The NERC Datagrid: Enabling Interoperable Climate Data Resources

Bryan Lawrence
Marta Gutierrez
Sue Latham
Ag Stephens
British Atmospheric Data Centre, UK

Ray Cramer
Siva Kondapalli
Roy Lowry
British Oceanographic Data Centre, UK

Neil Bennett
Kerstin Kleese van Dam
Kevin O'Neill
Andrew Woolf
CCLRC e-Science Centre, UK

Michael Burek
Don Middleton
National Center for Atmospheric Research, US

Dean Williams
Lawrence Livermore Laboratory US

Author Email:
b.n.lawrence@rl.ac.uk

Abstract

The NERC DataGrid (NDG) is a UK e-Science project that will provide discovery of, and virtualised access to, a wide variety of climate and earth-system science data. We present an overview of key elements of the NDG architecture from a number of perspectives, from an enterprise viewpoint, to the relationship to key ISO standards, and to the underlying metadata structures. The discussion stresses the interoperability characteristics of the design, and introduces the latest experiments of interoperability with the NCAR community data portal, and plans for wider interoperability with the climate data archives held in the Earth System Grid. Future initiatives, including the Big Data Analysis Network, are introduced.

1. Introduction

International climate science activities increasingly require international data management solutions. These must be built out of interoperating national activities. Major programmes, for example international climate modelling activities (e.g. IPCC [1] and climateprediction.net [2]), require access to and analysis of widely distributed, high volume datasets. Additionally, such projects involve verification studies, which compare model output with observational data from a range of sources and in a variety of formats. Many institutions archive and provide data in these and other contexts. While distributed paradigms are becoming prevalent, the distribution is often of relatively small granularity, and based on ad-hoc relationships; this is unlikely to change in the near future.

The NERC DataGrid (NDG) is a pilot project funded under the UK's e-Science programme to facilitate discovery, access and use of a range of environmental data across the UK and internationally. It aims to integrate a range of heterogeneous and distributed data resources, and provide a uniform interface accessible through desktop tools. The NDG is currently being built on data holdings from the British Atmospheric and Oceanographic Data Centres – BADC and BODC - and is expecting to expand to support other selected data centres (including the Southampton Oceanography Centre and the Plymouth Marine Laboratory). NDG has close links with other UK environmental grid projects such as the EcoGrid and the DEWS (Delivering Environmental Web Services). The NDG also liaises with a number of international partners:

- The Program for Climate Model Diagnosis and Intercomparison (PCMDI, Lawrence Livermore Laboratory, US)
- The Earth System Grid, ESG, US
- The MarineXML consortium, EU
- SEEGrid, Australia
- The Global Organisation for Earth System Science Portals (GO-ESSP)
- The World Data Centre for Climate (Germany)

The remainder of this paper is structured as follows: Section Two introduces some of the issues in dealing with distributed data, while Section Three provides an overview of the NDG architecture and Section Four addresses interoperability. A summary concludes the paper.

2. Generic Data Exploitation Issues

A consumer of such climate products is faced with three issues:

- **Data discovery**. Locating suitable data is the first step in exploitation. Data must be catalogued, and catalogues searchable in a manner that maximises dataset exposure. Federating catalogues - either through harvesting [3] or distributed searches [4] is an essential element. Effective federation relies on standard metadata schemas against which to query.
- **Data access**. Metadata typically provides a pointer to the location of data - either physical or digital. A consumer of climate data must then obtain access to it. Just as a library delivers a book to a remote customer through the postal service, so a network service is used to deliver digital data. And, as in the postal world, different classes of service are appropriate for different types of data product - some specialised to high-speed delivery, some to large volume; some provide value-added functionality (e.g. offering an extract of a book or dataset). Access restrictions may require proof of identity for delivery.
- **Data use**. Having discovered and obtained access to data, a consumer has to be able to use it. The large range of data formats in use raise a formidable challenge for climate data. An emerging paradigm focuses on characterising the semantics and structure of data (through formal data models) rather than file formats – 'content' rather than 'container'. Conventions may be established [5] for serialisation into different file formats, but primacy lies with the *a-priori* data model.

Developing interoperable systems which handle distributed data requires thus addressing all three issues in such a way that distributed systems understand the catalogues, metadata structures, and data models.

