THE METADATA MODEL OF 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

Abstract

The Natural Environment Research Council (NERC) DataGrid (NDG) aims to provide a framework for the discovery and use of data needed in NERC research. To support this, the NDG has developed a metadata and data model. The data model is directly concerned with the representation and use of the data, with the metadata model covering the discovery and higher-level considerations.

The metadata model has started by considering the data holdings of the BADC and BODC, but will be extended as data from new disciplines is incorporated. To provide a framework for this extension, five distinct metadata objects have been identified as core to the NDG, providing the ability to support basic relationships between these core classes, while using sub-classing to provide specialisations where required.

The schema will link with ontology systems, and will provide ways to express security policies, once these are established.

Keywords: Metadata, ISO19115, OGC, TC211, NDG

1. INTRODUCTION

NERC has a very 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 specialties. These data holdings are spread across a wide range of 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.

The data holdings are extremely diverse, including as they do not only model data, but also observations and data derived from both observation and model measurements.

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. As a by-product it is hoped that it will then be easier for scientists to contribute to and help maintain managed data holdings.

Key requirements are that the NDG should:

- Allow discovery and access of data without having to have a priori knowledge of details of storage characteristics, values or parameters;
- Be discipline specific, but provide functionality for users beyond that community;
- Allow discovery and access of relevant data by science beyond the discipline for which it was collected;
- Hide the heterogeneity of the data sources being queried, and combine the results into a single, consistent, result set;
- Allow the specification of pre-presentation processing, such as sub-querying, transformation, and consolidation, particularly where the data may be spread across several data sources;
- Deliver data to the desired place in the desired format;
- Optionally allow (limited) server-side processing of the data.

Given that the NDG is going to be built on pre-existing data holdings, with pre-existing metadata structures, the NDG will need to provide mechanisms to query metadata about the datasets and collate results, along with the means to declare profiles, into which a data holding can map its local schema to allow cross-holding queries.

It is intended to do this by providing a decoupled data and metadata infrastructure that will bring together developed versions of tools that either already exist or are under development within e-science-or the worldwide earth science community. Initially, Atmospheric and Oceanographic data held in the BODC and BADC will be made available, with data from other disciplines funded by NERC being added in due course.

In this paper we describe the development of the NDG metadata model, itself one component of the overall metadata environment.

2. OVERVIEW OF NDG METADATA

The overall NDG metadata environment is described in [1]. In brief, the key elements of the data metadata include (but are not limited to):

A [Archive] format and usage metadata.

B [Browse] superset of discovery, usage data, and contextual metadata.

C [Comment] annotations, documentation and other supporting material.

D [Discovery] metadata, used to locate datasets.

Type B is a superset of the Discovery metadata. In the NDG, Discovery metadata will initially consist of NASA Global Change Master Directory (GCMD) Directory Interchange Format (DIF) records [2]. However, a key tenet of the design philosophy is that NDG discovery will also support discovery of NDG holdings using the Dublin Core [3], the CCLRC scientific metadata format [4], and probably the GEO profile of Z39.50 [5] and Catalogue Interoperability protocol (CIP) records [6]. Other discovery protocols will also be supported where possible.

The key types are the "Type A" metadata, which is directly concerned with the use of the data, and the "Type B" core metadata. As explained in [1], we have implemented these as two different schema, the data model (type A, discussed in [7]) and the metadata model discussed here.

This categorisation has brought benefit by giving a clear split between discovery and use. Many disciplines have widely used, almost standard, data formats. Separation allows the discovery metadata model to be plugged into different data models in a manner that means that the underlying data model is transparent to the user. It also means that each model can tune the detail kept in it to that necessary to perform its task. For example, the data model must keep track of the actual data values and sufficient information to deliver the data to the user, if necessary transforming it from the original format to another, whereas the metadata model needs only a summary of the data values, but must hold detail of how and why the data was gathered. Thus, some data values are kept in both the data and metadata models, but their intended usages are very different.

