Fredericks and Botts *Open Geospatial Data, Software and Standards* (2018) 3:3
https://doi.org/10.1186/s40965-018-0048-5

Open Geospatial Data, Software and Standards

ORIGINAL ARTICLE Open Access

Promoting the capture of sensor data provenance: a role-based approach to enable data quality assessment, sensor management and interoperability

Janet Fredericks[1*] and Mike Botts[2]

Abstract

Sensor technologies and capabilities have an effect on observational data quality. Typically, data management begins, at best, when a data manager obtains the data and needs to describe it sufficiently to data consumers. Often, the sensing methods are not adequately described and the data manager does not know the appropriate questions to ask or where to direct questions about sensors, their configuration, and the deployment. Consequently, knowledge often remains buried in sensor manuals and field operator logs. Thus, most metadata requirements have been simplified to accommodate this gap in knowledge.

When information is captured where it is best understood and tools are created to easily capture this knowledge, machine-actionable descriptions can be provided to adequately describe the processes taken in generating observations. The information can be associated with the data and thus be accessible, discoverable and used in data quality control by data providers and in data quality assessment by the data consumers.

Here, we define actors and actions to promote role-based creation of fully-described, standards-based documents. These documents can be created in SensorML (OGC SWE) that includes links to resolvable term definitions (W3C Semantic Web), enabling the creation of associated mappings and ontologies to extend and resolve the meaning of each term.

Keywords: Sensor web, Data quality, Semantic sensor web, Provenance

Introduction

Since the inception of the world-wide web, geo-scientists have been putting data online for sharing and discovery. Machine-to-machine harvesting of data is enabled when data are sufficiently described in community-adopted standards frameworks. But with the discovery of data comes questions. To understand data, better metadata are needed. Most of the effort towards the generation of metadata has focused on who, what, when and where observations are made. However, little effort has been directed toward providing sufficient content to determine precisely how an observation was made. Information that would be meaningful in assessing discovered data would include sensor characteristics, sensor configuration that can significantly affect observational accuracy and precision, and sensor maintenance – or lack of maintenance - that can help explain otherwise inexplicable shifts or long-term trends in data. But how do we associate this information to the data and enable it to be discoverable and accessible in our machine-to-machine systems. We need to get this information out of the manuals and field logs and PDFs and into community-adopted, standards-based frameworks, which are machine-actionable.

The Open-Geospatial Consortium (OGC) Sensor Web Enablement (SWE) [1] provides a community-adopted framework that supports the ability to fully describe processes used in creating observations and the sensors used. It can also support the tasking of sensors and defines machine-actionable access to web-accessible observations. The OGC SWE standard Sensor Model

* Correspondence: jfredericks@whoi.edu
[1]Woods Hole Oceanographic Institution, MS#9, Woods Hole, MA 02543, USA
Full list of author information is available at the end of the article

Language (SensorML) provides a means to describe sensors and processes in machine-harvestable encodings [2]. In SensorML, all components are defined as processes. Some processes, such as sensors and actuators, are physical while others are computational. Additionally, these processes can be indivisible components such as a single detector or algorithm, or aggregate such as a complex sensor system or a process chain [3]. SensorML provides a general framework to describe such properties as inputs, outputs, parameters, components, capabilities and characteristics. It can also be used to describe procedures and history using its event list. The terms in SensorML are not domain specific, so tools and profiles can provide enhanced support for specific communities or sensor types. Because SensorML treats all components as processes, the over-arching SensorML model provides an appropriate framework that can describe the provenance of an observation including sensor characteristics, the deployment environment, as well as its data processing and quality control and assessment tests performed.

Background

In a NOAA-IOOS funded project called Q2O (QARTOD-to-OGC) [4], a SWE model was created to describe in SensorML all information needed in assessing data quality of an observation. The model describes the sensors, qc-tests, qc-flags and processing used to create derived products. A complete description of the content-rich model is described in [5]. From this activity, we recognized that the first step in the capture of the observational provenance is to encourage the manufacturer to describe the sensor in machine-harvestable SensorML. These descriptions must include capabilities (e.g., operational ranges), characteristics, input (observable properties), output (including units, accuracy and precision), as well as manufacturer contact information. The content can be used to enable automated QC, such as selection of appropriate precision and accuracy, as well as validating data using specified operational ranges. It is a small but significant step in automating the capture of observational provenance.

