



Building a Global Data Network Workshop, Kiel, May 2007

Plaintext to governed vocabularies: restoring order to anarchic metadata

Roy Lowry

British Oceanographic Data Centre



Presentation Outline

- **Introduction**
- **Interoperability**
- **Metadata Evolution**
- **Controlled Vocabularies and SeaDataNet**
- **Mappings and Ontologies**
- **The Vocabulary Entity Issue**



Introduction

- **What is an oceanographic data manager doing addressing a meeting on geological data management?**
- **I'm striving for interoperability, particularly semantic interoperability, between data systems**
- **The absence of this interoperability is a common problem across environmental science disciplines**
- **Common problems require common solutions across discipline boundaries**
- **This meeting is a promising sign that these boundaries are starting to break down**



Interoperability

- **Interoperability is the ability to share data from multiple sources as a common resource from a single tool**
- **Interoperability has four levels (Bishr, 1998)**
 - **System – protocols, hardware and operating systems**
 - **Syntactic/Structural – loading data into a common tool (reading each others' files)**
 - **Semantic – understanding of terms used in the data by both humans and machines**



Interoperability

- **Interoperability has been the elusive Nirvana for oceanographic data management since IODE was established in 1961**
- **Modern Semantic Web knowledge management technologies are bringing this dream within reach**
- **But without a foundation of published well-defined, well-governed and versioned base vocabularies the technology will achieve nothing**



Interoperability

- **Standards have become established that ease the syntactic interoperability**
 - **The Excel spreadsheet is a universally accepted currency for biological and chemical oceanographic data exchange**
 - **NetCDF, particularly standardised variants like CF, over protocols such as OpenDAP gives interoperable high volume data exchange**
- **Scientists without programming skills can visualise data from multiple sources in minutes**
- **But can they understand what it means?**



Metadata Evolution

- **In the beginning there was plaintext**
 - In the 1980s the enlightened few saw the value and potential of metadata
 - Most data suppliers saw metadata as an unnecessary waste of their valuable time
 - The enlightened in the MAST Data Committee and SeaSearch aimed to change this by making metadata creation as easy as possible
 - Plaintext is much easier to create than structured metadata so metadata formats based on plaintext fields such as EDMED were promoted and populated



Metadata Evolution

- Problem is that whilst humans have the intelligence to read and understand plaintext it is of very limited use to a computer
- Consider some example EDMED plaintext parameter descriptions from a knowledge management viewpoint:
 - A wide variety of chemical and biological parameters
 - CTD data
 - Amplitude de l'echo retrodiffuse
 - Cu, Zn, Fe, Pb, Cd, Cr, Ni in biota
 - MACRO-MEIOFAUNA, SED BIOCHEMISTRY, ZOOPLANKTON, CILIATES, BACT CELLS, BACT BIOMASS, LEUCINE UPT, PRIM. PROD, METABOL, COCCOLITH
- Consequently, the chances of these data sets being discovered by conventional parameter searching is virtually zero



Metadata Evolution

- **Plaintext kills interoperability stone dead**
 - In 2005 Taco de Bruin asked me to build Antarctic Portal DIFs from Dutch EDMED entries
 - DIF is a structured format representing each parameter thus:

```
<Parameters>  
  <Category>EARTH SCIENCE</Category>  
  <Topic>Oceans</Topic>  
  <Term>Salinity/Density</Term>  
  <Variable>Salinity</Variable>  
</Parameters>
```
- **How does one derive this automatically from the plaintext 'CTD data', 'T/S', 'temp+salin', 'temperature and salinity' and all the other variants in EDMED?**
- **Needless to say, Taco is still waiting.....**



Metadata Evolution

- **SeaDataNet has inherited thousands of EDMED dataset descriptions from SeaSearch**
- **SeaDataNet wants to establish interoperability with other metadata repositories**
- **So, something has to change....**



Vocabularies

- **The key to EDMED's evolution to parameter interoperability is to replace plaintext descriptions by keywords from a controlled vocabulary**
- **Controlled vocabularies**
 - **Ensure consistent spellings**
 - **Ensure entities are described using the same words in the same order**



Vocabularies

- **Controlled vocabularies featured in the legacy metadata inherited by SeaDataNet**
- **However**
 - **Content governance was total anarchy**
 - * Decisions made by individuals – even students
 - * Terms were set up and used with inadequate thought about their meaning and formal definitions were conspicuous by their absence
 - **Technical governance wasn't much better**
 - * No formal maintenance or versioning
 - * Vocabularies delivered on an ad-hoc basis as CSV files on FTP servers or web sites
 - * Data models differed from one vocabulary to the next



Vocabularies

- **All this has changed for SeaDataNet**
 - **Content governance**
 - * SeaDataNet internal vocabularies are governed by the Technical Task Team
 - * Vocabularies with wider implications are governed by SeaDataNet and MarineXML Vocabulary Content Governance Group (SeaVoX) e-mail list
 - **Data Model**
 - * Vocabularies built from entries comprising a key, a term, an abbreviated term and a definition
 - * Entries aggregated into lists, each corresponding to a real-world sub-class
 - * Lists aggregated into constraints corresponding to real-world classes
 - * Constraint entries populate one or more lists



