into account the satellite agencies. These are providing a host of different observations that are highly complementary in sampling strategy to in-situ measurements of the ocean and atmosphere. While JCOMM, for example, has initiated a satellite rapporteur cross-cutting DMPA and the two other Programme Areas (i.e. that for Observations and a third for Services and Forecasting Systems), satellite data systems have largely operated separately from the in-situ data systems leaving it up to users to figure out how to use both types of data in analyses or products.

2. DATA SYSTEM CHALLENGES

The challenges to a global data system are many. To list a few:

- The types of data collected are extremely varied ranging for example from classical measurements of seawater (surface and subsurface) temperature [6,7] to acoustics [8] to biology [9,10] to satellite systems [11]. Data systems need to manage these diverse forms and to permit easy inter-disciplinary studies.
- The many kinds of ocean data being collected are not and likely cannot be handled by a single data centre. The volume, diversity and expertise needed to manage all of these data into the future will not be found in a single institution.
- There are continual developments of new instrumentation. Many of the CWPs document improvements to existing instruments and advances in new instrument capabilities. This increases the importance of recording information (metadata) about the platforms and instrumentation so that systematic differences between different data collection methods can be known and compensated for in future data analyses (e.g. [6])
- New instrumentation often allows the electronic capture of greater data volumes. This is happening already, but the development of cabled observatories will significantly increase data handling demands.
- The number of agencies and individuals engaged in data collection is large and growing. It is difficult for a national data centre to be aware of all of the potential contributors. Many of these groups place data on web sites for public consumption (e.g. [9]) and these data sometimes reside nowhere else.
- There is an increasing desire for access to data in real-time. This demands cooperation from data collectors, efficient assembly and dissemination systems, and automated procedures to identify suspect or duplicated data.
- Calibration is important to extract the greatest value from measurements, but calibration information often arrives much later (if at all) than the data, particularly if the data are distributed in real-time. Different versions of the same original data challenges "duplicate" detection algorithms and can confound drawing reliable conclusions from data sets.

It has been suggested that the details of the data system should not be a concern for either data contributors or data users. As long as data can go into and come out of the system in a reliable, easy and timely way, there is no need to burden contributors and users with the details.

However, it is important for data providers and users to have some understanding of what goes on inside the data system to operate such a service. This document provides some details of the challenges, some of the solutions being used, and suggests targets and paths to the future.

The remainder of this document will be organized under the following broad categories:

- Assembly of the data from sources
- Quality control, duplicates checking and other procedures to verify the data
- Placement of the data into accessible archives where their existence and availability are advertised
- Timely delivery

At the end, there will be a summary that describes how many of the pieces that are needed are actually available in some form, and some suggestions of actions to pursue.

3. DATA ASSEMBLY

Reference [12] argues that oceanographic sensors need to be web enabled so that data can be distributed more efficiently rather than having to go through a few data centres. This is certainly technically possible, though the organizational mechanisms to do so are not yet in place. However, in order for data to be easily used, every data set would need to adhere to a set of community wide standards or the user would be bombarded with different attributes from the different sensors and would need to reconcile them all. The same comment applies to data that are placed on the many web sites from projects. Data centres provide precisely the service of consolidating data into a consistent format, as well as ensuring the preservation of the data beyond the life of a project, a researcher, or a technology. By providing this service the data centre removes a data preparation burden that many researchers are not prepared to support. In addition however, data centres need to make available authoritative metadata that can lead the user to recognize the provenance of the data source.

There are a number of reasons for data assembly centres (DACs) to bring data together. These centres often assemble the data for a specific need (e.g. Coriolis for operational oceanography, http://www.coriolis.eu.org/) and therefore carry out additional processing to check the consistency of the data set at basin scales that may allow them to detect additional anomalies [13]. Moreover producing an integrated product, such as an ARGO (Array for Real-time Geostrophic Oceanography) climatology, allows the detection of other problems.

The assembly process:

- brings uniformity of data structure or format to make the job of a user easier
- brings uniformity to data quality assessment to assist in making use of the data
- eases the work for users looking for data since there are fewer places to search
- imparts a standardization of terminology which helps users to deal with data from different sources
- adds value to the data by facilitating the merger of data from different sources
- provides documentation of what has been done to the data in the course of transforming data received from the collector
- eases the data management burden for collectors
- ensures the preservation of the data into the future

Data assembly is a different process when satellite data are considered. In this case data assembly is relatively straightforward in the sense that a single satellite produces data from one or more sensors that are downloaded to a single processing and archiving facility (though the processing and handling may be quite complicated). However those data are typically extremely voluminous (compared to in-situ data), which raises a host of different technical and archival issues. There are only a few archive facilities for satellite systems, and combining the data from these different systems is a challenge. However, the CEOS (Committee on Earth Observation Satellites, http://www.ceos.org/) community has inserted a degree of standardization to satellite data handling that helps explain the data when they are delivered to a user.

