INTEGRATING QUALITY ASSURANCE AND QUALITY CONTROL INTO OPEN GEOSPATIAL CONSORTIUM SENSOR WEB ENABLEMENT

Janet Fredericks[1], Mike Botts[2], Luis Bermudez[3], Julie Bosch[4], Philip Bogden[5], Eric Bridger[5], Tony Cook[2], Eric Delory[6], John Graybeal[7], Sara Haines[8], Anne Holford[9], Carlos Rueda[7], Jordi Sorribas Cervantes[10], Feng (Barry) Tao[11], Christoph Waldmann[12]

(1)Woods Hole Oceanographic Institution, MS 9, Woods Hole, MA 02543 USA, Email: jfredericks@whoi.edu
(2)University of Alabama in Huntsville, National Space Science and Technology Center, 320 Sparkman Drive, Huntsville, AL 35805 USA, Email: mike.botts@botts-inc.com; tcook@nsstc.uah.edu
(3)Southeastern Universities Research Association (SURA), 1201 New York Ave NW, Washington DC 20005-391720005 USA, Email: bermudez@sura.org
(4)National Coastal Data Development Center, NOAA (National Oceanic and Atmospheric Administration), Stennis Space Center, MS 39529 USA, Email:Julie.Bosch@noaa.gov
(5)Gulf of Maine Ocean Observing System, 350 Commercial St, Portland, ME 04101-5509 USA, Email: bogden@gomoos.org; eric@gomoos.org
(6)dBscale Sensing Technologies, C/ Leon y Castillo 25, Telde 35200, Canary Islands, Spain, Email:eric.delory@dbscale.com
(7)Monterey Bay Aquarium Research Institute, 7700 Sandholdt Rd, Moss Landing, CA 95039-9644 USA, Email:graybeal@marinemetadata.org; carueda@mbari.org
(8)University of North Carolina, Chapel Hill, NC 27599-3300 USA, Email: sara_haines@unc.edu
(9)University of Aberdeen, Newburgh, Aberdeenshire, AB41 6AA, UK, Email: a.holford@abdn.ac.uk
(10)Unidad de Tecnologia Marina, (CSIC), 08003 Barcelona, Spain, Email: sorribas@cmima.csic.es
(11)National Oceanographic Centre Southampton, University of Southampton, European Way, Southampton SO14 3ZH, UK, Email: bt@soton.ac.uk
(12)University of Bremen/MARUM (Center for Marine Environmental Sciences), P.O. Box 330440, 28334 Bremen, Germany, Email: waldmann@marum.de

ABSTRACT

As standards in best practices in data quality assurance and quality control evolve, methods for discovery and transport of information relating to these practices must also be developed. An observation's history, from sensor descriptions, processing methods, parameters and quality control tests to data quality flags and sensor alert flags, must be accessible through standards-based web services to enable machine-to-machine interoperability. This capability enables a common understanding and thus an underlying trust in the expanding world of ocean observing systems. For example, a coastal observatory conducts several tests to evaluate and improve the quality of *in situ* time series data (e.g. velocity) and then generate an oceanic property (e.g. wave height). Using content-rich web-enabled services, a data aggregation center will be able to determine which tests were conducted, interpret data quality flags and provide value added services, such as comparing the parameter with those from near-by observations. These additional processing steps may also be documented and sent along with the data to other participating ocean observing systems throughout the world. By utilizing standards-based protocol (Open Geospatial Consortium (OGC) frameworks) and well-defined community adopted QA/QC (Quality Assurance/Quality Control) tests and best-practices (Quality Assurance in Real-Time Oceanographic Data - QARTOD), information about the system provenance, sensor and data processing history needn't be lost.

Are data providers ready, willing and able to describe sensors and processing history? And can we transport the information using a framework that offers semantic and syntactic interoperability? The group developing this community white paper has demonstrated that it can be and is being done. A project called Q2O, QARTOD to OGC (Open Geospatial Consortium), bridges the QARTOD community with the OGC community to demonstrate and document best practices in the implementation of QA/QC within the OGC Sensor Web Enablement (SWE) framework. This paper describes this demonstration project and documents the existence of parallel related efforts. With adequate funding to enable the strengthening and broadening of these communities, a solid foundation for ocean observing systems will be built with the assurance that best-practices of data quality are communicated in a meaningful way.

