ICES CM 2006/M **Environmental and fisheries data management, access, and integration**

Semantic Interoperability: A Goal for Marine Data Management

Roy Lowry, Luis Bermudez & John Graybeal

Whilst the problems of syntactic interoperability between data sources have reduced significantly through the acceptance of standard formats such as NetCDF and Microsoft package formats, the problem of semantic interoperability remains. Semantic interoperability is like language translation for science—making sure the terms we each use in our data can be understood by people and computers, allowing data to be exchanged reliably and automatically. There are steps that those who are in the enviable position of building a marine data system from scratch can take to maximise the possibility that their system will achieve semantic interoperability. From defining terms, to using standards, to linking data and metadata, following simple guidelines will make a system's data available to many more researchers, much longer into the future. If users are to adopt vocabulary and content standards then those responsible for maintaining the standards need to deliver service that includes versioning, convenient access to current and past versions and timely response to requests from users for updated content. In most cases this has not been achieved to date. Those with either legacy data or strongly enforced local terminology will need vocabulary maps if they are to interoperate semantically. The Marine Metadata Interoperability Initiative (http://marinemetadata.org) has been developing high quality tooling to facilitate building this type of map. The time has now come for the marine science community to help evolve these tools and apply them to capture its domain expertise into semantic web resources that will form the basis for future semantic interoperability.

Contact author:

Roy Lowry, BODC, Joseph Proudman Building, 6 Brownlow Street, Liverpool L3 5DA, UK [tel: +44 151 795 4895, e-mail rkl@bodc.ac.uk].

Introduction

Interoperability is defined as the ability to share data from multiple sources as a common resource handled through a single tool. Four different levels of interoperability can be depicted: system, syntactic, structural and semantic [3]. System interoperability refers to protocols, hardware & Operating Systems, spatial data files and Database Management Systems interoperability. Syntactic interoperability refers to the ability to load data from multiple sources into a common tool. Schematic interoperability refers to the structural organization of the data (e.g. XML schema); and semantic interoperability refers to the ability of making sure the terms used in data can be understood by people and computers, allowing data to be exchanged reliably and automatically. An example of interoperability is a data discovery browser searching through the metadata catalogues of tens or even hundreds of oceanographic data repositories using a single structured query and returning results as if the query had been directed at a single database. A significantly more

ambitious example is a daily average global ocean salinity field built dynamically from moored instrument, thermosalinograph, CTD and towed undulator data combined with model output, again sourced from tens or hundreds of repositories.

Interoperability has been the elusive Nirvana for oceanographic data management since the establishment of the Intergovernmental Oceanographic Commission (IOC) International Oceanographic Data and Information Exchange (IODE) network in 1961[10] Much of the early work of IODE was targeted at the syntactic interoperability level. However, some work was also done to address semantic heterogeneity issues. With the current advent of the Semantic Web, creating and inter-relating controlled vocabularies this is made much easier, but there is still a need to define, publish, version and govern the base vocabularies. This paper discusses semantic lessons learned from working with GF3 codes, and recent approaches taken by BODC and the Marine Metadata interoperability to maximize semantic interoperability in the marine community.

Development of Syntactic Interoperability

In the early 1980s syntactic interoperability was a major issue. Virtually all data transfers involved the physical exchange of 9-track magnetic tapes. Different computer manufacturers encoded data using differing byte lengths, word lengths and character encodings. Each manufacturer also developed their own unique style for packing the bytes of data into physical files or groups of files onto the tape. As a result, it was not uncommon for it to take two or three days' work to get to the stage of listing a file. Beyond this lay deciphering the logic of the data encoding format, involving reading complex documentation or extensive detective work before the nuggets of data held within could be utilised.

The early interoperability work by IODE, such as the development of the GF3 format [7] focused primarily on addressing syntactic interoperability through standardisation of physical tape and character encoding, which succeeded in specifying tapes that could be read on any system with ease. However, what was gained here was lost with the demise of the 9-track tape. Logical formatting issues were addressed through the development of a Fortran API and provision of standard utilities. These could be used to access data, but required considerable programming skill. Contrast this with the current situation where the Excel spreadsheet is the universal currency for data exchange between biological and chemical oceanographers, and NetCDF [11]—particularly standardised variants such as the CF Conventions [5], transported over a protocol such as OpenDAP [13]—provides for interoperable large-volume data exchange.