In-situ data assembly (or assembly of data from remote sensing devices such as shore based radars) is a process that brings together measurements, sometimes by the same or similar instruments or platforms, but that have a variety of collectors. This is a more complex problem than for satellite data because of the variety of data sources, the variety of instruments and how the instruments operate and are used to gather measurements.

There are three types of data assembly processes. The most straightforward process happens at centres that are set up by a project to manage the data coming from participants. The second typically happens at national data centres whose responsibilities are to assemble all (usually more limited than all types) of the data that are generated by public and private sector data collectors from their country. The third kind, and usually the most difficult, happens when assembling the data after (sometimes long after) a project or the data collection was completed. This is sometimes called data rescue.

Project DACs (these may be operated by national data centres as well) have many advantages.

- Usually, they know all of the collectors of the data in the project and all are eager to contribute
- The data content and structures are usually agreed to at the start and all sources normally provide data to these specifications
- Because data content and structures are known early in the project, the assembly centres can build processing software that is tuned to project operations
- If data are to be delivered in both real-time and delayed mode, time frames are set and roles defined for data handling
- The appropriate metadata that describe the instruments, methods, etc. are all available during the project and so relatively easy to capture (as long as this is designed into the project)
- Quality control functions are agreed upon between the data providers and the DAC
- Projects can deal with highly complex data because of the established partnerships between the data producer and the DAC.

More recently some projects, but certainly not all, build in funding that addresses transferring of the resulting data to institutionalized archive centres. Typically, funding is at a 5-10% level of the total project, varying by the novelty of the data and information required to be managed. But it is rare that project proposals are funded at the requested level, and unfortunately it is common for data management funding to be disproportionately reduced. Data management is as important a component as others and needs an equivalent level of attention.

Good examples of developing project DACs include the planned US Ocean Observatories [14], and a similar cabled observatory on the west coast of Canada. It is expected that both of these projects will have long term funding and will develop data management activities that deal with the volumes and diversity of the data produced as well as the ability to adjust sampling schemes in reaction to recently received observations. This latter ability relies on fast processing of data streams and quick access to the resulting data.

NOAA (US National Oceanic and Atmospheric Administration) provides an example of one country's approach to assemble all data collected by national organizations. It operates discipline focused data centres including the National Climatic Data Center (NCDC) and the National Oceanographic Data Center (NODC). Because of overlaps in data responsibilities (for example of buoy data spanning meteorological and oceanographic disciplines) in some cases they share the same data holdings. Such national centres rarely have the advantages itemized above for DACs because only rarely are they consulted on long term archive needs at the start of the collection planning. A data centre may know well who are the researchers or potential data collectors, but they may be unaware of ongoing or developing projects and so they may not be aware of all of the data collected. If not aware of all of the data, they will certainly miss some in the assembly process.

In cases where data rescue is necessary, additional factors are that the storage medium may be subject to deterioration (such as paper), or the data are in electronic form but on obsolete media or in proprietary formats. This is complicated by inadequate documentation of the measurements or data formats, missing information on instrumentation, calibrations, methods, etc., and no one may be available to answer these questions. This argues again for adequately resourcing data management at an early enough stage to alleviate the need for future data rescue activities.

Many collectors, even though not in a partnering arrangement with a data centre, still contribute data to archives. Some do so very quickly, often with the arrangement that the data centre provides wider distribution such as over the GTS. Others hold on to the data they have collected until they have verified the measurements, analyzed them, written a scientific paper, or fulfilled other requirements. It is still common that researchers never turn over data to a data centre even if they know of the appropriate repository.

Currently, there are few incentives for a researcher to provide data to archives. Some journals in other disciplines are requiring submissions to include the data on which results are reported. Other initiatives are underway to provide a way for authors of scientific papers to cite data that they used, and thereby provide credit to the researcher for releasing the data to the public. At present and to any significant extent none of these are in place in marine meteorology or oceanography, but both merit consideration and development.

Often a collector considers that data sharing is met by placing data on a web site, though even today some data only sit on a personal computer's hard drive. Both of these are significant challenges to a data system striving to gather and maintain the data into the future. These challenges also affect the scientific community, particularly those studying climate which requires an historical record of ocean variables that is as complete as possible. It should be noted that [15] provides a suggestion for data release times that if adhered to would be helpful in ensuring data get to archives and to users. Also, [16] discusses the importance of immediate and widespread access to real-time data, and the often

remarked-upon success of the Argo (Global array of free-drifting profiling floats) program is strongly linked to its open data sharing policies.

Different views about sharing have existed historically (but with continuing influence) between oceanography and marine meteorology. The former is heavily influenced by the more proprietary research view, and the latter, having longstanding operational connections e.g. with numerical weather (and increasingly ocean) forecasting, currently is more reliant on near-real-time data delivery and also with the sharing of atmospheric data and services formalized by WMO Resolution 40, 1995.