1. INTRODUCTION

Over the past several years, as ocean programs have moved from sampling to long-time observations, funding agencies have stressed the need to move data

from local servers to integrated systems, enabling interdisciplinary research and re-use of data across political and disciplinary boundaries. Towards that end, various groups across the globe have begun work towards the development of common methods which will enable the discovery and transport of common ocean observing parameters. This paper describes the capability to also discover how a property in the ocean was sensed and processed into an oceanographic observation: What sensor was used? What QC tests were applied to the data stream? What measures were taken to assure good data quality? What events are relevant to the observation? If a data quality test was done, what flagging convention was used to reflect suspect data? Data quality is of primary concern when building trust in shared data. Without the means to discover and set limits on data quality for varied needs, ocean observing systems will not be built upon a strong foundation.

The Global Earth Observation System of Systems (GEOSS) is currently organizing a global effort towards the integration of earth observing systems, which must encompass QAQC. Thus, the GEOSS Common Infrastructure Initial Operating Capability task force has just released a provisional set of requirements for GEOSS registered components and services (also currently categorized as best-practice) among which content quality management is required under the form of documentation. Accuracy assessment and reporting of measurements uncertainty is essential to assure data products consistency and interoperability, implying that the instrument calibration and product validation need to be continuously monitored and traceable to standards [1].

As data are transported from origin (the sensor) to a data provider and on to an aggregation center (which may also serve as a data provider), knowledge of the data provenance, system configuration, and information about what has or has not happened to these data during processing may be lost. Through the development of relatively easy to implement, community-adopted tools and frameworks, this knowledge needn't be lost to the community and can be updated at each level of exchange. Information about data quality can be used to assess data and may also be used to notify a data provider of problems that require action, helping to assure a reliable stream of good data.

With support from the U.S. National Oceanic and Atmospheric Administration (NOAA), a team, now called Q2O (QARTOD to OGC), was formed to integrate QARTOD (Quality Assurance in Real Time Oceanographic Data) recommended quality control best practices into the OpenGeospatial Consortium standards (OGC), specifically the Sensor Web Enablement (SWE) framework. The Q2O project bridges domain experts in ocean observing systems with information technology experts.

QARTOD, a grassroots organization, currently funded through NOAA, has convened four times over the past five years. It has brought together private and governmental interests, with participants including data managers, scientists and sensor manufacturers, and has made significant strides in defining minimum requirements in QAQC for four oceanographic domains: waves, *in situ* currents, CTD (Conductivity-Temperature-Depth) and dissolved oxygen.

The OGC approach is a consensus-based development of publicly available standards to "geo-enable the web", providing a proper framework for a network of ocean observing systems. Because it is standards-based, a rapid, broad-based growth in interdisciplinary tools can result from demonstration projects. The selection of Sensor Observation Service (SOS), as part of the Sensor Web Enablement (SWE), seems a perfect match for *in situ* sensor networks, because it is specifically designed to describe and publish observation components, including the details of the processes involved in such observation [2]. SensorML (Sensor Model Language) can describe in detail how a physical phenomenon, such as sea-water pressure or particle movement, is transformed into an observation, such as wave height.

Sensor Observation Service (SOS) is an OGC standard for the retrieval of data and metadata from sensors and sensor systems, utilizing SensorML encodings. Whether from *in situ* sensors (e.g. water monitoring) or dynamic sensors (e.g. satellite imaging), measurements made from sensor systems contribute most of the data by volume used in geospatial systems today. The SOS is the intermediary between a client and an observation repository or near real-time sensor channel and is one piece of the larger OGC SWE initiative, enabling all types of sensors to be accessible and, where applicable, controllable via the web.

