The MMI Ontology Registry and Repository: A Portal for Marine Metadata Interoperability

Carlos Rueda
Monterey Bay Aquarium Research Institute
7700 Sandholdt Rd. Moss Landing,
CA 95039 USA
Email: carueda@mbari.org

Luis Bermudez
Southeastern Universities
Research Association
1201 New York Ave. NW Suite 430
Washington, DC 20005. USA
Email: bermudez@sura.org

Janet Fredericks
Woods Hole Oceanographic Institution
Woods Hole, MA 02540 USA
Email: jfredericks@whoi.edu

Abstract—To address important marine research and application challenges, there is a critical need for ocean observatories to share data in a way that is easy to use and integrate. Fulfilling this interoperability requirement is greatly facilitated by several standards efforts and recent advancements in Web technology. Although these infrastructures provide a fundamental capability for data sharing, there remains a gap from mere syntactic interoperability toward a more powerful, semantic interoperability, in which the diversity of terminologies, vocabularies, and conceptualizations across multiple observatories can be more effectively harmonized, exchanged, and fully integrated. In this paper, we describe the MMI Ontology Registry and Repository and associated semantic services developed by the Marine Metadata Interoperability Project. We present the architecture and the suite of key functions that allow data providers and users to include, use, and exploit semantic information in representative real world scenarios. We show how OGC Sensor Web Enablement services describing sensor systems and data products can be enriched with semantic references that client applications can readily resolve in corresponding semantic descriptions. Data discovery mechanisms enhanced with the provided semantic services are demonstrated with a Web data portal for integrated ocean observing systems. Services for resolution of Web identifiers, along with query and inference mechanisms, which can be accessed in a programmatic way, enable better data integration capabilities. We also demonstrate the ability to relate data quality flags across agencies, to fully describe processing tests and methods, as well as the ability to link definitions to bibliographic references, also registered using the MMI semantic portal.

I. INTRODUCTION

A fundamental infrastructure for data interoperability is provided by standards efforts including the International Organization for Standardization (ISO), the Open Geospatial Consortium (OGC), and the World Wide Web Consortium (W3C). However, there remains a gap from mere syntactic interoperability toward a more powerful, semantic interoperability, in which the diversity of terminologies, vocabularies, and conceptualizations across multiple communities can be more effectively harmonized, exchanged, and fully integrated. Semantic interoperability challenges in Oceans and similar sciences are described in various reports [1]–[3] and literature [4]. In particular, the "IT Roadmap to a Geospatial Future" [1] expresses the need to advance the methodologies to dynamically categorize geospatial data and support data concepts and categorizations from different geospatial communities. These data concepts and categorizations are usually domain-specific metadata values (e.g. phenomena, sensor type, quality flag) which are not defined by ISO, OGC or W3C, yet are the key elements that capture meaningful information specific to a domain [5].

Semantic interoperability issues are related to two intertwined problems: semantic heterogeneity and information overload. Semantic heterogeneity mainly occurs when there are different representations for a concept. For example, the phenomenon “sea surface temperature” could have different representations, such as “SST” and “water temperature.” Information overload occurs when a user or application searches and discovers so many results that they are difficult to adequately filter and analyze. Categorization using semantic web technologies provides a possible solution. Categorization is used by web companies such as Yahoo, Ebay, and Amazon to help users filter search results.

Solving semantic heterogeneity in ocean communities requires that the controlled vocabularies used in metadata annotations be published in a formal format to facilitate interoperability. The World Wide Web Consortium (W3C) has published a set of recommendations related to the semantic web [6] describing how to encode and relate concepts so that they can be understandable by distributed programs in the web. One exemplar of an approach involving ocean experts and the use of W3C technologies to solve semantic heterogeneity was the workshop “Advancing Domain Vocabularies” hosted in 2005 by the Marine Metadata Initiative (MMI). In the workshop marine experts created mappings across different vocabularies. The tool used for the mappings relied on the Web Ontology Language (OWL) as the underlying model to manipulate the vocabularies and perform the relations among them. A basic ontology repository and several tools [7,8] were developed to facilitate mapping and creation of ontologies.