We aim to promote the best practice of creating descriptions of the "how this observation came to be" by those who best understand each step in creating the data. A sensor manufacturer knows best the information about a sensor model. When a manufacturer describes a particular sensor model, this description can be referenced by anyone who deploys that particular model of sensor. Furthermore, when the sensor manufacturer specifies a unique id within a SensorML document, a specific instrument can also be described, as built, and tracked to enable:

- Sensor inventory and management in large programs;
- Data quality resolution in situations where a particular sensor was used to create a composite data set and was found to have issues;
- Operational changes in calibration or maintenance (such as cleaning faces, replacing components, etc.)

A recent NSF EarthCube Integrative Activity called X-DOMES (Cross-Domain Observational Metadata for Enviro-Sensing) was funded to develop a community of participants and to build the necessary tools to enable role-based creation of SensorML with links to resolvable terms. The tools are designed to be configurable to meet the needs of different users (manufacturers, field operators, data managers) who are not experts in SensorML and the Semantic Web [6]. The project is engaging sensor manufacturers (providers) and data facilities (consumers) across the geo-sciences in order to assess their usability to fully describe and harvest descriptions using the tools. An editor/viewer was developed which facilitates the creation and viewing of SensorML (Fig. 1); a SensorML Registry & Repository was created to enable SensorML to be registered and reference-able (IRI); and, an Ontology Registry & Repository (ORR) [7] was created to enable easy creation of resolvable definition of terms to be embedded in the SensorML documents using the Online SensorML editor, via a SPARQL query [6].

In "Definition of role-based content creation" section, we describe the roles of content creators and how the content can be associated. "Results and discussion" section describes the role and significance of building communities to promote the adoption of this approach in a consistent manner. The use of the various components of software by different members of the community are illustrated in Fig. 2.

Methodology

Definition of role-based content creation

Sensor manufacturer role

Typically, the sensor manufacturer has the most comprehensive knowledge of how a sensor works and what is important in assessing data quality. By assuming the role of describing the sensor in SensorML, accurate and complete creation of content is more easily implemented and knowledge is captured where it is best understood. By creating this content in standards-based encodings and registering the content for access from an authoritative, freely and openly available service, anyone who purchases a sensor described by the original equipment manufacturer (OEM) can reference a vetted, community-adopted description of a

