Parameter Vocabularies in the NERC DataGrid Project

Roy Lowry[3], Ray Cramer[3], Marta Gutierrez[2], Michael Hughes[3], Kerstin Kleese van Dam[1], Siva Kondapalli[3], Susan Latham[2], Bryan Lawrence[2], Kevin O'Neill[1], Andrew Woolf[1]

[1]CCLRC e-Science Centre
[2]British Atmospheric Data Centre
[3]British Oceanographic Data Centre

Abstract

The NERC DataGrid (NDG) project has been investigating vocabularies describing measurements or model outputs for use in its 'discovery' and 'use' metadata schema. It has been recognised that the two types of metadata have significantly different requirements of this type of vocabulary and therefore that more than one vocabulary is required. Automated construction of metadata records from data holdings requires that mappings exist between data labelling and the vocabularies used. At the current time whilst the problems relating to parameter interoperability through the usage of standardised vocabularies are recognised, they have not been completely solved. Work has begun to explore the potential of schemas developed for the semantic web (OWL and SKOS) for providing the necessary solutions.

1 Introduction

The essential function of a Data Grid is the delivery of data from distributed sources as an integrated package. Effecting this integration without the need for human intervention whilst still producing a scientifically valid dataset requires that the source data be associated with rich metadata that describe the measurements made or model outputs in some considerable detail. The controlled vocabularies associated with this 'use' metadata are termed 'Parameter Usage Vocabularies'.

Data integration is impossible unless the user is able to locate appropriate source datasets. The discovery process also needs metadata describing the measurands. However, in this case the detail required for usage is more of a hindrance than a help. Consequently, separate controlled vocabularies, termed 'Parameter Discovery Vocabularies', have been developed for this purpose.

2 Parameter Usage Vocabularies

Parameter usage vocabularies contain detailed information about an individual measurand. They were initially designed to label data streams, such as spreadsheet columns or data arrays, in a standardised manner but have proved equally suited to 'use' metadata applications.

During the past twenty-five years a large number of parameter usage vocabularies have developed in the oceanographic and atmospheric domains. Many are based on the GF3[1] (an early oceanographic standard) model where a key is used as a label inside data objects and externally defined through a parameter dictionary. The scope, content and quality of these definitions is highly variable. At best the dictionary terms are consistent, structured and unambiguous. At worst they are simply collections of the uncontrolled labels applied to the data by the originating scientists.

There is currently strong interest, particularly in the oceanographic community, in building distributed data systems. Parameter usage vocabulary interoperability is one of the main problems that needs to be addressed to realise this.

There has been some debate on the applicability of parameter usage vocabularies to 'discovery' metadata. Whilst this is feasible, consensus opinion considers that the quantity and complexity of the information delivered is too high.

3 Parameter Discovery Vocabularies

Parameter discovery vocabularies (otherwise known as parameter keywords or valids) describe the parameters measured in broad terms, often arranged as hierarchies of increasing specificity.

It is quite common for the terms in these vocabularies, even at the lowest level in the

hierarchy, to cover groups of measurements. For example the term 'winds' covers many measurands such as 'wind speed' and 'wind direction', 'gust wind speed', etc. Consequently, they cannot be employed for data labelling or 'use' metadata applications where descriptions of individual measurands are required.

4 Parameter Vocabularies and NDG

A primary objective of the NERC DataGrid project is to provide integrated access to the data holdings of the British Atmospheric Data Centre (BADC) and the British Oceanographic Data Centre (BODC). The former includes climate system model outputs, meteorological data and atmospheric chemistry measurements, whilst the latter is dominated by multi-disciplinary observational oceanographic data including physical, chemical, biological and geological data. This objective requires that parameter interoperability in both 'use' and 'discovery' metadata be achieved between the two data centres.

