EarthCube IA: Collaborative Proposal: Cross-Domain Observational Metadata Environmental Sensing Network (X-DOMES)

Janet Fredericks, M. Botts, K. Janowicz, C. Rueda, J. Graybeal, F. Gayanilo

1. MOTIVATION/VISION

The U. S. Commission on Ocean Policy [1] has challenged earth scientists funded with federal resources to document and share data. Global, interdisciplinary issues have spawned a realization that we will greatly benefit from shared access to data from around the planet. The NSF Task Force on Grand Challenges [2] recommends that we "manage the pipeline from field instruments to large-scale data analysis to end-user visualization" and we manage data "from instrument to (remote and local) computing resources to archiving and visualization". We envision a world where a field operator turns on an instrument, and is queried for information needed to create standardized encoded descriptions that, together with the sensor's manufacturers knowledge, fully describe the capabilities, limitations and provenance of observational data. The X-DOMES pilot project, proposed here, initiates the first steps needed in realizing these goals. By facilitating and demonstrating standards-based technologies within existing geoscience infrastructures, the community will be prepared to implement automated capture of knowledge and assess its impact on access and interoperability across the geo-science community.

The knowledge of how an observable physical property becomes a measured observation must be captured at each stage of its creation. Each sensor-based observation is made through the use of applied technologies, each with specific limitations and capabilities. Each instrument model provides observations differently and with disparate issues relating to quality. Environmental sensors typically provide a variety of options that can be configured differently for each unique deployment, affecting the observational results.

INTELLECTUAL MERIT

By capturing the information (metadata) at each stage of its generation, a more complete and accurate description of data provenance can be communicated. By documenting the information in machine-harvestable, standards-based encodings, metadata can be shared across disciplinary and geopolitical boundaries. Using standards-based frameworks enables automated harvesting and translation to other community-adopted standards, which facilitates the use of shared tools and workflows, with persistence of knowledge [3]. Through the use of community-adopted standards, the X-DOMES pilot project, proposed here, facilitates all these advances, as well as automated management of program resources (devices).

By providing common content in standards-based encodings, data managers and researchers will have access to shared resources that are discoverable and readily available, with associated data and metadata, thus reducing the time required to archive, re-use or integrate data. The environmental sensing community will be able to leverage common tools developed to create and utilize these standards. The establishment of a cross-domain network of stakeholders (sensor manufacturers, data providers, domain experts, data centers), called the X-DOMES Network, will provide a unifying voice for the specification of content and implementation of standards, as well as a central repository for sensor profiles, vocabularies, guidance and product vetting. The ability to easily share fully described observational data will provide a better understanding of data provenance and will enable use of common data processing and assessment workflows, fostering a greater trust in our shared global resources. The X-DOMES

Network capabilities will form the foundation for interoperable architectures designed to integrate and document observational data, thereby fostering their reproducibility.

Current efforts in interoperability are focusing on knowledge management after the data are distributed. Each decision made during these data-generation steps can affect data quality and how it can be reused appropriately. The X-DOMES pilot project will begin the first steps in enabling propagation of knowledge through the various steps of creating and disseminating data. X-DOMES methods will automate parts of the process and provide needed tools for others. These methods can be applied to new or legacy systems. Instead of requiring large data centers to piece together metadata, we will transform the point at which we capture knowledge to enable more complete, accurate information to accompany cross-domain research data.

BROADER IMPACTS

The project will

- Help large observational data producers automate and manage sensor and operational provenance
- Encourage small federally funded data providers to describe sensor data in ways that meet agency requirements for data management
- Facilitate common content and standards-based production of interoperable sensor documentation by environmental sensor manufacturers
- Enable data aggregation centers to build relationships across domains for integration of sensor-based observations
- Facilitate the ability to assess data quality and automate quality control, based upon manufacturers' descriptions of sensor provenance
- Generate registries for sensor and deployment metadata that can be utilized by building blocks of a layered architectural cyberinfrastructure
- Create open-source tools and libraries to discover, access, translate and associate sensor metadata
- Promote better documentation for data archival of federally funded assets
- Speed sensor network deployments, providing better, faster event response data
- Reduce data analysis effort and time for scientists and emergency managers
- Improve the reproducibility of research products by capturing relevant metadata at each stage of data generation

Documenting sensors in machine-harvestable frameworks will also enable the automation of sensor and knowledge management in long-term, large-scale observation systems. By providing knowledge in standards-based frameworks, tools can be developed to address each stage of device management. These tools can be used across domains, since they are based on sensor provenance and process lineage, rather than domain-specific terminologies.

With access to machine-harvestable, standards-based encoded metadata about sensor provenance and process lineage, quality assessment tools can be developed to not only flag bad data for downstream consumption, but also the same tools can be used to alert field operators of a failed instrument or notify consumers of an event.

In the past, sensor manufacturers and observation system developers have each deferred action, each expecting the other to take the initial steps toward the development of a full network (manufacturers waiting for demand, while developers wait for sensors). With this proposal, both will be engaged simultaneously with tools and guidance to make the transformative transition to

interoperable sensor network ecosystems. The results will be vetted through the collaborations of existing large-scale programs by testing the integration of these technologies, as well as through participation in existing EarthCube Building Blocks, such as BCUBE and GeoLink.