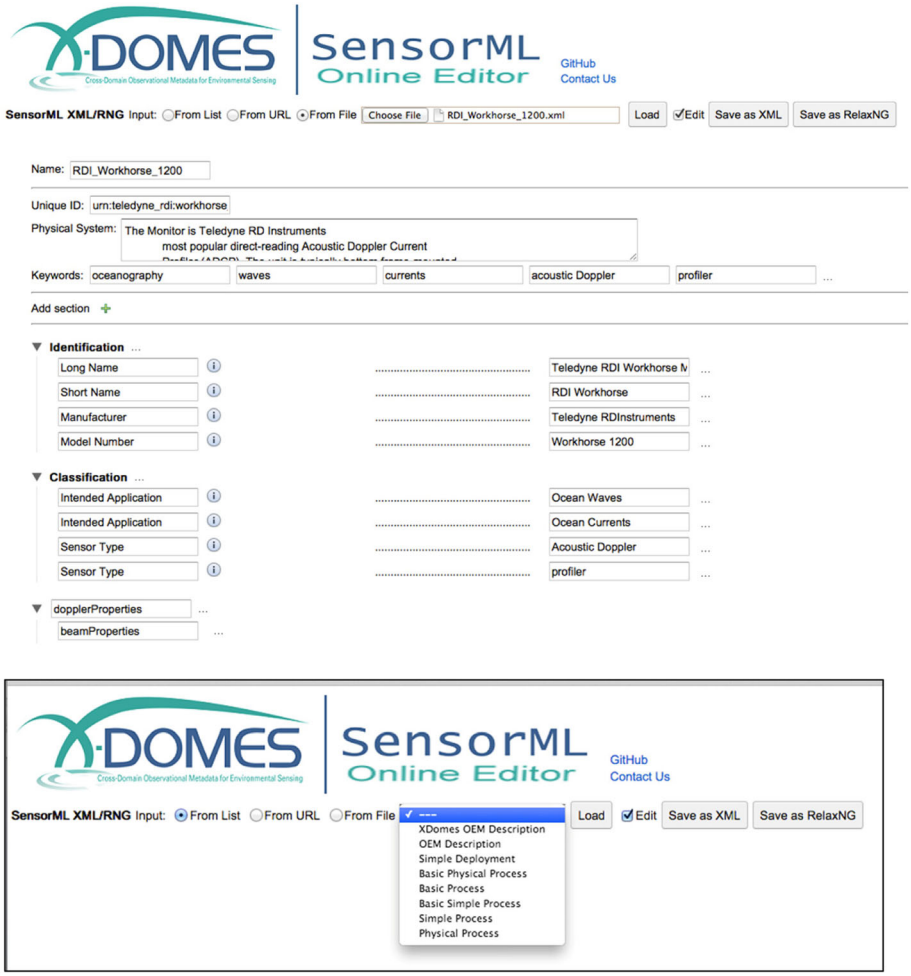

Fig. 1 The Online SensorML Editor has the look and feel of a technical specification sheet. A best-practice is to provide a urn as the SensorML Unique ID. When adding or editing links to controlled vocabularies by clicking on "…" on the right of the terms), a SPARQL query to a community adopted repository (e.g., X-DOMES ORR) facilitates the incorporation of links within the SensorML to community adopted controlled vocabularies. The second panel in this figure shows how to select RelaxNG profiles, which enables rules to guide in creation of content

particular sensor model. This can lead to more accurate and more fully-described information about the technology used in creating our observations.

Description of the sensor model A sensor model can be described broadly and will be referenced as a unique type of sensor. Some sensor models have options that will be defined as such. The SensorML file describing the Original Equipment Manufacturer (OEM) model is "owned by" the manufacturer, who can register it to provide a persistent, version-controlled, resolvable link to its content. The manufacturer has the responsibility to create, maintain, and register the OEM SensorML document that will have specific content, including at a minimum:

- A Uniform Resource Name (URN) to uniquely identify this sensor model
- Contact information of the sensor manufacturer

Additional information will include characteristics of the sensor (e.g. size, electrical requirements), capabilities (e.g. sensitivity, accuracy), identifiers (e.g. model number, sensor name), classifiers (e.g. intended application, sensor type), and other properties. By registering the sensor in a registry such as the X-DOMES SensorML Registry and Repository (SRR), the content is accessible and can be referenced via a URL (Uniform Resource Locator). The SRR also enables control of the content through authoritative ownership, version control and provides a mechanism for discovery.

Sensor manufacturer of a particular sensor The manufacturer will also create many instances of a sensor model that will be purchased and deployed. The manufacturer can create an Instance SensorML file that describes each particular sensor, as built and factory-