There are three "core" required SOS operations. GetObservation provides access to sensor observations and measurement data via a spatio-temporal query that can be filtered by phenomena. DescribeSensor retrieves detailed information about the sensors and processing. GetCapabilities provides the means to access SOS service metadata, providing a client with information about what is available.

The Martha's Vineyard Coastal Observatory (MVCO), owned and operated by the Woods Hole Oceanographic Institution (WHOI), provided the test bed for the first part of the Q2O project, returning real-time, quality controlled SWE offerings of waves. Wave parameters are computed using an acoustic Doppler current meter, deployed at the 12m isobath, continuously measuring pressure and horizontal

velocity at 2 Hz. The SensorML instances and SOS offerings describe the sensor characteristics, system provenance and lineage, and the computational steps used to derive wave height parameters. Quality control tests recommended by the QARTOD Waves Team were implemented with the test results reported through the SWE offerings. Multiple GetObservation offerings were encoded, each generated from one data stream and the associated SensorML process chain. Examples include offerings which provide all the data or one reporting only data that has a "passed" value for a data quality test. Vocabularies defining QC tests, processing methods and input, output and parameters for each of the tests have been registered with the Marine Metadata Interoperability (MMI) registry, enabling resolvable definitions to be linked within the SensorML instances. A paper recently presented at the IEEE (Institute of Electrical and Electronics Engineers) Oceans'09/Bremen conference presents the implementation in more detail [3].
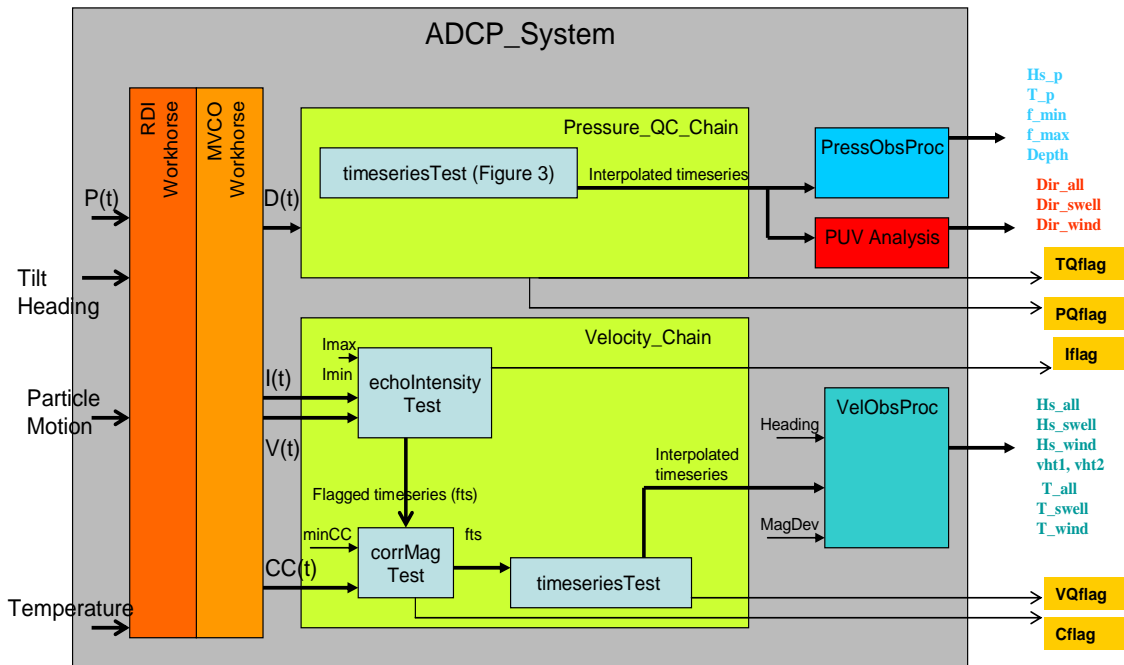

Figure 1. Each system documents input to the system, a description of the components (both the sensor and processing components) and its output.

