Integrating Standards in Data QA/QC Into OpenGeospatial Consortium Sensor Observation Services

J.J. Fredericks
Woods Hole Oceanographic Institution
Woods Hole, MA 02543 USA

Mike Botts, Tony Cook
University of Alabama in Huntsville
National Space Science and Technology Center
Huntsville, AL 35805 USA

Julie Bosch
National Coastal Data Development Center, NOAA
Stennis Space Center, MS 39529 USA

***Abstract*- As systems developers build the tools and methods for transport and discovery of marine data, we must carefully consider what information must accompany these data and how it is transported in order to enable adequate assessment and understanding in a machine-to-machine global exchange. Providing a common framework to communicate the history of a sensor, data processing and processing results can create a shared understanding which will build a solid foundation for the development of a trusted system of systems for sharing global environmental data. In this paper, we will present an oceanographic wave height offering, using the Open Geospatial Consortium, Inc (OGC) Sensor Web Enablement (SWE) framework, which documents not only the sensor characteristics, but also the processing steps and an assessment of the data quality based on recommended quality control tests.**

I. Introduction

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 source, system configuration, data provenance and information about what has or has not happened to these data during data 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 added to 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 [http://q2o.whoi.edu]), was formed to integrate QARTOD (Quality Assurance in Real Time Oceanographic Data [http://qartod.org]) recommended quality control best practices into the OpenGeospatial Consortium standards (OGC [http://opengeospatial.org/ogc]), specifically the Sensor Web Enablement (SWE) framework (http://www.ogcnetwork.net/SWE).

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 QA/QC for four oceanographic domains: waves, *in situ* currents, CTD and dissolved oxygen. Figure 1 provides a sample of some of the QARTOD recommendations.

TIME SERIES (Raw Calibrated Data)

Category	Criteria	Order	Flag	Action
Data Gaps	Consecutive N missing data. Maximum number of missing data.	1	Soft	N is user defined. Include in % count.
Spikes	User defined Points >= M*std with P iterations	2	Soft	Interpolate/extrapolate up to N points. N is user defined. M can be user defined, recommended M=4. Include in % count.
Range test	Location, instrument defined.	2	1. Soft 2. Hard	Max/min user defined. 1. Interpolate/extrapolate up to n points. N is user defined. Include in % count. 2. Instrument spec exceeded, reject.
Mean shift (segments)	A mean shift "P" occurs in this time series.	3	Hard	Reject entire record. P is user defined.
Acceleration test	User defined (a>M*g)	3	Soft	Recommended M<=1/2. Interpolate/extrapolate up to N contiguous points. N is user defined. Include in % count.
Mean test, variance test	User defined, location dependent	4	1. Soft 2. Hard	1. Flag unexpected values. 2. Reject unreasonable values.
Percent points good	Check for M% good data (based on above 6 criteria)	5	Hard	Recommended M>=90%

Figure 1. QARTOD recommended time series tests for waves processing from the Coastal Data Information Program (CDIP) website: http://cdip.ucsd.edu/documents/index/product_docs/qc_summaries/waves/waves_table.php

The OGC approach, a consensus based development of publicly available standards to "geo-enable the web," enables the development of a broad range of tools by many developers because it is standards based. The selection of SWE, and specifically Sensor Modeling Language (SensorML), seemed a perfect match for *in situ* sensor networks, because it is specifically designed to describe how observable properties (such as water velocity or pressure) are transformed into an observation (such as wave height).

Sensor Observation Service (SOS) is an OGC standard which enables retrieval of data and metadata from sensors and sensor systems. 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 geospatial 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. Clients can also access SOS to obtain metadata information that describes the associated sensors, platforms, procedures and other metadata associated with observations.

Sensor Observation Service (SOS) is one piece of the larger OGC Sensor Web Enablement (SWE) initiative. The goal of SWE is to enable all types of Web-accessible sensors to be accessible and, where applicable, controllable via the Web.

SOS has three mandatory "core" operations: GetObservation, DescribeSensor, and GetCapabilities. The GetObservation operation provides access to sensor observations and measurement data via a spatio-temporal query that can be filtered by phenomena. The DescribeSensor operation retrieves detailed information about the sensors and processes generating those measurements. The GetCapabilities operation provides the means to access SOS service metadata.

