METADATA MANAGEMENT IN GLOBAL DISTRIBUTED OCEAN OBSERVING NETWORKS

Derrick Snowden[(1)], Mathieu Belbeoch[(2)], Bill Burnett[(3)], Thierry Carval[(4)], John Graybeal[(5)], Ted Habermann[(6)], Helen Snaith[(7)], Hester Viola[(2)], Scott Woodruff[(8)]

[(1)] NOAA (National Oceanic and Atmospheric Administration)/Climate Program Office, 1100 Wayne Ave, Suite 1202, Silver Spring MD 20910, USA, Email: derrick.snowden@noaa.gov
[(2)] JCOMMOPS ((Joint World Meteorological Organisation (WMO)/Intergovernmental Oceanographic Commission (IOC) Technical Commission for Oceanography and Marine Meteorology/in situ Observing Platform Support Centre), 8-10 rue Hermès Parc Technologique du Canal, 31526 Ramonville St Agne, France, Email: belbeoch@jcommops.org; viola@jcommops.org
[(3)] NOAA (National Oceanic and Atmospheric Administration)/National Data Buoy Center, 1007 Balch Blvd., Stennis Space Center, MS 39529, USA, Email: bill.burnett@noaa.gov
[(4)] IFREMER (French Institute for Sea Research and Exploitation/Institut Français de Recherche pour l'Exploitation de la Mer) Centre de Brest, BP 70 29280 Plouzané, France, Email: Thierry.Carval@ifremer.fr
[(5)] NSF/OOI (National Science Foundation/Ocean Observatories Initiative) Cyber Infrastructure, 9500 Gilman Drive #0446, La Jolla, CA 92093-0446 USA, Email: graybeal@mbari.org
[(6)] NOAA (National Oceanic and Atmospheric Administration)/National Geophysical Data Center, 325 Broadway, Boulder CO 80305 USA, Email Ted.Habermann@noaa.gov
[(7)] National Oceanography Center, University of Southampton, European Way, Southampton SO14 3ZH UK, Email: H.Snaith@noc.soton.ac.uk
[(8)] NOAA (National Oceanic and Atmospheric Administration)/Earth System Research Laboratory, 325 Broadway, Boulder CO 80305 USA, Email: Scott.D.Woodruff@noaa.gov

INTRODUCTION

Many elements of the Global Ocean Observing System (GOOS) were designed to support weather forecasting, maritime safety, or other short-term operational requirements. Others evolved from research projects where the primary deliverable was a scientific manuscript rather than a sustained data stream. These data systems provide supplementary information, or metadata, designed to serve particular users and the detail and form of this metadata is typically only sufficient to satisfy the application for which the data system was initially designed. This has resulted in a collection of metadata that may be inadequate in scope or in level of detail to support broader user requirements, and generally does not conform to modern national or international standards. This makes it difficult to understand and use the data effectively and creates obstacles to meaningful data integration.

Climate data, on the other hand, needs to be used, and useful, for decades (if not centuries) into the future. Data sets collected today must be documented and described by accurate and complete metadata in order to ensure that they remain available and useful well into the future, and possibly for yet unimagined applications.

Other contributions to this conference have made a compelling case for the sustained ocean observations community to mature and modernize their approach to data stewardship so that data is available, discoverable and useful in perpetuity [1]. Metadata and metadata standards are an integral part of this long-term vision and it is a thesis of this paper that current metadata collection and management practices are insufficient. Significant uncertainties in the global ocean in-situ climate record can be traced to poor metadata. Continuing current practices risks invalidating or casting doubt on recent scientific discoveries (because they are not reproducible) and preventing new ones (because the data may not be decipherable in the future). Actions necessary to prevent this outcome are the responsibility of scientists collecting the data, program managers funding the campaigns, data managers distributing data, and archive centers preserving data. Furthermore, this responsibility for action is spread equally across these roles, and extends to international organizations responsible for managing many ocean data and metadata international, including the Joint WMO-IOC (World Meteorological Organization - International Oceanographic Commission) Technical Commission for Oceanography and Marine Meteorology (JCOMM).

Through the use of historical and current examples, we will highlight some of the risks due to poor metadata, show some recent improvements, and point to more improvements, including changes to operational procedures and new technologies that should be adopted over the next ten years.

