A specialised metadata approach to discovery and use of data in the NERC DataGrid

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

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

1. INTRODUCTION

The Natural Environment Research Council (NERC) has a wide range of data holdings, held in technologies from flat files to relational databases. These holdings are relevant to a wide range of scientific disciplines, despite often having been collected on behalf of quite narrow specialised disciplines. The data holdings are stored across a wide range of autonomous archives, ranging from specialist professional data curators and archivists, such as the British Atmospheric Data Centre (BADC) and the British Oceanographic Data Centre (BODC) to files held on the hard disc of an individual scientist's PC (See [Lawrence et al, 2003]). These, and the other NERC Data Centes are highly autonomous and are focussed on the needs of their various communities. Couple this with the fact that these all have well-embedded processes, then there is a need for a highly decentralised and distributed approach. The degree of implementation specification available to other data grids, e.g. EU DataGrid or GriPhyN, cannot be assumed.

The NDG vision is for the user to see these data resources as one entity, thus improving the ability of scientists to find and use data, with transitions between the discovery and use phases as seamless as possible. To enable this, the existing infrastructure, assets and resources of existing NERC Data Centres must be evolved to support a common framework. Initially, Atmospheric and Oceanographic data, respectively held at the BADC and BODC, will be made available. Data from other disciplines held in Data Centres funded by NERC will be available in due course. From this core, it is intended to extend the framework to include less formal "data centres", ultimately reaching down to the scientists' desktops. Also, provision is being made to allow interoperability with projects around the world, such as the Earth Science Grid [ESG].

A thorough process of requirements review and architecture specification, using RM-ODP as the methodology (see [Woolf et al, 2004]) was used in the design process. Key to the implementation of the result is a decoupled metadata infrastructure using, where possible, tools, standards, and mechanisms that already exist or are under development within e-Science and the worldwide earth science communities.

2. THE NERC DATAGRID METADATA TAXONOMY AND ITS USE

In trying to capture the entire metadata chain from discovery to use for the NDG within a single structure, it was found that this structure would be far too large to be easily managed or understood. Also, metadata values were found to have multiple semantics that may not sit together easily in the long run.

Thus, at the core of the NDG, there is a decoupled metadata infrastructure allowing specialisation of purpose, and enabling the linking of existing mechanisms and technologies such as search services and analysis tools, to perform required tasks.

The key types are the "Type A" metadata, which is directly concerned with the use of the data, "Type D" which is the metadata directly used by discovery services, and the "Type B" core metadata. These are linked together to form a flexible set of specialised "metadata modules" that allow specialist elements to be introduced for the broad range of disciplines to be covered.

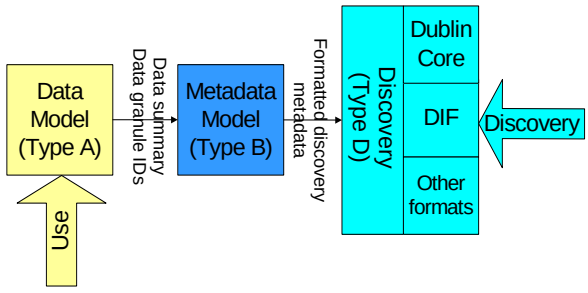

Figure 1 - Relation of major metadata types identified

2.1 Type A – the NDG Data Model

A, described in [Woolf et al, 2003], is directly concerned with the use of the data, its detailed content, and the location and assembly of the data for use. It represents the data in a format-neutral fashion, emphasising the semantics contained inside the data

itself and giving a foundation for the discovery of feature types represented by the data. This is referred to as the NDG Data Model, emphasising its inclination towards the data itself.