The systems described are built to enable machine-to-machine interoperability by providing both syntactic and semantic standards in accessing the observations and the information about the system, from observables to system output.

2. DEFINING YOUR SYSTEM

In this section, we will briefly describe how SensorML describes a system, using the MVCO waves implementation as an example. We will also demonstrate the purpose of the SensorML files, as they relate to enabling the web services (SOS). This is the core of what allows a system to explore data quality processes taken with associated input parameters:

What tests were done? What flags relate to which tests? What input parameters were used for each test?

In SensorML, all components are modeled as processes. This includes components normally viewed as hardware (e.g. transducers, processors), and sensors and platforms (which are modeled as systems). The building blocks of a SensorML description include: ProcessChain, System, ProcessModel and Component.

Processes in SensorML are conceptually divided into two types: (1) those that are physical processes, such as detectors, actuators, and sensor systems, where information regarding their positions and interfaces may be relevant, and (2) non-physical or "pure" processes, which can be treated as merely mathematical operations. These may include processes that have been applied to an observation (i.e. observation lineage) or can be applied on an observation (i.e. on-demand processing).

ProcessModel is used to define more or less atomic pure processes that are expected to be used within more complex process chains. An example is the minimum threshold test used by several of the QARTOD recommended time series tests. ProcessChain defines a collection of processes that are executable (or describe an executable) in a sequential manner to obtain a desired result. For example, the timeSeriesTest in MVCO ADCP_System (Acoustic Doppler Current Profiler) (Fig. 1) is composed of a number of processes: range check, a spike test and a minimum percent good test, minPointsGood, as seen in Fig. 2.

Any physical process can be modeled as a Component in SensorML, either if it cannot be subdivided into smaller sub-processes, or if one chooses to treat it as a single indivisible process. A Component can be considered as a real-world equivalent of a ProcessModel. As an example, VelObsProcess is treated as a single component. It is registered as an MVCO term with links to bibliographic references [4], which describe the computation of wave height from velocity in terms well understood by the waves community.

System is a physical equivalent of a ProcessChain. A System may include several physical and non-physical processes that all act to provide a certain set of System outputs, based on the System inputs and parameters. The MVCO ADCP_System is comprised of sensors and processing descriptions.

Figure 1 graphically represents the SensorML, as it was implemented for the MVCO waves instance. ADCP_System defines the relationship of the components, such as the RDI_Workhorse, containing the generic T-RDI ADCP Workhorse metadata and the MVCO_Workhorse, describing metadata specific to the MVCO implementation, along with the components describing the QC processes and computational methods used to calculate wave heights.
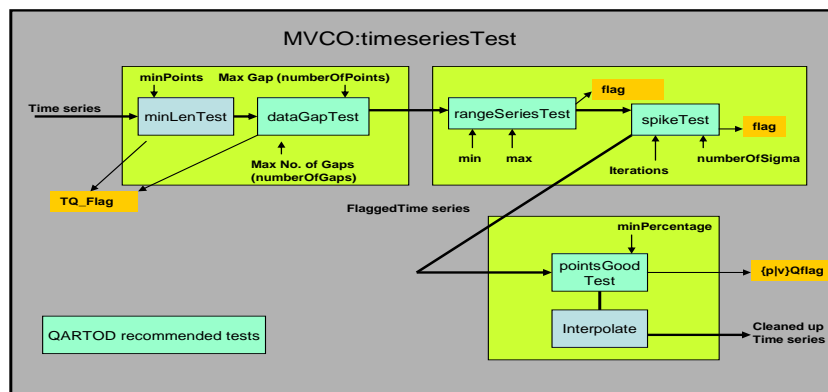

Figure 2. The timeseriesTest is a ProcessChain composed of several QARTOD recommended tests, such as data GapTest; range SeriesTest, and spikeTest. Each test has defined input (velocity or pressure), parameters and output, along with associated QC flag(s). Examples of parameters include: a minimum and a maximum for the range SeriesTest or number of the standard deviations from the mean and number of iterations for the spikeTest.