1. METADATA PRINCIPLES

Metadata describes a broad range of information that allows observations to be understood and evolved into information and knowledge. It provides a context for research findings, ideally in a machine-readable

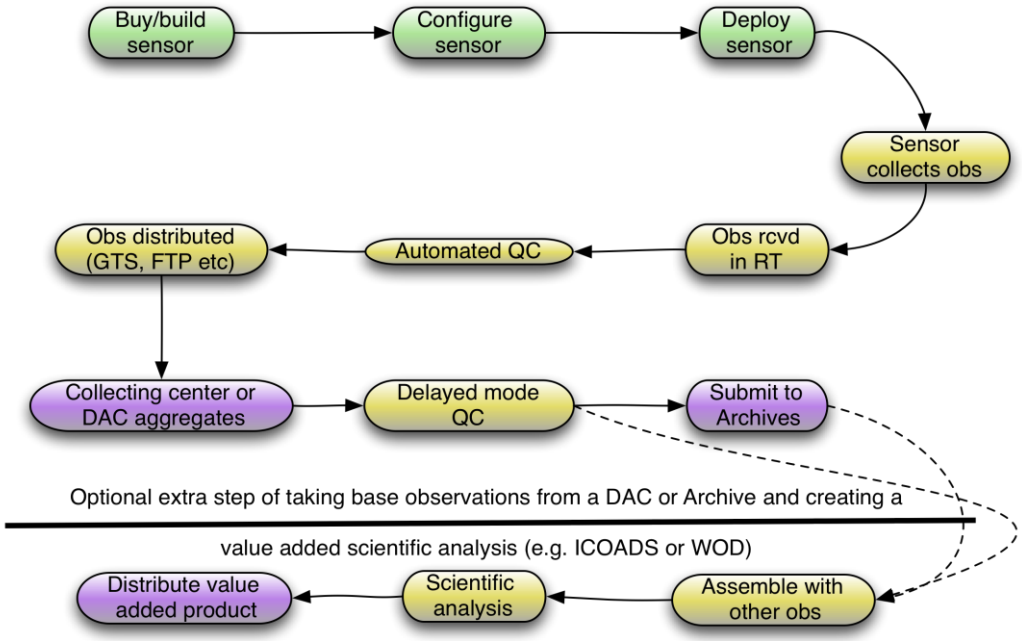

Figure 1: Generic data flow sequence of observational data from sensor design through to archive of delayed mode quality controlled observations. The sequence is meant to be simplified and may not encompass all possible variations. At each step of the sequence information or metadata about the observation is generated. The steps in green are associated with Instrument or Platform metadata; those in yellow are Provenance or Lineage metadata while those in purple are typically associated with Catalog or Discovery metadata. These associations are loose and meant as a general guide.

format. It enables discovery of data via an electronic interface, and correct use and attribution of findings.

In order to demonstrate the metadata collection and management process, we show the components of the process in a generic, linear data flow sequence (Fig. 1). In this process an observing requirement is identified, then a sensor is purchased or constructed to satisfy the observing need. The sensor is configured, deployed at sea and collects observations that are regularly transmitted to a shore-side receiving station. The receiving station reformats the data messages, possibly applying some automated quality control, and then distributes the data over one of several communication pathways like the Global Telecommunication System (GTS). Collecting centers monitor the GTS and assemble aggregations of similar data types, and finally the observations are incorporated into, possibly many, value- added data sets for subsequent scientific analysis. At each step in this data flow, information or metadata related to the observational data is generated or revealed. This model is not universally applicable but it does provide a general framework for discussion.

From the example it is clear that a broad definition of metadata can include virtually every piece of ancillary information ever associated with or contributing to a given observation, from the time a sensor was built, through routine data handling, and on to archive centers that must preserve and advertise data holdings. The volume of ancillary information is immense and designing a strategy to determine how such information should be identified, prioritized, collected, managed, distributed and preserved is a crucial element of observing system design and implementation.

Beyond these general definitions and desired characteristics, we differentiate between three major levels of metadata: Instrument-Platform (I-P), Provenance-Lineage (P-L), and Collection-Discovery (C-D). This classification is somewhat arbitrary, but we use it here since it provides a useful framework to identify the roles that must be filled in the future data management system.

2. INSTRUMENT-PLATFORM (I-P) LEVEL

This form of metadata is the most fundamental and ideally documents all the salient sensor and platform characteristics associated with an observation. From Fig. 1 we see the first three steps in the sequence (the green elements) are associated with sensor level metadata and importantly the very first step begins with the manufacturer. The sensor manufacturer is the

initial supplier of metadata and provides information about a sensor that is in some cases static throughout the lifetime of a sensor, such as make, model, and serial number.

