CONTROLLED VOCABULARIES

Roy Lowry
British Oceanographic Data Centre

Presentation Overview

- Controlled Vocabularies - What and Why
- Controlled Vocabularies - History
- Controlled Vocabularies - SeaDataNet
- Controlled Vocabularies - Mappings
- Controlled Vocabularies - Future

What and Why

- Controlled Vocabulary (CV)
  - A collection of concepts that may legally populate a given field in a data or metadata model.
  - A concept is an instance of the real world entity modelled by that field - e.g. Instrument, parameter.
- Concept Labelling
  - Machine readable label - code, URI (URN or URL)
  - Human readable labels - name, abbreviation, definition

What and Why

- Why?
  - Alternative to CV is plain language text which is subject to:
    - Spelling errors - e.g. Macoma baltica for Macoma balthica
    - Entity abuse - e.g. Sea-Bird SBE9-11+ in a parameter field
  - CVs and concepts may be incorporated into knowledge management infrastructure and linked semantically to build ontologies.
    - Smart discovery
    - AI-driven data aggregation

History - Beginnings

- In the 1980s there was IODE GETADE who developed the GF3 code tables
  - Thorough content governance with concepts well defined and their scope carefully considered
  - Published by IODE as a book in five languages
  - Result is a beautiful piece of work that cannot be maintained.

History - Dark Ages

- In the 1990s GETADE waned as funding squeezed
- Some vocabulary governance moved to individuals
  - Poor judgement on what new entries should be allowed leading to vocabulary abuse (e.g. Making a data model 1:1 into 1:many by adding a list as a vocabulary concept)
- Some vocabularies moved to local management
  - Like Galapagos finches they evolved into entities that were similar but significantly different
  - Unlike Galapagos finches many variants retained the same name!

History - Renaissance

- SEASEARCH - content governance delegated to individuals, but realisation that vocabulary management needed to be centralised with a master copy universally accessible 24/7.

- SeaDataNet/NERC DataGrid - developed the NERC Vocabulary Server at BODC to deliver this.
  - Accessible vocabularies with clear entity definitions
  - Every concept given a URN that resolves into a URL that delivers an RDF XML document
  - Basis of the Semantic Web

SeaDataNet Controlled Vocabularies

- SeaDataNet makes extensive use of CVs in its metadata models and data formats
- Each CV targets one or more fields in these models/formats
- List of SeaDataNet CVs may be found at http://seadatanet.maris2.nl/v_bodc_vocab_v2/welcome.asp
- Ignore the Mxx entries they are hosted by SeaDataNet on behalf of MEDIN, which just leaves 64!
- Common practice is to use the 3-character code in the 'Library' column as the CV name e.g. P01, P02, L05, L22

SeaDataNet Controlled Vocabularies

- SeaDataNet Controlled Vocabularies may be accessed in one of five ways:
  - Human readable forms
    - Maris client library
    - Maris client thesaurus
    - BODC thesaurus (concept scheme) best viewed in Chrome
  - Machine readable forms
    - RESTful interface to CV or concept (RDF XML)
    - SOAP interface

SeaDataNet Controlled Vocabularies

- RESTful Syntax
  - Base is http://vocab.nerc.ac.uk/collection/ (returns an RDF XML catalogue of all 263 CVs in NVS)
  - To this we add the 3-byte vocabulary ID plus 'current' e.g. http://vocab.nerc.ac.uk/collection/ P03/current/ (returns all concepts in that CV in RDF XML)
  - To this we can add
    - 'accepted' (returns all valid concepts in that CV in RDF XML)
    - 'deprecated' (returns all deprecated concepts in that CV in RDF XML)
    - 'all' (returns all concepts in that CV in RDF XML)
    - Concept code (returns concept document in RDF XML)

SeaDataNet Controlled Vocabularies

- RDF XML Concept Document (deprecated concept)

```
<skos:Concept
    rdf:about="http://vocab.nerc.ac.uk/collection/P01/current/PCONZZ01/">
    <ctskos:prefLabel xml:lang="en">Elecrical conductivity of the water
    body</skos:prefLabel>
<skos:altLabel xml:lang="en">WC_Cond</skos:altLabel>
<skos:definition xml:lang="en">This is an obsolete term for this definition. Use
    CNDCZZ01 instead.</skos:definition>
<dc:identifier>SDN:P01::PCONZZ01</dc:identifier>
<skos:notation>SDN:P01::PCONZZ01</skos:notation>
<owl:versionInfo>2</owl:versionInfo>
<dc:date>2014-01-22 13:48:35.0</dc:date>
<skos:note xml:lang="en">deprecated</skos:note>
<owl:deprecated>true</owl:deprecated
><dc:isReplacedBy
    rdf:resource="http://vocab.nerc.ac.uk/collection/P01/current/CNDCZZ01/"/>
<skos:broader
    rdf:resource="http://vocab.nerc.ac.uk/collection/P02/current/CNDC/"/>
<skos:related
    rdf:resource="http://vocab.nerc.ac.uk/collection/P06/current/UECA/"/>
<void:inDataset rdf:resource="http://vocab.nerc.ac.uk/.well-
    known/void"/></skos:Concept>
```

