Putting Meaning into Oceanographic Measurements

ROY LOWRY
BRITISH OCEANOGRAPHIC DATA CENTRE

Overview

An oceanographic data model

Parameter management

The P01 parameter mark-up vocabulary

The NERC Vocabulary Server

The P01 semantic model

Parameter discovery

Semantic interoperability

Some parameter semantics issues

Conclusions

An Oceanographic Data Model

Oceanographic data are made up of 'measurements'

- The thing that is measured is known as:
- Measured Phenomenon/Observable Property (Simon Cox in Observations and Measurements)
- Parameter (oceanographic data managers since the 1970s)
- Some measurements known as co-ordinate variables define where and when other measurements were made
- The co-ordinate variables form patterns known as feature types
- Point (bottle) - x, y, z and t all constant
- Profile (CTD) - x, y, t constant: z varies
- Time series (current meter) - x, y, z constant: t varies
- Trajectory (AUV) - x, y, z and t all vary

Parameter Management

Parameter management was founded in 1980s/1990s through format specifications

- GF3 (IODE)
- Medatlas (Ifremer)
- PXF/QXF (BODC)
- CF NetCDF

The parameter is represented by a character string called the parameter code that is ideally meaningless but might be a mnemonic

This code is linked to a description of what has been measured

The codes and descriptions are (or should be) formally managed by an authority known as the content governance

The codes and descriptions are (or should be) accessed by all from a single well-maintained source known as the technical governance

Parameter Management

In the 1980s there was some top quality parameter content governance

- IODE GETADE chaired by Meirion Jones developed the GF3 parameter codes
- Let down by dysfunctional technical governance (printed documents in 5 languages)

In the 1990s

- GETADE work was compromised by funding issues and what was learned about content governance in the 80s was forgotten
- Technology developments encouraged multiple locally-managed copies of vocabularies

In the 2000s SEASEARCH then SeaDataNet restored some order

Parameter Management

Problems faced by SEASEARCH

- Data misunderstandings caused by:
- Increasingly complex oceanographic measurements
- Parochial parameter labels – even shorthand labels (e.g. T1, T2 etc.) whose meanings were once known but are forgotten by the time data are archived
- Strong, knowledgeable content governance developed as the solution
- Evolving vocabularies
- Local copies made of a standard vocabulary (e.g. GF3)
- Each copy is edited to suit local needs
- Result is a set of similar but significantly different vocabularies
- Strong technical governance developed as the solution

Parameter Management

Solutions introduced by SEASEARCH/SeaDataNet

- A single parameter vocabulary, P01 developed by BODC, was adopted as the core standard
- Other vocabularies (e.g. MEDATLAS P09) were mapped to this standard
- Expertise was recruited to provide content governance
- Investment was made in technical governance technology
- Further vocabularies were developed collaboratively to provide a parameter discovery hierarchy (P08/P03/P02)

The P01 Vocabulary

P01 is home to >37,000 concepts that each represent a parameter

Each concept has a code, a label (plaintext description) and an alternative label

Many also have definitions (work in progress)

The labels are built systematically using a semantic model with terms from controlled vocabularies

- Ensures consistent spellings
- Ensures consistent information ordering
- Provides the basis for semantic search tools

All concepts are mapped to a parameter discovery vocabulary (P02) and a suggested units of measure vocabulary (P06)

The P01 Vocabulary

P01 content governance

- Provided by BODC
- 1988-2012 - Roy the benign dictator supported by scientific community (e.g. JGOFS)
- 2012 onwards - BODC Vocabulary Management Group chaired by Gwen Moncoiffe with Roy as consultant and continuing support from scientific community

P01 technical governance

- Based on a Vocabulary Server (NVS) and a web application toolkit
- Vocabulary Search
- Vocabulary Editor (some vocabularies but not P01)
- Semantic Model Search (for P01: in development)
- 'Bandit' Editor (for P01: in development)

The NERC Vocabulary Server

The NVS data model has concepts (e.g. TEMPPR01) organised into controlled vocabularies (e.g. P01)

Each concept is represented by a URI (first implemented in 2004!)

