



SeaDataNet 2007 Plenary Meeting, Trieste



Controlled Vocabularies, Web Services and Transport Protocols

Roy Lowry

British Oceanographic Data Centre



Presentation Outline



- **Metadata Evolution**
- **Controlled Vocabularies and SeaDataNet**
- **Web Services**
- **Transport Protocols**



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 of the MAST Data Committee and SEA-SEARCH 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>  
  <Detailed_Variable>Salinity of the water column</Detailed_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



- EDMED content re-engineering is something SeaDataNet will have to address
- BODC recently generated DIFs by manually building structured parameter descriptions for 72 EDMED entries
- This took 25 staff days (including learning overheads)
- There are over 3000 entries with over 2700 different parameter descriptions in the EDMED database



Metadata Evolution



- **The cloud may have a silver lining**
 - **The SEA-SEARCH EDMED XML schema follows ISO19115 describing parameters by a keyword list**
 - **This list is now to be populated from a standard vocabulary (BODC Parameter Discovery Vocabulary)**
 - **So EDMED started evolving towards parameter interoperability during SEA-SEARCH and the journey is being completed in SeaDataNet**
 - **Which just leaves the legacy content issue....**



Metadata Evolution



- **The cloud may have a silver lining**
 - **EDMED records supplied via the XML route should be relatively easy to recreate incorporating the standardised terms**
 - **Tools (XML database export tool and record validator) to support this are available or under development**
 - **Some EDMED records supplied via Access may be now be available in XML as partners develop local metadata systems**
 - **The scale of the required re-engineering may therefore be reduced to manageable proportions and we have many partners to help**
 - **Partial automation may also be possible**



Vocabularies



- **The key to EDMED's evolution to parameter interoperability is the use of 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 SEA-SEARCH and EDIOS**
- **However**
 - **Content governance was virtually anarchy**
 - * Decisions made by individuals – even students
 - * Terms were set up and used without due thought for their meaning or formal definition
 - **Technical governance wasn't much better**
 - * Vocabularies delivered as CSV files on FTP servers or web sites
 - * No formal maintenance or versioning



Vocabularies



- **All this has changed for SeaDataNet**
 - **Content governance**
 - * SeaDataNet internal vocabularies are governed by the TTT
 - * Vocabularies with wider implications are governed by SeaVoX e-mail list
 - **Technical governance**
 - * Vocabularies managed in an Oracle system with automated versioning and audit trail maintenance
 - * Vocabularies served through a Web Service interface
 - * Clients using this 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 in SeaDataNet progresses**



Web Services



- **‘Web Service’ is becoming a buzzword in SeaDataNet, but what is a Web Service?**
- **Definitions of Web Services are so loaded with computer science jargon that they are incomprehensible**
- **Essentially Web Services are programs that sit around waiting until they receive a question in the right type of XML document, do some work, then return the answer in the right type of XML document**



Web Services



- **To system builders Web Services provide wrappers that allow SAFE external access to primary data resources**
- **Web Services may be invoked from any type of application written in most of the current languages (Java, Perl, PHP, Python etc.)**
- **By calling Web Services any partner's web pages can include live data from databases anywhere on the SeaDataNet network**



Web Services



- **Web Services are currently deployed in SeaDataNet for**
 - **Vocabularies**
 - **SeaDataNet organisation address book (EDMO)**
- **Web Services are planned to provide access to**
 - **EDMED**
 - **EDMERP**
 - **Cruise Summary Report**
 - **EDIOS**
 - **CDI**
- **Maybe in SeaDataNet II some partners will deploy Web Services to expose their metadata to harvesting**



Transport Protocols

- The decision has been made by the TTT that for SeaDataNet I data transport will be based on physical files conforming to format standards
- For SeaDataNet II we will look at moving to ‘data virtualisation’ technologies
 - E2EDM developed by the Russian NODC for JCOMM ETDMP Pilot Project
 - CSML developed in the UK for NERC DataGrid
- Decision was not unanimous but the lower risks associated with simpler technology were considered prudent



Transport Protocols

- **Format standard comprises obligatory formats and optional formats**
- **All SeaDataNet data providers MUST be able to supply the obligatory formats needed to cover the data types they serve**
- **In addition they may opt to serve one or more of the optional formats**



Transport Protocols

- **Obligatory formats**
 - **CF-compliant NetCDF for data types stored as multidimensional arrays. Examples:**
 - * Gridded fields
 - * ADCP data
 - * Satellite images
 - **ODV ASCII spreadsheet input format for everything else**
- **Minor extensions to both these formats will be developed to provide storage for semantic interoperability metadata (parameter URIs)**
- **Note that NetCDF is only required for certain types of data so many SeaDataNet partners will only have to serve ASCII**



Transport Protocols

- **Optional formats – for exchange by consenting partners**
 - **Currently only one, but no doubt more will follow, which is the MEDATLAS ASCII format used by the Mediterranean and Black Sea community**
 - **Consideration needs to be given to developing SeaDataNet extensions to optional formats to enhance interoperability with obligatory formats**
 - **For example we could add standardised comments to the MEDATLAS format to store parameter URIs**



Transport Protocols

- **Information Loss**
 - **SeaDataNet data exchange will involve transforming source data into a SeaDataNet data object comprising a data file plus a CDI metadata record**
 - **Any information not included in these will not be transported**
 - **Some loss, limited to an extent considered acceptable by the TTT, is inevitable for SeaDataNet I**
 - **Ways of addressing the problem using more advanced technology need to be included in our considerations for SeaDataNet II**



That's All Folks

Thank you for your attention

Any questions?