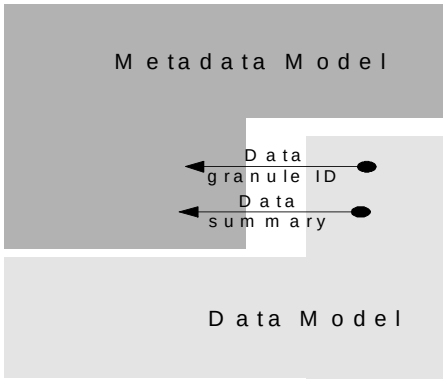

Figure 1 – Linking the data and metadata models

An ID generated by the data model links the data and metadata models. Once the data of interest is identified, by searching the metadata, the IDs of the data granules are passed to data browsing software for the user to identify and possibly process the

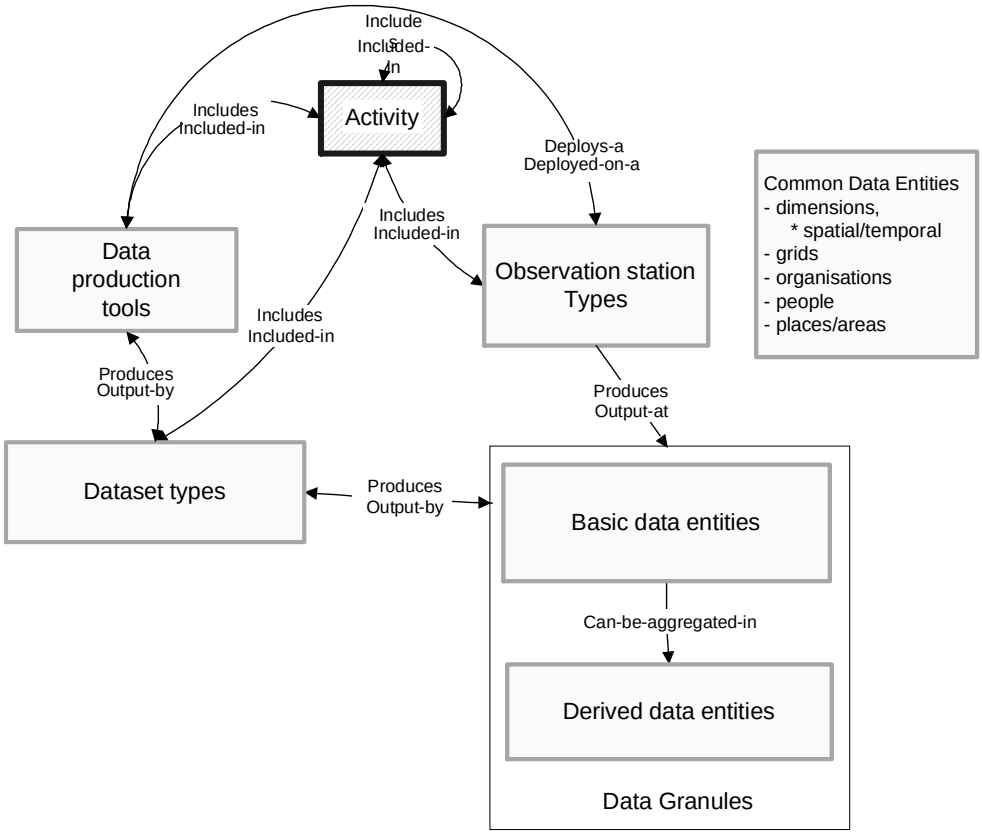

Figure 2 – Core entities and relations

actual portion(s) of data of interest. This process could include subsetting and aggregation of the data, in some cases producing new data granules that will be registered in the NDG, in others the result will be a temporary data set that will be discarded after use.

3. METADATA AND STANDARDS

The development of the NDG metadata and data models is being carried out in a standards compliant environment. In particular, the International Organization for Standardization (ISO) technical committee 211 and the 191xx series of standards are considerably influencing the development of both. This series includes more than 35 new or nearly released standards [8],[9] which cover geographic data, metadata and services.

The basic NDG architecture is consistent with the ISO domain reference model, and we aim to register the NDG metadata schema itself, or a subset, as an ISO19115 profile. Development of our metadata schemas and software will conform as far as possible as the standards are released. We are already making use of the standards published relating to spatial and temporal reference systems, metadata, data quality, and conformance requirements [9].