It is often hard to get data from commercial companies or academia. In the former case, there is the concern that releasing data that they paid to collect may allow a competitor to gain an advantage. The company may simply use what they collected as best they can, and then throw the data away. Considering the paucity of data from the oceans, this course of action is a great shame. Even if the company thinks the data might be valuable in the future and so spends some resources to preserve them, they may spend more in maintenance than they might lose by releasing the data to an assembly centre for safe keeping.

Academia is more focused on project outcomes and has concerns that often severely restrict distribution of the data they collect. In their case, the concern is intellectual property rights and first right to publish; i.e. if the data are released too soon, the originating researchers will not have completed their analyses and someone else will publish their work. Unless there are formal arrangements between data centres and academic institutions, it is seldom that data are provided to the data centres much less released to others. Data centres can work with the academic sector to provide tools that will facilitate a uniform standard of metadata entry, provide a data management and stewardship service, and publish data to an on-line repository. This will facilitate discovery and allow access to the vast amounts of marine data generated every year by universities thereby maximizing the collective investment made by researchers. There are movements, such as in the US, to ensure that data collected using federal grants are submitted to a national data centre.

Sustained international projects, joined to national data centres but also with strong cooperative research linkages, have demonstrated (e.g. [17]) the values of knitting together a fairly wide spectrum of ocean data in the ICOADS (International Comprehensive Ocean-Atmosphere Data Set**,** http://icoads.noaa.gov/**)** aiming primarily at creating more easily usable data and products for the research community. Moreover, JCOMM is seeking further modernization and streamlining of some existing facets of delayed mode data handling, linked with ICOADS, to achieve further benefits [18].

To have access to data collections and to gain trust, national data centres must become engaged in data collection projects. On the international scale, a consortium of data centres can undertake the data management activities. On a national scale, data centres need to make the connections to their data collectors. This can be particularly challenging if the number of national data collection activities are large. In that case, there may not be enough resources at a data centre to be involved with every activity, but the centre does need to connect with the leaders of these activities to ensure data flow smoothly to archives. A data centre must be active in working with data collectors and as early in the data collection process as possible. A data centre must provide a service of value to the data collector.

Most data centres operate as more than places where people can deposit data sets and get copies of others. Data centres accept data in a variety of structures but reformat the data to their own internal structure. Care is needed in doing this. First the internal structure must be rich enough to accommodate as much of the incoming information as possible, otherwise information is lost. In most cases, the measurements are easily accommodated, but the additional information (describing instruments, methods, calibrations, etc.) is harder to manage. Even if the data structure is rich enough, great care must be exercised to ensure the fidelity of the transcription. However, without this translation (e.g., from historical into modern scientific units) and reformatting of incoming data, subsequent users of the data would need to contend with myriad data structures and content.

As a support to these important user-oriented requirements, stronger archival policies also need to be considered (both nationally and internationally) regarding the preservation of "original" (as received) data to guard against inadvertent errors or omissions. This issue is discussed in the context of WMO's standard Binary Universal Form for the Representation of meteorological data (BUFR) format in [18] and referred to in the JCOMM Data Management Plan (http://www.jcomm.info/index.php?option=com_oe&task=viewDocumentRecord&docID=2877). The problems discussed are sometimes only discovered years (possibly even decades) after the fact and without the original data having been preserved, there is no opportunity to correct mistakes.

But this translation and reformatting exposes some additional weaknesses in current policies and technologies. First, data arrive typically without a "dictionary of terms". This is not so acute a problem for some measurements since "TEMP", or "temperature" or "T" are mostly self explanatory as long as the context is known. For example, a meteorologist would think of properties of the air, while oceanographers first think of the water. Where data are exchanged between these disciplines, it is important to explain what temperature is being talked about. The problem is more acute with the additional information that can accompany data, information that is necessary for proper interpretation. For example, a data set may contain a text field whose content is "SBE-19". Data centres need to know that this is the type of instrument that was used for the temperature measurement. There are many other examples, but they simply illustrate the need to use common vocabularies for describing data. There are the beginnings of these in the Climate and Forecast (CF, http://www.unidata.ucar.edu/software/netcdf/conventions.html) conventions, in the parameter code list of the Global Temperature-Salinity Profile Program (GTSPP, http://www.nodc.noaa.gov/GTSPP/gtspp-home.html), in the Global Change Master Directory (GCMD, http://gcmd.nasa.gov/) keywords, in the work done in projects such as SeaDataNet (Pan-European Infrastructure for Ocean & Marine Data Management) in Europe, the Marine Metadata Initiative in the US and in the International Organization for Standardization (ISO), etc. What there is not is community agreement on which vocabularies should be used followed by coordinated community implementation. This is an activity where satellite and in-situ data systems should strive to come together.