Categorizations for ocean observing data is a mechanism to filter and help a user retrieve the data of interest. An example of this approach is the Climate Data Library at Columbia University (IRI/LDEO), which provides categories for over 300 data collections from climate and other disciplines, using semantic web technologies [9].
tation of concepts requires that the concepts be expressed (or converted) to an homogeneous format (e.g., ontologies following W3C standards and recommendations) and that relationships among those terms be easily created. Furthermore, a one-stop shopping center to access these concepts and mappings facilitates community interactions and allows integration of the concepts and mappings into modern services-oriented architectures.

This paper presents an ontology service and ancillary tools, created by the Marine Metadata interoperability project, that facilitate the creation and access of ontologies and mappings, and services that facilitate dynamic categorization of observations. It also demonstrates the utility of these capabilities in cross-referencing data quality terms and flags between disparate institutional conventions.

To the authors’ knowledge, no other project in the marine domain is addressing the semantic issues explained in this paper. However, in other domains, such as biomedicine, such frameworks exist (e.g., BioPortal [10]). The best of those tools are being brought to the marine domain through the MMI project.

II. INTEROPERABILITY USING SEMANTIC WEB TECHNOLOGIES

The semantic web is about making the content on the web interconnected in a meaningful way for both humans and machines [6,11]. Key enabling technologies for the semantic web are formalized knowledge capture/representation (via ontologies) and global identification of web resources (via Uniform Resource Identifiers). Ontologies are used for knowledge representation via definition of concepts and their relationships in a particular domain [12]. Ontologies can be represented in many forms, with the Resource Description Framework (RDF) [13] and Web Ontology Language (OWL) [14] formats emerging as the most prominent in the context of the semantic web. Both concepts and relationships in an ontology are identified with corresponding Uniform Resource Identifiers (URI) [15], which, in general, do not have to be accessible or resolvable on the web. There are two URI subcategories: Uniform Resource Locators (URL), which, besides being identifiers, are in principle also intended to provide a mechanism to access the identified resource; and Uniform Resource Names (URN), whose main purpose is to identify but not necessarily to locate resources. Both ontologies and URIs provide the technological basis for a semantic framework as the one realized through the MMI ORR.

Most marine communities have a controlled vocabulary. A controlled vocabulary is a collection of terms (words, codes, etc.) that are defined, managed, and accepted in a specific community [16]. Vocabulary management involves careful selection of terms, maintenance of terms over time (i.e., addition, deprecation, modification), and presentation of the vocabulary in an accessible format.

Ontologies provide a formal and rich representation and accessibility mechanism for controlled vocabularies. For example, a vocabulary (ontology) about aquatic phenomena events may include terms for hurricanes, tsunamis, and harmful algal blooms (HAB). Rather that just being a collection of terms, the ontology representation provides a way to explicitly capture the fact that these terms correspond to aquatic phenomena events. In general, definitions often involve hierarchical relationships. As depicted in Fig. 1, we can indicate that Hurricane, Tsunami, and HAB are types of AquaticPhenomenaEvent (indicated with subClassOf relations). Particular events, eg., the Katrina hurricane, are captured via instanceOf relations to indicate their specific type. Several other aspects, eg., time and geographical extent of the event, can be captured by corresponding concepts and relationships using other complementary ontologies [17].

![Fig. 1. Example ontology.](image)

However, different communities may use different terminologies for the vocabularies and ontologies. For example, some vocabulary may use the term “red tide” instead of “harm algal bloom” to designate the same phenomenon. With no explicit association between these terms, a system will not be able, for example, to combine the corresponding instances while processing a search query for algal bloom events. Integration of information coming from these diverse sources require translation mechanisms to align the various concepts in a meaningful way. Various semantic web specifications, including the Simple Knowledge Organization System (SKOS) [18], provide mechanisms to link terms across multiple vocabularies with a variety of possible relationships. These alignments or mappings are also captured using an ontology representation.

