A Semantic Modelling Approach to Biological Parameter Interoperability

Roy Lowry[1], Laura Bird[1] and Pieter Haaring[2]

[1]British Oceanographic Data Centre Joseph Proudman Building, 6 Brownlow Street, Liverpool L3 5DA, UK
E-mail Roy Lowry: rkl@bodc.ac.uk, E-mail Laura Bird: labi@bodc.ac.uk

[2]Ministry of Transport, Public Works and Water Management Directorate-General of Public Works and Water Management (Rijkswaterstaat). National Institute for Coastal and Marine Management (RIKZ) P.O. Box 20907, 2500 EX, The Hague, The Netherlands
E-mail: p.a.haaring@rikz.rws.minvenw.nl

Abstract

The BODC Parameter Dictionary currently contains over 16,500 terms of which nearly 11,000 pertain to biological parameters. The Rijkswaterstaat database in the Netherlands covers over 10,000 types of measurement, most of which are either chemical or biological. A requirement to populate a metadatabase described in terms of the BODC dictionary from the Rijkswaterstaat database meant that parameter interoperability between these information sources needed to be addressed. One technique for approaching this is manual mapping, working term by term through one of the information sources then searching for matching terms in the other. However, whilst this may be feasible for dictionaries containing tens of terms, it is totally unrealistic when the counts run into the thousands and so an alternative, automated approach was required.

Automation was initially attempted using a semantic matching tool developed at Rijkswaterstaat to offer a restricted list of BODC terms (preferably a single term) as the possible matches for each measurement. However, this met with limited success because the BODC dictionary consisted of plain language terms that not been written with machine processing in mind and had no constraints on either syntax or vocabulary. To appreciate the problem consider the programming required to recognise that '*Calanus* abundance', 'Number of *Calanus'*, '*Calanus* count' and 'Abundance of *Calanus'* essentially mean the same thing. Further, no dictionary, especially a dictionary without vocabulary constraints, is perfect and there is a high risk that matches will be missed due to basic errors such as spelling mistakes.

The Rijkswaterstaat database is described in terms of a data model that qualifies measurements through associated attributes describing, amongst other things, what was measured and how it was measured. This is an example of a semantic model in which an entity is described in terms of discrete items of information, called semantic elements. Ideally, these elements are atomic, unambiguous and therefore ideally suited to machine interpretation. It was concluded that the only way a mapping could be achieved would be to develop a model along similar lines to describe the BODC dictionary and then map the two models.

A prototype semantic model based on three sub-models, each containing between 10 and 12 semantic elements, was developed to describe the biota, biota composition and chemical terms in the BODC dictionary and populated with approximately 13,000 terms. This was used as a basis for a two-stage mapping to the Rijkswaterstaat data model. The first stage was to set up a mapping

between the semantic elements in the two models. For example, it was established that the 'Parameter' element in the Rijkswaterstaat model was equivalent to the concatenation of the 'Param' and 'Param_Comp' elements in the BODC semantic model. The second stage was to produce a mapping between the vocabularies used in each set of matched semantic elements. For example, the Rijkswaterstaat compartment term 'Surface water' mapped to the BODC compartment term 'water column'. Once these mappings had been established an automated term generation procedure was used to translate sets of Rijkswaterstaat semantic elements into BODC terms and identify matches.

The result was an automated mapping for approximately 90% of the Rijkswaterstaat measurement description terms. Of the remainder, most were matched by straightforward extensions to the vocabulary mapping. However, a small number of problems remained that could only resolved by querying Rijkswaterstaat, including ambiguity caused by homonyms that only came to light through standardisation of the BODC model to the ITIS taxonomic database.

This exercise has shown that semantic modelling is a very promising technique for automating parameter interoperability between biological databases. However, without standardisation, particularly in the description of taxonomic entities, matches will be missed and there is a small but significant risk of false matches between parameters that are totally different.

Keywords: Parameters; semantic modelling.

Introduction

This paper documents the work done to develop parameter interoperability between the biological and chemical data holdings of the British Oceanographic Data Centre (BODC)[1] and the Dutch Rijkswaterstaat[2].