Common vocabularies are really just one example of a number of areas where standards are required to enable the interoperability of data. There is no agreement on what metadata should always accompany data. For example, should it be mandatory that the name of the instrument (perhaps other details) and the type and exposure of the platform used to make the measurement always be with the measurement? Is there a need to preserve information about who collected the data and why? What details of methods of collection or processing should be mandatory (this is especially important for chemical or biological measurements)? Should data uncertainties be recorded along with the measured values? One solution is to preserve together with the data as much ancillary information as possible [19] as suggested more specifically by [16]. To be successful, data collectors will need to cooperate with data managers to ensure the necessary information is transferred with the data.

The ocean data management community has been wrestling with these issues, including more precisely defining different levels of metadata, for some time, but as yet has not built community agreement. Experience in trying to interpret historical data with fragmentary metadata demonstrates that we must do a better job preserving this information for future use. Reference [20] differentiates between three major levels of metadata – collection or discovery metadata, provenance or lineage metadata, and platform or sensor metadata. Discovery metadata are needed to allow users to locate where data are held and provide ways of describing how those data can be accessed. By imposing a common method of describing the data, it is possible for people from different backgrounds to find data of interest. The International Standard ISO 19115 is becoming widely used to describe discovery metadata and the use of this standard (or a profile) by the marine community addresses the problem of describing data sets and defining their contents in a way that is widely understood. Through a project started by JCOMM and IODE (http://www.oceandatastandards.org), it is hoped that progress can be made to define and implement standards in data management practices [21]. Metadata are also important for satellite instruments and so this is a common issue. When comparing data collected from in-situ and satellite sensors, it is very important to have these metadata both on the instrument characteristics, but also methods and procedures of measurements.

SensorML (http://www.opengeospatial.org/standards/sensorml) is tackling part of this problem. If the technology is widely adopted, this has the potential to organize content (some impact on vocabularies) and also structure of the data streams. This should improve efficiency of data handling.

Some actions that will improve data assembly:

a. More data centres must take part in the design and planning of data collection activities. In order to have a seat at the table, data centres must contribute services that are valued by the activity.
b. Data centres must actively pursue the adoption of standards in as many aspects of handling data as possible. There is a start in the SeaDataNet project and through JCOMM and IODE, but much greater attention must be devoted to defining and implementing standards.
c. Data centres need to improve their internal data structures to be able to accommodate and preserve the variety of data and metadata that are being gathered in scientific and other data collection activities.

5. DATA PROCESSING PROCEDURES

The two topics of highest priority have to be the control of data quality and of duplications of data. Duplicates are not the same as replicate observations; a replicate is an observation made at the same time as another but with an independent instrument. Replicates are important checks on the functioning of instruments and are valuable to be preserved in archives.

Duplicates are troublesome when they are not recognized as different versions (see Fig. 1) of the same original data. Identifying duplicates be they exact or near exact is important but difficult today. Duplicates as seen at a data centre can arise in many ways. One version received may be the original data after conversion from instrument to physical units. But other versions of original data may also arrive. Several observing systems reporting over the GTS also provide data in delayed mode, such as VOS [5,18], and moored [22] and drifting buoys [23]. Dealing with real-time and delayed mode duplicates can be very complex, but necessary to obtain higher quality and more complete data. With scrutiny, a researcher may find errors in the positioning, or the time, or simply that some measurements are clearly unrealistic. Calibrations can alter measured values. Data transmission systems, including today's GTS, sometimes require data to be reduced in precision or resolution, or homogenized into uniform formats (e.g. tomorrow's GTS, including BUFR), with the accompanying loss of original information. Older data sitting on a shelf or a computer disk may be submitted twice. Simply copying data from one data centre to another and the subsequent processing has the potential to create duplicates. Data centres need to check submissions to see if they have arrived before and if so, determine which is of the greatest value to provide to users.

For purposes of identifying duplicates (and other QC processes such as platform track checking), there is currently in most data submissions no single, completely stable identifying field (even platform identifiers such as VOS call sign or WMO buoy number can be reassigned). However, the Argo program [24] provided from the start a solution where the platform identifier plus profile number is unique and is never reused.

Another solution that is being tried in GTSPP [7] matches lower resolution ocean profiles in real-time to higher resolution forms coming in delayed mode, both originating from the same instrument. It relies on attaching a unique identifier that is generated (without any external coordination requirements) early on in the data collection process, and does not use any controlled vocabularies, such as country codes or ship call signs. Because there is no other meaning to the identifier, there is no temptation to alter it when something in the data is changed. The identifier, once generated, lives forever. If this is attached to all versions of the same data that come to a data centre, the process of identifying duplicates is easy; it is a simple process of matching the identifier to one found in an archived record. If a match is found, the archive can simply store the new version with suitable identification of its source and choose some criteria for determining which version is first offered to a user.