The Martha's Vineyard Coastal Observatory (MVCO [http://www.whoi.edu/mvco]), owned and operated by the Woods Hole Oceanographic Institution (WHOI), provided the test bed for the first part of the Q2O project, returning the GetCapabilities, DescribeSensor and GetObservation responses for real time offerings of waves every twenty minutes. Wave parameters are computed using an acoustic Doppler current meter, deployed at the 12m isobath continuously measuring pressure and horizontal velocity at 2 Hz. SensorML instances and SOS offerings were developed, describing the sensor characteristics, system provenance and lineage, and the computation of the derived wave height parameters. Quality control tests recommended by the Waves Team of QARTOD (Fig. 1) were implemented and reported through the SWE offerings.

The implementation of the SOS GetCapabilities, DescribeSensor and GetObservation methods demonstrate a means for discovery of the sensor characteristics and lineage, the processing steps taken to produce the observations from the sensor output, and serve the observational data along with the QA/QC flags associated with the various tests implemented at the MVCO. Multiple offerings exist, each generated from one data stream and the associated SensorML process chain. Examples include GetObservation offerings which provide all the data, or one that only provides waves computed from pressure, or only waves computed from pressure and velocity (PUV), or only data that has a "passed" value for the relevant data quality flag in its corresponding observation. Terms and definitions for QC tests, processing methods and input, output and parameters for each of the tests have been developed and registered, enabling resolvable definitions to be linked within the SensorML instances.

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.

II. Defining Your System

Each system is composed of a number of SensorML files, including one that defines the relationships within the system. That top-level SensorML file (ADCP_Chain, in Fig. 2) links sensor and lineage descriptions, the process components (each represented by its own SensorML file), and the input and output of each of the processing steps. Data are created as csv text files and wrapped in the content-rich SensorML, as part of the return from GetObservation.

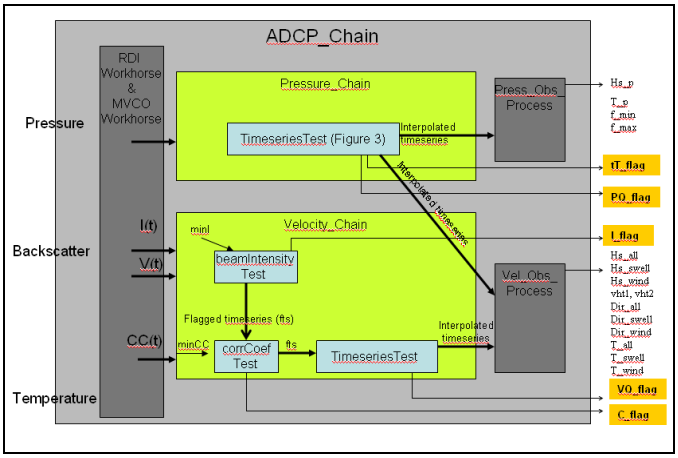

Figure 2. Each system documents input to the system, a description of the components (both the sensor and processing components) and its output.

In SensorML, all components are modeled as processes. This includes components normally viewed as hardware, including transducers, actuators, and processors. The building blocks of a SensorML description are ProcessChain, System, ProcessModel, and Component. ProcessChain and ProcessModel refer to non-physical composite and atomic processes, respectively. System and Component refer to actual physical instrumentation, where System is a composite system and component is an atomic sensor.

Below is a list of SensorML documents used to describe the MVCO system as shown in Figures 2 and 3.

- ADCP_Chain - The main ProcessChain, which pulls everything together to describe the system.

Sensor Descriptions

- RDI_Workhorse_1200.xml – A general SensorML description of the TRDI instrument, describing the sensor and its capabilities. It points to references from the manufacturer and can be used by anyone who is using an TRDI_Workhorse 1200.

- MVCO_Workhorse_1200.xml – SensorML description entailing specific details about the data, and the ProcessModels that operate on individual data points. It describes the set up at the MVCO and specifies particulars, like sampling frequency, reporting frequency, and burst length. It also refers to operational points of contact and time-stamped events that occur which may affect the quality of your observation (like a failed pressure port and its replacement, or a cleaned ADCP face).