III. MMI ONTOLOGY REGISTRY AND REPOSITORY

With the central goal of providing the marine community with supporting functions for semantic interoperability, the Marine Metadata Interoperability Project has developed the MMI Ontology Registry and Repository (ORR). The main components of ORR are shown in Fig. 2. With this system, users and software agents can find ontology concepts and associated annotations using semantic web based query mechanisms.

As shown in the right-hand side of Fig. 2, there are two main types of users of the MMI ORR.

- Data providers create vocabularies, term mappings, and ontologies, and register them in the ORR. As part of the process, semantic providers query, use, and map against already existing concepts in the repository. For annotation purposes, for example, a data provider can use this service as a first step to determine if any appropriate vocabularies already exist, either for reuse or extension to suit the particular application.
Consumers interact with ORR to issue queries, get terms, relationships, and categorizations. For example, upon a search for datasets or data providers involving a particular term (e.g., “algal bloom”), a data discovery tool can use the MMI ORR search service to enhance the search criteria to also consider related terms (e.g., terms that may be broader than or narrower than the specific term). These related terms are automatically obtained using mapping ontologies and semantic inference mechanisms through the MMI portal.

Although several sophisticated ontology development applications exist (see, e.g., [19]), the MMI semantic portal offers tools that facilitate the creation of vocabulary and mapping ontologies. Besides being fully integrated with the portal itself, these tools provide a user interface that is simpler and less demanding, in terms of ontology expertise, than existing ontology-related software tools. Once created, the ontologies can be seamlessly registered for subsequent maintenance and versioning in the repository.

The MMI ORR main components are:

1) **MMI Portal**: This is the entry point of the MMI ORR. The main page displays the list of the registered ontologies with options for search and grouping according to type of ontology (vocabulary, mapping), submitting user or organization, and ontology status. See Fig. 3. Browsing an ontology includes navigation options for all registered versions, individual classes and properties, and, as shown in Fig. 4 metadata attributes (properties associated with the ontology itself). Users can register a new version of the ontology and create new ontologies by uploading a local file or by creating a vocabulary or a mapping directly within the portal using the Voc2RDF or VINE tools, respectively, as described below.

2) **Voc2RDF**: This tool allows the creation of controlled vocabularies using a tabular based interface. It can also import content in comma-separated-values format. The defined terms are modeled as instances of a given class in the resulting ontology. Where appropriate, this class is associated with other classes in the registry to facilitate integration with other vocabularies.

3) **VINE**: This tool provides an intuitive interface where the user can easily navigate the existing entries in the repository and establish the desired associations between terms across vocabularies. Broadly adopted relationships in the semantic web, based on the SKOS vocabulary [18]—for example exactMatch, broaderMatch, among others—can be applied. An example of usage is given in Sec. V.

4) **SPARQL**: The ORR provides a SPARQL [20] endpoint for queries against the repository. SPARQL is the standard query language for RDF. The endpoint can be accessed via a simple form-based HTML page and also via HTTP requests that client applications can submit using common application programming interfaces. As an example, a simple yet powerful SPARQL query to obtain information about a particular term or concept in the repository is:

\[
\text{DESCRIBE } \langle \text{urn:ogc:def:crs:EPSG:6.5:4326} \rangle
\]

which produces all registered properties associated with the given URI. In this case, the term corresponds to a particular coordinate reference system as designated by the OGC. The output format can also be specified, with several options available to facilitate the integration of the response in diverse client applications.

5) **Reasoner**: The reasoner component allows ORR to make automatic inferences according to encoded rules in the underlying ontological model. For example, by using the available term mappings and the types associated to the terms, it can be inferred that “red tide” (see example in Sec. II) is of type...
AquaticPhenomenaEvent, a fact that was not originally stated in an explicit manner.

6) URI Resolution: MMI ORR provides support for automatically assigned and resolvable URIs, thus facilitating the exploitation of the registered information in the context of the web in general, and the semantic web in particular. Services for URI resolution, along with inferencing, enable better data discovery, integration, and processing capabilities. In the case of ontologies or terms whose URIs are not directly resolvable, i.e., because they are URNs or simply because the URLs are not resolvable, the ORR provides the mechanism for resolution via a suitable parameter in HTTP requests, for example, http://mmsw.org/ont?uri=urn:ogc:def:crs:EPSG:6.5:4326.

