THE NERC DATAGRID PROTOTYPE

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

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

Abstract

The NDG project began in September 2002. The aim of the project is to utilise the promise of grid technologies to allow members of a virtual organisation (the NERC community) to share environmental data held in disparate institutions. The NDG virtual organisation will begin with elements of the oceanographic and atmospheric community, but aims to expand to support science requirements of data sharing for earth system science.

At the timing of writing, the project has been running one year, and we have spent much of that time addressing design issues. In doing this we have tried to both remain standards compliant and to make the best use of existing knowledge and tools, particularly in the metadata arena. However, in the process we have found it necessary to design and implement our own data model and metadata schema and with these we have begun to build a simple prototype system. The process has identified a number of key challenges for the NERC DataGrid development. Two such challenges are access control and metadata acquisition.

In terms of access control, while a technical solution is not yet apparent, the real problems are social: getting data suppliers to trust grid-based access control, and getting them to describe their access control requirements. In terms of metadata acquisition, the holy grail of software agent mediated services requires a level of metadata that is difficult to generate automatically. A brief taxonomy of metadata is presented here, along with some of the issues involved in obtaining it. Two further papers in this volume describe the data and metadata models in more detail.

1. INTRODUCTION

There are three main constituencies in the NERC community: staff of the NERC institutes, staff at cooperative centres and institutes (not NERC employees, but where the majority are NERC funded), and individuals and their research groups in other institutions (primarily the universities). These three constituencies themselves consist of disparate groups whose identity (if it exists) is primarily discipline based (for example, oceanography, atmospheric science, geology). However, in recent years there has been a strong and consistent move across all disciplines to a more coherent, interdisciplinary view that all are inter-related as parts of "Earth System Science".

NERC science has always been based on a strong tradition of collecting and utilising observational data, both in its own right and in conjunction with simulations (themselves often involving vast quantities of data). Accordingly, NERC has long had a data policy [1], which has meant that NERC funded scientists are obligated to archive data in one of seven NERC designated data centres at an appropriate stage in the evolution of their work. Each of these data centres operates autonomously, and they implement different access policies (ranging from open internet based access to charged media delivery). In most cases, different access policies are implemented on a dataset by dataset basis within one data centre.

The data held within the designated data centres and the wider community are in a wide variety of storage formats, ranging from flat files to managed Oracle database systems, and are accompanied by varying amounts of metadata.

Earth system science requires the inter-comparison of datasets, but invariably when more than one dataset is being used, a sequence of complex steps are repeated for each dataset using different software tools

for each dataset and it is only at the processing and display step (if at all) that the digital data are combined for comparison. Users often need significant amounts of experience and data comparison inevitably requires manual migration of data from one place to another (often involving several user authentications at different sites). Large amounts of trained scientist time is spent on "reinventing wheels" in order to handle data, as each user needs to go through the process of learning about formats and/or SQL table structures etc before they can use the data scientifically.

The problem is compounded by the paucity of, and disparity between, discovery metadata formats in the different communities. Much observational data is simply not used because of the overheads of finding and handling it. The aim of the NERC DataGrid is to build a grid that makes data discovery, delivery and use much easier than it is now. Further we intend to make the connection between data held in the NERC managed archives and data held by individual research groups seamless in such a way that the same tools can be used to compare and manipulate data from both sources.

In the first instance, the NERC DataGrid will be built on the managed archives held in the British Atmospheric and Oceanographic data centres. Thus our prototype architecture is based on the necessity to provide seamless access to a carefully selected subset of data held in the two institutions.

This short paper outlines the key architecture behind our prototype development, our strategy, and identifies two main challenges to achieving our goals. In summary, it will be seen that our architecture depends on modularity coupled with two crucial underlying pieces of technology: a data model, and a metadata model. Two further papers in this collection describe these in more detail [2,3].

2. DATA METADATA TAXONOMY

In order to deliver the desired functionality, there is clearly a large amount of metadata that needs to be captured. In the desired grid, we will clearly need to categorise the data itself, the people who use it along with their roles, and the capabilities of the software.

In terms of the prototype we have concentrated on the data metadata. While there are many existing taxonomies of data metadata, we have found it useful to concentrate on a view of metadata which is based on where the metadata comes from, and its usage, rather than on usage alone.

When this is done, we find five major categories of metadata:

A. [Archive] Usage metadata, normally generated with the data, and always accompanying it. (For example, the metadata held in a NetCDF formatted file).

B. [Browse] Complete set of metadata which covers both semantics, and syntax, and includes both discovery metadata (D) and discipline specific metadata (E). This metadata is often built up over time.

C. [Comment] Ancillary metadata, such as annotations and publications. Usually provided after the ingestion of data into an archive.

D. [Discovery] The metadata needed to find datasets of interest. Usually produced by managed data centres.

E. [Extra] Discipline specific metadata that may exist as "institutional wisdom" within the community of (original) data users. Often fails to accompany the data on as it travels.

In practice, it is rare to find a dataset that has a complete set of this metadata, and yet all this metadata is needed to both produce a grid that allows the automation of data extraction and usage, along with meaningful interpretation by non-discipline specific experts.

With the possible exception of attempting inter-disciplinary work, this is not a new problem. In fact, this is the main reason why the NERC designated data centres exist. However, no data centre that we are aware of (world-wide!) deals with C well, and generally most produce B (including D and sometimes E) by hand. In practice there is no easy way of automating the production of these metadata, and as yet, no incentive for the data producer (and original users) to produce such metadata. Thus, the actual process of data ingestion into the grid will involve real people at least for the set-up, for the foreseeable future.