The fact that scientists with no programming skills are able to visualise data from many sources within minutes of receipt shows the degree to which syntactic interoperability problems have been overcome. However, this quick look at data is rarely accompanied by the information required for its complete understanding and reliable interpretation.

GF3 and Semantic Interoperability

The provision of information for understanding is the realm of semantic interoperability. GF3 did try to address this through the development of code tables [15] covering a range of subjects including countries, platforms, and instruments, but with particular emphasis on the description of measured phenomena. These controlled vocabularies provided standardised keys (the codes) to represent the entities and terms to describe them. However, their maintenance and distribution was based on distribution by paper publications, which made changes both expensive and infrequent. Consequently, the tables failed to keep up with scientists' developing needs and were largely ignored by most of scientific community. The GF3 code tables needed the internet, but it hadn't been invented. Although the wider scientific community did not adopt the GF3 code tables, they were used heavily by the IODE National Oceanographic Data Centre network. Sometimes this worked extremely well, but a number of problems and examples of bad practice have emerged in 20 years of use that provide valuable lessons for those currently working to develop semantic interoperability.

Initially, there was strong content governance for the GF3 parameter codes through the work of the IODE Group of Experts on the Technical Aspects of Data Exchange (GETADE). However, the problems of maintenance by printed publication ground this to a halt. There is no point specifying change if that change cannot be implemented and distributed. Once active content governance ceased, the vocabularies soon became stale and virtually worthless. For example, country codes were revised at a meeting of GETADE in 1981, but never since. Subsequent political change makes these codes not just useless but the cause of much ill feeling: Croatians are very unhappy when offered the country code for Yugoslavia to use in their metadata. If a controlled vocabulary is to effectively enhance semantic interoperability it must have active, strong and responsive content governance.

The breakdown of content governance led to proliferation of local copies of the code tables. These were usually created as short term fixes to maintain operations, but they soon became permanent components of the infrastructure. Local list maintenance in multiple locations is a process akin to evolution that produces differences of increasing significance through time. Re-establishing interoperability between these divergent entities through mapping and rationalisation requires significant effort. In spite of this weakness, there are examples of GF3 code table diversification on every scale from the oceanographic data centres of the world to the databases of a single data centre. Local list evolution is such an easy short-term fix that its long term consequences are ignored.

The GF3 code tables delivered keys together with terms but no definitions, which opened up the possibility for a term to acquire more than one meaning. Inevitably, this happened and even affects the most common words. For example, 'cruise' is understood by some to mean any data or sample collection activity, such as hand-picking mussels for analysis off rocks, whereas others understand it to mean the port-to-port operation of a research vessel. The key phrase in the previous sentence is 'understood to mean'. When it comes to semantics, the value of explicit, unambiguous definitions is paramount.

During the development of GF3, significant effort went into the design of the code syntax, loading them with semantics to make them memorable and designing them for use as sort keys. This was regrettable for two reasons. The first is that the practice is not scalable. Whilst it is possible to semantically differentiate tens of entities using four bytes, it is totally impossible when the entity count runs into thousands. The second is that such importance was attached to the codes that new lists were built for no reason other than changing the code syntax. As with local copies, these multiple lists diversified through evolution and it has required significant effort to bring them back together. Keys are items for low-level computer communication and as such should be both semantically neutral and not forced upon the human users.

Detailed examination of localised list developments revealed that many of the changes were made to use vocabulary population to either extend or correct errors in data models. For example, it has become common practice to extend one-to-one relationships in data models to one-to-many relationships by adding entries to vocabularies that are themselves lists, such as 'BOFS and JGOFS' or even the semantically useless 'multiple instruments'. Mapping lists rather than single entities is both more difficult and produces a less concrete result as there will rarely be a 'same as' relationship between terms. In other cases, the list entity definition has been manipulated resulting in semantically impure lists that are even more difficult to map. Controlled vocabulary terms should be restricted to a single entity conforming to a fixed definition and weaknesses in data models should be corrected by modifying the data model, not its vocabularies.

How to Achieve Semantic Interoperability (Ideal World)

Semantic interoperability is like language translation for science—making sure the terms we each use in our data can be understood by people and computers, allowing data to be exchanged reliably and automatically. There are steps that those who are in the enviable position of building a marine data system from scratch can take to maximise the possibility that their system will achieve this.