4.1 NDG 'Use' Metadata

The 'use' metadata schema developed for NDG is the GML-based Climate Science Markup Language (CSML)[2]. This requires that parameters, or phenomena in GML terminology, be defined in a GML phenomenon dictionary containing entries of the following structure:

```
<gml:dictionaryEntry>
<om:Phenomenon
gml:id="air_potential_temperature">
<gml:description>Potential
temperature is the temperature a
parcel of air would have if moved
adiabatically to sea level
pressure.</gml:description>
<gml:name
codeSpace="http://www.cgd.ucar.edu/c
ms/eaton/cfmetadata/">
air_potential_temperature</gml:name>
<gml:name
codeSpace="GRIB">13</gml:name>
<gml:name
codeSpace="PCMDI">theta</gml:name>
</om:Phenomenon>
</gml:dictionaryEntry
```

Two problems face NDG. The first is the construction of a phenomenon dictionary encompassing the data holdings of the two data centres. The second is the construction of maps between parameters in the data holdings and the phenomenon dictionary entries to allow tools to automatically populate CSML records by data set parsing.

At BADC the parameters in most observational data are labelled using non-standardised originator text strings. Other data have parameters labelled using terms (i.e. text strings) from the Standard Name List[3] of the Climate and Forecast (CF) content standard. All BODC data are labelled using keys defined in the BODC Parameter Usage Vocabulary[4].

From this it is clear that the basic phenomenon dictionary requirement is for a web service that returns a gml:DictionaryEntry XML structure for either a Standard Name term or a Parameter Usage Vocabulary key. In this way the mapping between data and dictionary entries is addressed, providing the data are labelled using one of the standards. However, if BADC and BODC data are to be interoperable the segment structures need to be consistent for both types of input and be identical for equivalent phenomena. This raises two issues:

- The definition of the phenomenon
- Equivalent phenomenon mapping

There are three possible definitions for the phenomenon. The first, as in the example shown, is to define a phenomenon for each Standard Name term. The second is to define a phenomenon for each entry in the Parameter Usage Vocabulary. The third is to define an independent set of phenomena and map the two vocabularies to them as code spaces.

This latter approach is being given serious consideration by a number of interoperability projects in the oceanographic domain, such as ESIMO[5], OceanSITES[6], and MERSEA[7]. It is an extremely seductive idea to projects whose interest is restricted to a small, clearly defined set of phenomena. It is an approach that can work providing the phenomena are clearly and unambiguously defined at the outset and then set in stone. Significant intellectual input is required into any subsequent maintenance to prevent changes to the semantics of the pre-existing phenomena. Consequently, the approach does not scale, particularly to data centres dealing in thousands of phenomena and whose portfolio is continually expanding. The only way it could work in NDG is by restricting the phenomena regarded as interoperable, which is not considered desirable.

NDG therefore faces the prospect of either providing a CF Standard Name for each of the 17,000+ BODC keys or a BODC key for each of the 500+ Standard Names. Neither is attractive and the former would be impossible without a significant, and totally counter-productive, relaxation in the quality assurance rules governing Standard Name allocation. Alternatively, ways to reconcile 'CF phenomena' with 'BODC phenomena' (i.e. accept differences in phenomenon dictionary entry structure) will need to be found.

Once the phenomena definitions have been established, the next issue to be addressed is the code space mapping. The gml:dictionaryEntry structure provides a container for such mappings. However, inclusion of terms from other code spaces implies that the terms are exactly equivalent and detailed examination of a number of parameter vocabularies reveals that this is rarely the case.

Work on parameter vocabulary mapping is in progress in collaboration with John Graybeal and Luis Bermudez of the Marine Metadata Interoperability[8] project using ontology development techniques. The method, due to be trialled at a workshop in Boulder in August 2005, is to translate each vocabulary into OWL format and then use bespoke tooling, driven by a domain expert, to build the map based on 'equal to','broader than' and 'narrower than' relationships. Encoding the maps resulting from this process into a GML phenomenon dictionary will not be possible without an extension to the structure to store the relationships.

The above discussion is concerned with those data that are labelled in a standardised manner. The non-standard name problem is being addressed by mapping the terms to terms from the BODC vocabulary, expanding this where necessary. Anyone who has experience of metadata created by others will realise the enormity of this problem. All too often parameters are labelled by strings laden with implied semantics whose full meaning is forgotten by the author as soon as work on the data has been completed.

Semantic analysis techniques are being used to reduce the problem to manageable proportions. Data files are mined for parameter description terms, which are mapped to originating campaign and scientists on the assumption that phraseology may be consistent within these contexts. Word matching, incorporating a refinable stopword list and fuzzy matching via a Levenshtein distance algorithm[9] can then be utilised along with the contextual grouping to aid what will finally be a manual mapping process. It may be possible to capture semantic rules that become evident during the process to decrease the size of the manual aspects, but this has not been put to the test.