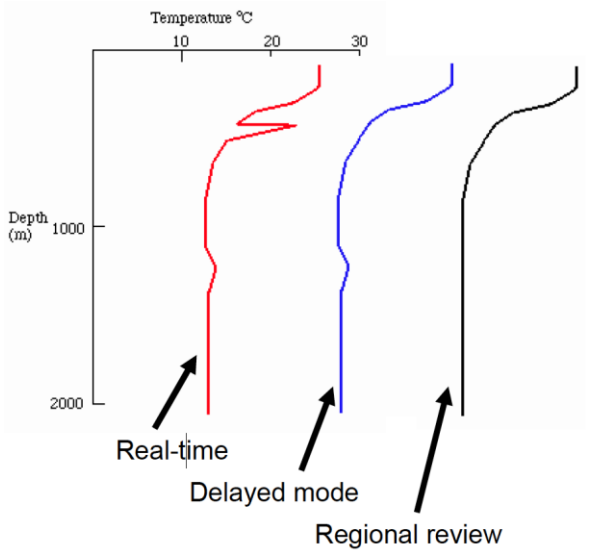

Figure 1: Data produced in real-time (red) may have errors that are corrected after delayed mode quality control (blue), or after comparison to other data in the same region (black) (note: horizontal axes for the blue and black curves are not shown). Sometimes locations and times are also adjusted through quality control and calibrations. These different versions need to be tracked.

For VOS, efforts have been initiated to explore the use of the International Maritime Organization (IMO) number (historically, the hull number) in place of, or in conjunction with, the ship call sign. Finding an alternative to the current practice of using just the ship call sign is becoming important because of the demands to mask ship identifiers to meet commercial and security concerns.

With all of the data that are exchanged around the world today, ideally only one version (and only one form of those data) should be considered as the "original." In reality most data undergo a variety of transformations for transmission (e.g. real-time vs. delayed mode formats) and for later archiving, so that a genuine original may no longer be available. Nevertheless it is

critical to seek to identify the most original possible information, because this should form the raw material for future re-analyses. In some cases, other versions of the original may be perfectly usable and more convenient to deal with. But there are times when nothing but the original will do.

It has been suggested that a system of identifying versions through some mechanism such as the satellite community uses (identifying "levels of processing") would be useful. Procedures for processing oceanographic and marine meteorological data are extremely varied and it will not be possible to capture the variants through some simple scheme. The best that this can do is to classify in a coarse way what has been done. However, if this is used in combination with finer detail of the processing history of the data [20 and 23], it can be an effective classification. Processing history serves the dual purpose of explaining what has been done, and also allows centres and users to backtrack as needed to identify problems. Both of these capabilities are valuable.

Data centres often assess the quality of the measurements seeking to ensure that possible data errors and inhomogeneities have been found. While there are many schemes to assess quality for data of a similar type [e.g. 25] only a few of them are well documented. Moreover, the procedures used by different agencies for the same ocean variable are often different as is the way to indicate data quality. This reinforces the importance of standards, except here they are for assessing and labeling data quality. Some groups have such schemes, but what is needed now is a community wide mechanism to propose, assess, recommend and implement standards.

Such a mechanism was recently put in place through the JCOMM / IODE Ocean Data Standards Pilot Project (http://www.oceandatastandards.org). The process is well described and supported by a joint committee of IODE and JCOMM. The committee is responsible for managing the review, assessment and recommendation of submitted standards from wherever they come. Whether it is successful in promoting standards that meet the pragmatic and skeptical criteria noted by [27] remains to be seen.

No matter how good the data quality assessment procedures may be, unless a user can know what tests were applied they are unlikely to accept that the data are well checked and so they are likely to repeat the verification with their own procedures. This is wasteful of everyone's time and resources. Reference [11] remarks that the satellite community has started to define, implement and operate a sustained SST validation program. It is their view that this needs to be done in cooperation with the in-situ measurement and data management community.

Even if well documented, it is to be expected that some users will not be satisfied with some of the procedures and will want to carry out checks of their own. They need to be able to easily identify what data need to be retested. This can be done by including, in the processing history or by some other mechanism, information that explains both what tests were performed and what were failed (or passed).

For some users, the tests performed at the data centre will be adequate, or they may want some derived product. In both cases, the data or product they receive should only use data that are considered to be good measurements. The data, therefore, should have information attached (often done now as a quality flag) which indicates the reliability of the measured value. Only those considered reliable would be delivered to these users, or employed in the generation of the product.

At the moment, in many data systems quality assessment is a mix of automated and manual procedures. The manual procedures are important since algorithms for detecting unreliable measurements are still not able to identify subtle (sometimes not so subtle) errors. However, the speed at which data volumes are increasing is driving all data centres towards increased automation. It will be important that the automated procedures used at data centres be developed through the cooperation of researchers and data management experts to characterize failure characteristics of the types of instruments and tune verification algorithms to exploit this knowledge.