Each URI resolves to an RDF document containing

- Labels and definition
- Governance information including version and date
- Mappings

Server data may be accessed through

- RESTFul interface
- SOAP interface
- SPARQL end point

NVS is a part of the Semantic Web

The NERC Vocabulary Server

Technical implementation

- Oracle back office holding concepts, internal mappings and mappings to external URLs
- Maintenance much easier than an XML database or triple store
- Master back office copy internal to BODC
- Master published each night (02:00 UK local) then a new web-facing copy and a Jena triple store (for SPARQL endpoint) are built
- RESTFul and SOAP interfaces are provided by Java Web applications
- Every P01 code represented by a URL
- URL resolves to an RDF XML document
- Search and edit tools run off the SPARQL endpoint

The P01 Semantic Model

The observable property 'Concentration standard deviation of phosphate per unit volume of the water body [dissolved plus reactive particulate phase]' may be broken down into

- Measurement property 'Concentration' (S06 vocabulary)
- Measurement statistical qualifier 'standard deviation' (S07 vocabulary)
- Chemical substance 'phosphate' (S27 vocabulary).
- Measurement-matrix relationship 'per unit volume of the' (S02 vocabulary)
- Matrix 'water body [dissolved plus reactive particulate phase]' (S26 vocabulary)

There are other possible semantic model elements such as biological entity, sample preparation, analysis and data processing protocol.

The P01 Semantic Model

- Semantic model used to develop a P01 search/edit tool based on a 'one-armed bandit' - BODC Semantic Model Vocabulary Builder.
- Elements of the semantic model form five bandit wheels:
  - Measurement property
  - Measurement statistical qualifier
  - Chemical substance
  - Measurement-matrix relationship
  - Matrix
- Choose one (e.g. a chemical substance), push the button and all combinations with that wheel fixed are displayed

Parameter Discovery

There are >37,000 P01 concepts

SeaDataCloud partners must map each local parameter name to a P01 entry

But which one????

There are several tools that can help

- NVS Search Tool (BODC)
- Hierarchical vocabulary search tool (Maris)
- NVS RESTful interface (BODC) - Linked data experience based on P02 URLs
- Semantic Model Vocabulary Builder (BODC) - work in progress that currently only covers part of P01

Parameter Discovery

NVS Search Tool

- Searches all vocabularies in NVS (over 300!)
- Runs off a SPARQL endpoint fed by a triple store that is built each night from the Oracle back office
- To limit searches to P01 use 'Simple search within a vocabulary'
- If you get too many hits bring in the advanced search option
- Extremely flexible and powerful with some learning overhead - but worth it
- My personal favourite

Parameter Discovery

Vocabulary hierarchy

- There are four vocabularies in the SeaDataNet parameter hierarchy
- P08 - Discipline: small number (11) of very broad concepts
- P03 - Parameter groups: narrower than P08 (70 concepts)
- P02 - Parameter Discovery Vocabulary (432 even narrower concepts)
  - Used to mark up CDI records in SeaDataCloud
- P01 - Parameter Usage Vocabulary (37364 VERY narrow concepts)
  - Used to mark up data in SeaDataCloud
- The hierarchy may be navigated using the Maris search tool
- Covers a subset of the vocabularies in NVS
- Runs off a cached copy refreshed over night from BODC
- Intuitive with low learning overhead
- Hits become unmanageable if large numbers of P01 concepts are mapped to a P02 concept

Semantic Interoperability

Discussion Use Case

- EMODNet chemistry wishes to create a product for cadmium concentrations in shellfish from SeaDataCloud data holdings
- To do this, some issues need to be addressed
- Define 'shellfish' to specify species that are within the product scope
- Define 'cadmium concentrations' to specify measurements that are within the product scope
- Specify the unit of measure for 'cadmium concentrations' in the product
- Somebody needs to make these decisions
- EMODNet chemistry has established a group of experts as a content governance for this purpose

Semantic Interoperability

The Group of Experts make the following decisions

- The species to include in ‘shellfish’ (Mytilus edulis, Mytilus galloprovincialis, Ostrea edulis, etc.)
- The scope:
  - Total soft parts analysed
  - Samples NOT subdivided by size, gender, etc.
  - Measurements on a wet weight basis
- Units of measure specified as micrograms per kilogram

The shellfish name list plus the scope specification may be represented as a list of P01 concepts

Semantic Interoperability

The following semantic interoperability infrastructure is then built

- An entry for 'Cadmium concentrations in shellfish' is made in the EMODNet chemistry product vocabulary (P35)
- The P35 entry is mapped to 'micrograms per kilogram' in P06
- The P35 entry is mapped to the list of P01 entries that represent 'cadmium concentrations in shellfish'

The ODV software ingests the data files and AUTOMATICALLY merges all P01 entries mapped to the P35 entry, converting them dynamically into the correct units

Some data are lost from the product that could have been included in a manual aggregation

- Dry weight data accompanied by water content
- Sample subsets that could be aggregated (e.g. data for males and females)

But it's fast!

