# Construction of Marine Vocabularies in the Marine Metadata Interoperability Project

**L. Bermudez**
MBARI

**J. Graybeal**
MBARI

**Anthony W. Isenor**
Defence R&D Canada – Atlantic

**Roy Lowry**
British Oceanographic Data Centre

**Dawn Wright**
Oregon State University

*Abstract* - **Data producers often overlook existing metadata descriptors and controlled vocabularies for describing their data, opting instead to create custom descriptors and vocabularies. One of the objectives of the Marine Metadata Interoperability (MMI) project is to reduce this vocabulary proliferation. The work is accomplished as community collaborations, which are supported via the content management framework of the MMI web site. Services, processes, techniques, and advice are all offered via the community web site supported by the MMI (http://marinemetadata.org). As part of the services, MMI has created and made available marine ontologies based on existing vocabularies. Ontologies are an explicit and formal specification of mental abstractions. Ontologies are being published using the Web Ontology Language (OWL), the ontology expression tool recommended by the World Wide Web Consortium (W3C), and made available using web services. By providing these services using common terminology, the MMI effort facilitates discovery, sharing, and markup of marine data.**

**The MMI methodology for creating the marine ontologies is composed of: identification, harmonization, alignment and mapping, and publication. First the marine vocabulary is identified. The namespace and the required transformation are then documented. Then the vocabulary is harmonized with the other vocabularies by transforming the vocabulary into a common structure, in this case OWL format. Having all the marine vocabularies harmonized in OWL allows alignment and mapping between the vocabularies. OWL allows the required mapping relationships such as “same As”, “narrower Than”, and “broader Than”. Finally, the ontologies are published via web services. The first three parts of this methodology, which form the foundation for the development of true semantic interoperability, will be discussed in this paper.**

## I. INTRODUCTION

In many subject communities, there is typically community-specific terminology that collectively represents a specialized vocabulary for that community. When these vocabularies are formally managed, they become a controlled vocabulary. Controlling the vocabulary is useful because it helps the community avoid misspellings and avoid the use of arbitrary words that cause inconsistencies. Also, use of controlled vocabularies helps to solve semantic incompatibilities among distributed systems [1-4].

A controlled vocabulary may have many functions. In ocean science, two common vocabulary functions are resource description and data discovery. The Parameter Usage Vocabulary (PUV) is used to describe resources. The PUV is made up of terms used to tag individual data values with information about what was measured, perhaps the sphere where it was measured, and how it was measured. These vocabularies usually consist of a numeric or alphanumeric key (sometimes termed a parameter code) that is used to tag the data, plus a lookup resource presenting the semantics for the tag. Examples of PUV related to the marine domain are presented in Table 1.

The subject matter of discovery vocabularies cover a wide metadata spectrum including descriptions of parameters, platforms, instruments, projects, geographic areas, etc. For this work, we deal with parameters and thus describe the vocabulary as a Parameter Discovery Vocabulary (PDV). The PDV contains lists of terms (sometimes called keywords) that are used to standardize the process of finding data. Examples of PDV related to the marine domain are presented in Table 2.

TABLE 1 PUBLIC FORMATS OF PARAMETER USAGE VOCABULARIES (PUV)

| **PUV** | **FORMAT** | **URL** |
|---|---|---|
| BODC | Comma Separated Value | http://wwwtest.bodc.ac.uk/data/codes_and_formats/parameter_codes/bodc_para_dict.html |
| U.S. JGOFS Dictionary of parameters | HTML | http://usjgofs.whoi.edu/datasys/param_master.html |
| IOC GF3 parameter codes | HTML | http://ioc.unesco.org/oceanteacher/resourcekit/M3/Formats/Integrated/GF3/GF3.htm |
| SEACOOS | Comma Separated value | http://twiki.sura.org/twiki/pub/Main/DataStandards/seacoos_draft_data_dictionary_v2.0.csv |
| CF | XML | http://www.cgd.ucar.edu/cms/eaton/cf-metadata/standard_name.xml |