Flags are used to convey data quality and are defined in a code-space which associates a flag's value with a particular definition. This enables mapping of QC flags with the same semantic meaning, even though they may have different numeric values, by using ontologies [5], as shown in Fig. 3.

For the MVCO ADCP_System, data are generated by the coastal observatory as ASCII (American Standard Code for Information Interchange) comma separated flat files. This makes it relatively easy for a local observatory to edit configuration files describing their system, while generating data in real-time separate from these content rich SensorML files. Data are "wrapped" in SensorML encodings, utilizing the ADCP_system SensorML ProcessChain, through the core operations of the Sensor Observation Service. The GetObservation operation can report the latest observation or respond to a time-based query (reducing the amount of XML (Extensible Markup Language) content required to report data records). The data returned from a GetObservation request are in the form of a sequence of "tuples", and may be either ASCII or binary. The DescribeSensor operation can describe the full system, including all the hardware components and the processing lineage used to generate the data. These capabilities enable the use of common technologies for

real-time systems, as well as archived data. SWE also provides the capability to add encodings at a server site, enabling the integration of sensor web technologies for bandwidth-limited sensing systems.

3. DEFINING YOUR TERMS

An important step towards achieving interoperability between different ocean observatory systems is to use a basic set of terms to describe the collected data. A number of vocabularies already exist. However, they have been defined having different applications in mind. For instance there are definitions on the format of electronic data sheets that are described with the standard IEEE 1451 or the CanOpen protocol that address similar specification needs but are not identical in format. A harmonization between these arrangements appears necessary for instance by defining a basic set of terms that can be easily mapped between the different standards. Furthermore it has to be agreed upon what reference document should be used in regard to describing the outcome of measurements. In regard to a unique description of the measurement process, an ISO (International Organization for Standardization) standard exists that resulted from an agreement between different national standards laboratories- the ISO/IEC (International Electrotechnical Commission) Guide 98-3:2008 Guide to the expression of uncertainty in measurement. It has to be evaluated whether this standard could also be used as a base for harmonizing vocabularies in the ocean observation realm.

Semantic mediation is a required mechanism to enable system interoperability and data integration. This mechanism comprises a set of key operations including controlled vocabulary definition and maintenance, terminology mappings, and inference, among others. The Marine Metadata Interoperability Project, MMI, is a community-based organization advancing a semantic framework that incorporates these various operations in an integrated system, the MMI Ontology Registry and Repository, ORR. This framework is complemented by a large body of guidance materials developed by MMI staff and participants, which includes tutorial and best practices documents for both data providers and data users.

Q2O and MVCO controlled vocabularies were created using the MMI Voc2RDF Web tool. Voc2RDF greatly facilitates the creation of controlled vocabularies by using a simple and intuitive form-based user interface. The terms were then registered in the MMI repository, ORR, providing machine-to-machine and/or human readable URLs for the Q2O and MVCO registered vocabularies.

The Vocabulary Integration Environment tool, VINE, allows the user to map terms across multiple vocabularies. Various relations can be established between the terms, for example exact, narrow, and related matching relations, as defined by the Simple Knowledge Organization System specifications, but others can be used.

4. NEXT STEPS

Over the next two years, the Q2O team will continue efforts to encode QAQC into SWE and to develop guidelines and best practices for implementers. We have begun work on *in situ* current offerings from an ADCP. In fall 2009, we will begin work on dissolved oxygen sensors, continuing into 2010. We have begun work to demonstrate real-time (on-the-fly) processing using SensorML to apply range checks on the wave parameters, as recommended by QARTOD.