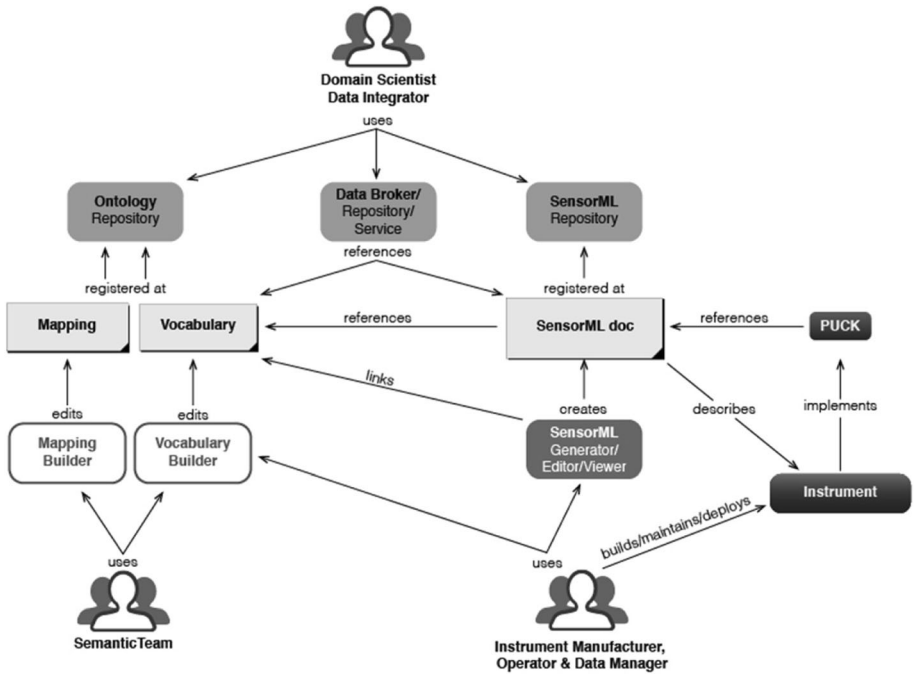

Fig. 2 Tools have been created to work together and to be used by different users. The ontologies can be developed by a community and mapped by cross-domain scientists and referenced and associated within the SensorML Online Editor. A PUCK [11] enabled sensor can hold content, which could be the contents of the SensorML describing the Instance and referencing the OEM. The vocabularies, ontologies and SensorML documents can be referenced within brokers, data repositories and web services, associating data with these descriptions

configured. The content of this SensorML description includes at a minimum:

- A URN to uniquely define this instance of the sensor
- a declaration that this is a typeOf the OEM model (Fig. 3), thus enabling inheritance of the SensorML descriptions in the OEM model documents, and
- Contact information of the sensor owner

Additional information will include, for instance, a serial number, calibration curve or date of last calibration, configuration settings, and more.

A URN for a model could be "urn:mfgr:model" and the instance "urn:mfgr:model:serNum", providing them each with a discoverable unique id.

The SensorML description of the instance is transferred to the owner of the instrument who becomes the curator of the document. It is the responsibility of the sensor owner to deploy a mechanism to register and reference the document (via a registry or a web service) and to associate it with the data that it produces.

Sensor user roles

Configuration of the sensor Each sensor typically has options available to the user to specify in order to meet specific deployment requirements. These options are also important to associate with the data, as they often affect operational range, accuracy and other parameters that affect how data can be interpreted. Currently, the X-DOMES project is exploring mechanisms to enable field operators to be able to add these descriptions to their sensor descriptions using the SensorML Editor/Viewer. Unless parameters are specified in the Instance SensorML, the operational descriptors will be inherited from the OEM descriptions.

Deployment of the sensor Once an instrument is purchased and configured, it is ready for deployment. There are standard descriptions that must be included such as location, orientation, station IDs, feature of interest, data ownership, etc. But, there are also many things that happen that can significantly impact data quality that have not been typically included in metadata. In particular, sensor preparation with regard to assuring quality measurements should be noted. For example, calibration of a sensor may happen in situ or at some point in time in a data stream. A sensor face may periodically be cleaned or fouling may be noted in the sensor deployment description. All this information must be fully described in machine-harvestable frameworks. The SensorML Editor/Viewer provides an easy way for the operators to insert