2. BACKGROUND

2.1 Web services and the OGC Sensor Web Enablement Framework

The Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) suite of standards [4] provides a community-developed and adopted framework that enables interoperability across the earth science community. OGC works with existing standards bodies [5] and has developed numerous standards to deal with web-based, geo-enabled services. Most earth observations have three things in common: they are by definition geospatial, a sensor collects them, and they are processed to become meaningful to researchers. The use of the OGC Sensor Web Enablement (SWE) Sensor Modeling Language (SensorML) [6] will enable interdisciplinary research through the use of standards-based encodings that describe the sensor, processing and provenance, while maintaining secure and authoritative controls.

OGC's Sensor Model Language (SensorML) [6] defines community-adopted standards that enable the encoding of sensor and process descriptions across disciplines. The SensorML standard, part of the broader framework (SWE), enables the sensor and process lineage (metadata) to be associated with the observations (data). Figure 1 depicts the association of metadata and data services and also the ability to link the SensorML terms to encoded terms that enable the development of broader ontologies [7]. Both the metadata and the data can be offered through the OGC SWE Sensor Observation Service (SOS) to provide specific observations (GetObservation) with its associated description of sensor provenance and process lineage (DescribeSensor).

Past NSF support has also provided significant steps toward interoperability through the development of tools and has fostered the growth of communities to share these tools. We will build upon the success of a number of these past efforts.

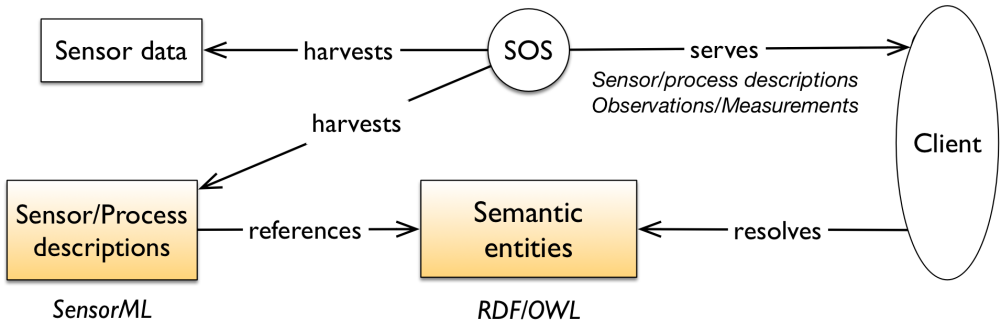

Figure 1. Web services (e.g., SOS) can associate data with sensor provenance and process lineage metadata and can include references to semantic entities enabling Semantic Web technologies [8].

2.2 Semantic Interoperability

Recent efforts promoting interoperability have focused on informatics that enable mapping of disparate terms for discovery and shared access. The efforts provide significant advancement by making volumes of data discoverable across institutional and disciplinary

boundaries. These efforts include the creation of sensor registries where instruments are registered and associated with particular locations, platforms and programs. Standard vocabularies are often minimal in their description, limited to a label and a collective definition, such as "a high precision CTD … SeaBird 911". Typically, there are no encodings that enable interoperable queries for specific content (like "What is the model number?" "What is the serial number?" or "What is the accuracy of the instrument and when was it calibrated?"). This content may provide answers as to how a data set may be used or reused.

The Marine Metadata Interoperability (MMI) project [9] enables participation in professional networks to share and discuss current interoperability efforts, and provides metadata guidance for data managers and scientists. The MMI project also publishes tools to help data providers easily create and register terms and for consumers to build semantic relationships across disciplinary and geopolitical boundaries [7]. The MMI-developed Ontology Registry and Repository (ORR) [10] provides operational semantic capabilities for the earth science community, supporting controlled vocabularies, term mappings, query processing (including inference support) and web resolvability of semantic information. The X-DOMES project leverages the services and open source software provided by the MMI project and will deploy updated ORR capabilities to advance X-DOMES objectives.

Key complementary tools to the ontology registry are the MMI Mapping Builder, which allows mapping of terms across multiple vocabularies according to various SKOS-based relations [11] (e.g., *exactMatch*, *closeMatch*, and *relatedMatch*), and the SPARQL [12] endpoint enabled at the ORR, which provides HTTP-based access to all semantic information in the repository, including reasoning capabilities that exploit properties in the mapping relations (e.g., the transitive property of *exactMatch*) to automatically derive more associations.

2.3 Semantic Web and Linked Data

Similar efforts have been carried out by the Semantic Web and Linked Data community [8] where different vocabularies, i.e., ontologies, are put into relation by ontology alignment techniques [13], while observational data are made available to humans and machines via the Linked Data paradigm [14]. A number of activities have focused on the transparent Semantic Enablement [15] of OGC's Sensor Web Enablement specifications and infrastructure. Examples include transparent RESTful proxies that seamlessly translate back and forth between Sensor Observation Services (SOS) and Linked Data [16] and add-ons to existing infrastructure [17]. A wide range of sensor and observation ontologies has been developed in the Semantic Web context; see [18]. One example that tries to strike the balance between sensors, deployments, and observational models is the Semantic Sensor Network ontology of the W3C-XG SSN group [19]. Focused more on the semantic lifting of observational data and measurement types, the OBOE ontology [20] is gaining traction in DataONE and beyond. Work on the detailed vertical documentation and semantic lifting of sensor metadata is not yet available on the Linked Data Web and the required ontologies and infrastructure are largely missing. The need for them has been recently recognized within the EarthCube context and the Building Block 'GeoLink' is tasked to develop a minimal instrumentation ontology design pattern and to publish the Oceanographic BCO-DMO and R2R data against it following Linked Data principles.