The general tests include:

- TimeContinuityTest - Test to make sure data meet time continuity requirements. This test operates on a time series of data.

- RangeSeriesTest - Test to determine if a data point lies between an upper and lower bound, operating on a data series

- RangeTest - The atomic ProcessModel for range checking a single point

- MinThresholdSeriesTest - Like RangeTest, but only operates on a lower bound, operating on a data series.

- MinThresholdTest - The atomic ProcessModel for testing if a data value exceeds a lower bound

- SpikeTest - ProcessChain for a SpikeTest

- InterpLinear – The atomic Process for interpolating over-flagged values in a time series of data

The Process Chain includes:

- Pressure_QC_Chain - General ProcessChain for Pressure QC on time series before computing waves statistics

- Velocity_QC_Chain - General ProcessChain for Velocity time series data before computing waves statistics

- Pressure_QC_Chain_Values - ProcessChain for Pressure time series data with parameters (values) configured for MVCO setup (how we get the parameters such as min/max pressure values in the range check, etc.)

- Velocity_QC_Chain_Values - ProcessChain for Velocity time series data with parameters (values) configured for MVCO setup

- Pressure_Obs_Process - Chain that generates a number of observable properties (waves from pressure record, using linear wave theory [1]) using the cleaned, interpolated time series that is output from Pressure_QC_Chain

- Velocity_Obs_Process - Chain that generates a number of observable properties (waves from PUV analysis [2]) from the cleaned, interpolated time series that is output from Velocity_QC_Chain

- TimeSeriesChain - ProcessChain composed of several individual processes that perform time-related QC checks on the Pressure and Velocity Series (Figure 3)

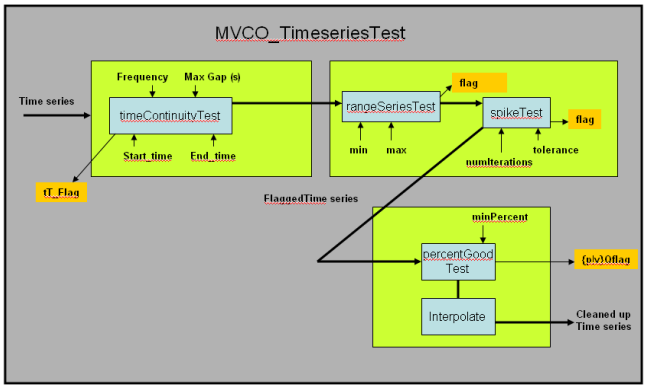

Figure 3. Each test must be defined in SensorML with input and output and test parameters and output Boolean flags, when appropriate.

The process components must each define inputs, associated processing parameters, and outputs with QC flags generated by the recommended tests. For example, the MinThresholdSeriesTest has input of a time series, either velocity or pressure in the MVCO instance. It also has an associated parameter, the minimum percent good allowed to pass the QC test; and it has output of the original flagged data and a flag (pass or fail) indicating the Boolean result of the test. The output of this process (PercentGood test) is the input to InterpLinear.

Each of the process components describes something that happens to the observations. The ability to harvest the information that describes the sensor characteristics and its lineage, the processing steps, the input parameters associated with the processing, and QC flags relating to the tests enables a system aggregating the observations to evaluate the data quality and what further processing may be required.

Flags are used to convey data quality. Flags are defined in a code-space which associates a flag's value with a particle definition. In the Q2O Project, QC flags are based on a simple pass or fail result of a test or series of tests, and the pass or fail flags are encoded by a 1 or 0, respectively. Compared to many QC flag sets for oceanographic data, this is only a basic reference, but it is designed to be related to the specific tests and test parameters, as they relate to real time data. The evaluation of how the test results may reflect the overall quality of the data can be defined through program or project specifications.