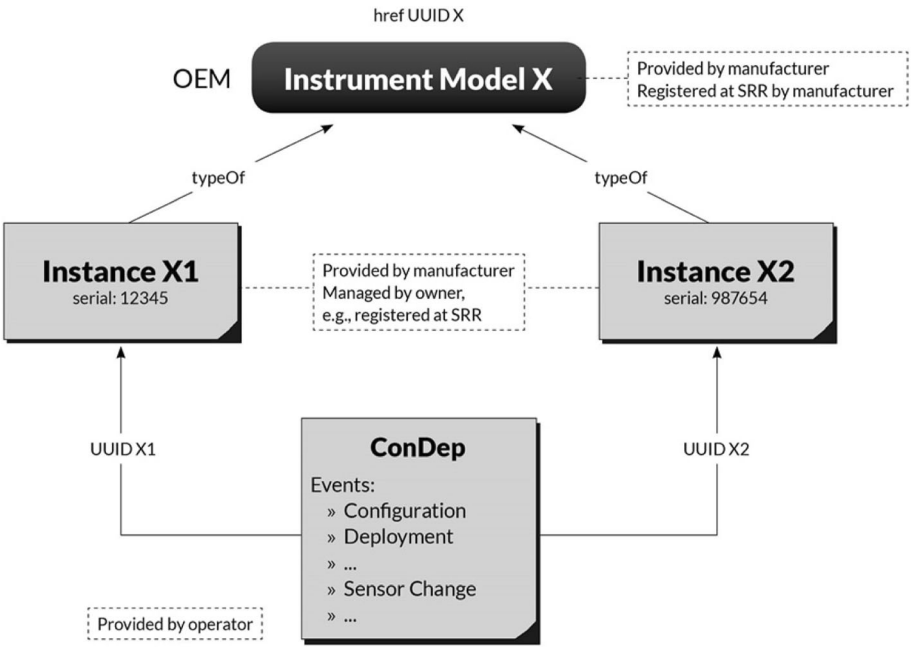

Fig. 3 Sensor manufacturers can take content out of their manuals into actionable, standards-based content by describing each sensor model in SensorML. These documents must also include references to registered vocabularies describing terms included in the documents. The SensorML model description documents are to be owned and managed by the sensor manufacturer and should be registered with a repository to enable access via an URL. The manufacturer can also create a SensorML document identifying a particular sensor, when distributed to an owner. The manufacturer is responsible for creating the Instance SensorML (and the associated URN) but transfers management of the document to the sensor owner. This document references the URL (HTTP UUID) of the sensor model and inherits descriptors as a typeOf the sensor model

history events into the combined ConDep (Configuration/Deployment) information SensorML document.

Processing for quality control and derived products Quality control tests are typically applied after data are collected and derived products are often computed from collected data. In the Q2O project [4], examples of SensorML process descriptions demonstrate how to describe quality control tests and processing along with associated URLs for further describing computational methods. This enables a data provider to provide options for data services (SOS) and provides a data consumer the option of requesting, for example, only ingesting data that passed particular QC tests or the unfiltered raw observations. Access to complete data processing descriptions enables a consumer to better understand how to interpret data that they receive. For example, if one is looking for extreme events, one might look at the processing descriptions to learn what threshold values were used to determine out-of-range data and whether these data were removed or replaced. If threshold value used is below the threshold value of your definition of 'extreme event', you cannot assume there were no occurrences of these events, but you can only note that they would not show up in this particular data set. The ability to automate descriptions of the processing chain while implementing it is important to assure accurate and complete descriptions. The authors encourage exploration of methods to automate the capture of process descriptions, as processing steps are being defined by those who best understand the methods being implemented. The processing descriptions must also include authoritative references to established methods of computation. And any seasonal or situational changes in parameters used in each processing step should be time-stamped and noted as an event in the process descriptions. Fig. 4 shows a SensorML description of a common quality control test. Like the OEM and Instance SensorML documents, this will become a component describing the system that created an observation.

The SensorML online editor enables the user to specify the type of SensorML component such as a physical component (OEM document) or process model selection (procedure), that is being created. The software uses RelaxNG [8] profiles to guide SensorML development, leading to more consistent content (Fig. 1 lower panel). Profiles can be created to standardize metadata for a particular community or manufacturer, the description of a particular sensor type, or the inputs, outputs, and parameters of a particular QA/QC test component. These profiles both

Fig. 4 Example of a quality control test SensorML description

simplify and standardize the creation of a SensorML documents to describe the various components. The X-DOMES community strives to encourage the development of profiles that enable non-experts, such as sensor manufacturers and data managers, the ability to create content from templates. Thus, through the development of example profiles for OEM, Instance, processing, deployment, platforms etc., a community-adopted set of templates can be registered, managed and assessed for use by the cross-domain enviro-sensing community.

Bringing it all together within an sensor observation service

Thus, the full description of the provenance of a particular observation can include the OEM description of the sensor model, the description of the particular configured sensor Instance, the list of QA/QC tests that have been applied along with their configuration, and the explicit chain of processes applied to obtain the collected observations. When the SensorML files are complete, the provenance for a given observational offering can be accessed using the DescribeSensor request defined within the OGC Sensor Observations Service (SOS) standard.

The Q2O project demonstrated how to pull these documents together though one should note that the descriptions in Q2O were SensorML 1.0 while SensorML v2.0 provided significant improvements to support inheritance of associated SensorML through use of the 'typeOf' association. Within XDOMES, SensorML version 2.0 was used, which led to a better separation of the OEM sensor model description from the description of the sensor Instance.

The content is created as stand-alone documents that need to be registered, so that they are web-accessible and persistent. The X-DOMES SensorML Registry and Repository (SRR) can be accessed through the xdomes.org site. The SRR enables content to be maintained and provides a Uniform Resource Locator (URL) for access. The documents can be referenced and harvested as part of a DescribeSensor response and included as a component of a SensorML process-chain. The links can also be referenced in the O&M encodings. Or, the documents can be included in non-SWE data management systems, since the content is registered and accessible via its URL. For example, if a data provider associated a set of data with a particular sensor and sensor model, a data facility could harvest the referenced SensorML document to incorporate information needed within their specific data management system automatically, since it is based upon a standards-based, community-adopted standard (SensorML).