2.4 Implementation of a Model Enabling Quality Assessment

In 2006, the MMI hosted a workshop called the "Sensor Metadata Interoperability Workshop" (SMI) [21], finding that "the availability of meaningful sensor metadata—indicating

instrument models, configurations, and calibrations, among many other parameters—is a key criterion for scientific usefulness of the data. These descriptions must be sufficiently interoperable that they persistently accompany representations of the data, particularly as it is indexed for search and access in national and international data clearinghouses." Towards that end, we have developed a model [22] for using OGC® SWE to communicate sensor and process lineage (NOAA Award NA17RJ1223, 2008-2011) to enable data quality assessment. The project is called Q2O (QARTOD-to-OGC®) [23]. The model is based upon development of SensorML role-based documents that provide the means to fully-describe a sensor type, the sensor's deployed environment and configuration, and the quality control (QC) tests that were applied to the data, with associated parameters and results (QC-flags). The QC tests used were those recommended by the grass-roots, NOAA-funded group called QARTOD (Quality Assurance for Real-Time Oceanographic Data).

The model was developed in close collaboration with a sensor manufacturer, field operators and IT specialists, as well as domain specialists. A collection of SensorML v1.0 process descriptions (including sensor provenance, preprocessing and QC tests applied) are fully documented and linked through a "connection list" in the SensorML system document. For a given set of observations, two SensorML documents can fully describe the instrument (sensors) and deployment(s) [3]. One is an original equipment manufacturer (**OEM**) file that applies to any instrument of that model. The other is called the Configuration and Deployment (**ConDep**) file that has information about a particular instrument, its configuration and the specific parameters associated with the setup, deployment(s) and operational events affecting the data collection.

The SensorML descriptions contain metadata relating to contact information, sensor characteristics and capabilities, as well as deployment and maintenance history (an event list) and operational parameters and modes. The encodings enable machine-to-machine queries of sensor and process provenance: capabilities, characteristics and parameters, as well as events.

An **OEM** file fully describes a generic instrument, and therefore a manufacturer can create one for each model. By registering the **OEM** document on-line, it will be openly and freely available to anyone who uses the particular instrument model. An **OEM** document will include information that is important to enable quality assessment. For example, a sensor model may have a recommended field deployment length, accuracy or operational range. The SensorML document will also reference online instrument documentation, using encodings that enable the link to be discoverable and machine-harvestable.

A **ConDep** file documents how a sensor was configured and includes a time-stamped history of deployment(s) and maintenance events. The **ConDep** document is specific to a data stream but can have time-stamped descriptions added as things change (such as a sensor is swapped out or cleaned). As the sensor is being prepared for deployment, the parameters may be modified that affect data quality. Generation of the **ConDep** documents will be the responsibility of both the manufacturer and the operator. For example, a manufacturer typically generates a header file (often in binary) or has a response that is generated when an operator queries the instrument about its configuration. The manufacturer best understands this information and should be able to communicate it in using the interoperable technologies. But where it is deployed and any maintenance is best known by an operator and the **ConDep** file must be updated to reflect the important information. If the information is generated in machine-harvestable, standards-based encodings, the metadata can be discovered and shared across-

domains and data management systems, as well as having the ability to be translated between technologies through brokering.

When a manufacturer implements the OGC SWE PUCK [24] standard sensor interface, the instrument can either provide or reference relevant SensorML documents in the instrument's "PUCK payload", and the documents can be subsequently retrieved by the observing system through standard OGC PUCK protocol [24]. Various use scenarios of PUCK protocol were explored in the OGC Ocean Interoperability Experiment [25].

Semantic mediation is a required mechanism to enable system interoperability and data integration. In the Q2O SWE implementation, terms in the SensorML (and O&M) instance, from the input observables to the output definitions reference a meaningful, on-line resolvable definition (Figure 1). The encodings include specific content for each term, such as: ID, definition, relevant authoritative references, images, and equations. The Q2O model is built upon the use of resolvable links to individual RDF encoded terms. Content-rich vocabularies were created using the MMI Vocabulary Builder tool. This tool facilitates the creation of controlled vocabularies by using a simple and intuitive form-based user interface or by importing tabular information from typical spreadsheet applications. The terms are registered in the ORR, providing machine-to-machine and/or human readable URLs for each vocabulary and each term within the vocabulary.

2.5 ESIP: A Community of Practice [26] for the Earth Science Community [27]

The Federation of Earth Science Information Partners (ESIP Federation) is a community that was founded by NASA in 1998. It has grown over the years to include many governmental agencies and membership is also open to individual entities. From their Constitution: "The goal of the ESIP Federation is to establish and continuously improve science-based end-to-end processes that increase the quality and value of Earth science products and services throughout their life-cycle for the benefit of the ESIP Federation's stakeholder communities." In 2009, ESIP formed new communities with members whose interest centers on data preservation and stewardship, information quality and data visualization.

One cluster group of the ESIP Federation is the "Environmental Sensing" group, which is working on the development of best practices for sensor networks and sensor data management for on or near surface point measurement with the purpose of long-term or "permanent" environmental data acquisition [28]. The group attracts a large number of participants and has expressed a cross-agency interest in development of interoperable solutions to metadata describing environmental sensing observations.