Closely allied with the ISO work is the industry-based OpenGIS Consortium (OGC). Consisting of over 250 companies, government bodies and HEIs, it is concerned with developing standards for interoperable commercial geographic information systems. As such, OGC both influences and draws upon ISO work. OGC has developed web-service specifications for rendering and retrieval of geographic data [10]. A number of vendors are now supporting these specifications in commercial off-the-shelf products. Our development is being influenced by an

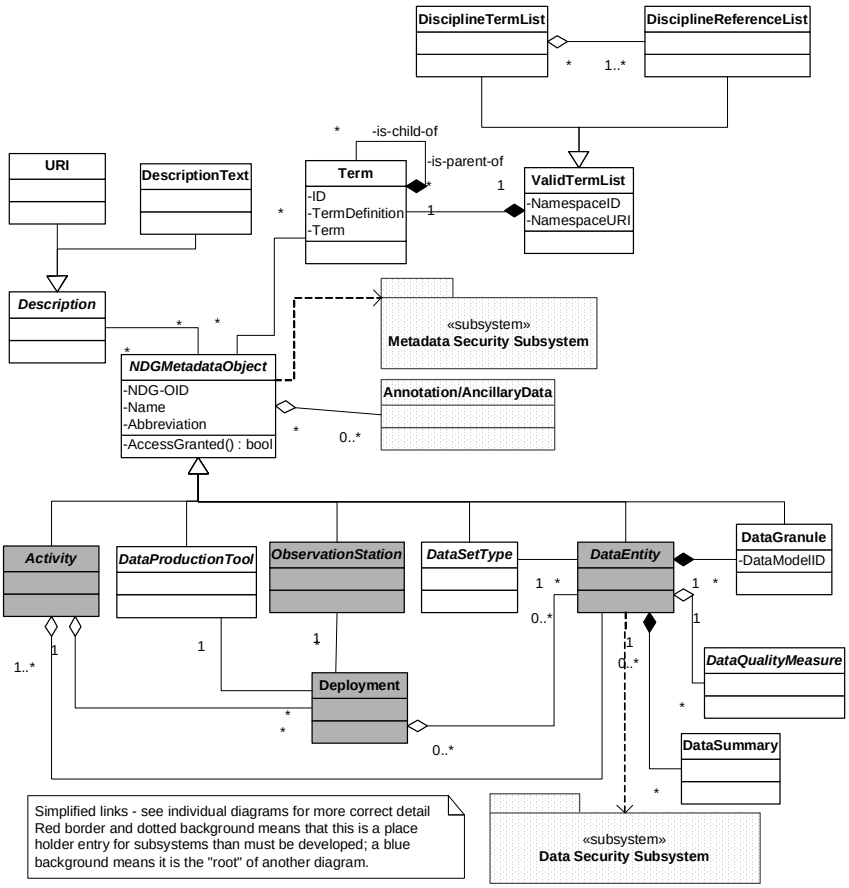

Figure 3 – Representation of top-level metadata model

objective to ensure our data delivery systems are OGC compatible.

Both the NDG metadata and data models have been constructed first as conceptual models in the Unified Modelling Language (UML), and then transformed into an XML schema. This procedure is compliant with draft ISO specifications on conceptual modelling, application schema and encoding [8].

In the terminology of the ISO TC211 series of standards, the NDG Data Model represents an application schema.

4. METADATA MODEL CONCEPTUAL OVERVIEW

Much effort has been spent in determining the key entities that must be explicitly represented. Major difficulties have been the need to keep separate things that, to a non-domain scientist, appear to be the same, and to avoid the use of terms that carry different meanings in different disciplines. The latter in particular caused much thought, as some of the original assumptions about discipline-neutral vocabularies were demolished. The emphasis has been to encompass the existing situation rather than propose a neutral format and migrate the communities towards it.

4.1 Relations represented

A keynote of the relations required is that there is always one more. The ones represented in Figure 2 are a starter set, based on those felt to be fundamental to the NDG from the start. It is recognised that the list will grow, almost arbitrarily, as time goes on. However, this list of entities forms the basis of what can be supported in

the query schema (Q, see [1]), and so it is necessary to have a defined list to lay the foundation for the eventual addition of intelligent search mechanisms.

4.2 Entities

The key premise in the development of the NDG metadata model is that it must support discovery and data identification (including differentiation between similar data sets). A list of typical queries made on the holding of the existing data centres has been used to guide the development of our entity list, and domain scientists have driven the process of analysis and categorisation of test cases. By combining these entities with the appropriate relations, it will be possible to represent the inter-linked hierarchies that characterise the data holdings.

It has also been necessary to produce a broad and extensible framework, capable of representing the diversity to be found within the purview of NERC-related projects.

The principal entities identified are:

1) Activities: these range in scope from entire programmes of work down to individual data gathering exercises, such as flights or cruises.

2) Data Production Tools: these are broadly classified into Instruments and Models.

3) Observation stations: these include permanent establishments, temporary moorings, or a moving platform such as a ship, aircraft, or satellite.