Integration of the technology solutions described in this paper into tools that expedite publishing of Sensor Observation Services is planned. In particular, the solutions will be incorporated in the toolkits developed by the OOSTethys project. OOSTethys is an open source collaborative project, lead by SURA (Southeastern Universities Research Association) and mainly advanced by the SURA Coastal Ocean Observing and Prediction (SCOOP) Program and the MMI. The OOSTethys community has developed reference implementations (open source, standards-compliant software components) in JAVA, PERL (Process and Experiment Automation Real-Time language) and PYTHON with detailed how-to guides and provides a test bed for testing end-to-end software system components. The MVCO PERL reference implementation will be updated to handle QA/QC SWE components and posted with a guide at OOSTethys. The OOSTethys client, OpenIOOS is also being adapted to provide a searching functionality, which will allow discovery of data collections taking into account the QA/QC information. These standards are being adopted globally. The Spanish Research Council and the Politecnics University of Catalonia (Spain) have adopted the Sensor Web Enablement for their oceanographic vessels and observatories data management systems.

Currently, there are efforts underway to enable integration of NetCDF/OpeNDAP (Network Common Data Form/ Open-source Project for a Network Data Access Protocol) data systems into the SWE framework. The Q2O project will investigate whether and how QA/QC can also be transported through these technologies by working with the OOSTethys transport team, who is already providing NetCDF-to-SWE services.
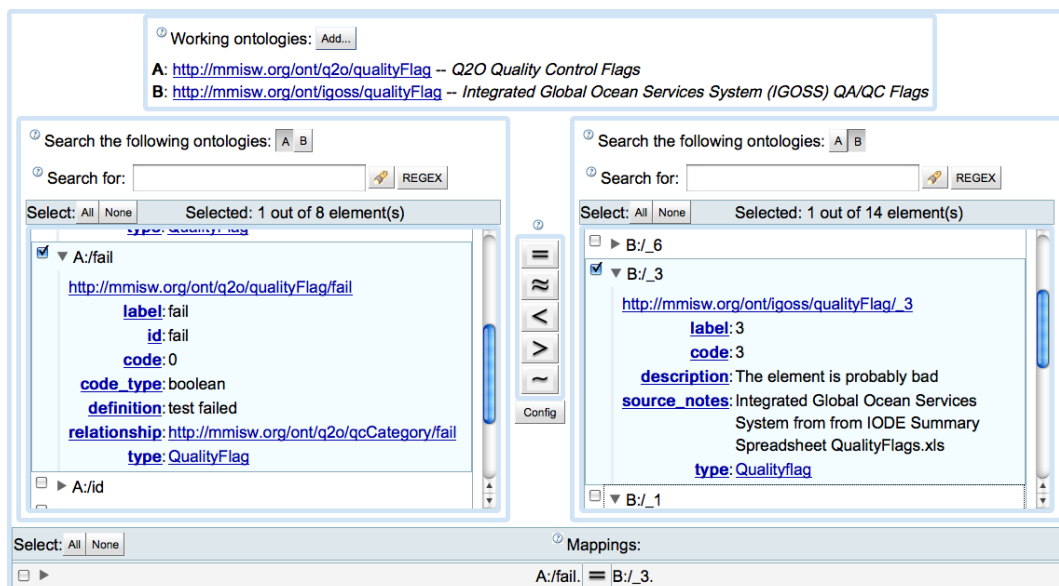
Figure 3. Using MMI Vine tool to generate relationship between Q2O Boolean QC flag and IGOSS (Integrated Global Ocean Services System) flagging convention.

5. COMMUNITY DEVELOPMENT

As development of these capabilities progress, its utility can be expanded to meet the needs of the evolving ocean observing community. The NOAA Integrated Ocean Observing System (IOOS) Program Office has recently adopted and begun implementation of the OGC SWE framework for several core variables. The U.S. National Science Foundation Ocean Observatories Initiative (OOI) has the need to address data processing and QAQC within the data management infrastructure. As part of its mission, the recently renewed OGC Oceans Interoperability Experiment, Oceans-IE, also defines integration of data quality and control as an objective. OGC based standards adoption encourages the development of common tools and interfaces, enabling interdisciplinary access to data. Through the development of an active, international community developing and adopting standards in data processing and sensor related best-practices, well-defined expectations in data quality will evolve. On the semantic front, the existence of a preliminary MMI repository is promising, but this can be only advanced with more resources and by incorporating community vocabulary development and maintenance into the mission. In the UK, a Collaborative Ocean initiative has been set up in National Oceanography Centre, Southampton (NOCS) to work on technologies that support semantic web management activities in the oceanography domain. This includes marine ontology management, semantic annotation, storage, and reuse (inter-connect, navigate,