2.6 NSF EarthCube Activities

In a past EarthCube brokering [29] "hack-a-thon", the Q2O model SOS provided data and metadata through brokering services. The Data Access Broker (formerly the EuroGEOSS Broker) was able to populate its catalogues with metadata from the SOS GetCapabilities operation directly, rather than the data provider filling out a form. Another brokering service (NOAA-ERDDAP) translates the SOS metadata to ISO11915 and FGDC and the observations from the SWE framework (O&M) to netCDF or JSON. As brokering capabilities develop, providing common content in standards-based encodings becomes paramount to the brokers' success in providing fully described data and sensor provenance.

Current building block projects are poised to collaborate with the proposed X-DOMES project, including BCUBE and GeoLink. The work conducted through this pilot project will also

be aligned with the efforts of the nascent working groups of the Technology and Architecture Committee and Science Committee, such as the Standards Working Group and Gap Analysis groups.

3. PROPOSED WORK

The X-DOMES pilot implementation seeks to automate the generation of sensor provenance by providing tools and guiding communities in best practices to facilitate the generation of critical information about sensor capabilities. The EarthCube End User Workshop Reports [30] highlight the lack of associated metadata and provenance as a critical concern. We will engage end users, where this need was noted, to assure alignment of X-DOMES development with the EarthCube end user community.

The project focuses on fostering the adoption of existing standards that will be implemented using a consistent approach that provides content-rich metadata to accompany data as it is moved along its lifecycle. These technologies will be assessed with regard to their ability to enable automated discovery, quality assessment, reuse, integration, processing, brokering and automated sensor and data management and archival to meet the needs of the geo-science community. Building on a social network of environmental sensor manufacturers and stakeholders, we will research and document opportunities and impediments for enabling automated generation of cross-domain common content and adjust the pilot implementation in response to the feedback. Stakeholders are those who will benefit by the capabilities, such as data managers, aggregators and curators, who can utilize the common content, in standards-based, machine-harvestable encoding to automate access and processing, as well as end-users who can benefit from a more complete description of data. The project will promote the use of accurate, consistent and content-rich descriptions of provenance by developing tools for manufacturers to create, edit and ingest the documents. Within the 24-month period, we will develop and deploy the tools for use by sensor manufacturers, which are to be selected by our collaborators (CUAHSI, R2R, BCO-DMO, NTL-LTER). The standards-based encodings of vocabularies and sensor profiles of the commonly used environmental sensors will be discoverable, openly and freely available via the WWW. The products will be integrated into respective data management systems of each program, as well as into DataONE or through brokering technologies. This will enable hundreds of other users of the sensors to reference the documents. The documents can be referenced as semantically enriched stand-alone documents or as an integrated part of an Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) framework or as a component of a Linked Data network. The pilot implementation will be assessed from the perspective of the sensor manufacturer, data provider, data manager, brokering team and the end-user geo-scientist.

We will harmonize and ontologically align the implementation with other existing frameworks in the Semantic Web context, such as PROV-O, the W3C-XG SSN ontology, OBOE, and ongoing work on instrumentation in other EarthCube projects.

Just as manufacturers moved from printed documents to on-line information access (HTML), with proper tools, guidance and leadership, they will be able to provide machine-harvestable and machine understandable documentation that will promote interoperability, automation and reproducibility. This knowledge enables automated data quality assessment.

3.1 X-DOMES Activities Overview

This pilot project focuses on developing and assessing methods to facilitate automated generation and tailored editing of the **OEM** files, so that manufacturers can easily generate and edit the sensor/observational provenance documents without knowledge of the underlying technologies. It is critical to have the robust tools in place while engaging manufacturers to facilitate implementation easily and consistently. And, utilizing the products, assess the ability to transfer and ingest the content within data management and brokering systems. Towards that goal we will:

- Engage sensor manufacturers representing disparate sensing technologies and geo-science domains;
- Upgrade the existing model for conformance to current standards (e.g. SensorML 2.0) and to promote the implementation of the common content in the structure of a role-based (layered) model with community input to assure cross-domain adoption;
- Develop software to provide on-line forms which will be used to create the SensorML documents (with links to RDF encoded, registered terms from ontologies), fostering common content and validation of the document syntax for Original Equipment Manufacturer (**OEM**) content;
- Enhance the MMI Ontology Registry and Repository software as required to enable X-DOMES operations;
- Develop and populate a registry for **OEM** sensor descriptions, with discoverable information about its capabilities and technologies.
- Enhance, and customize existing tools (with the role-based architecture in mind) used in the editing and viewing of the sensor description documents for the environmental sensing community (e.g., the SensorML editor and PrettyView);
- Provide the sensor manufacturers with guidance and user-friendly tools to create and register their terms and documents in resolvable, semantic encodings;
- Develop content requirements and a methodology to automate creation of **ConDep** (configuration/deployment) content, enabling the ability to update the documentation with history of the sensor and its deployment; and, define requirements for APIs for manufacturers' use for the creation of updated content at configuration
- Generate APIs for workflow interfaces to sensor documents with an eye towards automated quality control;
- Determine how an OGC PUCK-enabled instrument can be most effectively used to store and provide interoperable access to the **OEM** and **ConDep** SensorML documents; and
- Test and evaluate interoperability mechanisms.