4) Data Entities: from the point of view of the metadata, these are arbitrary objects defined as a single data granule by the data model. They can and will range from individual measurements, to profiles, sections, or Lagrangian paths, through to complete data sets.

   There are two categories: "basic", whether produced by observation or computation; and "derived" which are the result of processing from basic data objects, to produce such as climatologies, time series, and integrations.

5) Dataset types: these are measurements from instruments, simulations produced by model runs; and analyses that are combinations of measurement and simulation data, such as weather forecasts.

In addition, there are common data entities. These entities recurred in relation to all the entities identified above, but are not primary scientific objects. They include units, people, organisations, and places.

It is expected that there will be considerable extension and refinement of the various entities, with some developing towards being taxonomies in their own right.

5. MOVING TOWARDS IMPLEMENTATION

5.1 Supporting inter-disciplinary searching

A major reason for the NDG is the need to access and integrate data across disciplinary boundaries. This implies that the scientist use the terms familiar to **them**. Many

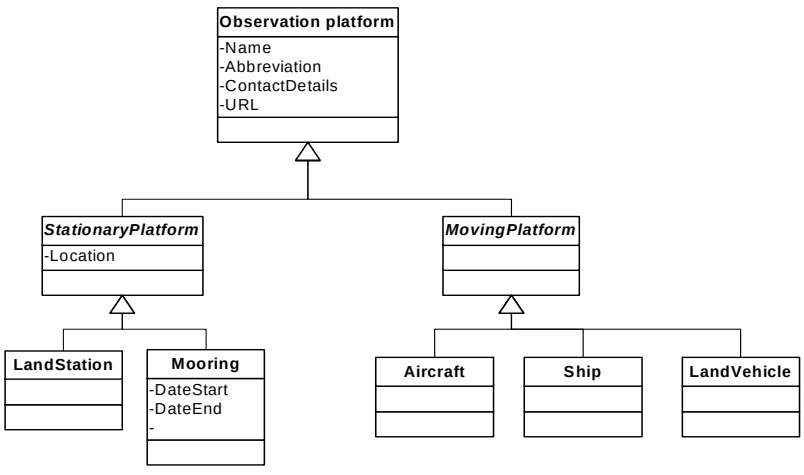

Figure 4 – Observation Stations

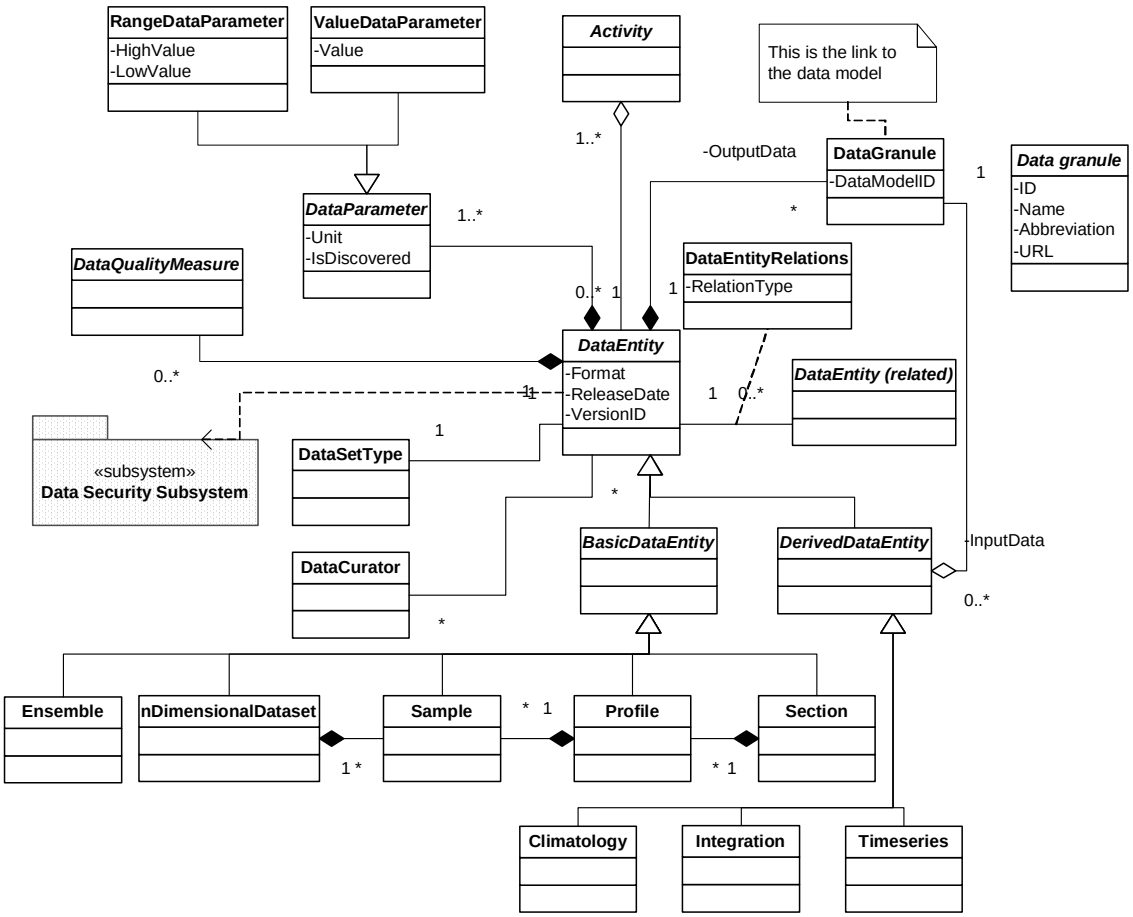

Figure 5 - The Data Entity

controlled vocabularies exist, but these often disagree over basics such as the definition of geographic area. Also, there can be a number of informal dialects, so that the meanings of terms may not be the same as in the parent vocabulary.

At first, this will be handled by indexing the metadata against specific existing structured vocabularies from relevant disciplines. Unfortunately, this is likely to mean that a single metadata record will initially be catalogued several times, against different vocabularies. In the longer run, a project will be spun-off to investigate the feasibility of creating and maintaining a NDG "reference" ontology, into which we can map different "industry-standard" vocabularies etc. This will not be easy, as shown by the fact that two disciplines that appear closely allied cannot agree on the meaning of the word "westerly"…

5.2 Elaboration and development of the basic concepts

Having identified the essential concepts, the next stage was to "sketch" the metadata in UML to allow scrutiny by a wider audience and to provide an implementable set of classes. This is not a complete schema, and is unlikely to ever be complete (unless scientists stop thinking of new concepts). It is meant to provide sufficient detail to start populating and an extensible framework that will be refined and added to.

Figure 3 shows the top-level of the metadata. Some sub-systems such as security are indicated, but these should be regarded as place-holders.

As example of how these entities are developing, a UML diagram of observation stations is included (Figure 4). The entities have reached the stage at which the domain