The X-DOMES project has focused on tools to create the OEM/Instance content. The tools currently can also be used to create process descriptions. Tools are also needed to more easily connect the components into a process-chain that can fully describe how they connect

to show process lineage from observable property to observation, thereby enabling quality assessment.

Results and discussion

A network of stakeholders has been established to further develop the model and to encourage the adoption of these best-practices. With the funding from the NSF-EarthCube X-DOMES project, the authors are updating and creating the suite of tools enabling manufacturers to create accurate, consistent description of sensors in SensorML, while providing registered and linked vocabularies developed by the community of users. The authors are also working with the Ocean Data Interoperability Platform (http://www.odip.org) and towards the development and adoption of standardized SWE Marine profiles through a recently organized working group [9].

Interested communities can participate by joining the Earthcube X-DOMES Network (http://earthcube.org/group/x-domes) or through the Earth Science Information Partnership (ESIP) Enviro-Sensing Cluster (http://wiki.esipfed.org/index.php/EnviroSensing_Cluster). The communities provide the development team with cross-domain, cross-agency input to guide in our development of tools and give us the ability to test the integration of products into existing and emerging data management systems through our collaboration with a few data facilities, such as the NSF-funded R2R (Rolling Deck to Repository), BCO-DMO (Bio-Chemical Oceanographic-Data Management Office) and Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI).

Access to the tools are available on the xdomes.org site (http://xdomes.org) and links to tutorial videos can be found on the ESIP X-DOMES website (http://esipfed.org/earthcube-xdomes).

Conclusions and future work

These tools and the associated standards they promote were designed to be for international adoption and are applicable across domains. As the process of registering observational sensor models descriptions in machine-harvestable frameworks grows, manufacturers will feel compelled to participate since they would lose exposure to their potential market if their products are not in the registry.

The capture of knowledge is more difficult the further one gets from where the decisions are made. The more automated the capture of information and the earlier this information is captured, the more accurate and complete it will be. The social barriers of getting sensor manufacturers and field operators involved in the generation of this information can be overcome by providing access to simple tools that enable them to contribute without being aware of the technologies involved in enabling interoperable solutions for capturing metadata. By implementing community-adopted standard frameworks (OGC/W3C), we can easily broker [10] across other adopted standards (e.g., ISO/FGDC). As data are shared across a broader community and as data are less associated with those who created it, it is imperative to be able to understand and assess it for reuse by expectation of a broader set of descriptions of how the observation came to be.

Abbreviations

FGDC: Federal Geographic Data Committee; IOOS: U.S. Integration Ocean Observing System; ISO: International Organization for Standardization; NOAA: National Oceanic and Atmospheric Administration (USA); NSF: National Science Foundation (USA); OEM: Original Equipment Manufacturer; OGC: Open Geospatial Consortium; ORR: Ontology Registry & Repository; Q2O: QARTOD to OGC; QARTOD: Quality Assurance of Real-time Oceanographic Data; SensorML: Sensor Markup Language; SRR: SensorML Registry & Repository; SWE: Sensor Web Enablement; URN: Universal Resource Name; W3C: World Wide Web Consortium; X-DOMES: Cross-Domain Observational Metadata for EnvironSensing

Acknowledgments

The authors gratefully acknowledge Dr. Carlos Rueda (Monterey Bay Aquarium Research Institute) and Felimon Gayanilo (Texas A&M – Corpus Christi) for their leadership and contributions to the X-DOMES project tool development. The model was developed with the help of X-DOMES Team Members and workshop participants including Darryl Symonds, T-RDI, and Bob Arko, LDEO.

Funding

Background efforts for the Q2O project, defining the model for providing information about data quality in a OGC SWE framework, was provided under NOAA's Cooperative Agreement FY 2007 Regional Integrated Ocean Observing System Development (NOS-CSC-2007-2000875), 2008–2011. The development of tools and registries for the capture and delivery of SensorML and community-adopted vocabularies is funded by the National Science Foundation (NSF) as an EarthCube Integrative Activity called X-DOMES (Cross-Domain Observational Metadata for EnviroSensing). EarthCube is a collaboration between the Division of Advanced Cyberinfrastructure (ACI) and the Geosciences Directorate (GEO) of the US National Science Foundation (NSF). For official NSF EarthCube content, please see: http://www.nsf.gov/geo/earthcube/ (Award #1541008).