3.2 Community Development (X-DOMES Network)

We will promote and document sensor manufacturer and stakeholder participation throughout the entire two-year period. In the supporting documents, a number of organizations and individuals have provided letters of commitment to target sensor models and assess the proposed work for their own data and sensor management systems. During the 24-month period, we will engage manufacturers to develop content-rich vocabularies and sensor documents for 1-2 sensors per collaborator. This will enable the test implementation of a dozen or so sensor models representing cross-domain environmental sensing systems, providing common content to both their respective data managements systems and participating data systems, such as the EarthCube building blocks and DataONE. During this phase of researching geo-science community needs,

capabilities and interests, we will provide opportunities for development and assessment by operating within the ESIP Environmental Sensing Cluster [28]. This will provide a sustainable community that can persist beyond the X-DOMES pilot project.

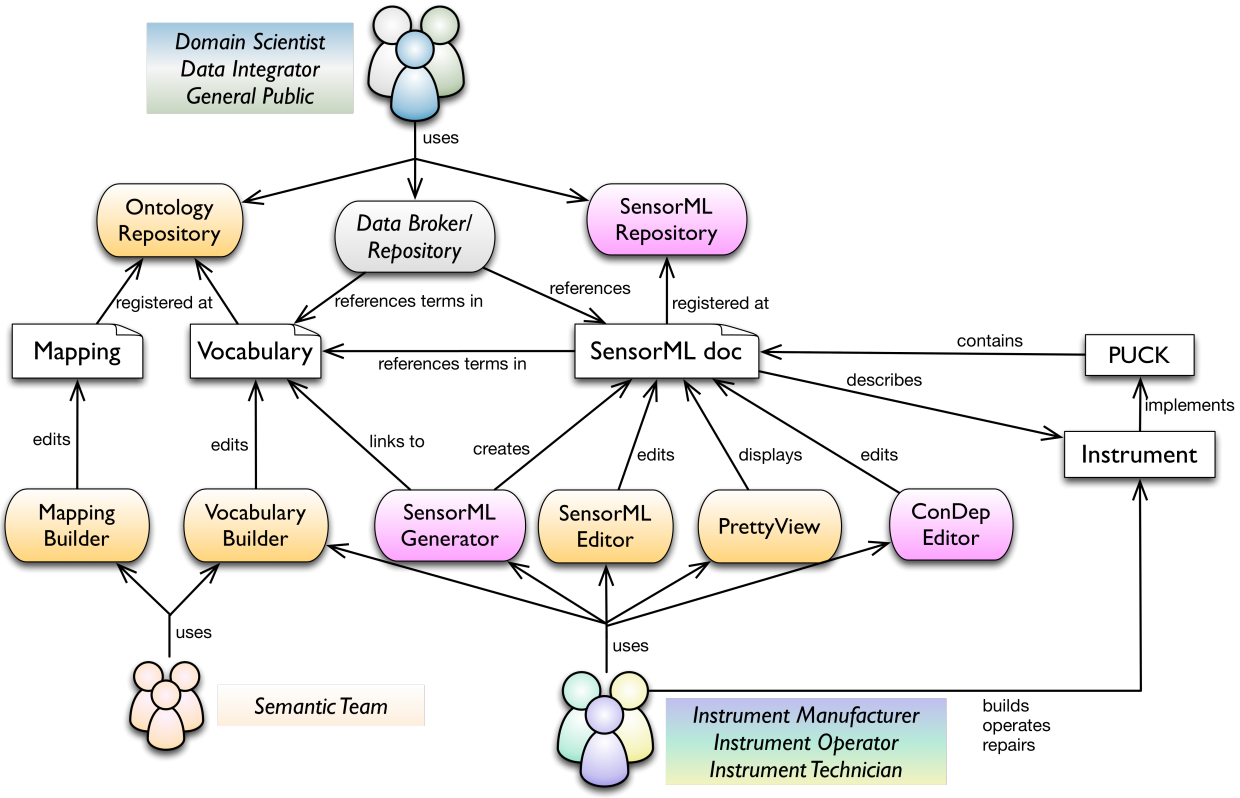

Figure 2. Software tools that will be updated (yellow) and developed (pink) to enable easy-to-use tools encouraging common content in OEM (developed and demonstrated) and ConDep files (requirements assessment). The resulting SensorML document can be referenced online or loaded in a PUCK-enabled sensor. The SWE implementation that utilizes SensorML documents is shown in Figure 1. The documents can also be referenced as Linked Open Data.

3.3 X-DOMES Pilot Project Activities

A name in parentheses designates the person(s) with primary responsibility for each activity.

3.3.1 ***Management Plan*** *and Organizational Support (Fredericks, Graybeal, Rueda)*

Fredericks will be responsible for organizing X-DOMES Network workshops at the ESIP Summer Meetings annually, working with Dr. Gries, and holding X-DOMES weekly team meetings to coordinate development and collaborative activities and face-to-face team meetings to develop strategies and review progress. Fredericks and the project team will prepare the annual assessment reports with input from collaborators. Fredericks and Graybeal will be responsible for engaging instrument manufacturers and science stakeholders representing cross-domain participation through direct contact with select environmental sensor manufacturers and stakeholders.

Fredericks will establish a team web site (workspace) to provide a means of communication and coordination. An X-DOMES Network workspace through the ESIP Commons will be created as a community resource that enables social networking for sensor manufacturers and other stakeholders. Fredericks will oversee the implementation of the site along with ESIP Environmental Sensing Cluster organizers (Gries).