Description of the Problem

BODC and Rijkswaterstaat have large marine databases holding a wide range of physical, chemical and biological measurands (the Open GIS Consortium (OGC) term for something that has been measured). Both organisations participated in two EU projects, EDIOS[3] and SEA-SEARCH[4] that developed pan-European metadatabases, which use a measurand discovery vocabulary[5] developed by BODC. As part of the vocabulary development, a mapping was built to the BODC measurand mark up (the BODC Parameter Usage Vocabulary[6]) but no such mapping existed for the Rijkswaterstaat measurand mark up. There were two possible solutions to this problem:

- Mapping the Rijkwaterstaat measurand mark up and the BODC discovery vocabulary
- Mapping between the measurand mark ups of the two organisations, allowing the BODC discovery vocabulary mapping to be used.

It was realised that whilst the second approach was more difficult and would involve more work, especially enhancement of the BODC Parameter Dictionary, the resulting parameter interoperability offered significantly greater reward. Resources for BODC

dictionary development were available through the NERC EnParDis[7] (Enabling Parameter Discovery) project. Consequently, the full mapping approach was taken.

Measurand Mark up in BODC and Rijkswaterstaat

There are significant differences between the measurand mark up strategies used by BODC and Rijkswaterstaat. The BODC system has its roots in the GF3 data model[8] in which measurand instances are linked to a key (termed the parameter code) defined by an entry in a parameter dictionary. This specifies one or more items of information about what the measurand is and how it was obtained. As the BODC database expanded from physical into biological and chemical data, limitations of legacy data formats resulted in a significant increase in the parameter code information load.

The Rijkswaterstaat database was designed around the DONAR[9] data model. This has the measurement as the primary entity, which is linked to a set of attributes containing specific atomic items of metadata describing the measurand and where, when and how the measurement instance was made. Each item of metadata is populated from a controlled vocabulary. The DONAR Parameter Dictionary is therefore simply a catalogue of valid combinations of metadata information items pertaining to the identity of the measurand and how it was made linked to a key.

The Starting Position

At the start of the mapping exercise in October 2003 the BODC dictionary described the mark up code through two plain text fields containing up to 200 bytes each. These had been populated over the 25 years of the dictionary's development in a less than consistent manner. Certain information categories were sometimes in one field and sometimes in the other. The grammatical structures were inconsistent and consequently the fields could not be concatenated sensibly. Whilst this situation was acceptable for interpretation by a human, it was totally inadequate for use by software agents.

In contrast, the DONAR presented Rijkswaterstaat with a particular type of information in a consistent and readily identified field within the data model. Furthermore these items of information, which we will term semantic elements, could be concatenated to provide comprehensible measurement descriptions in both Dutch and English. The system could therefore be used both by software agents and for presentation through a user interface.

Dictionary Mapping

At its most basic level, the mapping between the DONAR catalogue and the BODC Parameter Dictionary involves the following steps:

- For each entry in the Rijkswaterstaat catalogue (delivered as a spreadsheet with one column per semantic element):

- Use BODC dictionary search tools (Microsoft Access Filter by Form) to locate the entry having the same meaning as the combination of semantic elements
- If found: Copy code from Access form and paste into the DONAR spreadsheet
- Else: Manually prepare a dictionary entry record and submit to for quality assurance and loading.

This process is tedious, error prone and pushes the limits of human endurance for dictionaries with more than a couple of hundred entries. This mapping exercise involved thousands and it became obvious that a mechanism for automating the procedure was required.

The first attempt at automated mapping was developed by Pieter Dekker (from Xi advise bv) and used semantic analysis on the two BODC dictionary plain text fields to identify the DONAR element combinations that were a close match. The user then selected which one to use. This failed to work in practice for two reasons. First, the system had no mechanism to expand the population of the BODC dictionary. Consequently, if the Rijkswaterstaat element combination wasn't covered, there was no way in which it could be mapped. Secondly, the vocabulary and syntactic structure of the BODC dictionary plain text fields were not standardised. Human intelligence can recognise that '*Calanus* abundance', 'abundance of *Calanus*', 'number of *Calanus* per unit volume' have the same meaning, but artificial intelligence cannot without an extensive domain thesaurus or ontology.