NDG is currently at the stage of recognising the 'use' metadata parameter problems it faces, but has a significant way to go before they are fully resolved. The 'working demonstrator' requirement of NDG Phase 1 will be addressed by using a few carefully chosen phenomena. However, an operational and scalable solution to the phenomenon dictionary issues will be required for the operational phase of NDG. One of the aims of exposing these issues through this poster and paper is to invite input from other domains in the e-Science community who have experience in this area.

4.2 NDG 'Discovery' Metadata

The NDG 'discovery' metadata strategy is to develop an over-arching repository from which discovery records corresponding to a range of schemas may be generated using XQuery and XSLT transforms. A schema, termed Metadata Objects for Links in Environmental Sciences (MOLES), has been developed, which recognises two important issues relating to parameter discovery vocabularies. First, the discovery vocabulary required for a particular type of discovery record is usually dictated by the discovery record schema. Secondly, discovery vocabulary terms can be a hierarchy built from sets of sub-terms of increasing specificity. The MOLES schema allows each parameter to be described using as many vocabularies as required, each of which may be hierarchical.

For its first phase NDG adopted the Global Change Master Directory (GCMD) Directory Interchange Format (DIF)[10] for its discovery metadata, which incorporates a parameter discovery vocabulary (GCMD Parameter Valids)[11] covering a wide range of Earth Science domains. Experience working with this vocabulary revealed some problems, particularly with the variability in the granularity of the terms, which caused particular difficulties building the mapping required to automatically generate discovery records from usage metadata and hence data.

These issues may soon be addressed, at least for the oceanographic domain. Co-operation between BODC and GCMD to incorporate the relevant entries from the EDMED[12] metadata catalogue into the GCMD Antarctic Master Directory portal by metadata record interoperability will enhance the GCMD Parameter Valids though incorporation of a parameter discovery vocabulary[4] designed from scratch by BODC for the EU SEA-SEARCH[13] project. This will take with it a mapping to the BODC Usage Vocabulary and potentially the CF Standard Names and therefore permit automatic generation of discovery metadata parameter information from datasets labelled using either data convention.

5 The Synonym Issue

It is inescapable that data users and creators use many different words to describe a single phenomenon, even when restricted to a single language. We need to be able to assemble a dataset from components labelled 'PCB28' as well as '2,4,4'-trichlorobiphenyl' without requiring the whole community to be fluent in IUPAC.

Presently, little work has been done to address the synonym issue in the domains represented by the NDG project beyond recognising that there is a problem. The CF Standard Name List incorporates the concept of aliases, but their usage is far from comprehensive. The BODC Parameter Usage Vocabulary is underpinned by a semantic model built from elemental facets of parameter-relevant information. Each facet is described using a controlled vocabulary and it has been recognised that the synonym issue affects both individual terms and combinations of terms from these vocabularies. However, the problem of how to manage this information in such a way that datasets are discovered despite variations in query phrasing has yet to be addressed.

The potential of RDF-based semantic web technologies such as OWL[14] and SKOS[15] has been recognised, but real work to exploit this has yet to begin. Input from and collaboration with other domains in the e-Science community who are working in this area would therefore be welcomed.

6 References

[1]A General; Formatting System for Geo-Referenced Data. Volume 2: Technical description of the GF3 Format and Code Tables. IOC Manuals and Guides 17, UNESCO, 1987.

[2]ndg.nerc.ac.uk/csml

[3]www.cgd.ucar.edu/cms/eaton/cf-metadata/standard_name.html

[4]www.bodc.ac.uk/data/codes_and_formats/parameter_codes/bodc_para_dict.html

[5]www.meteo.ru/nodc/Project_e/progr.html

[6]www.oceansites.org/OceanSITES/

[7]www.mersea.eu.org/

[8]marinemetadata.org

[9]www.merriampark.com/ld.htm

[10]gcmd.gsfc.nasa.gov/User/difguide/difman.html

[11]gcmd.nasa.gov/Resources/valids/gcmd_parameters.html

[12]www.bodc.ac.uk/data/information_and_inventories/edmed/

[13]www.sea-search.net

[14]www.w3.org/2004/OWL/

[15]www.w3.org/2004/02/skos/