Fredericks and Janowicz will coordinate this effort with the broader NSF EarthCube community and EarthCube Building Blocks such as GeoLink and BCUBE, to ensure maximum engagement and reuse across communities. Fredericks will serve as liaison to relevant working groups and standing committees of the EarthCube governance.

Rueda will provide support for vocabulary development and registration for the X-DOMES Network participants throughout the project duration.

Unfunded collaborators (Supplemental Documents) will be engaged throughout the process and will be supported in their travel, as needed, to encourage participation.

Leaders of DataONE, CUAHSI, R2R and BCO-DMO have provided letters of collaboration to confirm their participation in the incorporation of the technologies into their data management systems and assessment of the functionality and usefulness of the tools and resultant documents (vocabularies and sensor descriptions). CUAHSI (Hooper), R2R (Carbotte), NTL-LTER (Gries) and BCO-DMO (Chandler) will develop the list of sensors that will be targeted in the X-DOMES pilot project.

Engagement with EarthCube GeoLink is through the association with Dr. Janowicz, who is also funded through the GeoLink building block. EarthCube BCUBE is collaborating in testing the interoperability of the pilot implementation, as an unfunded participant.

3.3.2 Software Development (Figure 2)

Updates and documentation of all software components will be open source and openly and freely available through repositories such as GitHub. Fredericks will oversee the schedule of development, modifications, validation and releases to assure timely and accurate coordination of capabilities across components.

SensorML 2.0 Editor and PrettyView (Botts)

The SensorML editor and the SensorML viewer (PrettyView) will be updated to enable changes, validation and visualization of the common content defined in the Q2O model, as it is refined throughout the X-DOMES project period. The Q2O model was developed before the adoption of version 2.0 of the SWE standards. Some of the improved features in SWE 2.0 were a result of the work in Q2O and enable essential improvements to the model. The improved capabilities include improved inheritance models that allow better separation of **OEM** and deployed sensor descriptions, as well as better support for real-time streaming data. Throughout the X-DOMES project, the editors will be enhanced to accommodate the specific content that will be needed to describe the **OEM** documents.

MMI Ontology and Repository Update (Rueda/Graybeal)

The MMI Ontology Registry and Repository (ORR) currently serves over 300 marine science and related vocabularies and mapping ontologies, including those authored by organizations such as Ocean Observatories Initiative (OOI), U. S. Integrated Ocean Observing System (IOOS), International Coastal Atlas Network (ICAN), Rolling Deck to Repository (R2R), CUAHSI, ODM2, and Unidata. It also holds several vocabularies developed by the MMI

project itself, including ontologies for oceanographic platforms and devices, and an ontology version of the Climate and Forecast (CF) standard names parameter vocabulary.

The X-DOMES project relies on these capabilities to provide embedded links to on-line resolvable RDF-encoded semantic definitions. This enables us to offer knowledge in familiar terms useful in collaborative environments. Through support from this project, we will update the tools to current technologies and provide ready access to the registry for all of the sensor manufacturers.

Although successfully providing basic capabilities for many communities, to properly support this operational network the MMI ORR software requires some updates. These will make the ORR more capable and align its underlying technologies with current practices. The main elements involved in this general update are:

- Component upgrades: For the ontology database the ORR currently uses the MySQL-based backend system of an old version of the BioPortal technology [31]. The current BioPortal release is a major upgrade, adding new features that the MMI ORR could leverage. Similarly, the ORR user interface is based on a very old version of the Google Web Toolkit (GWT). As part of the proposed project, we will evaluate current BioPortal and GWT technologies to determine and implement effective upgrade strategies bringing reliability, usability, and maintainability to X-DOMES required levels.
- Deployability: The customization of the ORR system (including configuration, branding, and dependencies resolution) will be streamlined. This will facilitate its targeted use as requested for specific domains (in this case, for the X-DOMES Network).
- Issues and enhancements: Throughout the project, we will examine, prioritize, and address the most critical of the ORR issue and enhancement requests to meet the needs of the X-DOMES Network. For example, a REST-like HTTP interface would enable X-DOMES applications to more readily integrate with the ORR content; this will be evaluated against other needs and implemented if appropriate.

Sensor Registry and Repository (Botts)

A Sensor **OEM** registry will be developed and deployed. The manufacturers will be able to register and update (with versioning) complete profiles of each instrument model. The registry will utilize the MMI ORR capabilities to enable discovery of the Sensor profiles using well-known terms, such as sensor technology (“Doppler”), sensor manufacturer (“Nortek”), capabilities (“accuracy”) and observed property (“wave height”).

Create forms, guiding content and consistency while minimizing errors (Fredericks, Rueda)

We will develop a form-based tool to generate and validate the common content of the OEM documents. We will build upon existing systems such as the MBARI, IOOS and the ESONET Network of Excellence SensorML form tools (platform-specific generation) to implement the X-DOMES model of role-based creation of content.

Development of application programming interfaces for standardized workflows (Gayanilo)

Throughout the project, we will develop a series of Application Programming Interfaces (APIs) in Python and Matlab to provide community-ready modules that can be employed to create and read SensorML content, demonstrating its potentials. This collection of modules will include input and output elements and their semantic definitions, and also options to drill through the components to expose system characteristics and capabilities.