TABLE 2 PUBLIC FORMATS OF PARAMETER DISCOVERY VOCABUALRIES (PDV)

| **PDV** | **FORMAT** | **URL** |
|---|---|---|
| GCMD | HTML | http://gcmd.gsfc.nasa.gov/Resources/valids |
| BODC Discovery | Comma Separated Value | http://wwwtest.bodc.ac.uk/data/codes_and_formats/parameter_codes/bodc_para_dict.html |
| AGU Index Terms | HTML | http://www.agu.org/pubs/gaplist.html |
| MEL | HTML | https://mel.dmso.mil/docs/metadata_guide/section_6.htm |
| NOAA CoRIS Thesauri | PDF | http://www.coris.noaa.gov/backmatter/keywords/discovery_thesaurus.pdf |

Historically, vocabularies have evolved within organizations. When larger communities form, the issue becomes the relating of terms used in the different vocabularies. The process of formally relating terms between vocabularies is called mapping. Mapping exercises can be complicated, time consuming, and require considerable subject matter expertise. This is due in part to the detailed definitions that often accompany terms in the vocabularies.

The mapped vocabularies must then be represented in a formal system. The process of harmonization is the representation of vocabularies in a system that allows the formation of relationships among the terms. The harmonization process thus creates a single repository for mapped vocabularies.

## II. MARINE METADATA INTEROPERABILITY PROJECT (MMI)

The Marine Metadata Interoperability Project was created to address metadata management in the oceanographic science community. Operating since late 2004, the Project initiated several activities to enhance scientific data management for marine projects. Many of the techniques, tools, and services provided by the MMI Project will be equally applicable to other scientific domains; all are welcome to use and contribute to the project's activities, which are open and public.

MMI devoted its initial efforts to creating a web site that would host its many activities and resources, and form a central presence around which the marine metadata community could organize its collaboration. The collaboration has grown steadily, with 140 members from over 10 countries. The first activity provided and organized references about current metadata activities in the marine sciences; over 500 references are documented and searchable on the site. As the site matures, more time will be allocated to evaluating these resources and providing guidance to all levels of marine scientists and data management professionals.

The project also provides tools, processes, and services to enhance the availability and use of metadata in the marine domain. A vocabulary-mapping workshop scheduled for August 9-11, 2005 will illustrate the approaches described in this paper, and give hands-on experience to domain experts in the process. Services being developed by the MMI team will enable the application of these concepts by a variety of data applications, which in turn will enable demonstrations of scientific data access using advanced metadata concepts.

The science and data management community, both via the MMI Steering Committee and by the many national and international contributors, guides all of the MMI work. The project strongly encourages membership and participation of anyone with related technical interests or needs.

## III. ONTOLOGIES AND CONTROLLED VOCABULARIES

Controlled vocabularies can be represented in different organizational systems. Hodge’s [1] classification of knowledge organization systems includes: term lists, glossaries, dictionaries, gazetteers, classification and categories, taxonomies, relationship lists, thesauri, semantic networks, and ontologies (Fig. 1). A further classification of these vocabularies [2], presented a distinction of these systems based on the level of conceptualization .

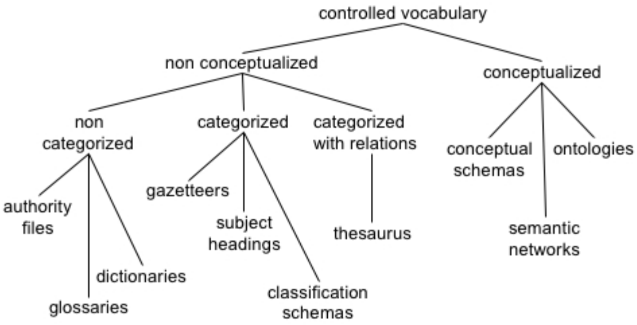

Fig. 1 Knowledge Organization Systems [1, 2]

Typically, marine vocabularies live inside databases or inside software programs. They are publicly available in HTML pages, in XML documents and ASCII files (Table 1 and Table 2). One of the goals of MMI is to bring all these vocabularies together using a single system, with automatic conversion from their original format and with the possibility to create relations among them. We select ontologies as the vehicle to represent controlled vocabularies and the Web Ontology Language (OWL) as the expression medium.