BODC Dictionary Development

Following a presentation of the DONAR model and the semantic mapping tool at Rijkswaterstaat in December 2003, it became apparent that the BODC dictionary required drastic improvement if the mapping was to succeed. The approach taken was based on the extension of the DONAR design principles to the scope covered by the BODC Parameter Dictionary. In particular, the ability to combine semantic elements into meaningful text descriptions was enhanced.

The DONAR model per se was not adopted because there was already evidence in its usage at Rijkswaterstaat of 'shoehorning' where multiple items of information were forced into a single semantic element because they were needed and there was nowhere else for them to go. The scope of the BODC dictionary would make the problem much worse. For example, some BODC zooplankton data includes development stage information that would have had to be included in the same element as the taxon name.

The model developed for biological dictionary entries currently contains the following semantic elements:

- Parameter (Abundance, Biomass)
- Taxon_code (Integrated Taxonomic Information System (ITIS[10]) code)

- Taxon_name
- Taxon_subgroup (gender, size, stage)
- Parameter_compartment_relationship (per unit volume of the, per unit area of the)
- Compartment (water column, bed, sediment)
- Sample_preparation
- Analysis
- Data_processing.

Element content is governed by a controlled vocabulary, with any elements that are not relevant to a particular dictionary entry coded as 'not specified'.

The elements may be combined into text descriptions like:

'Carbon biomass of Urotricha (ITIS 46243) <20um per unit volume of the water column by optical microscopy and abundance to carbon conversion using the equation of Putt & Stoeker (1989)'

This is one of three sub-models currently being developed to cover the scope of the BODC dictionary, the other sub-models being for contaminant in biota data and a 'chemical' sub-model that seems to cover everything except biology. Once the model population has been completed these sub-models will be combined into a single element superset. Further atomisation of the model will also be undertaken at this stage where 'shoehorning' has been observed, such as division of the biological sub-model taxon_subgroup element into 'gender', 'size', 'development stage' and 'taxon_subgroup'.

Semantic Model Mapping

Mapping between two semantic models is a two-stage process. The first stage is to produce a mapping between the semantic elements in the two models. For example, the DONAR 'parameter' semantic element contains entries such as 'biomass per surface area unit' and 'number per volume unit', which are concatenations of the BODC model elements 'parameter' and 'parameter_compartment_relationship'. Note that it is by no means certain that this mapping will be a simple one-to-one relationship, particularly if shoehorning has occurred during population of the model instances.

The second stage is to produce a mapping between the vocabularies for the mapped elements. For example, the DONAR 'compartment' element maps to the BODC element of the same name. A subset of the vocabulary map is as follows:

Rijkswaterstaat	**BODC**
soil/sediment	bed
suspended solids	suspended particulate material
surface water	water column
porous water	sediment pore water

This two-stage process normalises the mapping procedure, cutting down the number of comparisons required by at least an order of magnitude. Furthermore, as the semantics of the element vocabularies are simple, mapping automation becomes an achievable goal. In practice, once BODC dictionary population issues had been addressed, over 90% of the final map was achieved by running a single SQL statement. Manual expansion of the vocabulary maps to deal with instances of different names meaning the same thing brought the level of completion to over 99%. Develop of thesaurus servers would allow automation of this part of the process as well.

There were a small number of DONAR combinations that required significant manual effort to achieve a mapping due to unclear or ambiguous semantics in the element vocabularies. For example, after an e-mail exchange to clarify semantics the term 'residual beta' was mapped to 'beta emitters other than 3H and 40K'. It would therefore seem that fully automated mapping is currently not an achievable goal.

An obvious, but important, point is that if a complete map is to be produced then one of the models must be the superset of the other. Until the Nirvana of an all-encompassing model is achieved, this will inevitably mean that the population of one of the models will need to be expanded as part of the mapping exercise. Adding records as part of a manual mapping exercise is a long and tedious process. However, the dictionary expansion requirement from semantic model mapping involves either adding new combinations of existing vocabulary members or a vocabulary extension. This was achieved quickly and relatively easily for the mapping exercise documented here in a semi-automated manner using a basic general-purpose tool (Oracle's SQL*PLUS). Bespoke tools are currently under development that will make the job easier still.

Checking the map produced revealed some errors due to homonyms such as Branchiura. The taxon identifier fields used contained names with no qualifying information such as a reference or a taxonomic database key. The exercise emphasised that this standard of labelling is insufficient to support totally reliable automated interoperability between biological databases. The BODC semantic model now includes an ITIS key element as a result of the lessons learned.

