APPLICATION OF THE NERC DATA GRID METADATA AND DATA MODELS IN THE NERC ECOLOGICAL DATA GRID

Neil Bennett[1], Rod Scott[2], Mike Brown[2], Kevin O'Neill[3], Mandy Lane[2], Andrew Woolf[3], Kerstin Kleese-van Dam[1], John Watkins[2]

[1]CCLRC – Daresbury Laboratory, Daresbury, Warrington, Cheshire, WA4 4AD, UK.
[2]CEH - Lancaster Environment Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP, UK.
[3]CCLRC – Rutherford Appleton Laboratory, Chilton, Didcot, Oxon, OX11 0QX, UK.

Abstract: The Centre for Ecology & Hydrology (CEH) holds various databases collectively representing a valuable environmental research resource. However, their use inside and outside CEH is constrained by lack of data accessibility and interoperability. This project has focused on three test-bed datasets held at the Lancaster Environment Centre which form a good example of the diversity of CEH terrestrial and freshwater data. Metadata systems and access tools have been constructed in collaboration with the NERC Data Grid (NDG) to provide users with Grid services linking data discovery to dataset delivery. This paper focuses on the modelling of CEH's data and metadata using the NDG models.

1. Introduction

CEH is the leading UK body for research, survey and monitoring in terrestrial and freshwater environments. The main aim of this project is to demonstrate that a diverse subset of CEH's data can be made more accessible by integration into the NDG. Three datasets have been used for the EcoGrid project: the Countryside Survey (CS) vegetation database [7]; the Environmental Change Network (ECN) database [8]; and, the Lakes database [9].

NDG has defined a detailed metadata model [1], MOLES (Metadata Objects for Links in Environmental Science) and a data model [2] known as CSML (Climate Science Modelling Language). Both models are standards-based and represented as XML schemas. They are intentionally generic so they can accommodate data from a wide range of scientific disciplines.

MOLES has been designed to allow production of the various 'industry standard' discovery formats, such as DIF [10], ISO 19115 [11], FGDC/GEO [12] and SensorML [13]. MOLES allows a smooth link from data browse to data usage.

CSML provides information about the data that processing and visualisation services need. It describes the data in semantic terms and virtualises it, removing the need to know the actual format of the data.

To integrate the EcoGrid data into NDG requires mapping to the NDG models [6]. Often the distinction between data and metadata is blurred. Theoretically, metadata are descriptive data whereas data are physical measurements. However, in reality, measurements can be part of a descriptive coding system, e.g. to search datasets by species of interest requires the individual species codes recorded in each data set to be in the discovery metadata. Thus, species will feature in both MOLES records (from which discovery metadata is generated) and CSML.

For discovery metadata such as species, it is important that EcoGrid recognises synonyms. Otherwise, a user searching the portal may not retrieve all relevant records. EcoGrid is collaborating with the producers of a data dictionary, using a common taxonomical classification, to solve this problem.

Also, a significant proportion of EcoGrid data relates to freshwater chemistry where various parameters are measured and values recorded in appropriate units. To help describe datasets fully and aid comparison with other datasets both inside CEH and outside, it is important to reference units and parameters to standardised definitions in widely accepted dictionaries. CSML incorporates references to unit and 'phenomenon' dictionaries so EcoGrid's use of CSML will help with the issues of accessibility and interoperability.

2. Securing Sensitive Data

Much of the testbed data is sensitive so it is important that users do not receive any more information than they need. NDG security takes care of user authentication and authorisation [5]. Its development used the CCLRC Data Portal Authorisation Architecture [3] as a starting point.

Where appropriate, CEH have implemented a two tier data structure so that authorised users can access raw data whereas summarised data is available to everyone. For both ECN and Lakes, the data has been summarised temporally.

For CS, the raw data is accessible but the precise location of the survey is not disclosed. Thus, datasets are available at the CS square level but the location shown is just the government office region. Hence, only one authorisation role is required.

3. The Elements of MOLES

To gain an understanding of how the test data has been modelled and the issues involved, it is important to introduce the models themselves.

There are four main elements within MOLES:

- Activity, e.g. a whole project or a field survey in a particular location.
- Data Production Tool. Used to make measurements and collect data, e.g. something as simple as a quadrat.
- Observation Station. A site of data production tools , e.g. a nature reserve.
- Data Entity. This contains metadata about the data itself and is linked to usage metadata (in a CSML file).

These elements are linked by IDs to create deployments (the use of a data production tool at an observation station to produce a data entity on behalf of an activity). Much of CEH's data is collected by scientists undertaking surveys in the field. MOLES handles this by classifying the surveyor as an observer within the deployment.

4. Introduction to CSML

There are several components within CSML.

Phenomenon. This is a property or variable that an activity sets out to measure. This can be referenced to a standard dictionary (such as Climate Forecast standard names [14])

Unit. The unit of measure of a phenomenon. Units can be referenced to a standard dictionary such as Unidata's 'udunits' [15].