An ontology provides the structure for the controlled vocabulary similar to a dictionary or a thesaurus. The difference is that an ontology includes relationships such as whole-part, cause-effect, or parent-child relationships and rules and axioms (e.g., stating that a relation is transitive).

Ontologies are also important in the context of marine Geographic Information Systems (GIS). In GISs the specifications of relationships between feature data sets, feature classes, and object tables are critical in the formulation of customized data models, upon which the exchange of data may be based, as well as a foundation for geospatial analysis [3, 4]. In terms of marine metadata, it is often critical for the metadata of a data set to be related to the metadata of an expedition, a mobile survey, an instrument, etc., each of which should be an important part of an ontology as well.

OWL provides a means to accurately formulate these relationships. OWL is recognized as a core component of the Semantic Web [5]. The Semantic Web may be described as a universe of metadata and ontologies, expressed in machine-readable format along with software tools that allow the understanding of semantic relations among heterogeneous and distributed resources in the Web [6]. It is based on technologies recommended by the World Wide Web Consortium, such as the eXtensible Markup Language (XML), Resource Description Framework (RDF), and Uniform Resource Identifier (URI).

## IV. IDENTIFICATION

The first stage for mapping is identification of vocabularies. The identification process takes advantage of the MMI collaborative framework, where contributors freely add vocabulary references to the MMI web site. Then, using the MMI discussion lists, strategies are developed to map the vocabularies to one another.

## V. HARMONIZATION

### *A. Definition*

Two or more vocabularies are harmonized if they are represented in a unique and similar system, which allows for the creation of relations among the terms in the vocabularies (Fig. 2). The various systems allow different relationships to be modeled. For example, if two vocabularies are harmonized in a system that is based on the tree-type model, then the only relations allowed among the terms are the parent-child relationship.

A relational database schema can be another harmonization mechanism [7], allowing the terms to be expressed following a unique schema and stored in table records. Relations among the terms depend on the design of the schema. For example, a table could have an attribute parent, which points to another record in a table where the parent term resides. This is common in marine GIS, where most of the applications are going to be implemented in either a relational or object-relationship database management system.

OWL is an example of a harmonization system that utilizes the ontology relationship model. In particular, OWL, allows for the creation of classes (that could be considered categories) and individuals (members of categories). Individuals are related to each other via properties. Properties such as "hasParent" can be created and used to relate individuals in the vocabularies. The idea of presenting controlled vocabularies in an ontology form is not new (e.g., RDF encoding of controlled vocabularies in [8-10]).

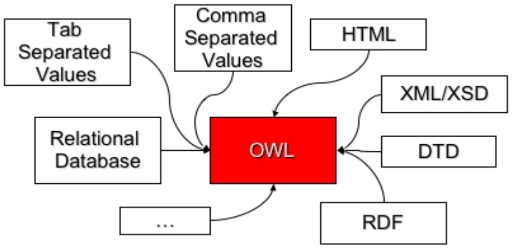

Fig. 2 Harmonization of vocabularies using OWL.

As noted above, vocabularies are found in many different formats. Sometimes their terms are organized (e.g., in a hierarchy) and sometimes they are not. When term organization is minimal, selection of what should be a class (a general group of terms) or an individual (a specific term) is often nontrivial.

An OWL ontology is basically composed of classes, properties and individuals. The questions that the following sections will answer in the context of MMI include what should be a class, a property and an individual when constructing marine ontologies from existing vocabularies.