For autonomous, expendable platforms, such as profiling floats or drifting buoys, the majority of the I-P metadata is known at the time of deployment (or earlier) and does not change through during the finite life of the platform. For other platform types, such as research and volunteer ships (VOS), or moored buoys, the I-P metadata changes every time an instrument is altered. For example, each tropical moored buoy is reconfigured approximately once a year at which time each sensor is swapped out for a replacement. Similarly, the barometer on a VOS is calibrated annually. These events must be captured in a metadata management system and propagated through the data flow, so that subsequent data archives record the all changes in the platform configuration along with the observed data. I-P metadata is largely the responsibility of platform operators, though metadata experts can help with utilizing metadata standards and tools for creating and managing this data.

2.1. PROVENANCE-LINEAGE (P-L) LEVEL

This metadata describes the processing and history of observations, including information about their source, version, quality assessment and control, history and accountability. The information included in this category of metadata has not usually been made available in real-time.

This metadata has not been a priority in the past for the marine weather or ocean observing community and requirements have changed over time too quickly for the current processes to keep up. Consequently, there is a very strong need to improve the management of this metadata from groups such as climate and ocean service providers, and researchers, data providers, program managers, platform operators and platform manufacturers, to realize how vital the metadata is to data users and how it can assist in their own data management and ensure proper versioning, so that the sources of good quality data can be recognized. It is particularly important to be able to identify the latest, best copy of data, whilst still tracking changes to that data and links back to the original.

2.2. COLLECTION-DISCOVERY (C-D) LEVEL

Many scientists are familiar with the C-D metadata that forms the foundation for several data discovery portals (e.g. NASA (National Aeronautics and Space Administration) Global Change Master Directory, Geospatial One-Stop, GEOSS (Global Earth Observation System of Systems) Registry, the NERC (Natural Environment Research Council) Data Grid and INSPIRE (Infrastructure for Spatial Information in Europe)[i]). It includes information that identifies a collection of files as a coherent data set, supports discovery of that collection using full-text, keyword, and spatial/temporal searches, and describes where the collection can be obtained. Further, the format of the data (i.e. instructions for software to be able to read the data files) is specified in C-D metadata. C-D metadata can also include more detailed information, but that is generally optional and less common.

The discovery portals mentioned above provide searchable interfaces on top of catalogs of Collection-Discovery metadata submitted by data providers. A casual perusal of each portal demonstrates some inconsistencies. For example, some data sets are described in one portal but not another. Or, when a data set is discoverable in multiple portals, it is unlikely that the records are identical and in some cases directly contradict one another. The true promise of these discovery portals will be realized when there are automated harvesting procedures that enable searching across multiple portals from a single interface. This capability is theoretically available, though not tested in an operational sense, from the Open Geospatial Consortium, Catalog Services for the Web service (OGC CSW). Once fully developed, this functionality will enable data discovery services that are wide reaching, and analogous to Google for geospatial data, but still retaining the control needed for precise and directed searches. This type of federated searching capability depends on data set documentation, C-D metadata, being provided in one of a select few, documented and tested standard forms.

3. METADATA STANDARDS

Standards are used to manage complexity. Standard models for data content and data representation bring some level of homogeneity to the data management enterprise and allow for effort to be focused on data services rather than data decoding. Data standards are being increasingly recognized as necessary to enable data integration both within and across disciplines [2]. As the sustained ocean observing system evolves from a primarily physical science endeavor to a truly multidisciplinary effort, standard methods of communicating data and metadata across disciplines will become even more crucial. Adopting standards for data content and data encoding will enable interoperability in several ways. First, standards encourage the development of make it more likely that

[i] These registries are available at www.gcmd.nasa.gov, www.geodata.gov, and www.geossregistries.info, http://ndg.nerc.ac.uk/ , http://www.inspire-geoportal.eu/ respectively.

common tools, which can easily work with the data, and encourage the development of those common tools (because they apply to many different more data sets, so they are more useful). As an example, consider the wide adoption of NetCDF (Network Common Data Form) as a data encoding standard, made possible, and perhaps caused by, the ubiquity of well documented, tested and reliable software available for reading NetCDF.

Examples of standards for electronic data formatting include NetCDF, HDF (Hierarchical Data Format), XML (Extensible Markup Language) and even proprietary forms such as Microsoft Excel. Adopting one of these standards allows individuals to utilize common software tools, reference materials and build on a whole community of experiences. However, simply adopting a data representation standard does not facilitate interoperability. Given a NetCDF file, most savvy data analysts will have no trouble reading the file. Understanding the data within the file requires more information, often codified in data content standards. The data content standard describes the elements comprising the data model and how they relate to each other. Standards help to impose structure on information so that users (and software) know what to expect. They make it possible for a naive user to access and understand data (because the data is well described and references to terms and