inference and query) of semantically enriched oceanographic resources such as sensors and datasets. The semantic web infrastructure will allow organizations such as BODC (British Oceanographic Data Centre) and MMI to apply their standards in a real environment where resources need to be semantically enriched to improve data provenance and interoperability. Once the method and infrastructure are adopted by more oceanographic projects to describe their resources and move their metadata onto the semantic web, enhanced interoperability capabilities will facilitate multiuse of ocean data, as defined in Oceans 2025 [6], as well as make more Earth Observation data available on the semantic web [7].

Projects such as Q2O bridge the expertise of the domain specialists and can build a community of trust and also move the developing standards forward. The ability to map terms mandates a knowledgeable community engaged in developing meaningful ontologies, all pointing to the need for an international MMI-type initiative.

There is this unique opportunity to combine the individual national efforts to a world-wide initiative on ocean observations. Argo already demonstrated that this is possible. It is mainly a matter of identifying domains where the need is obvious to have an international cooperation. Surely quality assurance and control belongs into that category. This general idea can not be implemented as a whole. However, as QARTOD suggests certain parameters shall be selected

that have a high importance for instance in the framework of judging about the impact of climate change on the ocean environment. This is the case for carbon dioxide measurements and according quality management procedures are in place. It has been suggested to include the parameter dissolved oxygen into a priority list that has high relevance for the integrity of the ocean biosphere [8].

To bring this content-rich framework into individual ocean observatories, internationally adopted models for communicating the details of our sensor systems must be agreed upon, engaging not only information technologists but also domain specialists. International, interdisciplinary workshops can facilitate the harmonization of parallel efforts, broadening the value of each and expediting the development by sharing the tasks at hand. Once common models are adopted, resources must be allocated for the development of tools to facilitate the implementation of the framework for any observing system, large or small. Just like common tools are available for web documents (html), instrument manufacturers and systems developers can build tools which will enable the integration of web services with specifics of the individual system. Training and international workshops for sensor manufacturers and data managers will assure the development of a global system of ocean observations with a solid foundation of quality data.

REFERENCES

1. Ungar, S., Campbell, P., et al. (2007). Data Quality Guidelines for GEOSS Consideration. The CEOS Working Group on Calibration and Validation (WGCV), Geoscience & Remote Sensing Symposium, IGARSS, IEEE International.

2. Bermudez, L.E., Bogden, P., Bridger, E., Galvarino, C., Graybeal, J., Forest, D., Cook, T., and Creager, G. (2009). Web Feature Service (WFS) and Sensor Observation Service (SOS) comparison to publish time series data. Presented at International Symposium on Collaborative Technologies and Systems - Workshop on Sensor Web Enablement Baltimore, Maryland.

3. Fredericks, J.J., Botts, M., Bosch, J. (2009). Integrating Standards in Data QA/QC into OpenGeospatial Consortium Sensor Observation Services, IEEE Xplore Oceans2009.

4. IAHR Working Group on Wave Generation and Analysis (1989). List of Sea State Parameters. *Journal of Waterway, Port, Coastal and Ocean Engineering,* **1156**, 793-808.

5. Noy, N.F. & McGuinness, D.L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Knowledge Systems Laboratory.

6. Oceans 2025, *www.oceans2025.org*.

7. Tao, F., Campbell, J., Pagnani, M. and Griffiths, G. (2009). Collaborative Ocean Resource Interoperability: Multi-Use of Ocean Data on the Semantic Web. In, *The 6th Annual European Semantic Web Conference (ESWC2009).* Heidelberg, Germany, Springer. (Lecture Notes in Computer Science).

8. Eds: Rabalais, N.N. & Turner, R.E. (2001). Coastal Hypoxia: Consequences for Living Resources and Ecosystems, American GeUnion, *Coastal and Estuarine Studies*, **58**.