Feature Type. An abstract representation of the measurement of a phenomenon, e.g. a measurement taken at one point over a time series is a PointSeriesFeature.

Array. Sometimes, a temporal or spatial coordinate series is used repeatedly. It is then more efficient to declare an Array once and make references to it than explicitly write out the series every time.

Reference System. These provide efficiencies when there is a systematic pattern in (say) spatial coordinates or a time series. This means that every time or spatial coordinate need not be specified explicitly. Instead, we can specify a starting point, an interval and an end-point or total number of elements.

5. Modelling data using CSML

5.1 Water chemistry & meteorological data

CSML was developed by the Atmospheric Science and Oceanography communities and as such copes very well with water chemistry & meteorological data. Such data is characterised by chemical and physical measurements taken at single points in space over time series.

5.2 Species

Each CEH dataset may contain data for hundreds of different species. These species names could be rolled up into phenomenon names e.g. 'count of bat species A', 'count of bat species B', etc. However, this would require huge numbers of different phenomena within a dataset that can not be matched to standard phenomenon dictionaries. The solution implemented is to create an EcoGrid dataset for each species.

5.3 Categorical Data

CEH has a significant amount of categorical and free-text type data (e.g. habitats), which should clearly be represented as phenomena. Unlike most phenomena, these do not possess units and so will not possess a 'unit' attribute.

5.4 Missing values

A common scenario is that measurements are made over a time series. However, for CEH, occasionally there are breaks in the series, perhaps due to equipment being shut down for maintenance, etc. Potentially, this could result in more times in the series than there are measurements recorded. CSML copes well with this by allowing us to explicitly state when a "missing value" occurs.

6. Issues & Potential Solutions

6.1 Species

One of the main limitations of CSML is that it has no obvious way of modelling biological concepts that make up much of CEH's data. Often the number of individuals of a species is counted and this can be subdivided by gender, morph, stage of development, etc in places. There are two possible solutions.

Solution 1

The various combinations of these attributes could be rolled up into phenomenon names. This is conceptually simple but creates a few problems. Firstly, a very large number of phenomena would be required. Secondly, these phenomena would be so fine grained that they can not be referenced to a standard dictionary, and other researchers would find it difficult to compare their results with CEH's. Finally, the phenomenon name used in the MOLES file would be a simple name (such as number of individuals) and as such would not match up to the phenomenon name in the corresponding CSML file.

Solution 2

Another solution is the use of composite phenomena [4], a concept inherited by CSML. This consists of another phenomenon (e.g. count of individuals of species A) to which a vector is applied. In this example, the vector could be gender (e.g. male and female). It is also possible to nest composite phenomena, so that a composite phenomenon consists of another composite phenomenon (and a vector) which in turn consists of a simple phenomenon (and another vector). So, extending the example, could result in:

Composite phenomenon 1 (count of individuals by species by gender) = composite phenomenon 2 (count of individuals by species) * vector 1 (gender)

AND

Composite phenomenon 2 = Simple phenomenon (count of individuals) * vector 2 (species)

An advantage of this approach is that the root phenomenon is simple and standardised. However, a major problem is that the 'Composite Phenomenon' construct of OGC is under development [4], and CSML's associated tools do not currently support it. An additional problem is that it is not clear how the relationship between data values and vectors would be illustrated. For example, if a phenomenon applies across a range of locations then the string of values measured will be of the form "x1 x2 x3". If the phenomenon is complex and contains a vector (say gender) applied to a simple phenomenon then the string could be of the form "(x1 y1), (x2 y2), (x3 y3)". It is not then clear whether y1 is the phenomenon witnessed for gender 1 at location 2 or whether it is for gender 2 at location 1. The problem gets worse as the level of nesting increases.

For now, this issue has been circumnavigated by making data available only at the 'number of individuals' level. In the future, ontologies may be used to cope with such situations better.

6.2 Profile Series Features

CSML has a feature type called ProfileSeriesFeature to represent a measurement recorded at various points along a directed line in space over a time series.

CEH has lake data where measurements are taken at various water depths over a time series. Here, the ProfileSeriesFeature type would be ideal for modelling such data. However, since the measurement depths may not be constant from one time to the next, ProfileSeriesFeatures can not be used. This situation is encountered in other domains (e.g. observational oceanography) and a new feature type will be introduced to cope with such eventualities in a future revision of CSML.

6.3 Observations & Measurements paper

CSML inherits part of its structure from the schema detailed in the Open Geospatial Consortium Observations & Measurements paper [4]. This paper contains an example of how this model could be applied to an ecological survey example, and has been reviewed to see whether it provides solutions to the issues faced when using CSML. However, the data modelled in the example is far simpler than what CEH has so it has not provided a complete solution.

6.4 Transects

Some of CEH's data (e.g. butterfly surveys) is collected via transects. Transects are just routes

that the observer walks along. As the observer follows the transect, they make observation notes at various points. CSML has a Trajectory feature type that might be used to model transect data. However, data is repeatedly collected from a transect forming a time series. Unfortunately, trajectories can not be used to model time series.