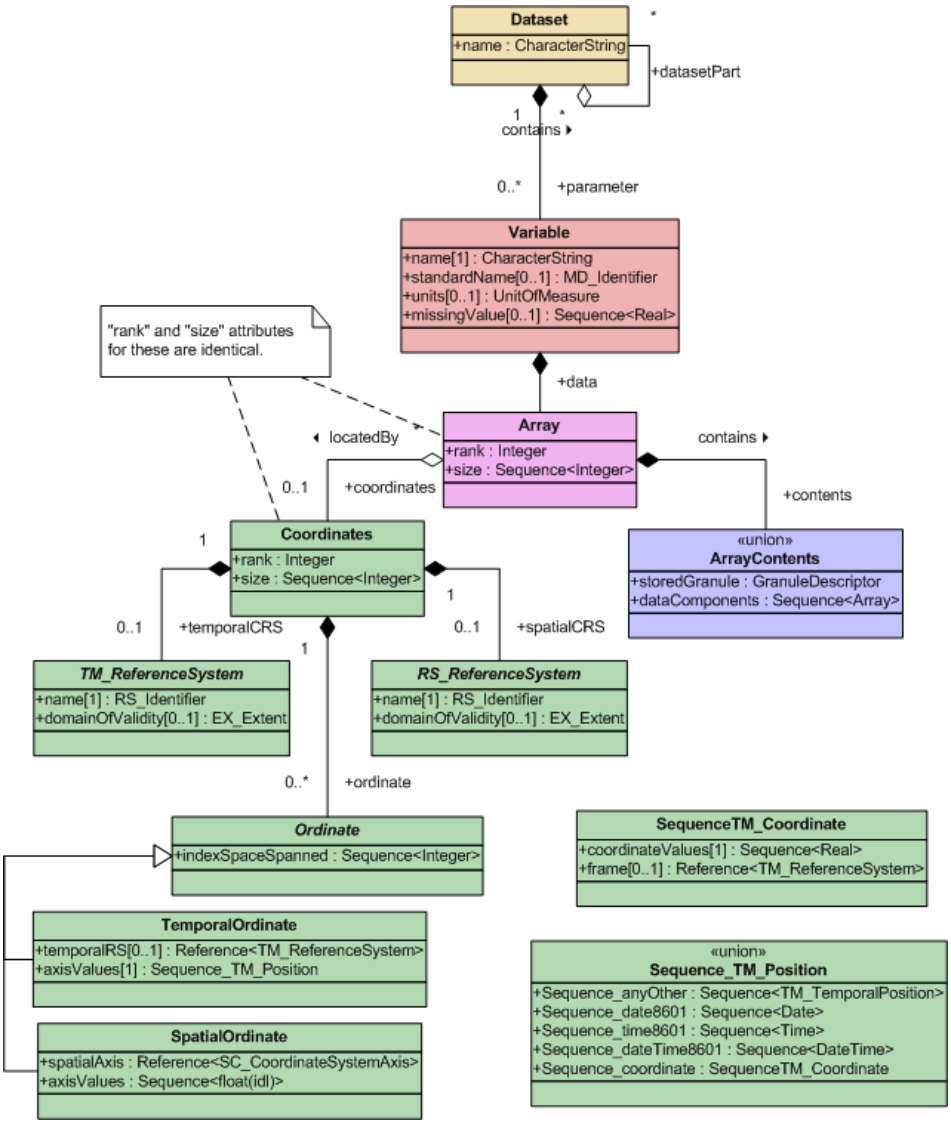

Figure 2 - Core of the NDG Data Model

Representing the data via a generic “array of arrays” model with provision for specifying spatio-temporal reference systems and allied metadata, it draws strongly on the work of ISO TC211 (see [ISO/TC211]) and the Open GIS Consortium, in particular the concept of “feature types”, which correlate the dimensionality and value constraints of the data to semantic objects, and the data types and representation systems specified.

By use of this metadata it is possible for the user to extract and use the data in a number of formats without having to be aware of the original format.

2.2 Type B – The NDG Metadata Model

B is a superset of the Type D Discovery metadata and is described in [O’Neill et al, 2003]. It identifies the major data feature types of interest, supports the metadata required for high-level data discovery, and relates the entities identified therein. It is used to generate different “D Type” discovery formats from a single corpus of metadata, e.g. Dublin Core, GCMD DIF, or FGDC Z.39.50 “GEO” profile. As B is a superset of these formats, it provides a metadata resource more comprehensive than the “raw” discovery formats that can be accessed prior to accessing more detailed metadata. It is referred to as the NDG Metadata Model.

It is implemented using a “relational XML” approach, in which major metadata entities have their own records that contain information relevant to them only (see [O’Neill et al, 2003]). These are related by using XQueries to operate on the keys to build the complete metadata record, using the concept of a “Deployment” as the relating entity.

In a Deployment, the important elements that produce the data are brought together. It gives a simple, yet meaningful, mechanism for navigating the metadata.

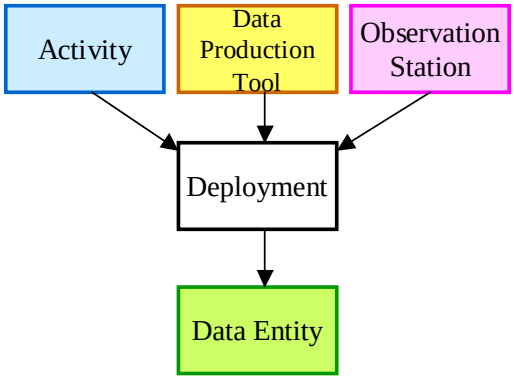

Figure 3 - A "Deployment"

Much as the Data Model looks inwards at the actual data, the Metadata Model looks out to show how the data fits in the wider world. It links the data to the entities that produced it and will provide the "hook" to link to ancillary systems that will be added in the future, such as publications databases and annotation systems. However, its major task is to allow the generation of discovery formats commonly used in existing search engines in the Earth Science world, such as GCMD's DIF (see [DIF]) or FGDC (see [FGDC]).