To build the NDG we need to encompass this information in a machine-readable, machine understandable way, which also allows updates by suppliers and maintainers and annotation by users. It also needs to be done in a discipline independent manner, and in a way that it can easily allow expansion as new disciplines join the NDG. We have found no existing metadata schema(s) that do this for all the categories of information required, and so we have chosen to implement our own. However, one major aim is to be able to export metadata in a variety of existing and familiar formats.

There is plenty of experience in the community that shows the benefit of a clear distinction between "discovery" metadata and "usage" metadata. We have followed that concept, by defining a metadata model [2] and a data model [3] based on XML schema. Broadly these correspond to B (without E) and the A categories above. However, we have made the clear assumption that these are linked and that there is an element in common. To reinforce that, we have further defined:

S [Summary] The overlap between D and A type metadata. We will be generating S automatically from the A type metadata.

We assume that the procedure of data acquisition can include three clear stages: dataset discovery (mediated by D), data browsing (mediated by B and S), and data extraction (mediated by A).

For an eventual goal of allowing software agents to proceed via this sequence, we need to define clearly what sort of queries and usage our metadata will need to support. To that end, we will eventually need to characterise:

Q [Query] The complete set of query types that we expect upon our metadata.

In the longer term we will also be supporting the concept of "server-side" processing, which itself will require metadata description.

3. ACCESS CONTROL

In the process of scoping the requirements of our metadata, it has become apparent that the full NDG will have complex access control requirements that will need to support control over the metadata records and the data independently – some users will have access to both, some to the data, and some to the metadata. Examples of the latter two are the requirement for users to see metadata before purchasing data, and the requirement to allow users to access land-use data without seeing actual geographical metadata for privacy reasons.

In the full NDG we must then support all these criteria both within the metadata and the software. However, an analysis of the available grid software suggested that we should wait before trying to implement anything sophisticated. Further, just trying to understand the existing access control policies at the BADC and BODC was a larger job than anticipated. Access control to some datasets is mandated by human decisions, and it will be difficult to define and codify these existing arrangements.

NDG planning has also had to deal with considerable resistance to trust in grid-based authentication. Like all new tools, familiarity is required before trust can be established.

Accordingly, the initial implementation of the NDG will simply authenticate users and compare them to an authorisation list, but not implement access control beyond that.

4. PROTOTYPE ARCHITECTURE

The NDG prototype will provide a discovery service, a browse service, and a file delivery service. In the first instance these will be based on web services, but we intend to migrate to GT3 as soon as practicable.

The prototype architecture envisages the situation where there are **n** (initially 2) data archives each of which provide two services: a data delivery service and a metadata service. There will also be a portal discovery service not necessarily located with any data.

The metadata which will underlie the metadata service is described in more detail in [2], but the key aspect for grid operation is that it allows both the remote harvesting of D metadata, and the local browsing of B metadata. It is a key facet of the design philosophy that the metadata service will support the harvesting of a variety of D metadata formats.

As the initial step, the schema is being mapped into the data holdings of the BADC and BODC. Once this mapping is complete, XSLT scripts will be produced to map the metadata into NASA Global Change Master Directory (GCMD) Directory Interchange Format files (DIFs) [4] as our initial D metadata description. This will allow the data to be discovered using tools already existing within the Atmospheric and Oceanographic disciplines. At this stage, both BADC and BODC will be able to publish DIFs for their holdings, which we can then harvest to our prototype portal to allow searching and querying. It is a design criterion that it should be possible for these records to be harvestable by other engines as well.

Once we have automated DIF production from our metadata, mappings to other discovery formats will be made (for example to the GEO profile of Z39.50 [5], allowing support for the existing NERC Metadata Gateway [6]). This will deal with immediate requirements. However, in the long run, it is intended to expose the entire structure of the NDG schemas for query and develop the appropriate search tools as the schemas are more expressive than alternatives which were either produced for a very specific area, e.g. model data or were intended to be globally applicable for discovery of datasets alone. (At some point it should also be possible for disciplines to construct search engines that can yield both our B metadata and their own E metadata).

The output from the searches will be a set of Data Model IDs, which will be handed over to the data processing software. This will invoke the data delivery service mentioned earlier. This service supports both browsing of the A metadata for the desired dataset (including data location information), and retrieval of a specified portion of the data in a format required. Alternatively, the service supports retrieving the raw data files themselves, in their original format. Visualisation of the data will be enabled through a dynamically generated Live Access Server (LAS) [7], or the Climate Data Analysis Tools (CDAT, [8]) software.

5. SUMMARY

The NDG is in the process of developing an initial prototype that will exploit the newly developed data and metadata models and inform future migration from web services to grid-services.

REFERENCES

[1] NERC Data Policy Handbook http://www.nerc.ac.uk/data/documents/datahandbook.pdf

[2] O'Neill, K.D., et.. al., 2003: The metadata model of the NERC DataGrid, UK e-Science All Hands Meeting, 2003.

[3] Woolf, A., et. al., 2003: Data virtualisation in the NERC DataGrid, *ibid*.

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

[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 NERC Metadata Gateway.

http://www.nmp.rl.ac.uk

[7] The Live Access Server

http://ferret.pmel.noaa.gov/Ferret/LAS/ferret_LAS.html

[8] Climate Data Analysis Tools

http://esg.llnl.gov/cdat