For example, the ARGO Program uses the following quality flags:

0 = No quality control was performed
1 = QC was performed; good data
2 = Probably good data,
but value may be inconsistent with statistics
3 = Probably bad data
(spikes, gradients,…if other tests passed)
4 = Bad data, impossible values (out of scale, …)
5 = Value modified during quality control
8 = Interpolated value
9 = Missing value

The Q2O flags can be related to those broader flag sets through the development of relationships mapped in ontologies [3]. Through this development, the Q2O flag for "pass = 1" can be mapped as equivalent to the ARGO flag "1 = QC was performed; good data" and may be mapped as closely similar to "2 = Probably good data, but value may be inconsistent with statistics." The use of codespace offers the opportunity to automate the data quality mapping, where common definitions are used, e.g., the MVCO failed test can be mapped to the Q2O failed test value through ontology.

Additonal SWE protocols include the Sensor Planning Service (SPS), which defines an API for tasking Sensor Systems. The Sensor Alert Service (SAS) is an API for publishing and subscribing to alerts from sensors, for example a failed test. The ultimate vision of SWE is to define and approve the standards foundation for "plug-and-play" Web-based sensor networks. Used in conjunction with these other OGC specifications, the SOS provides a broad range of interoperable capability for discovering, binding to and interrogating individual sensors, sensor platforms, or networked constellations of sensors in real-time, archived or simulated environments.

III. Defining Your Terms

Each term referenced in the SensorML, from the input observables to the output definitions, should reference a meaningful, resolvable definition. The initial Q2O and MVCO vocabularies are registered at the Marine Metadata Interoperability Project (MMI [http://marinemetadata.org]) vocabulary registry (http://mmisw.org/or). The registry provides a unique resolvable Uniform Resource Locator (URL) for each term. Existing terms and definitions are referenced wherever possible. We have encouraged manufacturers to register the terms that define their sensors and instruments, and have created descriptions to provide guidance on content and style. Caution must be taken when utilizing existing terms to be assured of consistency in meaning. An example of the misunderstanding that can occur with conflicting terms is as follows: the registered climate forecast (cf) definition of water pressure includes a definition in db, while the output of the MVCO system is in cm. This points out the need to determine whether it is appropriate to have units of measure in a definition, when they can be defined in the implementation. However, the main point here is to make certain your definitions really match your implementation.

In developing the Q2O and MVCO vocabularies, an attempt was made for each term to include the same components, listed below:

> Id, LongName, ShortName, Definition, Symbol, Reference, Figure(s), Category, Relationship(s), Equation(s), Note(s)

Each component may be populated, but some may not apply. However, those that are essential are the ID, ShortName, LongName and Definition. Some components may contain links to URLs, such as links to figures or equations.

The 'relationship(s)' component is included with the terms to assist in further relationship (ontology) building. The MMI is developing tools to assist in implementing these capabilities and is engaging the oceanographic community to build meaningful mappings amongst the vocabularies.

The Q2O team developed vocabularies which define the QARTOD recommended tests as well as input parameters, QC flags and bibliographic references that may be required to fully describe its process chain. The use of some of these is demonstrated in the MVCO instance. For example, the definition of the Vel_Obs_Process is specific to MVCO. It defines how the significant wave height (Hs) is computed at the MVCO. However, this definition includes a link to a URL that is registered as a Q2O reference. In particular, it points to the IAHR [2] paper defining how MVCO computes Hs.

For each of our sets of vocabularies, we set up categories which are reflected in the URLs, once they are registered. Table 1 and Table 2 demonstrate this for the MVCO instances and the Q2O common definitions, respectively.

TABLE 1. MVCO Categories

Category ***An Example*** **URL**
Platform MVCO, 12m Node http://mmisw.org/ont/mvco/platform/12mNode
Sensor ADCP http://mmisw.org/ont/mvco/sensor/ADCP
Property Seawater_pressure, beamIntensity, … http://mmisw.org/ont/mvco/property/beamIntensity
Test checkBeamIntensity http://mmisw.org/ont/mvco/test/checkBeamIntensity
Parameter minimumBeamIntensity http:// …/parameter/minimumBeamIntensity
Flag beamIntensityFlag http://mmisw.org/ont/mvco/flag/beamIntensityFlag
Process Velocity_Obs_Process http://mmisw.org/ont/mvco/process/VelocityObsProcess