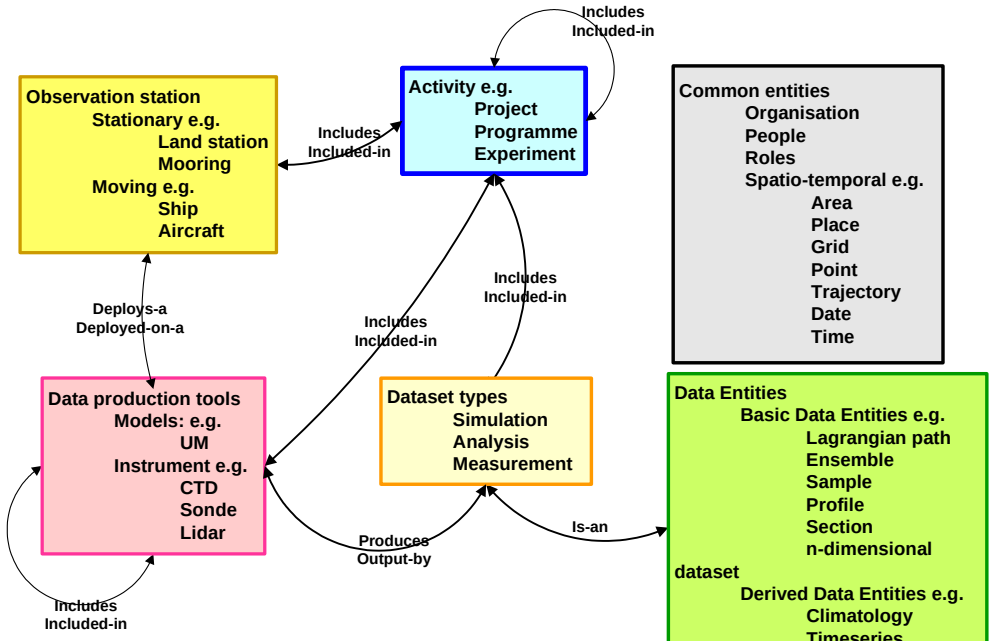

Figure 4 - Overview of NDG Metadata Model

3. LINKING THE DATA AND METADATA MODELS TO EACH OTHER AND THE WORLD

While each is useful in its own right, for the Metadata and Data Models to reach their full potential the of the two sides should be linked. This linkage is achieved via two key features, the use of identifiers and the summarisation of the Data Model into the Metadata Model.

3.1 Identifiers

It is vital that the "A" and "B", and hence the "D", metadata be able to reference each other. Two identifiers are generated, one by the Data Model and one by the Metadata Model.

The identifier generated by the data model links the data and metadata models. This identifier allows the Data Model to "rebuild" the relevant data set on demand.

The Metadata Model allocates unique IDs to all metadata objects, based on the OAI identifier standard (see [OAI Identifier v2]). This allows local repositories to allocate their own IDs, whilst ensuring that the IDs are unique across the NERC DataGrid. These IDs give a permanent reference for external systems that may wish to reference this metadata. They will be the basis of meaningful URIs to allow direct access to the metadata. Finally, these identifiers will allow the harvesting of discovery metadata by OAI compliant repositories.

3.2 Type S metadata - the data summary

The "B" metadata contains a summary of the data contents ("S" metadata). This contains details that are of use to the discovery services, such as parameters represented in the data, but that are dealt with in more detail in the "type A" data use metadata. This summary will be based on the feature types that are defined in the B metadata. Typically, this will include the spatio-temporal coverage of the data, the parameters contained in the data, and summary of the values of those parameters. Other details useful to the end user may also be included in the summary.

The parameters are not simple names, but identify real scientific "objects", being based on vocabularies being developed and maintained by the Earth Science community. As such, there is a strong linkage with projects such as the NetCDF CF convention (see [CF convention]), the BODC parameter dictionary (see [BODC dictionary]), and the EnParDis project (see [EnParDis]). Also, once they have been enumerated, feature types identified by the Data Model will be reflected here to allow users to search for the semantic objects thus represented.

The summary may gloss over details that are important in actual use, for example, an oceanographic profile can be regarded, for discovery purposes, as being a set of measurements taken at regular depths from a particular x,y,t co-ordinate in the sea; whereas the reality is that there is a time lapse between samplers, and the samplers may drift.