IV. Enhancing Sensor Observation Services with Semantics

Sensor Observation Services (SOS) [21] are being advanced by the Open Geospatial Consortium (OGC), as part of the Sensor Web Enablement (SWE) initiative [22]. SOS provides the interface for discovering, binding to, and interrogating individual sensors, instruments, platforms, and systems. SOS servers use the following standards:

- SensorML, Sensor Model Language: A model to describe sensors and sensor platforms.
- O&M, Observations and Measurements: A specification for encoding observations and measurements
- SweCommon: A model to provide encoding of data (observation results) across SWE technologies.

The general sequence of steps to obtain sensor metadata and observation data is as follows. As an SOS service, the data provider first returns a capabilities document upon a GetCapabilities request by a client. This document includes the identification of the provider itself and the description of the available streams in the system, which are organized in the form of observation offerings. An offering includes information about the period of time for which observations can be requested, the phenomena being sensed, and the geographic region covered by the observations. Once a client is interested in a particular sensor data stream, it may submit a DescribeSensor request to the provider. The corresponding response is a SensorML document describing the sensor and process lineage that generates the data stream. Next, the client requests the actual observation from the sensor system. This is done by submitting a GetObservation request, whose response is an O&M document.

OGC Ocean Science Interoperability Experiment (OSIE) [23], initiated by the OOSTethys community, is advancing the use of SOS standards to publish marine data. OOSTethys is an open source collaborative project intended to accelerate the pace at which ocean observations and associated technologies become more broadly and publicly available. OOSTethys and the OSIE experiment have defined the following set of use cases [24]:

1) Find sensors/systems in a region of interest, including proximity to user-defined location, from a heterogeneous and distributed network of sensors/systems.

2) Find sensors/systems with observations in desired time range.
3) Return data within a user-defined time range for a specific sensor or system of interest.
4) Get latest observation for a specific sensor/system.
5) Get descriptions of the sensors/systems used to obtain the measurements.
6) Return a description of the measurement processes, which could include quality control procedures.

The OSIE is experimenting in using semantic web technologies to categorize sensor observations. SOS offers a means to provide content rich semantics. Specifically, the different components of both sensor systems and observations can be identified by using semantic web technologies. Fig. 5 illustrates an example observation offering with embedded URIs. Each URI corresponds to a concept defined in an ontology and available through MMI services. When resolved, these URIs provide rich semantics to the corresponding elements in the offering.

Fig. 5. Example of a SOS observation offering with embedded URIs for semantic augmentation of the described service.

The categorization of SOS services provide the discovery functionality. This categorization requires that the portal and the data providers publish their controlled vocabularies in an homogeneous way. This is accomplished by using MMI tools. The data providers that participated in the OSIE experiment used recommended Climate and Forecast (CF) terms [25] while the MMI ORR used main NOAA/AOOS variables. The general process can be summarized as follows:

- Create an ontology representing the portal categories. In the experiment, the NOAA/AOOS variables were used.
- Data providers create an ontology representing the concepts used in their services. The use of specific vocabularies is encouraged, but not required.
- Create an ontology with mappings between the service providers terms and the portal categories. The VINE tool was used in the experiment to create the mappings.
- Register the ontologies (vocabularies and mappings) in the MMI ontology registry and repository.
Query the ontologies. The OpenIOOS portal client queries the MMI ORR using the SPARQL service, and retrieves the mappings of the portal terms to the services, allowing the portal to categorize the observations accordingly. See Fig. 6.

- Provider B has indicated via an SOS DescribeSensor response that an echoIntensityTest has been conducted. The service also describes a test failure by setting the echoIntensityFlag, but using the convention that 0 is a fail value and 1 is a passed value.

A semantic provider can map providerA:beamIntensityTest to be the same concept as providerB:echoIntensityTest (using the SKOS exactMatch relationship). The VINE tool generates and registers the corresponding ontology with the mapping. See Fig. 7. The disparate data quality flags may also be mapped. But, if the flags are integrated utilizing code space to specify the numeric or Boolean codes or values representing the corresponding conditions (pass/fail), a client may utilize a data stream of QC flags for their application, even when they have differing codes for pass/fail.