TABLE 2. Q2O Categories

Category ***An Example*** **URL**
Test *percentGoodTest* http://mmisw.org/ont/q2o/test/percentGoodTest
Parameter minimumPercentage http://mmisw.org/ont/q2o/parameter/minimumPercentage
Flag Pass mmisw.org/ont/q2o/flag/pass
Reference IAHR_1989 http://mmisw.org/ont/q2o/reference/IAHR_1989

IV. Accessing The MVCO Waves Instance

The released versions are available from the Q2O website Activities/Deliverables page: http://q2o.whoi.edu/node/116. From the Q2O site, the vocabularies can be browsed and the SensorML files viewed directly, or by using the SOS methods:

GetCapabilities provides system metadata and notifies a client of the observation offerings available. The MVCO instance provides six offerings all from the listed set of SensorML files and one data stream.

DescribeSensor provides sensor characteristics, provenance and lineage, including links to processing and parameter SensorML.

GetObservation for each of the offering provides the data along with a description of the data.

Some offerings include all data, including data that have not passed QA/QC, and the flags that triggered the QA failure. Other offerings include only data that have passed all QA/QC, and do not include the flags in the data stream. The former can be utilized in house by the data provider to trace where data have not met QA/QC, while the latter may be more suitable for public dissemination. How to group and offer the data is at the full discretion of the DataProvider, and not limited by the SOS architecture.

UAH has developed a basic web application, PrettyView for displaying SensorML files in a tabular form (http://vast.uah.edu/SensorMLforms/upload.jsp). This application is still in beta, but supports most of the SensorML constructs in its current form. In addition, the SensorML Editor is a beta application for assisting humans in the construction of SensorML documents:

http://vast.uah.edu/index.php?option=com_content&view=article&id=149&Itemid=103

V. Next Steps

Over the next two years, the Q2O team will continue efforts to encode QA/QC into SWE. We have begun work on *in situ* current offerings from an ADCP. In fall 2009, we will begin work on dissolved oxygen and conductivity, temperature and depth (CTD) sensors, continuing into 2010.

The MVCO instance describes processing of the ADCP time series data for the waves processing. Currently, other data providers are working with the UAH members of our team to implement real-time (on-the-fly) processing using

SensorML to apply range checks on the wave parameters, as recommended by QARTOD.

Integration of these capabilities into the cookbooks of the OOSTethys/OpenIOOS project (http://www.oostethys.org) is planned. Through the OpenIOOS project, we will demonstrate mapping of different flagging conventions, as described above.

The NOAA IOOS Data Information Frameworks team has also implemented a version of SWE framework implementation for several key oceanographic parameters. Our products are demonstrations for implementation of QA/QC and are being reviewed by the NOAA IOOS office, with whom we meet periodically.

We are open to suggestion and opportunities to share our results. We will exercise our implementations and once confident in the model, will post the SensorML files and guidance on implementation.

VI. Conclusions

Through well thought out, community-based adoption of processing standards and best practices (QARTOD) and the development of common standards for the transport of relevant information (Q2O), a better understanding of data will be accomplished. The application of OGC SWE standards to real time marine data processing not only enhances existing global networks of data sharing through use of standard methods of transport, it also extends the value of long-term data sets by enabling providers to serve well documented sensor and processing history with their offerings.

Acknowledgments

Thanks to Sara Haines (University of North Carolina) and Eric Bridger (Gulf of Maine Ocean Observing System) for their participation in and contributions to this project. Also, we thank John Graybeal and the MMI staff for their guidance in development of our vocabularies and the development of the vocabulary tools and “The Registry.”

References

[1] Dean, R.G. and R. A. Dalrymple, 1984. *Water Wave Mechanics for Engineers and Scientists*, Englewood Cliffs, N.J., Prentice-Hall, Inc., 353 pp.

[2] IAHR working group on wave generation and analysis, 1989. “List of Sea State Parameters”. *Journal of Waterway, Port, Coastal and Ocean Engineering* 1156, pp. 793-808.

[3] Noy, N. F. & McGuinness, D. L. *Ontology Development 101: A Guide to Creating Your First Ontology*. Knowledge Systems Laboratory, March, 2001 (http://www.ksl.stanford.edu/KSL_Abstracts/KSL-01-05.html)