An aspect often overlooked but one that is important to the integrity of the archives is the interplay between data providers, data users and the archives (e.g. [24]) as well as downstream projects (e.g. [17]). As data centres examine the data that have arrived, it is common for questions to arise about the instruments used, about the methods, about the validity of some of the measurements, or other attributes. Data centres need to pose these questions to the providers to verify content and improve metadata captured with the data. Likewise, users of the data may identify missing or suspect information in the data sets they receive. They should pose questions to the archives so that these anomalies can be either verified or corrected.

Reference [17] notes that data assimilation models, including for operational Numerical Weather Prediction and re-analyses, are another source of largely untapped QC information (e.g. differences between observed values and model analyses). While information of this

type holds much potential to provide feedback about the quality of the archived data, and thus eventually lead to improved data (including for subsequent re-analyses), no very effective and widely usable feedback mechanisms to all the communities of interest currently appears to exist. In the VOS context for example, [5], the quality of VOS reports is monitored by several major meteorological centres, but primarily the UK Met Office. While results of this monitoring are regularly distributed to Port Meteorological Officers, who are expected to take follow-up actions to correct deficiencies, effective sharing of those results could provide additional benefits feeding down to research products (as discussed e.g. in [18]).

Actions that will improve data processing:

a. Data systems and centres should devise and use unique data identifiers as a solution to resolving duplicates.
b. Data systems and centres should adopt the practice of preserving processing history so that data users can be informed of the details of the steps in transformation of the data from measurements received at the data centre to the form the user receives.
c. Documentation of quality control tests applied, their results and a measure of data quality should be included with all data. Standards for presenting this information need to be adopted.
d. Data centres need to develop effective communication with those who provide data to the archives and with those who take data from the archives so that problems can be identified and corrected in the data.

6. DATA ARCHIVING

It is not just data volumes that are straining present data systems. A more significant factor is the increasing diversity of data and the increasing and detailed information that accompany measurements. This has at least two impacts. First, data centres need to "tool up" their software to manage the new kinds of data and may need to make changes in archive data structures to accommodate the measurements or information. Second, data centres may not have the necessary expertise to manage the new data in ways that will protect them in the long term.

The first impact can be addressed in different ways, but two contributions to the solution are going to be more abstract data models and the use of concepts such as appear in table driven codes (such as BUFR). The latter concepts have been used in a number of data systems to advantage. The former are starting to appear in forms such as Geographic Information System (GIS) data models, and Climate Science Modeling Language (CSML). The use of these more abstract data models (Fig. 2) is not widespread, but they will need adoption in the future.

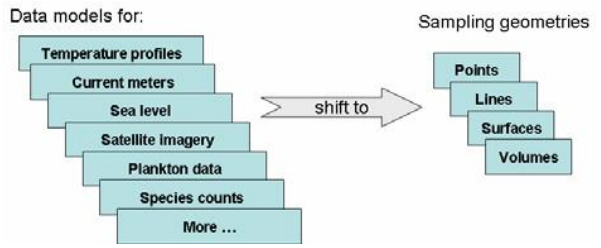

Figure 2. Existing data models exploit characteristics of the type of data being collected. More abstract forms are based on the sampling geometry of the returned data (see [4]).

Even if a data centre has a robust data model, and capable staff, there will always be some kinds of data that are not well understood by data managers at a particular centre. It is not possible for a single data centre to have people with expertise in all the kinds of data that are now collected. The only solution is to develop a distributed network of expertise. Members of this network need to operate on a reciprocal benefits basis. That is, one member undertakes some work on behalf of another with the understanding of receiving in turn some help in a weak area. Members may be found in a centre's own country, but could also be at centres in another country. There is already a basis for this in the old Responsible National Oceanographic Data Centre / Specialized Oceanographic Centre (RNODC / SOC) system of IODE, in the Global Collecting Centres dedicated to delayed mode VOS data [18], and in the more recent Global DAC [24] and GHRSST [11] models. In these examples, data centres volunteered to manage all data of a certain type, or all operations of a certain kind. This is also embodied in a number of the international projects such as ICOADS [17]. To be effective, though, the standards work mentioned earlier must be developed and implemented.

As a final consideration, we need to look at the developing World Data System (WDS) that is replacing the ICSU World Data Centre system. Originally, the WDCs had a main focus to manage the long term preservation of all oceans and other physical science data. There are a number of such data centres, often segregated by kind of data. No one centre does everything for the same reasons no one national centre can handle all kinds of data.

There are a number of national data centres who provide the long term stewardship of their national data. There are also some data centres acting as international data assembly or archive centres that also preserve data into the future. In neither of these cases is it necessary for

the new WDS to act as a primary archive. However, a WDS is still important to provide redundancy as secure archives to guard against accidental loss.