3.3 Model "daisy-chaining"

Most disciplines have widely used, almost standard, data formats encapsulating discipline-specific semantics. To try and encapsulate all of these or the many possible viewpoints that drive discovery

within a single metadata or data model is infeasible. However, by taking advantage of the modularity of the NDG Metadata Taxonomy, we can produce a series of Metadata and Data Models that can be

While it is hoped that the Metadata and Data Models will be able to cope with the vast majority of requirements, this mechanism means that each model can link to another, more appropriate, version of its partner should it be necessary. In this manner, separate data models could be kept under the same discovery metadata umbrella, as all NDG-compatible Data Models will have to provide the standard summary and be able to honour requests based on the accompanying identifier. Thus, the UK e-Science project EcoGRID (see [EcoGRID]) is using the NDG Metadata and Data Models as starting points, but expect these to be able to evolve as required, whilst retaining the ability to interoperate with the core system.

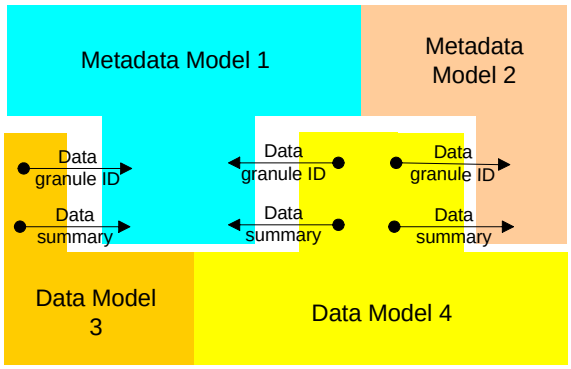

Figure 5 - Linking multiple Metadata and Data Models

4. CONCLUSIONS

The NERC DataGrid has developed a flexible and extensible standards-based framework for enabling both discovery and use of the data held at the various NERC Data Centres. While there is some apparent repetition of data values, there is no repetition of the values in a semantic sense. It forms a firm platform for expected evolution of the system.

5. REFERENCES

[Lawrence et al, 2004] Lawrence, B.N., R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, R. Lowry, K. O'Neill, and A. Woolf. 2004: The NERC DataGrid: "GOOGLING" secure data. Proceedings of the U.K. e-Science All Hands Meeting, 2004

[Woolf et al, 2004] Woolf, A., R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, B.N. Lawrence, R. Lowry and K. O'Neill. 2004: Enterprise specification of the NERC DataGrid. Proceedings of the U.K. e-Science All Hands Meeting, 2004.

[Lawrence et al, 2003] Lawrence, B.N., R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, R. Lowry, K. O'Neill, and A. Woolf. 2003: The NERC DataGrid Prototype. Proceedings of the U.K. e-Science All Hands Meeting, 2003

[Woolf et al, 2003] Woolf, A., R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, B.N. Lawrence, R. Lowry and K. O'Neill. 2003: Data Virtualisation in the NERC DataGrid. Proceedings of the U.K. e-Science All Hands Meeting, 2003.

[O'Neill et al, 2003] O'Neill, K, R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, B.N. Lawrence, R. Lowry and A Woolf. 2003: The Metadata Model of the NERC DataGrid. Proceedings of the U.K. e-Science All Hands Meeting, 2003.

[Lawrence et al, 2004] Lawrence, B.N., R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, R. Lowry, K. O'Neill, and A. Woolf. 2004: The NERC DataGrid: "Googling" secure data. Proceedings of the U.K. e-Science All Hands Meeting, 2004

[OAI Identifier v2] Open Archives Initiative, Specification and XML Schema for the OAI Identifier Format, Version 2002/06/21T21:48:00Z, http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm

[EnParDis] EnParDis, a NERC funded project run by the British Oceanographic Data Centre, http://ndg.badc.rl.ac.uk/public_docs/RKL_IOC_March2004.ppt

[EcoGRID] Ecological Data Grid, UK e-science project, http://www.escience.clrc.ac.uk/web/projects/hydrology_data_grid

[BODC dictionary] British Oceanographic Data Centre, http://www.bodc.ac.uk/cgi-bin/framer?http://www.bodc.ac.uk/documents/bodc_params.html

[CF convention] NetCDF Climate and Forecast Metadata Convention, http://www.cgd.ucar.edu/cms/eaton/cf-metadata/

[ESG] Earth System Grid project, http://www.earthsystemgrid.org/

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

[DIF] Global Change Master Directory DIF Writers' guide, http://gcmd.gsfc.nasa.gov/User/difguide/difman.html

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