### B. *Namespaces*

Namespaces are used to uniquely identify each term. To facilitate the process of publishing, MMI created a namespace to hold the different ontologies created, as well as to keep tracking of the different versions. The format of the namespace is:

**http://marinemetadata.org/yyyy/mm/abbreviation**

Examples of namespaces are:

**http://marinemetadata.org/2005/02/ioos**

**http://marinemetadata.org/2005/02/gcmd**

### C. *Selection of classes*

A class is a term that represents a category of individuals. For example, Marine-Variables may be considered a class that categorizes marine terms by segregating these terms from other spheres (e.g., atmosphere). When harmonizing vocabularies, a strategy was developed to identify class names from the various vocabulary encodings (Table 3). As examples, in the Global Change Master Directory (GCMD) vocabulary the term 'variable' is a category, while in the British Oceanographic Data Centre (BODC) vocabulary the term 'parameter' is a category. In the OWL ontology, these category names are expressed as a class.

TABLE 3 SOURCES OF LABELS FOR CLASS NAMES

| **Encoding** | **Source for class** |
|---|---|
| Plain list | Title (heading) of the list |
| Table | Title of the table |
| XML file | Element tag |
| RDBS | Name of the table (If the term and its attributes are stored in a table) |
| UML | Name of the class (If the term and its attributes are stored in a class) |

### D. *Selection of individuals*

The terms in a vocabulary that are not the main categories are said to be individuals. You can ask the following question to determine whether an individual is the appropriate type: Is the proposed individual *a member* of the class? For example, in chemistry we can say, "iron is a *member of* the class elements". In this case iron is an individual, while elements is a class.

### E. *Selection of properties*

Each individual has its own collection of properties in the ontology. For example, properties of an individual may include short-name, id, definition, date of creation, author, and other descriptive characteristics. These properties may be either data-type properties or object properties (Fig. 3). A data-type property (owl:datatypeProperty) is used when the range of property values can be represented using a number or string. The object property (owl:objectProperty) is used if the range of property values is a resource (e.g., a URI).

Attributes of the individual terms in the original vocabularies will be mapped to the properties expressed in OWL.

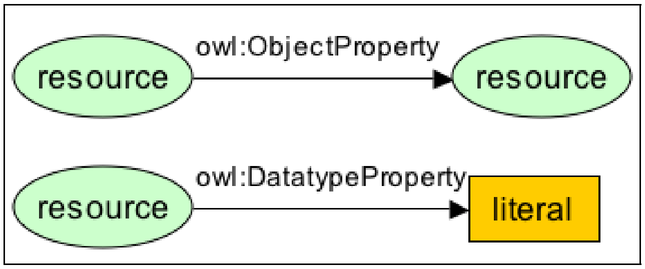

Fig. 3 Owl Properties

To select the properties to be included in the OWL representation, we identify three potential sources: 1) Owl built-in properties, 2) Dublin Core properties, or 3) Other properties. OWL built-in properties (which include RDF properties) are used for attributes that label a term (e.g., the OWL built in property rdfs:comment, see [10] for complete list of OWL properties). Dublin Core properties are also useful because of their wide acceptance, tool support (e.g., OWL and JENA) and compatibility with RDF. The third source, denoted other properties, includes any arbitrarily assigned label. If using OWL, these properties will have a unique namespace to avoid semantic conflicts.

One expeditious methodology for creating class properties from an available vocabulary is to mimic the attributes of the vocabulary term as data-type properties in OWL (the third property selection approach). However, for harmonization of multiple vocabularies this may produce an assortment of similar, yet discrete, properties. At MMI, we recommend a minimum set of properties predominately based on OWL and Dublin Core.

The minimum suggested set of properties for the harmonization of marine parameter usage and parameter discovery vocabularies is presented in Table 4. Note that an *x:* denotes a user defined namespace for an arbitrary property.