3.3.4 Formalize use of PUCK-enabled sensors (Rueda)

The OGC PUCK standard [24] defines a standard 96-byte "instrument datasheet" to be stored within the instrument; the datasheet includes a universal "serial number" unique to each PUCK-enabled instrument, as well as manufacturer and model identifiers. We will determine how best to use the datasheet. For example, the datasheet's universal serial number could provide a unique lookup key into sensor registries. PUCK also provides an optional, variable-sized "PUCK payload" within the instrument that can contain arbitrary content. We propose to formalize use of the payload, specifically as a container for relevant SensorML 2.0 documents. We will determine the specific SensorML content, taking the limited payload space available to some instruments into account. Depending on our findings, we may recommend specific changes to the PUCK standard, for example taking into account a minimum required payload size. We envision that manufacturers will ultimately provide PUCK payload content with their instruments, and will work with manufacturers such as those in the Smart Ocean Sensors Consortium (SOSC) [32] to ensure that our recommended PUCK use is practical from their standpoint. We will collaborate with the SOSC to jump-start manufacturer engagement in the sensor manufacturer workshops and formation of an earth science sensor manufacturers network.

3.3.5 Organize annual sensor manufacturer workshops to advance the formation of a network of sensor manufacturers and stakeholder (Fredericks)

Two workshops will be held at each of the ESIP Summer meetings to

- Educate manufacturers and geoscientists on the value of interoperability to promote the use of standards-based encodings (consistent with SensorML and the Semantic Web and Linked Data paradigms in general) in describing sensor provenance and process lineage and how they can be utilized in an interoperable infrastructure
- Engage manufacturers and geoscientists to define requirements and assess X-DOMES models, tools and resources.

3.3.6 Harmonization and Alignment (Janowicz)

X-DOMES will harmonize and align X-DOMES efforts with commonly used ontologies such as W3C SSN-XG's Semantic Sensor Network ontology, OBOE, the W3C provenance ontology PROV-O and the instrumentation ontology design pattern currently developed within the EarthCube Building Block GeoLink. These alignments will ensure that machine-readable and understandable schemata are available at all vertical levels from sensor metadata to observation results. Besides harmonization and improved interoperability, this alignment will also ease the integration of sensor provenance metadata with the existing work on the Semantic Sensor Web and the EarthCube effort in general. The alignment will be realized as axioms specified using the Web Ontology Language (OWL) and will establish formal relations between classes from different ontologies. The harmonization will be achieved by keeping X-DOMES developments in line with the aforementioned work and by Janowicz's involvement in the SSN-XG standardization that will be carried out by the W3C *Spatial Data on the Web* working group. This will also ensure that X-DOMES products can be made available as Linked Data in the future.

3.3.7 Outreach (Fredericks, Gayanilo, Janowicz, Graybeal)

Fredericks and Janowicz will coordinate this effort with the EarthCube community, working with GeoLink, BCUBE and other newly forming groups.

Graybeal, Janowicz and Gayanilo will work with Fredericks in coordinating our implementation with existing large integrative programs (e.g., DataONE) and soliciting feedback.

Gayanilo will coordinate with DataONE to create a common standard extending their current implementation to use SensorML-encoded metadata. Other outreach activities include publishing project results throughout the two-year project, e.g., in AGU conference proceedings. Other opportunities for outreach will be obtained through teams such as the OGC TC meetings, Research Data Alliance (RDA) working groups, and the Ocean Data Interoperability Platform (ODIP).

4. TIMELINE

Year 1:

Community Development and Outreach: Contact sensor manufacturers of large NSF-funded programs; establish X-DOMES Network through ESIP. Begin work on vocabulary development and registration; building on existing relationships, introduce project to other working groups, such as EarthCube/DataONE/RDA; and, raise awareness of the pilot project through conference participation in AGU. Set up web presence for X-DOMES team and X-Domes Network.

Software Development: Complete updated SWE Model (2.0); preliminary updates to the MMI Vocabulary Builder and ORR; updated SensorML Editor/Viewer for OEM files; alignment and harmonization of SensorML model with the W3C-XG SSN ontology and OBOE; and, definition of functional requirements for Python and Matlab APIs/modules and faceted Linked Data interface.

Evaluation Materials: Report on manufacturers' participation and feedback; assessment of synergies with existing communities; reviews of updated software, models, and harmonization efforts.

Year 2:

Community Development and Outreach: Conduct second annual X-DOMES Network workshop at ESIP Summer Meeting introducing tools/guidance for sensor model (**OEM**) development and registration, as well as continued guidance on vocabularies; continue outreach (EarthCube/RDA/AGU/DataONE); Demonstration of API/modules for **OEM** documents; continued sensor manufacture outreach. Assess opportunities for student engagement. Develop requirements and strategies for the automated creation of **ConDep** content.

Software Development: Launch forms to generate content (SensorML generator); create SensorML registry; develop Linked Data compliant interface to registry; release updates to the MMI Vocabulary Builder and ORR; alignment and harmonization with standardization results from the W3C *Spatial Data on the Web* working group as well as EarthCube GeoLink Instrumentation ontology design pattern; and integration and inter-linkage with Prov-O provenance ontology.

Evaluation Materials: Report on sensor manufacturers participation and feedback, on data management facilities' response to new content, and to brokering experience with BCUBE; report on Linked Data and SSN alignment activities; and lessons learned.