Not all kinds of data have a WDS, global data centre, or even a national one. The WDS centres can help to organize storage and consolidation to ensure the long term availability of these data. For example, any centre aiming to store data in perpetuity needs to consider rigorous and often expensive requirements for adequate data security both physical (e.g., redundant remote backups and media refreshment) and logical (e.g., enduring standards for file formats, software, etc.). These are areas in which the WDC system has long and valuable experience.

WDS centres have a role to play in generating products that are not in the interest or mandate of national centres. A good example is the World Ocean Atlas generated by the WDC for Oceanography (at the US NODC). This WDC has been a leader in the Global Oceanographic Data Archaeology and Rescue (GODAR) project, one that has been very successful in identifying data at risk of loss, digitizing old records and providing easy access to these data in association with consolidated global records for certain kinds of data. On the marine meteorology side, work under the umbrella of the ICOADS project [17] has undertaken many similar activities.

Data centres, both governmental and non-governmental, can respond to the new criteria being set forth for World Data Systems and seek accreditation. Providing they meet the criteria to qualify, they would receive broad recognition of the role they play, and would have a forum within ICSU to connect to the governmental systems, both national and international. This can provide an opportunity for sharing expertise by having specialized archives for different kinds of data.

Actions that will improve archives:

a. Data centres need to upgrade their data models and structures so that they are more robust to handle new kinds of data and additional metadata.
b. Data centres must develop an international strategy to share expertise to manage the many types of data they are faced with archiving.
c. Data centres must work with the new WDS to identify archives for all of the types of data collected, and ensure these data reach the designated archives.

7. TIMELY DELIVERY

The demand is increasing for ocean measurements, among other kinds, to be delivered from instruments to modeling centres very rapidly [28]. It is technically possible to move data directly from instruments to modeling centres with many of the automated quality control procedures being done right at the instrument and using web enabled instrumentation. However, much of the infrastructure to accomplish this is still not in place and so this will take some time to develop. And when or if it does, the role of archive centres will change. Instead of having the time pressures to carry out automated processing, or needing to worry about comparing real-time and delayed mode records, these issues would no longer be present (assuming all users were able to access web enabled instruments). The complication for archives will be staying up-to-date with the data returned from these reporting sensors and ensuring their data also come to the archives for long term stewardship and use in other studies or analyses carried out later. The advantage of dealing in real-time data is not simply in rapid access to the data, but even more in knowing who is collecting what kind of data and when. Because of data transmission limitations, the data that come in real-time transmissions should appear later at data centres as delayed mode submissions.

Finally, it is not enough to place data in archives; it is absolutely necessary that the data also become available to others. Thus, it is important for archives to develop means for data discovery, for browsing of the archives, and of course, for downloading data and products. These mechanisms must also include feedback to archives when data appear wrong, or needed information is missing. A data delivery system should be a value-adding process that provides access to authoritative sources of data and is able to exploit the data. Collective data discovery can be achieved through the use of distributed web service architecture. However this calls for conformance to standards and protocols. The use of agreed standards by data providers will promote interoperability between distributed services and allow the integration of data from disparate data sources. This conformance, however, does not necessarily require standardization of in-house technology. Data providers remain free to choose their own business solutions and to control their internal technical environment. Data dissemination is treated in another paper.[3].

8. SUMMARY

This paper has provided significant detail about the considerations in moving data from collectors to archives, and it is evident that there are many different solutions employed. It should be clear that the differences derive from a number of factors such as when they were developed and hence the technology that was available at the time, and because different organizations developed systems in isolation.

Some of the CWPs have addressed the broader vision of what data systems should be (e.g. [1 and 21]). Reference [1] presents a view of the system that concentrates on the necessary and general attributes of a data system that preserves data and information for future use. The discussion of the Ocean Data Portal (ODP) [21] is based on a distributed set of data providers all linked through appropriate technology and with certain information standards met to allow for data discovery and interoperability. But the ODP view goes beyond simple dispensing of data, to allowing fusion of different kinds of data into on-line products. The idea is attractive, but hides the many details that must be agreed to in order for the system to work.

Reference [26] discusses yet another level of detail and lists many recommendations about aspects of the functioning of a data system. Other authors focus on different components, sometimes explaining how these issues have been tackled in particular cases, and sometimes suggesting new solutions to existing problems. All of these are instructive.

Many of the elements that are needed to build a well coordinated, international data system are to be found already. The list of relevant work includes the following.

- The significant work on both development and implementation of standards at local or regional levels, such as in Australia, the US and Europe.
- The existence of international bodies such as IODE, JCOMM and ICSU that can be used to coordinate data management activities.
- The prototyping of ways to manage duplicates identification, data versions, and more extensive metadata by projects such as GTSPP and Argo.
- New ways to handle biological data through work carried out under the Census of Marine Life and illustrated by OBIS.
- Integration of diverse data (physical, chemical, biological) through on-line access that will allow data to be visualized and analyzed through a single portal.
- Initiatives to implement interoperability arrangements for exchanging data through such applications as ODP and GEOSS (Global Earth Observation System of Systems).