- Take note of the work of others and take on board the lessons learned by the operators of existing data systems, such as those documented in the previous section.

- Avoid the overuse of free text fields in metadata. Whilst there is a role for free text in metadata formats to provide human-readable dataset summaries such as the ISO19115 abstract field, there are many data and metadata formats that have free text for critical metadata fields, such as spatio-temporal co-ordinates. Free text may be easier to generate than structured metadata, but its significantly lower value overwhelms any savings.

- Wherever possible populate metadata fields from a pre-existing standard controlled vocabulary. Whilst the adoption of standards might require more work at the outset and some degree of compromise, it greatly facilitates interoperability.

If there is no suitable vocabulary available then generating a new one should be done in collaboration with others so as to lay the foundation for new standards that may be used to enhance future semantic interoperability. For example, Bermudez and Piasecki [1] propose a methodology to create dynamic community profiles.

- Ensure that your data and metadata are as fully integrated as possible (that is, your data is self-describing). The worst-case scenario is a data file with rows of neat columns of text, but no other descriptive information. Beyond the simple expedients of descriptive headers (at least including column names!), data files should contain embedded references to their descriptive metadata, and data packets in an observing system should include pointers or indices to similar descriptions maintained elsewhere in the system. Ideally, your data resource should be usable by someone who stumbles across it without knowing how, when, or by whom it was created.

Adoption of vocabulary standards is a recommendation that has been made in oceanographic data management more often than it has been adopted. There are a number of reasons for this, including the low priority assigned to interoperability in the past and an unwillingness to compromise or even communicate. However, much of the blame can be laid on an issue raised in the GF3 discussion, namely poor governance. Governance includes two aspects: technical governance and content governance. Technical governance covers the storage, maintenance and delivery of a vocabulary. Content governance is the process by which changes to the vocabulary, adding new terms or modifying existing terms, is managed. The totally inadequate technical governance of printed publications was followed by distribution based on regularly updated files on FTP sites such as the recently superseded BODC system, web query interfaces such as the MEDS system [6] and CF Standard Names [4] and downloadable lists such as the Global Change Master Directory (GCMD) Keywords [12]. Content governance has been undertaken by individual 'dictators', closed committees, e-mail discussion lists or even an anarchic free-for-all. E-mail lists seem to work best, with the CF discussion list an excellent example of how content governance should be managed.

Consultations with potential users and metadata specialists from a range of domains established that vocabulary governance should:

- Provide convenient access to an up to date version of each vocabulary served
- Provide rigorous versioning for each vocabulary served
- Provide access to all previous versions of each vocabulary served through maintenance of maintenance audit trails
- Provide a guarantee that maintenance will not break existing vocabulary usage
- Provide semantically neutral identifiers (keys) for vocabulary entries
- Provide rapid but considered and well-managed response to requests for vocabulary extension or modification

None of the examples quoted above satisfy all of these requirements, with version management and audit trails being particularly poorly represented. To address this, BODC developed a vocabulary technical governance system [16] for the NERC DataGrid and SeaDataNet projects with fully automated version management in the back end database and a Web Service-based front end. This is supported by a content governance e-mail discussion list within the European SeaDataNet project, which will be expanded onto a global scale under the auspices of the IOC MarineXML Steering Group and possibly into other domains such as CF. Such a broad community infrastructure should provide a reference model for developing standard vocabularies that are truly worthy of adoption.

How to Achieve Semantic Interoperability (Real World)

The most common use case for semantic interoperability is the requirement to seamlessly combine data or metadata from multiple repositories into a single application. This raises this issue of legacy data, which will need vocabulary maps if they are to interoperate semantically. The Marine Metadata Interoperability Initiative [8] has been developing high quality tooling to facilitate building this type of map. The time has now come for the marine science community to help evolve these tools and apply them to capture its domain expertise into semantic web resources that will form the basis for future semantic interoperability.

The Marine Metadata Interoperability (MMI) Project received its first funding in 2004, for the purpose of creating more advanced, interoperable, and community-based metadata practices. The project quickly attracted an international community of technical and scientific contributors, who collaboratively document and create solutions for common metadata problems.