Vocabularies

- **All this has changed for SeaDataNet**
 - **Technical governance**
 - * Vocabularies managed in an Oracle system with automated versioning and audit trail maintenance
 - * Vocabularies served through a Web Service API
 - * Clients using this interface are available
 - <http://vocab.ndg.nerc.ac.uk/client/vocabServer.jsp> (BODC)
 - http://seadatanet.maris2.nl/v_bodc_vocab/welcome.asp (Maris)
 - Maris client accessible from SeaDataNet web site



Vocabularies

- **The importance of controlled vocabularies to SeaDataNet will increase as we stitch together disparate data sources for real**
- **BODC client currently exposes 69 lists**
- **Maris client exposes 27 of these that are particularly relevant to SeaDataNet**
- **These numbers will grow as vocabulary harmonisation within SeaDataNet and between SeaDataNet and the wider community progresses**



Vocabularies

- **Emphasis in SeaDataNet is to learn from the past and adopt (preferably) or develop a robust set of metadata and data standards**
- **Initially, these standards will be applied to new content as it is produced**
- **Retrospective application of standards to legacy content will need to be addressed as the project develops from the pilot to the operational stage**
- **Good job we've got a lot of partners!**



Mappings and Ontologies

- **SeaDataNet standards need to:**
 - **Link into established partner legacy systems**
 - **Provide the basis for semantic interoperability with other metadata and data systems**
- **Consequently, building maps between legacy, internal and external lists forms an important part of the work**
- **Currently working on maps between parameter vocabularies from BODC, GCMD and CF using SKOS relationships**



Vocabulary Entities

- A ‘Statement of the Obvious’
 - Each vocabulary entry describes an instance of a ‘thing’ or entity in the real world
 - These ‘things’ should share a common set of attributes making them members of a consistent level in a class hierarchy
 - In other words, it should be possible to create a common formal definition covering all the real-world ‘things’ described by the entries in a list
 - This formal definition should be compatible with the definition of the metadata field populated by the entries from that list
- Unfortunately, the obvious wasn’t obvious when some of the vocabularies commonly used in oceanographic data management were being developed



Vocabulary Entities

- **Some examples of past sins**
 - **Entity definition inconsistency**
 - **Data model 'repairs'**
 - **Shoe-horning**



Vocabulary Entities

- **Inconsistent entity definitions**
 - **Cruise Summary Report (ROSCOP) 'parameters'**
 - **Atmospheric chemistry: a domain**
 - **Phosphate: a chemical species**
 - **Grab: a sampling device**
 - **Bottom photography: a shipboard activity**
 - **Sea level: a phenomenon**
 - **Bathythermograph: a sensor class**
 - **GCMD Instrument Keywords**
 - **AA Spectrophotometry: a sample analysis technique**
 - **Neuston Net: a sampling device**
 - **SEAWIFS: a platform**
 - **Altimeter: a sensor class**
- **Makes mapping much more difficult or even impossible**



Vocabulary Entities

- **Data model ‘repairs’**
 - **One-to-one relationships turned into one-to-many by creating vocabulary entries that are themselves lists**
 - * **Examples from a ‘data type’ vocabulary**
 - **Moored profiling CTD + acoustic current meter**
 - **Multiple data types – aircraft**
 - **Multiple data types – ship**
 - **Makes mapping much more difficult or even impossible**



Vocabulary Entities

- **Shoe-horning**
 - A shoe-horn is an implement for forcing a large foot into a small shoe
 - In metadata terms, it means using a structure or record to describe something for which it wasn't originally designed
 - This is generally accomplished through the 'imaginative' usage of vocabularies, which inevitably corrupts the entity definition
 - Makes mapping much more difficult or even impossible



Vocabulary Entities

- **Shoe-horning example**
 - **Cruise Summary Report is a metadata description of a cruise**
 - * A cruise can be defined as a ship leaving port, doing science, then returning to port
 - * What if we have an instrumented ferry?
 - **Addressed(?) by :**
 - * Generating a CSR for each route
 - * Setting 'ship name' field to ferry route
 - **Consequences**
 - * Controlling vocabulary entity definition destroyed
 - * 'Route: Folkestone - Boulogne' enters the standard vocabulary of ship names
 - * Looks crazy when it appears in the 'ship name' drop-down list in a metadata population system
 - * Temporal metadata are nonsense (show ferry at sea 24/7 for a couple of years!)



Vocabulary Entities

- **Shoe-horning alternative**
 - **Redesign the data model with an activity class**
 - **Activity class has subclasses such as**
 - * Cruise
 - * Instrumented ferry sailing
 - * Littoral zone sampling visit
 - **The real world is complicated and metadata models attempt to describe that world**
 - **Experience has shown again and again that attempts at simplification end in tears**



Some Conclusions

- **Plaintext is a destroyer of interoperability**
- **Multiple controlled vocabularies covering a common domain topic whilst undesirable is repairable by mapping/ontology building**
- **Controlled vocabularies need rigorous management for operational mapping to be feasible**
- **Mapping issues are not confined to vocabulary entries but also crop up in vocabulary topics making mapping between entries much more difficult**



That's All Folks

Thank you for your attention

Any questions?