TABLE 4 MINIMUM SET OF PROPERTIES

| **Property** | **OWL property** |
|---|---|
| Unique identifier | rdf:ID |
| Original Unique identifier | x:originalID<br>*Note: it could be any namespace and any local name.* |
| Definition of the term | Three choices:<br>dc:description, rdf:comment or x:description<br>*Note: the last one could be any namespace and any local name.* |
| Units | See next section |

### F. *Units Strategy*

There are two possible strategies for the incorporation of units into the ontology (Fig. 4). Units may be encoded as string values, or assigned unique ids. In the first case, a datatype property (owl:datatypeProperty) is used to store the string value in the property that indicates the unit.

In the second case, objectProperties are used to link the particular vocabulary individual to a unit expressed as a resource. The units expressed as resources are individuals of a new class (e.g. x:UnitTypes).

Fig. 4 presents the previous two cases. A box represents a literal (e.g. String values) and an oval represents a resource.

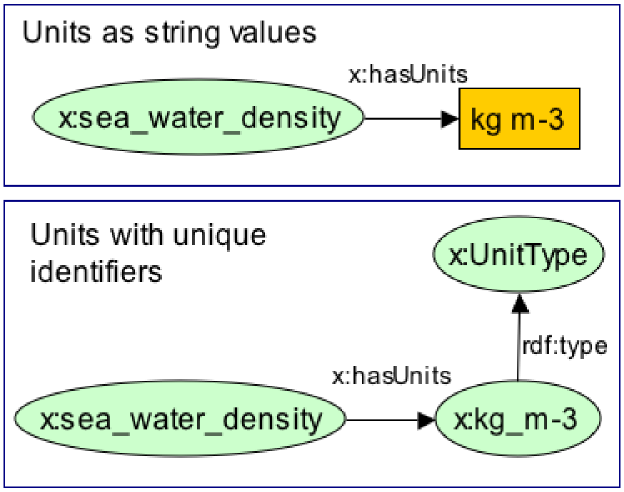

Fig. 4 Representation of units in an ontology

## VI. ALIGNMENT AND MAPPING

### G. *Definition*

Vocabulary alignment and mapping is the process of relating the terms contained in two or more vocabularies. Different kinds of relations can be depicted depending on the system used to harmonize the vocabularies. Some examples are:

- Taxonomy relations: Specific and general relations (e.g., A is more general than B).
- Tree relations: (e.g., A has a leaf B).
- Graphs relations: Node A is connected to Node B
- Thesaurus relations: Narrower-than, broader than, same as, similar-to (e.g., A is broader than B).
- Conceptual model relations: Class-subclass relations or class-individual relations (e.g., A is superclass of B).

The relation representation is also predetermined by the model or system where the mapping is talking place. If the representation mechanism is a table, writing the two terms in the same table row could mean that they are the same. If a graph is used instead, a line connecting two terms could also represent the "same as" relation. The difficult part of this process is how to represent the relations in a format that a computer program could understand and utilize. The World Wide Web Consortium has a project called Simple Knowledge Organization System, (SKOS, [8]) that seeks development of specifications and standards to support the encoding of controlled vocabularies to be used in the Semantic Web. MMI is taking a similar approach as SKOS and aims to be interoperable with SKOS recommendations.

### H. *MMI Strategy*

To simplify the integration of vocabularies, the MMI strategy uses three relations: *sameAs*, *narrowerThan* and *broaderThan*. The first relation is an OWL property, while the other two are Dublin Core types. If needed, these properties can be set equivalent to SKOS defined properties *exactMatch*, *narrowerMatch*, *broaderMatch*, respectively. Note that *sameAs* is a symmetric relation (A = B -> B = A), while *narrowerThan* and *broaderThan* are inverse to each other (A > B -> B < A) and are transitive (A>B and B>C -> A>C). This will allow inferring of vocabularies. An example of mapping the term bioluminescence among three controlled vocabularies using *owl:sameAs* is shown in Fig. 5.