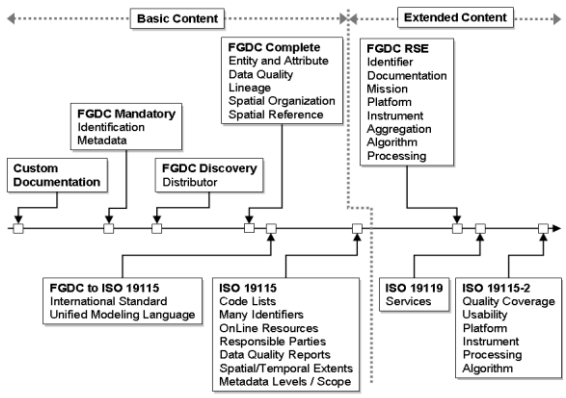

Figure 2. Metadata content standards spectrum. On the left are simpler forms, mostly derived from the US Federal Geospatial Data Committee standards that facilitate basic search and discovery. Toward the right are more complex forms, largely from the ISO family of standards that enable more complete data processing elements and history to be described.

formats are widely available) and they encourage good practices by setting appropriate minimal criteria that users must follow.

Common metadata content standards (and their corresponding representation standards) have evolved significantly. Early standards included content elements that were minimal in scope and only really enabled the basic discovery functions that C-D metadata should facilitate (Fig. 2). As more users implemented these standards more functionality was demanded of them, so they had to evolve in complexity. For example, describing where a data set is archived (a C-D metadata element) is much easier than describing the sequence of events and quality control that led from raw sensor data to a derived product (a P-L metadata element).

Choosing a data content standard (and its related representation standard) depends on several factors, but most importantly on the services the metadata will support. For example, the simpler data content standards, such as Dublin Core, NASA Directory Interchange Format, and US FGDC (United States Federal Geographic Data Committee) Content Standard for Digital Geospatial Metadata (CSDGM), support basic search and discovery services. Extensions to FGDC have enabled the description of some sensor characteristics and more complete standards such as ISO (International Organization for Standardization) 19115/19139, have recently been developed that describe the metadata elements for all three of the levels of metadata described previously.

Several wide-ranging data management enterprises have adopted ISO as the standard of choice for metadata delivery, including the European INSPIRE and NERC Data Grid efforts, the Australian Integrated Marine Observing System, and the WMO. The US FGDC has recently approved an ISO extension as the eventual replacement for the FGDC CSDGM data content standard.

The Climate and Forecast conventions for (CF) metadata are another data content standard that is designed to promote the processing and sharing of files created with the NetCDF API (Application Programming Interface)[ii]. The CF conventions are increasingly gaining acceptance and have been adopted

by a number of projects and groups as a primary standard. The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities. Another key element of the CF conventions is the adoption of a standard names table.

[ii] http://www.cfconventions.org

This standard names table is an example of a controlled vocabulary that greatly enhances interoperability and machine readability of data. Further, through the use of ontologies, controlled vocabularies in use by one community can be mapped to those in use by a separate community to allow for machine-based translation between separate conventions. This level of interoperability, facilitated by metadata and data standards, is the goal for all of the JCOMM Observing Programs.

The CF conventions focus mostly on providing a data model and associated metadata that allow data in NetCDF files to be understood and less on preserving the data provenance or providing discovery capabilities. However, communities such as GHRSST (Global High Resolution Sea Surface Temperature) have adopted a two pronged approach that encodes the data and selected metadata in a CF NetCDF file and encodes extended metadata covering data provenance in a XML file encoded using FGDC with Remote Sensing Extensions conventions[iii].

A final data content (and representation) standard, particularly focused on describing sensor level metadata, is the OGC Sensor Markup Language or SensorML (Sensor Model Language). SensorML excels at describing sensor and platform characteristics precisely. Additionally, SensorML can describe procedures or sequences of steps in an algorithm that can be read by a machine and directly converted to executable code. This feature makes SensorML an excellent candidate for describing quality control chains. Efforts such as QARTOD (Quality Assurance of Real-Time Oceanographic Data) and Q2O (*QARTOD* to OGC (Open Geospatial Consortium) [iv] are experimenting using SensorML to process ocean observing system data automatically and to document the entire process using an internationally accepted standard. While these features are still experimental they represent significant investments and are likely to influence future data management systems.

4. A DISTRIBUTED SYSTEM

Data and metadata management is an evolving discipline and at any point in time the current "best practice" is subject to the influence of advances in information theory, hardware design, and software development practices among other disciplines. What is a best practice today will likely be an antiquated notion in ten years; suggesting a specific design for a data management infrastructure is impossible. Further, the exact implementation that would yield the best data management infrastructure is hotly debated even today. Nevertheless, we can elucidate some characteristics of a data and metadata management system that would be a vast improvement over the system in place today, and would ameliorate some of today's known problems. Against that general framework, we can judge some of metadata management practices in use today and begin to identify specific areas where investment of resources is warranted.