5. DELIVERABLES

Community Development: Establishment of an ESIP working group called the X-DOMES Network with a dedicated workspace to further the goals & mission of the newly formed group; an organizational structure (with scope, goals & mission); presence in and alignment with other

existing communities, such as EarthCube Working Groups and Building Blocks, DataONE and the Semantic Sensor Web community.
Open-source Software: Tools specifically developed to facilitate the documentation of sensor provenance (**OEM**). The tools will be used to generate, edit, validate and view registered sensor documents exposed to the Linked Data web; underlying code for online registries for referencing vocabularies and sensor provenance documents; application interfaces to access content from commonly used working environments, such as python and Matlab. All software will be available through open-access repositories (such as GitHub).
Guidance Documents: All workshop materials and X-DOMES Network resources will be openly available.
Vocabularies and Sensor Documents: Registry sensor profile content provided by the X-DOMES Network participants will be openly available on-line. Recommendations for PUCK specification improvements will be provided to that community. The developed ontologies and alignments (mappings) will be published at the publicly available MMI ORR, as well as ontohub.org.
Assessment Documents: Surveys (providers/consumers), survey results and reports from discussions with all stakeholders regarding requirements, impediments, usability and value of tools and products will be made available. Requirements and methodologies for **ConDep** content creation and updating by manufacturers software and field operators will be compiled through conversations with the X-DOMES Network participants and made available by the end of the 24-month project.

QUALIFICATIONS OF TEAM

Ms. Fredericks, as the principal investigator (PI) on the Q2O project, led the development of the quality assessment model that will be used by X-DOMES. She also brings with her many years of research experience, both as an operational oceanographer and a systems programmer. She is manager of the cabled observatory called the Martha's Vineyard Coastal Observatory, which hosts several research projects each year and has been serving a suite of real-time sensor data since 2001. She served as a liaison to the Inter-Agency Ocean Observation Committee DMAC-ST and is currently on the U.S. IOOS Quality in Real-Time Oceanographic Data Board of Advisors and a participant of the EU-Australia-US Ocean Data Interoperability Platform project. She was involved in EarthCube in its initial phase as a member of the Brokering, Data Access and Semantics & Ontologies community groups and a funded participant in the Layered Architecture Concept Award and is currently participating in the Technical Architecture Committee and the Gap Analysis Working Group of the EarthCube test governance.

Dr. 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 Editor and PrettyView, the Space-Time Toolkit visualization package and a variety of open-source libraries in support of SWE. He 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. He also was a Co-PI in the Q2O project. Dr. Botts was an elected member of the OGC Architecture Board (2008-2014).

Mr. Graybeal co-founded the Marine Metadata Interoperability Project in 2004, and continues to serve as the Project Lead. As part of his role he co-hosted several workshops, including the 2006 Sensor Metadata Interoperability workshop. He led MBARI's development of the Shore Side Data System, and guided Data Browser features for Marinexplore/PlanetOS. He was one of three principal co-developers of the Conceptual Architecture for OOI, served on the W3C's Semantic Sensor Network Working group, serves as the co-chair of the ESIP Federation's Attribute Conventions for Data Discovery, and wrote the NetCDF Climate and Forecast Conventions Frequently Asked Questions. He continues to develop and refine vocabularies served by MMI's Ontology Registry and Repository.

Dr. Rueda has been the technical lead for the MMI project since 2008. He coordinated an international group toward the development of a marine device ontology (elements of which were adopted by the W3C Semantic Sensor Network ontology effort) and has been the main developer of the MMI ORR system. He assists IOOS, OOI, ICAN, and other communities in the development of controlled vocabularies and has co-hosted international technical meetings with ontology repository developers, promoting the need to define inter-repository standard interfaces. He has participated in the DataONE Semantics and Integration Working group and the Open Ontology Repository initiative.

Mr. Gayanilo is the Systems Architect for two major scientific data portals in the Gulf of Mexico. The first is the Gulf of Mexico Coastal and Ocean Observing System (GCOOS) that is nested in a National Backbone of coastal observations to aggregate and disseminate the region's near real-time oceanographic data in OGC Sensor Observation Service (SOS) to facilitate interoperability among coastal and ocean observing sensors. The second is the Gulf of Mexico Research Initiative Data and Information Cooperative (GRIIDC), a 10-year multi-disciplinary and multi-institutional research effort in response to the 2010 Deep Horizon oil spill disaster. Mr. Gayanilo is an active member of the DataONE User Group Steering Committee and was a member of the original team that drafted the bioassay ontology (BAO) and the publication of the corresponding semantic search engine for medical research while with the University of Miami.

Dr. Krzysztof Janowicz is an Assistant Professor for Geographic Information Science at the Geography Department of the University of California, Santa Barbara, USA as well as one of the two Editors-in-Chief of the Semantic Web journal. He is also the community leader of the 52°North Initiative for Geospatial Open Source Software GmbH semantics community that develops open source solutions that bridge OGC's Geo-Web and the Semantic Web. Janowicz is a member of the W3C SSN-XG that developed the Semantic Sensor Network ontology (SSN), and was responsible for the development of the Stimulus-Sensor-Observation ontology design pattern (SSO) that forms the core of the SSN ontology. Besides ontologies, Janowicz has developed software and specifications for sensor mediation, Semantic Enablement of Spatial Data Infrastructures, and Restful Linked Data proxies for the OGC Sensor Observation Service. He published large Linked Data sets such as the ADL gazetteer and tools for their exploration.