AGU Fall Meeting, San Francisico, December 2008

25 Years of Controlled Vocabularies in Oceanographic Data Management

Roy Lowry

British Oceanographic Data Centre

The Journey

- The GF3 Era
- The Dark Ages
- SeaSearch – A New Hope
- SeaDataNet – Vocabularies Strike Back
- The Road Ahead

The GF3 Era

- **International Oceanographic Data and Information Exchange (IODE)**
  - **Established in the 60s**
  - **Objective was to promote the sharing of oceanographic data**
- **Group of Experts on the Technical Aspects of Data Exchange (GETADE)**
  - **Active in the 1980s**
  - **Objective was to provide the technology to facilitate oceanographic data exchange**

The GF3 Era

- GETADE efforts focussed on developing a standard format (GF3) for data exchange on magnetic tape
- GETADE identified the need for common terminology and provided content governance for controlled vocabularies
- In 1987 GETADE published 7 oceanographic domain controlled vocabularies (code tables) as a book in English, French, Spanish and Russian

The Dark Ages

- Reprinting multi-lingual books is prohibitively expensive technical governance
- Consequently:
  - Content governance had no purpose
  - Vocabularies couldn't develop
  - Vocabularies fulfilled part of the needs of some potential users
  - Vocabularies were not widely adopted

The Dark Ages

- In the late 80s and 90s there was significant systems development activity in the IODE network
- This included implementation of digital controlled vocabularies
- Centralised vocabulary content and technical governance existed through IODE RNODC (Formats) but was very weak
- Consequently local vocabulary management became the norm
- This led to widespread vocabulary abuse

The Dark Ages

- **Vocabulary abuse**
  - **Source vocabulary (often GF3) taken and extended locally**
    - New term granularity varies from place to place
    - Like Galapagos Finches the results were similar but significantly different
  - **Shoe-horning**
    - Entities redefined through mission creep or to patch data models
      - We have ‘ships’ called ‘helicopter’ and ‘Dover-Calais’
      - We have terms in use like ‘multiple instruments’

The Dark Ages

- **Vocabulary Abuse**
  - **Content degradation**
    - GF3 vocabularies included high quality term definitions
    - These mysteriously disappeared from some of the clones
    - If they survived, new term definitions were either absent or poor e.g. simple term repetitions
    - Vocabularies developed into collections of terms that mean different things to different people
- **Vocabulary abuse makes it much harder to implement semantic interoperability**
- **Our challenge is to treat its victims without changing the meaning of legacy data**

SeaSearch

- SeaSearch was an EU project at the turn of the century aimed at providing interoperable metadata across Europe and the Mediterranean
- SeaSearch developed a digital library of 10 vocabularies to provide the backbone for semantic interoperability
- This was a significant step in the right direction

SeaSearch

- **But**
  - **Content governance was delegated to individual ‘volunteers’**
    - No effort to define terms so terms misunderstood
    - Extension suggestions never challenged so internal inconsistencies developed
  - **Technical governance was weak**
    - Local vocabulary copies allowed to develop
    - Metadata records were created from these
    - Things broke when these records encountered the ‘master’ vocabularies

SeaDataNet

- **SeaDataNet is a current EU project running from 2006 to 2011 to provide interoperable data across Europe and the Mediterranean**
- **At the outset**
  - **SeaDataNet recognised that semantic interoperability is essential**
  - **SeaDataNet recognised that effective vocabulary governance is essential for semantic interoperability**

SeaDataNet

- **Content Governance**
  - **Based on moderated e-mail discussion**
    - Within SeaDataNet Technical Task Team (12 members)
    - SeaVoX - joint with IOC MarineXML SG (50 members)
- **Technical Governance**
  - **Based on NERC DataGrid Vocabulary Server**
    - Oracle back-office storage and version management
    - Export as dynamically created RDF documents

SeaDataNet

➢ Vocabulary Server Access

- **Resource URLs**

  * Every concept has a URL delivering a SKOS document describing the resource and its relationships to other resources e.g. http://vocab.ndg.nerc.ac.uk/term/P021/current/TEMP

  * Every vocabulary (concept grouping) has a URL delivering a SKOS document describing the component resources e.g. http://vocab.ndg.nerc.ac/list/P021/current/

SeaDataNet

➢ Vocabulary Server Access

- **'Method-based' APIs provide controlled access to resources and their mappings driven by method input parameters**
  - HTTP-POX version documented at http://www.bodc.ac.uk/products/web_services/vocab/methods.html
  - SOAP version WSDL (plus demonstration clients) available at **http://vocab.ndg.nerc.ac.uk**
- **User-friendly client for SeaDataNet lists (50% of public lists) at http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx**

SeaDataNet

- **Vocabulary Server Content (2008-12-03)**
  - **126 vocabularies**
  - **123701 concepts**
  - **79012 mappings (RDF triples)**
- **2008 Vocabulary Server Usage (to 2008-12-03)**
  - **4,293,779 total hits**
  - **~125,000 ‘human’ hits**

SeaDataNet

➢ Vocabulary Server Applications

- **Semantic crosswalk**
  - Requirement to create GCMD DIF discovery metadata from SeaDataNet EDMED V1.0
  - For parameters
    - GCMD use GCMD Science Keywords
    - EDMED uses BODC Parameter Usage Vocabulary (PUV)
    - Traditional metadata cross-walks circumvent this problem by leaving the DIF parameter section out
    - A mapping is maintained in the Vocabulary Server between BODC PUV and a version of GCMD Science Keywords
    - Semantic cross-walk with a dynamically translated parameter section is therefore possible

SeaDataNet

➢ Vocabulary Server Applications

- **Metadata field content verification**
  - ∗ Content encoded as URNs in metadata documents
  - ∗ Schematron schema extensions specify valid URNs for field
  - ∗ Schematron automatically updated when vocabulary is updated
  - ∗ Metadata documents have structure AND CONTENT verified using generic XML tools like Oxygen or a dedicated SeaDataNet validation service

The Road Ahead

- **Fixing known bugs and facilitating external content governance in Version 1**
- **Vocabulary Server Version 2**
  - **API changes**
    - Consider concepts and lists as http resources
    - Pure RESTful API making full use of http syntax to both serve and maintain resources
  - **Payload document changes**
    - Limitations discovered with current SKOS schema
    - Consulting with computer scientists at Manchester University to specify a new schema (updated SKOS or OWL)

The Road Ahead

- Version 1 developments currently underway
- Seeking funding for development of Version 2
- Breaking clients hitting a resource >100,000 times per annum isn't the way to win respect in cyberspace…..
- Versions 1 and 2 will be maintained in parallel for the foreseeable future

That’s All Folks

Thank you for your attention

Any questions?