Semantic Model versus Parameter Dictionary

It is clear from this exercise that the semantic model is a much more powerful interoperability tool than the parameter dictionary (a vocabulary describing measurands through a plain text description). Furthermore, mapping measurands across databases uses only part of their potential. The map generated in this exercise can be used to determine that a given measurement in the Rijkswaterstaat database is exactly the same thing as a measurement in the BODC database. Such measurements may obviously be safely combined into a composite data set. However, what about cases where measurement descriptions are nearly the same, or where 'fit for purpose' criteria determine the measurements that may be safely combined?

Consider the following example for chlorophyll. The following are two entries from the BODC dictionary generated by concatenation of elements from the chemical semantic sub-model:

Concentration of chlorophyll-a {chl-a} per unit volume of the water column [particulate >30um phase] by filtration, acetone extraction and fluorometry

Concentration of chlorophyll-a {chl-a} per unit volume of the water column [particulate 0.6-5um phase] by filtration, acetone extraction and fluorometry

A user wishing to determine the timing of the spring bloom would happily merge data corresponding to both of these descriptions. A user with a compartmentalised biogeochemical model would not as the two measurements represent two completely different phytoplankton communities. If our first user was given control over which semantic model elements were used to build the description both the above could be reduced to:

Concentration of chlorophyll-a {chl-a} per unit volume of the water column

In this way we have provided a simple mechanism for user control over the scope of data interoperability. This same mechanism could also be used as the basis for a data set discovery interface. By turning elements on or off the user is able to control the level of information detail used in the subsequent search. Note that in this scenario it is not always necessary to specify precise semantic elements values for a search. In many cases it will be sufficient to simply say that the element should be other than 'not specified'.

It can therefore be seen that semantic models provide far more than just a tool for automated parameter mapping.

Conclusions

The following conclusions were drawn from this work and other associated activities in the EnParDis project:

- Manual mapping of measurand metadata is only feasible on the smallest of scales and that automation is not possible if the metadata is encoded as unstructured plain text
- Automated mapping becomes feasible if the description is encoded as atomised semantic elements, but it could be further improved by the availability of domain-specific thesauri and ontologies
- Standardisation of vocabularies, especially the utilisation of keys from published reference sources, renders automated mapping both easier and more reliable
- 99% of a map is completed in 10% of the time required to produce a complete map
- Semantic models may be used for purposes other than parameter mapping

- The conversion of the 16,500 term BODC parameter dictionary from plain text descriptions to a semantic model has significantly enhanced its value as a tool for database federation and data interoperability.

Acknowledgements

The authors are indebted to many colleagues in their parent organisations, the international marine XML community (the ICES/IOC SGXML group and the EU MarineXML project) and the MMI project for stimulating discussions that have helped shape this work and for the encouragement that has helped it progress. The work at BODC was supported financially by the UK Natural Environment Research Council through the EnParDis project.

Notes

[1]British Oceanographic Data Centre: http://www.bodc.ac.uk
[2]Rijkswaterstaat: http://www.verkeerenwaterstaat.nl/?lc=uk or
http://www.rijkswaterstaat.nl (Dutch language)
[3]European Directory of the Ocean-observing System: http://www.edios.org/
[4]SEA-SEARCH: http:// www.sea-search.net
[5]The BODC Parameter Discovery Vocabulary:
ftp.pol.ac.uk/pub/bodc/jgofs/datadict/new/parameter_group.csv
http://www.bodc.ac.uk/data/codes_and_formats/parameter_codes/
[6]TheBODC Parameter Usage Vocabulary:
ftp.pol.ac.uk/pub/bodc/jgofs/datadict/new/parameter.csv
http://www.bodc.ac.uk/data/codes_and_formats/parameter_codes/
[7]Enabling Parameter Discovery (EnParDis): http://www.bodc.ac.uk/projects/uk/enpardis/
[8]GF3, A General Formatting System for Geo-Referenced Data, IOC Manuals and Guides 17, UNESCO 1987
[9]DONAR: http://www.waddensea-secretariat.org/news/symposia/Demowad/RIKZ.html
[10]Integrated Taxonomic Information System: http:// www.itis.usda.gov