One of the essential purposes of rich metadata is to enable and ensure the long-term preservation of data. As such, the metadata itself must be collected and managed within a system designed for long term preservation or archival (Conkright-Gregg et al, 2010 [1]). One key element of preserving metadata is storing it in a format that is not at risk of media degradation. Additionally, enabling the preservation of large quantities of data and metadata requires that it be stored in a format that is machine-readable. Formats such as word processing software outputs fail both of these measures. A better alternative is one of the aforementioned ASCII (American Standard Code for Information Interchange) XML formats such as SensorML or ISO.

The various observing systems information at all stages should be discoverable, identifiable and served by a connected information system, even if the information is distributed.

5. HISTORICAL PRACTICES AND LESSONS LEARNED

5.1. Argo

Argo is a global array of 3,000 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean with all data being relayed and made publicly available within hours of collection. One of the successes of Argo, in addition to being an excellent example of international coordination, is the underlying data management system [3]. Data collection and management was planned very early on in the campaign, which began in 2000, and metadata was emphasized from the beginning. The result was a system of data assembly centers (DAC)[v] working together to document everything from the sensor manufacturer specifications to the results of delayed mode quality control procedures. The documentation of data provenance in the Argo system is among the

[iii] The GHRSST Long Term Storage Facility at the US NODC: http://www.nodc.noaa.gov/SatelliteData/ghrsst/

[iv] http://q2o.whoi.edu

[v] The Global DACs are run by the US Navy (http://www.usgodae.org/cgi-dods/nph-dods/ftp/outgoing/argo) and IFREMER (http://www.ifremer.fr/cgi-bin/nph-dods/data/in-situ/argo)

most thorough in the JCOMM Observation Program Area (OPA), despite the shared responsibilities being distributed over approximately 15 countries, each with multiple centers.

The data are distributed along with metadata using an Argo specific NetCDF file format. Within each NetCDF file, pertinent sensor and provenance metadata are encoded as global attributes or variable attributes and many elements use a common naming convention based on the BODC/GF3 parameter vocabulary. The data content is well documented in a user manual that is regularly updated and within the Argo community. The Argo standard is evolving and the community is looking to embrace the NetCDF Climate and Forecast conventions. Further, experiments with moving the metadata into an international metadata standard such as SensorML and ISO are ongoing.

5.2. OceanSITES

OceanSITES (Ocean Sustained Interdisciplinary Timeseries Environment observation System) (the international, multidisciplinary time series and long-term, deepwater reference station network) is a newer effort in the JCOMM OPA and benefitted from the lessons learned during the Argo. OceanSITES has adopted the DAC structure and creates detailed NetCDF data files.[vi] The data content is based on the NetCDF CF conventions and the parameters within the files are named according to controlled vocabularies. OceanSITES also standardized a set of quality control procedures that are applied consistently to every platform in the program and recorded in the NetCDF data files.

OceanSITES documents detailed information about the physical configuration of each platform in the program. Additionally, provenance metadata such as deployment information, sensor calibration and maintenance, and sensor location on the platform (e.g. the exact location of each temperature sensor relative to the center of the surface buoy) are collected and distributed along with the data. Initially the information was collected in the form of Microsoft Word documents but more recently the metadata is being collected, and managed, in the form of SensorML XML documents. Data managers at the DACs currently create the SensorML documents manually. Ideally this should be implemented smart sensors to essentially "describe themselves" through a hardware interface, though this is a long-term goal. Developing smart sensors and the data systems to dynamically capture the sensor description is at the heart of the OGC Sensor Web philosophy. SensorML will likely play a role in the evolving Sensor Web and any investment in this IT infrastructure should pay off, both in terms of capturing crucial metadata and in enabling a range of web services and dynamic interaction with sensors deployed at sea.

5.3. Volunteer Observing Ships

Shipboard marine meteorological observations have been recorded for hundreds of years, coordinated today under the JCOMM Voluntary Observing Ships (VOS) Scheme. The VOS observations form a baseline data source (e.g., Fig. 3) for many analyses of marine climate [4]. However, data quality varies dramatically over this long period so biases and random errors in the data need to be assessed—making use of I-P metadata—so that consistent estimates of long-term climate variability and change can be made.