An illustrative use case is as follows:
- Provider A has indicated via an SOS DescribeSensor response that a beamIntensityTest has been conducted. The service also describes that a QC flag is generated in the data stream that is set to 1 upon failure of the minimum threshold (that is also obtainable via an SOS) and 0 if the test has passed.

V. EMBEDDING DATA QUALITY IN WEB SERVICES WITH SEMANTICS

As the national and international efforts for integrating disparate data sources progress, access to data quality information must be demonstrated and harmonized. Through the registration of provider specific terms, a parameter such as wave height may be described in a way that fully describes the source of information and its associated processing. For example: what was the source of data (pressure and horizontal velocity or surface-tracking and horizontal velocity; was the measurement made using an ADCP or a wave buoy)? What processing was used and how can it be referenced? What tests were applied to assure good data quality? Once the sensor and its lineage is fully described in SensorML, SOS can serve the information in a standards-based offering. If the responses contain embedded definitions referencing a registered URI, a client may better determine the differences and similarities in two or more sets of observation offerings.

An illustrative case is as follows:

- Provider A has indicated via an SOS DescribeSensor response that a beamIntensityTest has been conducted. The service also describes that a QC flag is generated in the data stream that is set to 1 upon failure of the minimum threshold (that is also obtainable via an SOS) and 0 if the test has passed.

VI. CONCLUSION AND FUTURE WORK

We described the MMI Ontology Registry and Repository (ORR), a set of services developed by the Marine Metadata
Interoperability Project to provide the marine community with supporting functions for semantic interoperability. The services allow data providers and users to include, use, and exploit semantic information in real world applications. We showed how OGC Sensor Web Enablement services can be enriched with semantic references that are resolvable against the repository, and along with powerful query mechanisms, such that users and client applications can easily exploit the registered information.

An initial experiment from the OOSTethys / Ocean Science Interoperability Experiment community demonstrated how semantic heterogeneity and information overload can be addressed, while separating the content model from the detailed semantics. The soft typing of SOS allows service providers to include their own semantics, thus enabling interoperability in the content model. Data portal categories and mappings between service providers terminologies and portal categories can be created using MMI services. The OpenIOOS portal then is able to present a categorization of SOS services using standards-based semantic web technology.

Examples of a SWE implementation of waves, describing tests and associated flags, can be found on the QARTOD to OGC (Q2O) website (http://q2o.whoi.edu). The Q2O project [26] implements the tests recommended by a grass-roots, inter-agency community of oceanographers, data managers and data aggregators, who are interested in developing minimum requirements in data Quality Assurance for Real-Time Oceanographic Data (QARTOD, http://qartod.org). The project will continue its work by also implementing instances for In situ currents, dissolved oxygen and CTD data.

Although the OOSTethys applications described in this paper use categorizations based on phenomenon types (salinity, water temperature, currents), we envision a more comprehensive system by also including categories like the following: Earth Realm (ocean, river, atmospheric observations), Platform type (research vessel, buoy, satellite), Quality control (quality levels), Discipline (biology, oceanography, chemistry), Scientist, Organization (company, agency, project), Problem (coastal hazard, climate change, sea level rise). The methodology described here can be used to incorporate these and other facets.

ACKNOWLEDGMENT

The authors would like to thank the MMI team, in special John Graybeal and Paul Alexander, for their contributions during the development of this project. This activity is supported by the National Science Foundation under awards ATM-0447031, OCE 0607372; Monterey Bay Aquarium Research Institute and the David and Lucile Packard Foundation; Office of Naval Research, Award N00014-04-1-0721; by NOAA Ocean Service Award NA04NOS4730254. Funding for demonstration of integration of data quality was funded under NOAA’s Cooperative Agreement FY 2007 Regional Integrated Ocean Observing System Development (NOS-CSC-2007-2000875). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect those of the funding agencies.

REFERENCES