The bullets above are all necessary components of a more coherent data system with a global perspective, but the work is still happening by developers working largely in isolation. This is often driven by the pressing need to develop a solution for a particular project rather than choosing a more global solution. The limitation is that there is no effective and ongoing communication between data system developers. A starting point would be to organize a meeting of the representatives of all major data systems to compare lessons learned, capabilities developed and develop strategies to develop common solutions to common problems.

What is missing is the framework by which these components can be assembled. The international organizations can assist, but all of these rely on volunteers from nations to do the necessary work. The message is clear. If a better global data system is wanted, nations will have to provide the commitment and the resources for an internationally coordinated effort.

A consolidated list of seven actions that should be undertaken is presented below. The actors to carry out these actions are varied. However, both JCOMM and IODE are the bodies with international coordination functions for data management. Both could use their offices and reporting mechanisms to monitor and report progress on these.

8.1 Actions for a way forward:

1. Convene a meeting of data system developers and maintainers from remote sensing, as well as the different disciplines of the in-situ oceanographic community. The objective is to discuss strategies employed, lessons learned, and to seek common solutions or common developments needed. Follow on meetings will be needed to address specific components. This could begin under the auspices of the JCOMM or IODE.
2. All projects, national and international, must contain a component that addresses how the data resulting from the project will be managed and migrated to long-term archives. The component should be developed jointly with the archive and should be funded at the 5-10% level. This will recognize the importance of preserving the data for the future and will provide needed resources to the often underfunded activity of data management.
3. managers must engage with their relevant national organizations, international partners and scientific journals to provide a career enhancing mechanism to recognize the value of making data available to archives and therefore to other researchers both now and in the future.
4. Data managers must make use of the IODE / JCOMM Standards Process to submit suggested standards, participate in the assessment of their suitability, and implement recommended ones in a timely way. This has the potential to act as a focal point for all of the good solutions and work carried out in agencies, but only if it is used and is effective.
5. Data centres must address the many technical details that appear earlier in the document. IODE and

JCOMM should both encourage this and monitor progress.

6. Data centres, IODE, and JCOMM need to be better connected to the evolving WDS. This can be accomplished by having formal representation in the ICSU WDS governing structures.
7. IODE and JCOMM need to provide a well publicized reference site for data management information, standards, etc. There are the beginnings of this in the JCOMM Catalogue of Best Practices and this should be expanded.

REFERENCES

1 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.

2. 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

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

4. Hankin, S. & Co-Authors (2010). "Data Management for the Ocean Sciences - Perspectives for the Next Decade" in these proceedings (Vol. 1), doi:10.5270/OceanObs09.pp.21.

5. Kent, E. & Co-Authors (2010). "The Voluntary Observing Ship (VOS) Scheme" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.48.

6. Rayner, N. & Co-Authors (2010). "Evaluating Climate Variability and Change from Modern and Historical SST Observations" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.71.

7. 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.

8. Dushaw, B. & Co-Authors (2010). "A Global Ocean Acoustic Observing Network" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.25.

9. Boehme, L. & Co-Authors (2010). "Biologging in the Global Ocean Observing System" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.06.

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

11. 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.

12. Brainard, R. & Co-Authors (2010). "An International Network of Coral Reef Ecosystem Observing Systems (I-CREOS)" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.09.

13. Guinehut, S., C. Coatanoan, A.-L. Dhomps, P.-Y. Le Traon and G. Larnicol (2009). "On the use of satellite altimeter data in Argo quality control", *J. Atmos. Oceanic Technol.* 26: 395-402.

14. Brasseur, L., Tamburri, M. and Pluedemann, A., (2010). "Sensor Needs and Readiness Levels for Ocean Observing: An Example from the Ocean Observatories Initiative (OOI)" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.10.

15. 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.

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

17. Worley, S. & Co-Authors (2010). "The Role of the International Comprehensive Ocean-Atmosphere Data Set in the Sustained Ocean Observing System" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.94.

18. Woodruff, S. & Co-Authors (2010). "Surface In Situ Datasets for Marine Climatological Applications" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.93.

19. Heimbach, P. & Co-Authors (2010). "Observational Requirements for Global-Scale Ocean Climate Analysis: Lessons from Ocean State Estimation" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.42.

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

21. 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.

22. McPhaden, M. & Co-Authors (2010). "The Global Tropical Moored Buoy Array" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.61.

23. 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.

24. 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.

25. 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.

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

27. 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.

28. Eyre, J. & Co-Authors (2010). "Requirements of Numerical Weather Prediction for Observations of the Oceans" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.26.