Some Parameter Semantics Issues

Content governance knowledge limitations

Language

User education

Data model limitations

Rigidity

Some Parameter Semantics Issues

Content governance knowledge limitations

- Parameter semantics cover an extremely broad and forever expanding range of topics in the marine science domain
- Content governance decisions inevitably involve detail
- Should we allow a parameter code for ‘DDT’?
  - No – DDT has many different meanings (ortho, para, DDE, DDD etc.)
  - Ambiguous code could cause product ‘bullseyes’ if one lab includes DDD/DDE but others don’t
- Do we need separate parameter codes for dissolved inorganic carbon (DIC) and TCO2?
  - No – they are different names for exactly the same measurement
  - A bet on this issue once gained me a very nice bottle of grappa!
- Do we need separate parameter codes for significant wave height and average height of the highest one third of waves?
  - No – again they are the same thing.
  - We got that wrong in P01!

Some Parameter Semantics Issues

Content governance knowledge limitations

- No one person can know enough so we either need groups of experts or open community governance to provide the necessary expertise
- Both governance models have problems
  - Everybody is very busy
  - Some people ‘Shoot from the hip’
  - Experts can have different opinions
  - Skilled moderation required to turn deadlock into consensus
  - Apathy (or exhaustion) can result in bad decisions
- Knowledge issue is moderated to some extent by the Internet
  - In 2000 the only chemistry reference I had was the ‘Rubber Book’
  - Now we have Wikipedia, WoRMS, ChEBI, CAS, etc., etc.
  - Just need to remember that the Internet isn’t error free!

Some Parameter Semantics Issues

Language

- Advanced English is the language of parameter semantics
- I’m lucky and have English as my first language but many P01 users do not
- NVS does have a multilingual capability
- Harnessing this isn’t straightforward, but projects are being proposed

User education

- Many find difficult what I find easy – Why?
- Small nuggets of knowledge can make a big difference
- A recent request was for a new P02 code to match a given P01 code
- But EVERY P01 code is already mapped to a P02 code
- Need to include parameter semantics in SeaDataCloud training courses
- Those trained need to actively share the knowledge they have gained

Some Parameter Semantics Issues

Data model limitations

- Parameter code is the only 'hook' we have to define what has been measured
- Creates pressure for increased information content in the parameter code label
- Adding more information increases the number of parameter codes required
- For example, contaminant in biota
  - Concentration of cadmium in fish - 1 metal, one code
  - Concentration of cadmium in 10 named species of fish - 1 metal, 10 codes
  - Concentration of cadmium in 10 named species of fish by one of five named measurement techniques - 1 metal, 50 codes
  - Ten metals, specified organs, 100 organic contaminants…….
- Presents content governance with difficult decisions
- How much information should be included?
- How much change is needed before a new parameter code is assigned?

Some Parameter Semantics Issues

Rigidity

- Once a P01 code is published it cannot be deleted or change meaning
- Forces content governance to be perfect
- But of course it never is…..
- Content governance imperfections include
- Accidental creation of duplicate codes
- Inconsistencies
  - Some dissolved metals have ‘no method’ codes, others don’t
  - Mixed organic chemical naming conventions
- Deprecation (soft deletion) is an option
- Limited usage (353 codes) to eliminate exact duplicates
- P01 RDF document always gives the replacement code to use

Conclusions

Oceanographic data are made up of measurements of parameters that need to be described

Parameter descriptions could be ungoverned plaintext but:

- Data would be lost as meanings get forgotten
- Multi-laboratory data aggregations like EMODNet would be impossible

Parameter description governance is difficult and errors are possible

- Lots of knowledge required
- Knowledge needs to be managed and organised

Parameter description governance is helped by Internet technologies

- Internet resources provide access to knowledge
- Semantic Web provides knowledge management tools

Thank You for your Attention

Questions?