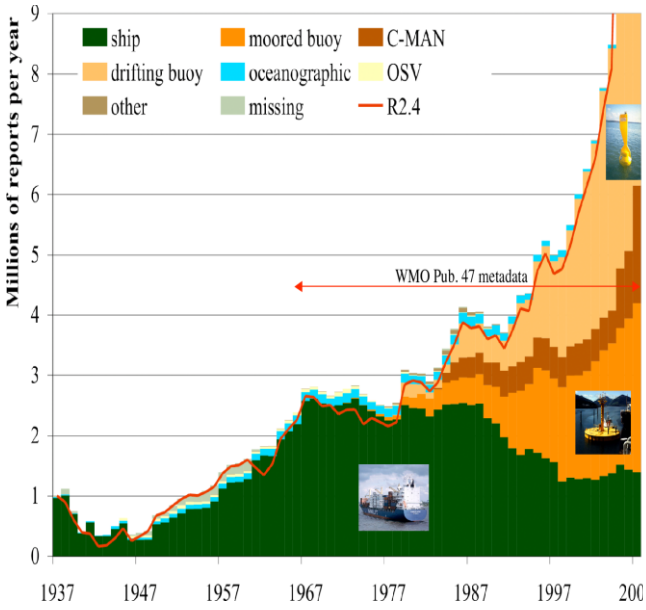

Figure 3. Annual distribution (1937-2007) of major platform types in the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5 shown as millions of reports per year. Selected WMO Pub-47 metadata are being blended with the ICOADS ship reports for 1966-forward. For clarity the vertical scale is truncated at 9M; years 2005-07 have 13M, 15M, and 16M total reports (not visible) in Release 2.5, respectively. The red line curve shows the Release 2.4 annual counts. VOS = ships, buoys are self explanatory, Ocean R/V = oceanographic, Ocean permanent station vessel = OSV, Coastal-Marine Automated Network = C-MAN, ocean drilling rigs and other small entities = other, and unidentified platform types = missing (after Fig. 5 in Worley CWP; ship photo courtesy of www.ShipPhotos.co.uk).

Since the mid-20th century, extensive I-P metadata has been regularly gathered by WMO from "directors of

[vi] http://www.oceansites.org

national meteorological services concerned" and recorded in Publication No 47[vii] [5]. Originally a paper-only annual publication, a digital version became available in 1973. The metadata fields have gradually evolved to reflect temporal changes in VOS instrumentation and observing practices and since 1998 the digital metadata have been issued approximately quarterly, though with some significant lags. The NOAA Climate Database Modernization Program (CDMP) has digitized the earlier 1955-72 editions [6].

The basic WMO Pub-47 is distributed as a single data file containing entries for all known VOS ships, past and present. I-P metadata files have been used to generate an I-P metadata "attachment" to the International Maritime Meteorological Archive (IMMA) format, used for ICOADS and being developed under JCOMM (Woodruff 2007). This attachment contains a selection of metadata of primary user interest that is delivered along with the observed values, thus conveniently enabling improved climate research applications. For example, instrument type and placement are now being used to create products for ocean surface fluxes of heat, water, and momentum [7] and [8], and information about sea surface temperature (SST) measurement method can be used to create analyzed SST products to evaluate historical climate variability [9].

Additional earlier ship I-P metadata exist in sources including commercial organizations, such as Lloyd's Register containing ship size and type information on many worldwide merchant ships back to 1764, and in the archived logbooks of US, UK, and other early national collections (e.g., Fig. 4). Modern efforts to digitize additional historical marine data (e.g. [10]) seek to also capture as much as possible of the scientifically relevant I-P metadata. However, owing to early technological limitations many collections digitized decades ago are imperfect or incomplete, and recognizing and addressing omissions and biases in existing digital collections also needs to be considered an important effort [11].

While developed many years ago, WMO Pub-47 metadata represent a successful existing I-P standard, as we have described generally in sec. 4, which was developed by a large community and serves a broad user base. Looking toward the future, however, we need to consider the possibility of expanding Pub-47 to other ship-based (or manual) systems, and addressing some weaknesses in the timely delivery and storage of the metadata (including exploring possible additional convergence, as appropriate towards other recognized modern metadata standards).

5.4. Ships Of Opportunity: Expendable Bathythermographs

In the early 1990's the operational XBT community launched a series of field experiments to verify whether the parameters in the fall rate equation that was supplied by the manufacturers were correct. Based on these tests, changes were made to the fall rate equation used in operational data dissemination via the GTS. Recent investigations indicate that the fall rate may have again changed indicating that this metadata is inherently time dependent and should be tracked accordingly [12]. Unfortunately many profiles in the historical database, especially prior to 1995, do not have any information about the fall rate equation that was employed. Without this knowledge it is difficult, if not impossible, to correct or adjust the XBT temperature data and this uncertainty has propagated into several scientific analyses. In the future, all parameters used to convert raw data (i.e. the ten hertz temperature observations) into geo-referenced physically useful data (i.e. the depth temperature pairs at a given latitude and longitude) must be preserved, as metadata entries, along with the data itself. Both ISO 19115 and SensorML have the capability of encoding this sort of sensor and provenance metadata, though neither is currently used within the XBT community.