Availability of data and materials

Not applicable

Authors' contributions

This article was written equally by Ms. JF and Dr. MB. Dr. MB also participated in the conceptualization of the model which led to his fostering of the adoption of updates to OGC SWE 1.0 to OGC SWE 2.0 that enable the implementation of this model. Both authors read and approved the final manuscript.

Authors' information

Janet Fredericks is an information systems specialist with the Woods Hole Oceanographic Institution, where she has been responsible for systems programming, data management and operational oversight of meteorological and oceanographic observatories. She has served as a liaison to the Inter-Agency Ocean Observation Committee DMAC-ST and on the U.S. IOOS Quality in Real-Time Oceanographic Data Board of Advisors. She is currently serving as an at-large member of the NSF EarthCube Leadership Council.

Mike Botts is the author of SensorML and has served as the chair of the OGC® SWE Domain Working Group since its conception. He received the 2008 Gardels Medal for his role in leading the SWE standards activities in OGC®. He was also the lead for development of the current SensorML Online Editor Viewer and is currently managing a project to develop an open-source SensorHub to support easy deployment of sensors with immediate access and tasking through SWE 2.0 standards.

Competing interests
Both authors declare that they have no competing interests. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Publisher's Note

Author details
[1]Woods Hole Oceanographic Institution, MS#9, Woods Hole, MA 02543, USA. [2]Botts Innovative Research Inc., 9668 Madison Boulevard, Suite 103, Madison, AL 35758, USA.

Received: 28 July 2017 Accepted: 20 March 2018
/ Published online: 02 April 2018

References

1. Reed, C., M. Botts, G. Percival, J. Davidson (2013). OGC sensor web enablement: overview and high level architecture, OGC 07-165r1.
2. Botts, M. and A. Robin (2014). OGC SensorML: model and XML encoding standard, OGC 12-000.
3. Examples of both physical and computation components in SensorML. Available online: http://www.sensorml.com/sensorML-2.0/examples/index.html. Accessed 26 Mar 2018.
4. Q2O Landing Page, Available online: http://q2o.whoi.edu. Accessed on 31 Mar 2017.
5. Fredericks J, Botts M, Bermudez L, Bosch J, Bogden P, Bridger E, Cook T, Delory E, Graybeal J, Haines S, Holford A, Rueda C, Sorribas J, Feng B, Waldmann C. Integrating Quality Assurance and Quality Control into Open GeoSpatial Consortium Sensor Web Enablement. In: Hall J, Harrison DE, Stammer D, editors. Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, Vol. 2, Venice, Italy, 21–25 September 2009, ESA Publication WPP-306; 2009. https://doi.org/10.5270/OceanObs09.cwp.31.
6. World Wide Web Consortium, Semantic Web. Available online: https://www.w3.org/standards/semanticweb/. Accessed 26 Mar 2018.
7. Rueda C, Bermudez L, Fredericks J. The MMI Ontology Registry and Repository: A portal for marine metadata interoperability. OCEANS 2009, MTS/IEEE Biloxi – Marine Technology for Our Future: Global and Local Challenges. 2009. Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5422206&isnumber=5422059. Accessed 26 Mar 2018.
8. Jirka S, et al. Marine Profiles for OGC Sensor Web Enablement Standards. 2016. Available online: http://meetingorganizer.copernicus.org/EGU2016/EGU2016-14690.pdf. Accessed 26 Mar 2018.
9. Clark J, Murata M, editors. RELAX NG Specification. OASIS, 2001. Available online: http://relaxng.org. Accessed 26 Mar 2018.
10. Fredericks J. Persistence of knowledge across layered architectures. In: Diviacco P, Fox P, Pshenichny C, Leadbetter A, editors. Collaborative knowledge in scientific research networks. Hershey: IGI Global; 2015. p. 20. http://www.igi-global.com/book/collaborative-knowledge-scientific-research-networks/110007.
11. OGC PUCK Protocol Standard. Available online: http://www.opengeospatial.org/standards/puck. Accessed 26 Mar 2018.