MMI pursues multiple strategies to maximize its benefit to the community. At the most basic level, MMI provides an information clearinghouse that captures and organizes information about key metadata projects and solutions. This clearinghouse, like most MMI work, is hosted at the MMI web site, http://marinemetadata.org. All of the other efforts described in this section are managed through public pages at the MMI site.

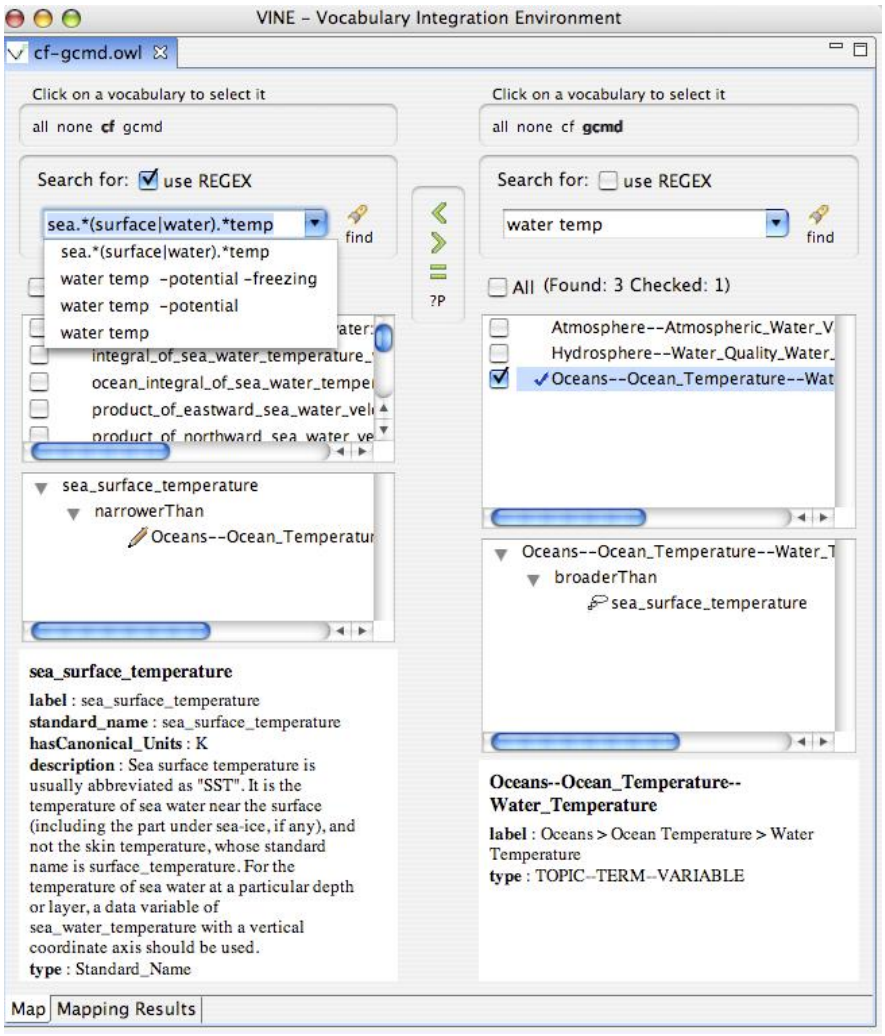
A fundamental conceptual framework, created in the first few months of the project, separates metadata standards into three categories: transport protocols; content standards; and vocabularies. While real-world solutions sometimes mix these three concepts, the MMI team has found them extremely effective divisions of areas of concern. The transport protocol section addresses how metadata moves from one system to another, and is relatively static. The content standards specify what fields can be described in a metadata document; the standards produced by FGDC, and now ISO, are the best-known examples. Given a list of concepts to document (for example, sensor type), a truly interoperable solution indicates what values are acceptable in many of the fields. These acceptable values are usually specified by means of a controlled vocabulary, the third type of standard documented on the site. For each type of standard, MMI provides some

introductory descriptive material, and then lists instances known by the site's contributors.

A number of other topics are directly addressed by the site. Sections on metadata tools, guides, and events bring together key information at a single place. To date MMI has provided a small amount of context on each external reference, but in the future will provide more detailed analyses or user reviews of the referenced material.