5.5. Drifting and Moored Buoys and ODAS Metadata

Both the drifting buoy program and the tropical moored buoy programs have been in existence far longer than Argo or OceanSITES. As such the data systems for these data buoys are based on older technology. Data and metadata management for these systems are stovepipe systems that, while they may satisfy the requirements under which they were designed, they are not built on standards and do not satisfy modern requirements for data and metadata exchange. As such, information about the sensors on each platform is unavailable, and detailed information managed by operators may be at risk of loss due to media degradation. Each of these programs could benefit by leveraging some of the work done by OceanSITES and Argo in metadata collection.

In recognition of the lack of a centralized international I-P metadata archive, like the WMO Pub-47 for drifting and moored buoys, and other Ocean Data Acquisition Systems (ODAS), JCOMM, at its first session in 2001 adopted an ODAS metadata standard. In 2002 the National Marine Data and Information Service (NMDIS, China) offered to host an ODAS Metadata Service (ODASMS; http://www.odas.org.cn/) using that standard as the basis. Many metadata records have now been gathered from the DBCP and the China

[vii] The Pub-47 data content format is described at http://www.wmo.int/pages/prog/www/ois/pub47/pub47-home.htm

Argo Data Center for profiling floats, drifting buoys and moored buoys, and some fixed platforms. NMDIS has been working on electronic versions of the standard. Also, an XML Schema for ODAS metadata has been proposed and is also available for download on the website.

Retrospective buoy/ODAS I-P metadata (e.g., sensor characteristics, buoy-hull types, and links to documents and photographs) reside mainly at different national buoy centers, or other scattered locations, in a variety of formats, with a long time frame required for availability of much retrospective metadata. Some of the earliest retrospective metadata might also be in danger of loss due to media degradation or personnel changes. Work was proposed in 2006, but not funded, with these goals: (a) To accelerate the JCOMM effort, and help populate the ODASMS by gathering retrospective ODAS metadata from primary US and selected international buoy arrays. (b) To blend key selections of the metadata with ICOADS (following the method used for VOS Pub-47 metadata as described in sec. 5.3), for the convenience of users and immediate benefits to climate research.

Looking toward the future, a more unified strategy for the archival of I-P metadata for ocean moored and drifting buoys urgently needs to be developed (noting that coastal arrays may have somewhat different issues). The ODASMS represents a useful undertaking and should ideally form a continuing, but more clearly targeted, component within the future international system. Possibly the future international system could seamlessly consolidate I-P metadata for ocean buoys plus other closely related ODAS (including rig and platforms), even if the underlying technological solution spans multiple physical archive locations. Related to that general goal, different approaches to proposed convergence with water temperature metadata suggested by the Meta-T Project also need to be considered. Moreover the unfulfilled goals outlined above for the rescue and availability of retrospective buoy metadata arguably now are even more critical, and should be strongly endorsed by the ocean community with a solution devised at the earliest opportunity, and tied in appropriately with future metadata requirements


Name of vessel, Holywood
Rig, Ship
Wood, iron, composite? Iron
Steamer or sailing vessel?
Length, 280 Beam,
Name and address of owner,
Commander,
Address,
Nationality of register, British
Screw or paddle,
Depth of hold, Tonnage,
Names of the observers who have filled this Journal,

INSTRUMENTS:

MERCURIAL BAROMETER.

By whom made?
How high is it carried above sea-level?
What is the diameter of the bore of the tube?
What is the diameter of the cistern?
Measure carefully the inch-divisions of the scale—are they EXACT inches? Yes
If not, what portion of an inch are the divisions that are marked such?
Where is it carried?
When and where compared with standard? New York
Give the comparisons with the standard.
Does the mercury pump much when ship has motion? Slightly
Of which—a Mercurial or an Aneroid—are the readings recorded in this Journal?

ANEROID BAROMETER.

By whom made?
How high is it carried above sea-level?
When and where compared with standard?
Where is it carried on board?
What are its comparisons with standard?

THERMOMETERS.

By whom made?
Are the dry and the wet-bulb thermometer mounted in a lattice-work case such as that recommended in the Introduction to this Journal? Yes
If not, state how they are arranged and where kept.
What is the method of taking the temperature of the sea-water?

Figure 4. Early ship platform and instrumental metadata recorded in the US Marine Meteorological Journal of ship Holywood, 1886.