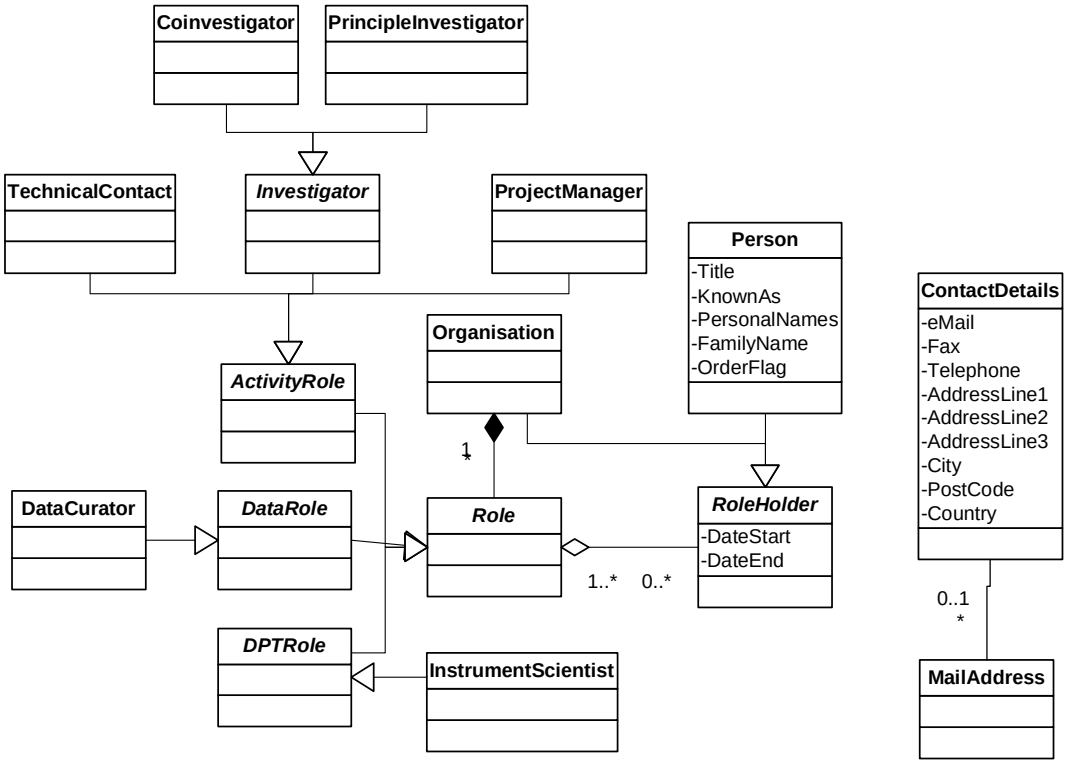

Figure 6 – Roles, organisations, and people

scientists are looking at them in co-operation with the data scientists to ensure that all significant "leaf nodes" are represented. The queries range from the general ("I want measurements from stationary platforms in this area") to far more particular examples ("I want data for the area in this time frame that has been captured by a Dobson spectrophotometer"). These examples imply that various instances are organised into a variety of overlapping hierarchies, and that these will have to be developed. In addition, each leaf node itself will have its own characteristics and attributes.

5.2.1 The Data Entity

The core of the schema is the data; Figure 5 describes the data entity. The data parameter will be a summarised form of the data held in the data model [7]. For example, multiple values may be reduced to a single entry giving the unit and the range of values. To allow these to be searched, we are investigating the feasibility of converting to and from a reference set of units, to allow the user interface to use units familiar to the user, yet bring back data held in other units. The UDUNITS library [11] is providing the starting point for this activity, but its set of supported units will have to be extended.

Again, it is expected that the types of data entity will proliferate.

5.2.2 Roles, people and organisations

The metadata requires a proper model of people and their roles, not only to help users of the data, but eventually to help mediate modification and annotation of metadata, and deliver access control with implied and inherited rights and responsibilities. Although not yet used for access control, Figure 6 displays our structure for this.

5.3 Presentation to external applications

In this schema, there are four distinct types of metadata record, "dataset type" being

implemented as an attribute of the data entity. All records carry a unique ID to allow inter-record referencing and joining.

The metadata will be made available using XML syntax according to a XML schema that has been developed from the intermediate UML diagrams. The XML will be generated from the existing cataloguing systems, and either generated "on-the-fly" or harvested and cached elsewhere, depending on the practicalities as discussed briefly in [1]

6. SUMMARY

The NDG metadata model is one component of the NDG metadata hierarchy, but will lie at the heart of inter-disciplinary data discovery. As such, it is unlikely ever to be a finished work, and so development is being driven by standards compliance and extensibility. The initial development is nearly complete, and is finding acceptance by domain experts. As the initial definition phase concludes, the focus is likely to move towards interoperability with other activities.

REFERENCES

[1] Lawrence, B.N., et. al., 2003: The NERC DataGrid Prototype, UK e-Science All Hands Meeting, 2003.

[2] Directory Interchange Format Writer's Guide, Version 8, http://gcmd.gsfc.nasa.gov/User/difguide/difman.html, Sept. 2001

[3] Dublin Core Metadata Element Set, Version 1.1: Reference Description - http://www.dublincore.org/documents/dces/, June 2002

[4] Matthews BM, Sufi SA. The CCLRC Scientific Metadata Model - Version 1, 2002.

[5] Z39.50 Application Profile for Geospatial Metadata or "GEO" version 2.2, http://www.blueangeltech.com/standards/GeoProfile/geo22.htm, May 2002.

[6] The Catalogue Interoperabilty Protocol.

http://www.dfd.dlr.de/ftp/pub/CIP_documents/ics/local_attributes_tn.pdf

[7] Woolf, A. et. al., 2003: Data virtualisation in the NERC DataGrid, UK e-science All Hands Meeting, 2003.

[8] ISO TC 211, outstanding work programme as of 2003-08-02 - http://www.isotc211.org/pow.htm

[9] ISO TC 211, publications as of 2003-05-13 - http://www.isotc211.org/publications.htm

[10] Open GIS Consortium Inc. (website), http://www.opengis.org/

[11] UDUNITS, http://my.unidata.ucar.edu/content/software/udunits/index.html (online), University Corporation for Atmospheric Research, Boulder, Colorado, USA.