An alternative is the ProfileSeriesFeature type. However, a profile must be a straight line with a specified direction, whereas a transect may be a curved path. This problem is avoided by summing data at a higher level (e.g. site).

6.5 Sublocations

For some of its data, CEH uses several different subdivisions to express location more precisely, e.g. for spittle bugs there are sites comprising locations that comprise quadrats. This situation is not handled by CSML well. One option is to use reference systems but this would be very complex to implement and understand. Again, the solution adopted has been to sum data at a higher level (e.g. site or location code).

6.6 Date type phenomena

In some cases, there is date-type data that does not reflect the time at which measurements were made but instead is one purpose of the survey itself. For example, one survey measures when frogs are first seen congregating, hatching and leaving the pond. This data must therefore be modelled as phenomena. It is possible to do this if we do not specify a related unit. However, this implies that the date is just a text string when in fact it carries more meaning than that.

If CSML allowed us to specify the corresponding unit as 'date' then a user would be able to make sensible comparisons between data e.g. to see where frogs are hatching first.

6.7 Date Type Parameters

Typically, a measurement is made at a point in space over a time series. However, there are cases, particularly for water chemistry, where samples are taken from a river and then at some subsequent time/date, analysis is done. There can be several stages for the analysis and the time/date of each is recorded, e.g. pH measurement, filtration, and completion of analysis.

There is currently no suitable place for these within CSML. They are not phenomena because they are not something a scientist sets out to measure. On the other hand, it is possible to record one time/date as an input parameter but not several. Such data is omitted from CSML files.

6.8 Replicate measurements

For water chemistry measurements, CEH sometimes uses multiple test tubes at the same location at the same time to measure the same phenomena. There is no mechanism within CSML for modelling this.

Composite phenomena could be used with a vector containing the replicate IDs. However, these are not currently supported in CSML. At this stage the only possible solution is to store mean or modal values of these replicates.

6.9 Protocol metadata

To study freshwater invertebrates, CEH uses nets to collect samples from rivers and lakes. The properties of a net are significant in determining which species it yields. Technically, the net is a data production tool and as such it would normally be described in the corresponding MOLES record. However, several different nets can be used within one dataset or study so this information must go into the CSML record instead. However, there is currently no suitable place for such information within CSML.

7. Future

It would be desirable to extend EcoGrid to cover all CEH data and also incorporate data held by the National Biodiversity Network which focuses on Sites of Special Scientific Interest and thus is complementary to EcoGrid's data.

UK ecologists would like to collaborate with ecologists worldwide. The Knowledge Network for Biocomplexity [16] is trying to solve similar problems and has developed Ecological Metadata Language (EML) to describe ecological data. Creation of EML records should open up CEH's data to a much wider audience.

Environmental data always has a spatial component. The first version of NDG will allow spatial searching of datasets. However, nothing more complex is permitted. The second stage of NDG, which began in late 2005, may introduce improved spatial capabilities with consequent benefits for EcoGrid.

8. References

[1] A specialised metadata approach to discovery and use of data in the NERC Data Grid. K O'Neill et al., Proceedings of the UK e-Science All Hands Meeting 2004. (http://www.allhands.org.uk/2004/proceedings)

[2] Climate Science Modelling Language: Standards -based markup for metocean data", 85th meeting of American Meteorological Society, San Diego, Jan 2005.

[3] Grid Authorisation Framework for the CCLRC Data Portal. A Manandhar et al., Proceedings of the UK e-Science All Hands Meeting 2003.

[4] Observations and Measurements, OGC Discussion Paper 05-087r3, Simon Cox editor. http://portal.opengeospatial.org/files/?artifact_id=14034

[5] NERC Data Grid Authorisation Architecture. N Bennett et al., Proceedings of the UK e-Science All Hands Meeting 2005. (http://www.allhands.org.uk/2005/proceedings)

[6] NERC Ecological Data Grid. N Bennett et al., Proceedings of the UK e-Science All Hands Meeting 2005. (http://www.allhands.org.uk/2005/proceedings)

[7] CEH Countryside Survey. http://www.cs2000.org.uk

[8] CEH Environmental Change Network. http://www.ecn.ac.uk

[9] CEH Lakes Database. http://www.ceh.ac.uk/sections/eaf/EAFCumbriaLakesDatabase.html

[10] Directory Interchange Format. http://gcmd.nasa.gov/User/difguide/difman.html

[11] ISO/TC211. ISO activity in the Geographic Information/Geomatics domain including the ISO191xx series of standards. http://www.isotc211.org/

[12] FGDC. Federal Geographic Data Committee Standard for Digital Geospatial Metadata (FGDC-STD-001-1998). http://www.fgdc.gov/metadata/metadata.html

[13] Open Geospatial Consortium – SensorML. http://www.opengeospatial.org/

[14] CF standard name table. http://www.cgd.ucar.edu/cms/eaton/cf-metadata/standard_name.html

[15] udUnits. http://www.unidata.ucar.edu/software/udunits/

[16] http://knb.ecoinformatics.org/software/eml/