Mapping to Controlled Vocabularies

- Mapping strategy depends upon workflow order
  - What comes first - CDI record or Data file?
  - CDI record first
    - Parameters and instruments for CDI assigned by manually mapping local vocabularies to P02 and L05
    - Parameters and instruments for data file assigned by manually mapping local vocabularies to P01 and L22
  - Data file first
    - Parameters and instruments for data file assigned by manually mapping local vocabularies to P01 and L22
    - Parameters and instruments for CDI automatically obtained using P01/P02 and L05/L22 mappings in NVS

Mapping to Controlled Vocabularies

- Manual Mapping Techniques
  - Library Text Search
    - Input a string into the 'Free search' box and press 'Search'
      - Wildcard character is '%' for 1 or more characters
      - Search is case-insensitive
      - Wildcard automatically added to start and end of string
        - 'Microzooplankton taxonomy-related biosurface area per unit volume of the water column' found by search for 'zooplankton'

Mapping to Controlled Vocabularies

- Manual Mapping Techniques
  - Library Text Search
    - Hardest vocabulary to map to is P01 because it's big (currently 30500 concepts)
    - Planned construction of search strings can help
      - P01 concept labels can be long and complex
      - BUT they are constructed using a semantic model so information is always presented in the order
        - What
        - Substance name then synonyms
        - What to where relationship
        - Where
        - How

Mapping to Controlled Vocabularies

- Manual Mapping Techniques
  - Consider a search for PCB183 in 'standard' fine sediment (<63um)
  - The following will fail to find anything
    - 'PCB183 concentration'
    - '63um sediment' (63um%sediment gets false hits)
    - 'sediment <63um%dry weight'
  - But this is right on the money
    - 'con%PCB183%dry%sediment%<63'

Mapping to Controlled Vocabularies

- Manual Mapping Techniques
  - String has all components in the right order
    - What - con for Concentration
    - Substance synonym - PCB183
    - What to where relationship - dry for dry weight
    - Where sediment%<63 for sediment <63um
  - The resulting hit
    - Concentration of 2,2',3,4,4',5',6-heptachlorobiphenyl {PCB183 CAS 52663-69-1} per unit dry weight of sediment <63um

Mapping to Controlled Vocabularies

- Manual Mapping Techniques
  - Thesaurus Search
    - For parameters entry point is P08 (Disciplines), P03 (Agreed Parameter Groups) or P02 (Discovery Parameters)
      - Pressing a '+' in P08 opens up P03
      - Pressing a '+' in P03 opens up P02
      - Pressing a '+' in P02 opens up P01
    - Works well for finding P01 in cases where small numbers of P01 terms are mapped to each P02
    - In other cases the list may be too long for comfortable scanning and library string searching will work better

Mapping to Controlled Vocabularies

- Automated Mapping Technique
  - To automatically find the P02 code for a given P01 code
    - Obtain the RDF XML document for the P01 code
    - Look for <skos:broader rdf:resource including the URL for a P02 concept which in this example is:
      - <skos:broader rdf:resource="**http://vocab.nerc.ac.uk/collection/P02/current/NTRI/**"/>
    - Job Done - all that's needed is a bit of software to do the job programmatically
    - In BODC we store P01 codes in data and automatically convert to P02 to generate CDI. SHOULDEXCLUDE CO-ORDINATE VARIABLES

Mapping to Controlled Vocabularies

- Parameter mappings
  - The biggest problem with mapping local parameter vocabularies to a standard vocabulary is understanding EXACTLY what is meant by the local term.
  - Consider the parameter 'Particulate Zinc'
    - This could mean:
      - The concentration of zinc per unit dry weight of the residue of a filtered sample
      - The concentration of zinc contained in the particles per unit volume of a body of water
      - The concentration of zinc contained in the particles per unit mass of a body of water
    - Each of these has a different P01 code.
  - Think carefully and ask many questions!

Controlled Vocabularies - Future

- P01 mappings based on semantic model
  - Expose elements of the semantic model
    - Concentration of
    - 2,2',3,4,4',5',6-heptachlorobiphenyl {PCB183 CAS 52663-69-1}
    - per unit dry weight of
    - sediment <63um
  - User selects combination that maps to their local parameter like one-armed bandit wheels
  - System returns appropriate P01 code or automatically generates a new P01 code

Controlled Vocabularies - Future

- P01 mappings based on semantic model
  - Creates the risk of 'Green Dog' syndrome
    - User is free to select any combination of elements
    - Some combinations may be valid, others are not
      - Consider lists of animals plus colours
        - GREEN + LIZARD - good choice
        - GREEN + DOG - not such a good choice
    - Consequently, quality control of user-selected semantic model combinations is essential
    - Places latency in new code assignment cycle but I am convinced this is worthwhile - others disagree

Controlled Vocabularies - Future

- Semantic aggregation
  - Set up a vocabulary of aggregated parameter concepts - P35 for EMODNET chemistry
  - Map each P35 concept to P01 concepts that may be validly included in the aggregation
  - Aggregation software issues RESTful call to NERC Vocabulary Server for P35 concept
  - Software then parses returned RDF XML document to identify P01 concepts that may validly be aggregated
  - This functionality is currently being written into ODV