6. CONCLUSIONS AND RECOMMENDATIONS

Without an active effort to manage the metadata describing ocean observations, much of the current research and operations expenses may be considered wasted in the worst case or suspect at best, because there will be no useful access to the data that results (where useful means: discoverable; recoverable; manipulable (understandable syntax); deconstructable (understandable provenance); and meaningful (understandable semantics). This will make the conclusions developed from that data scientifically marginal and less useful than they could be, because there will be no way to evaluate or test them by re-analyzing the data.

The successful characteristics of some of the current systems combined with some new technological developments on the horizon give us a fair picture of how metadata management in the JCOMM OPA

should evolve.

a) Automation is paramount: Creating metadata can be tedious. Wherever possible the process should be automated. Good candidates for near to medium term developments include smart sensors that automatically report salient sensor characteristics when they are deployed on a platform. Additionally, automated data processing steps such as quality control and real time dissemination over the GTS are prime candidates for automatic documentation.

b) Unique identifiers: Develop a robust means to identify platforms and sensors uniquely. The current system of generating WMO IDs to identify platforms at sea is inadequate for long-term climate applications. WMO IDs are reused and the lack of uniqueness inhibits matching a raw observation to the rich provenance metadata generated throughout the data lifecycle.

c) Planning: Incorporate metadata planning at the earliest stages of the data collection campaign. Requirements for documentation should be identified early as adjustments in projects are much easier and cheaper in the planning stages than they are midway through the implementation.

d) Open access: The JCOMM OPA partners have made great strides in making raw data available freely and openly. The same effort should be afforded to the metadata that must accompany the data.

e) Timely access: The primary source of ship metadata, Pub-47, is typically a year or more out of date. Data and metadata needs require real time access to that data and management of the database should switch from the WMO to an agency more able to manage it effectively.

f) Web services: The web and advanced web services will be fundamental to future ways of disseminating data and information. Once we can uniquely identify platform information, web services will allow us to access only the information needed for a given application. This will relieve the burden on individual programs and DACs and allow for more distributed responsibilities and processes.

g) Real time dissemination: This can be achieved by capitalising on comprehensive data formats, such as BUFR, SensorML, ISO and CF-NetCDF, which allow for collection and real-time communication of as much metadata as necessary. This is possible with newer telecommunications systems where restrictions on file sizes and formats are no longer a constraint.

h) Documenting quality control: Documenting quality control procedures allows for data to be unambiguously described using common terminology.

i) Adoption of standards: Standards enable interoperability. Every effort should be made to judiciously adopt standards if one exists that satisfies the observing requirements. If not, working within the standards process is more effective than designing from scratch.

j) Distribute the system appropriately: There are several efforts to develop system wide procedures such as ODAS and META-T. These efforts should be focused more toward the unifying data elements they are attempting to describe. The more successful metadata projects are those where metadata management is tightly coupled with the data management. The structure and heterogeneity in the metadata described here is complex and rich. It is unlikely that one center or one database will have the breadth to address documentation needs of every platform type. That is not to say that global efforts to define data content or collect requirements do not have merit however.

The system envisioned here is one in which incremental changes to the current state will likely be most successful. Revolutionary advancements will likely come from smaller more focused efforts and like all research ideas, will take thorough testing and vetting before they become an operational reality in a global system. There are however emerging and established standards that can be applied successfully in the very short term with positive effect. We underscore however that thorough documentation leading to good data stewardship is as much a social or organizational commitment. The technological hurdles are easy compared with changing an operational agencies procedures and commitments.

6. 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. 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.

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

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

5. WMO (1955). *Publication 47: International List of Selected, Supplementary and Auxiliary Ships*. World Meteorological Organization.

6. Kent, E.C., S.D. Woodruff, and D.I. Berry (2007). *Metadata from WMO Publication No. 47 and an Assessment of Voluntary Observing Ship Observation Heights in ICOADS.* Journal of Atmospheric and Oceanic Technology, **24**(2): p. 214-234.

7. Fairall, C. & Co-Authors (2010). "Observations to Quantify Air-Sea Fluxes and their Role in Climate Variability and Predictability" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.27.

8. Kent, E.C., P.K. Taylor, and S.A. Josey (2003). *Improving Global Flux Climatology: The Role of Metadata*, in *Advances in the Applications of Marine Climatology: The Dynamic Part of the WMO Guide to the Applications of Marine Meteorology*. p. 89-97.

9. 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.

10. Brohan, P., et al. (2009). *Marine Observations of Old Weather.* Bulletin of the American Meteorological Society, **90**(2): p. 219-230.

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

12. Wijffels, S., et al. (2008). *Changing expendable bathythermograph fall rates and their impact on estimates of thermosteric sea level rise.* Journal of Climate, **21**(21): p. 5657-5672.