The MMI Project has progressed beyond the information clearinghouse stage, now providing strategic solutions for a number of common metadata challenges. To enable common access to different controlled vocabularies, MMI provides tools to convert vocabularies to a common format (Voc2OWL), and to map the terms in different vocabularies to each other (VINE—See Figure 1). These tools were demonstrated and heavily used in a vocabulary mapping workshop [9] produced by MMI, and have been released to the public via SourceForge [14].

Figure 1. VINE – Tool developed by MMI to search and map terms across vocabularies



Having developed tools to capture knowledge about vocabularies in the science domains, MMI also created some prototype services to use the results, and some demonstrations of the technologies within an overarching architecture. These projects continue, with modifications based on community input and contributions.

Finally, MMI has tried throughout its history to enable the building and collaboration of communities, so that narrow "stovepipe" developments are avoided and common, interoperable solutions are emphasized. There are many marine metadata standards in use in the marine community; MMI is helping users to be aware of the range of choices, identify choices that are most useful and tasks that need to be addressed, and nudge the existing efforts toward consolidated, reusable, and interoperable approaches. With its recent 3-year grant from NSF, MMI is poised to continue these objectives.

Conclusions

It is now widely recognised in many scientific domains, particularly oceanography, that significant benefits result from sharing distributed data and information resources. Delivering interoperability across these resources presents a significant technical challenge, and whilst significant progress has been made in the delivery of syntactic interoperability, the development of effective semantic interoperability is in its early stages.

The work to date in oceanographic data management and semantic interoperability has generated both a firm foundation and a challenge. The foundation comprises lessons learned that may be used to provide guidance to those building new marine data systems so that they may build in semantic interoperability from the start. The challenge is accessing the vast information resource currently locked away in isolated stovepipe systems. Meeting this challenge requires high-quality leading edge technical solutions, community building and copious amounts of hard work populating information management tools. MMI has provided technology, guidance and community building tools through its website and list servers. It is now the responsibility of the oceanographic scientific community to work together to produce the Semantic Web [2] resources that will deliver effective semantic interoperability to the benefit of all.

References

1. Bermudez, L.E., Piasecki, M.: Metadata Community Profiles for the Semantic Web Geoinformatica 10 (2006) 159-176
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, Scientific American 184 (2001) 34-43
3. Bishr, Y.: Overcoming the Semantic and Other Barriers to GIS Interoperability, Geographic Information Science 12 (1998) 299-314
4. CF Standard Names: 2006, http://www.cgd.ucar.edu/cms/eaton/cf-metadata/index.html

5. Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S.: Netcdf Climate and Forecast (Cf) Metadata Conventions, http://www.cgd.ucar.edu/cms/eaton/cf-metadata/CF-1.0.html
6. Fisheries and Oceans Canada: Marine Environmental Data Service (Meds), http://www.meds-sdmm.dfo-mpo.gc.ca/meds/About_MEDS/standards/login_e.asp
7. GF3 Format, http://www.ices.dk/ocean/formats/gf3.htm
8. Graybeal, J., Bermudez, L.E., Bogden, P., Miller, S., Watson, S.: Marine Metadata Interoperability Project: Leading to Collaboration, in Proceedings of the Local to Global Data Interoperability - Challenges and Technologies Symposium, Sardinia, Italy (June 19-24, 2005)
9. Graybeal, J., Watson, S., Bermudez, L.E., Galbraith, N., Stocks, K., Subramanian, V.: Marine Metadata Interoperability Workshop: Advancing Domain Vocabularies Workshop Report, MBARI, Moss Landing, CA, (2006)
10. IODE Oceanographic Data and Information Exchange, http://www.iode.org/
11. NetCDF (Network Common Data Form), http://www.unidata.ucar.edu/software/netcdf/
12. Olsen, L.M., Major, G., Leicester, S., Shein, K., Scialdone, J., Weir, H., Ritz, S., Solomon, C., Holland, M., Bilodeau, R., Northcutt, T., Vogel, T.: Global Change Master Directory (Gcmd) Earth Science Keywords, http://gcmd.gsfc.nasa.gov/Resources/valids/keyword_list.html
13. OPeNDAP, http://www.opendap.org/
14. VINE, http://sf.net/projects/vine
15. GF3: A General Formatting System for Geo-Referenced Data. Volume 2. Technical Description of the GF3 Format and Code Tables. 1987 UNESCO.
16. The NERC DataGrid Vocabulary Server, http://www.bodc.ac.uk/products/web_services/vocab/