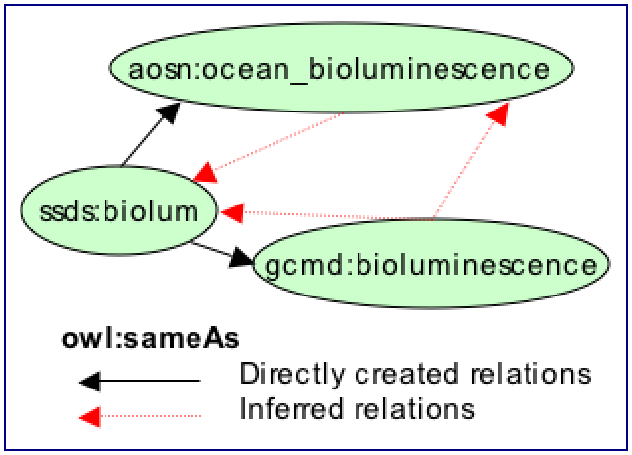

Fig. 5 *sameAs* relations from an MMI mapping exercise (http://marinemetadata.org/ontMapSession2)

These are basic relationships that, although allowing further inferences, do not fully describe the relationships between different terms. As time permits, in the mapping workshop and thereafter, we intend to expand the list of supported relationships (e.g., A is a *typeOf* Phenomena).

We will also need to develop processes by which these mappings can be maintained and validated by the community. In this way, the initial MMI efforts will continue to sustain community credibility, while continuing to maximize interoperability.

## VII. CONCLUSIONS

MMI facilitates the discovery, sharing, and markup of marine data. This paper has outlined an important means to this end: a specific methodology for creating and sharing marine ontologies based on controlled vocabularies. When published via web services, these ontologies ultimately make scientific data easier to distribute, advertise, reuse, and combine with other data sets.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the National Science Foundation, MBARI and the David and Lucile Packard Foundation, the MMI Steering Committee (Robert Arko, Philip Bogden, Julie Bosch, Ben Domenico, Daphne Fautin, John Graybeal, Steven Hankin, Roy Lowry, Stephen Miller, Mark Musen, Lola Olsen, Michael Parke, Stephanie Watson, Robert Weller and Dawn Wright ), the MMI Topic Leaders (Dru Clark, Roy Lowry, Robert Arko, and Julie Bosch), and the other members of MMI.

## REFERENCES

[1] G. Hodge, Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files, The Digital Library Federation, Council on Library and Information Resources, Washington, DC, 2000.

[2] L.E. Bermudez, Ontomet: Ontology Metadata Framework, Ph.D. Thesis, Drexel University, 2004.

[3] D.J. Wright, M. Blongewicz, P.N. Halpin, J. Breman, A new object-oriented data model for coasts and oceans, 6th International Symposium on Computer Mapping and GIS for Coastal Zone Management, Arberdeen, Scotland, 2005.

[4] M. Zeiller, The ESRI Guide to Geodatabase design, ESRI Press, 1999.

[5] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American 184(2001) 34-43.

[6] M.J. Egenhofer, Toward the Semantic Geospatial Web, Tenth ACM international symposium on advances in geographic information systems, ACM Press, McLean, Virginia, USA, 2002.

[7] Fisheries and Oceans Canada, Marine

Environmental Data Service (MEDS), http://www.meds-sdmm.dfo-mpo.gc.ca/meds/About_MEDS/standards/login_e.asp

[8] World Wide Web Consortium, Simple Knowledge Organization System (SKOS), http://www.w3.org/2004/02/skos/

[9] A. Magkanaraki, S. Alexaki, V. Christophides, D. Plexousakis, Benchmarking RDF Schemas for the Semantic Web, International Semantic Web Conference, Sardinia, Italy, 2002.

[10] D.L. McGuinness, Ontologies Come of Age, in: D. Fensel, J. Hendler, H. Lieberman, W. Wahlster, (Eds), Spinning the Semantic Web, The MIT Press, London, England, 2003.