3. Overall Architecture

Figure 1 presents a schematic overview of the NDG architecture that was developed in response to the issues outline in Section Two.

A number of *policies* govern NDG activities on matters such as security (authentication, authorisation and accounting), resource usage, and quality of service. As a consequence, the following *roles* are identified:

- **User**: participates in a broad range of NDG activities related to discovery and access of data

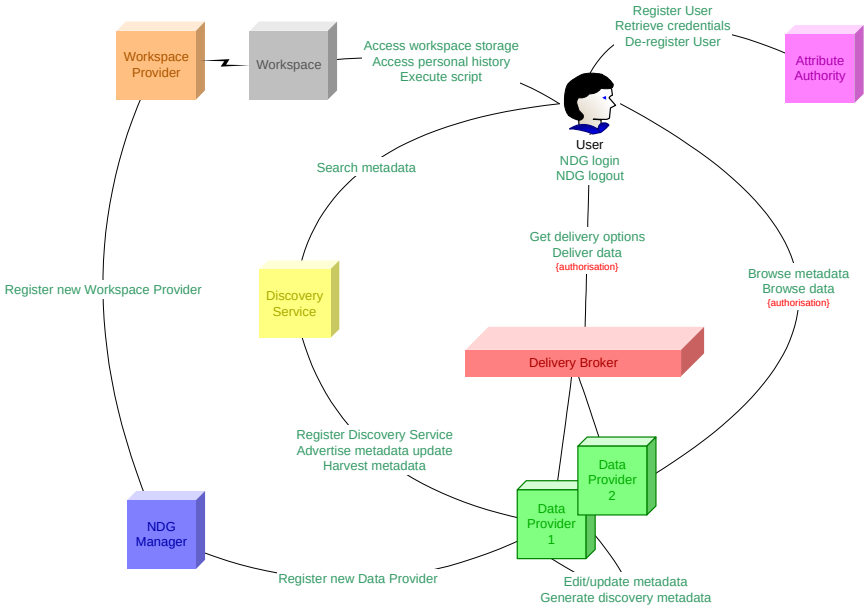

Figure 1: Key components of the NDG Architecture

- **Data Provider**: supplies data and metadata through NDG infrastructure
- **Discovery Service**: allows searching across metadata harvested from Data Providers
- **Delivery Broker**: mediates and fulfils requests for data delivery across one or more Data Providers
- **Attribute Authority**: assigns security attributes to Users to control access to data and metadata.
- **NDG Manager**: maintains registries of Data Providers, Discovery Services etc
- **NDG Workspace**: provides logging, storage, and scripting facilities for Users
- **NDG Workspace Provider**: supplies resources used for Workspaces

The spectrum of *activities* within the NDG includes:

- **Search and discovery**: searching over discovery metadata and browsing of detailed metadata
- **Data browse and delivery**: browsing of dataset structures, selection of data subsets, and delivery.
- **Workspace management**: provision of resources for a Workspace and interaction of a User with Workspace facilities
- **Metadata management**: updating of metadata and datasets, and harvesting of discovery metadata; registration of Data Providers and Discovery Services
- **User administration**: logging in and out of NDG, and assignment and retrieval of security credentials

While Figure 1 describes the NDG architecture, the main concepts are generic, and any system providing distributed access to data would have to provide similar components (with the possible exception of the three components explicitly labeled as NDG: the workspace

provider, the manager, and the workspace). From an interoperability point of view, one has a number of points at which interoperability needs to be considered: at discovery, security, browse, and delivery. It may not be necessary for all points to be interoperable with all groups, as differing levels of interoperability may make political, funding, or technical sense.

In addition to the overall viewpoint, one needs to consider the information environment supporting these components. Leaving aside the security information structures for the moment, the NDG has chosen to follow the 'Domain Reference Model' [6] of ISO Technical Committee 211 for geographic information, shown in Figure 2. This sets out a high-level view of the structures required for interoperability of geographic information (whether distributed or not). It underlies an entire series of emerging standards for geographic data, metadata and services (the 'ISO 19xxx' series of standards). Conformance to the abstract Domain Reference Model is a prerequisite for standards-compliant distributed data infrastructures.

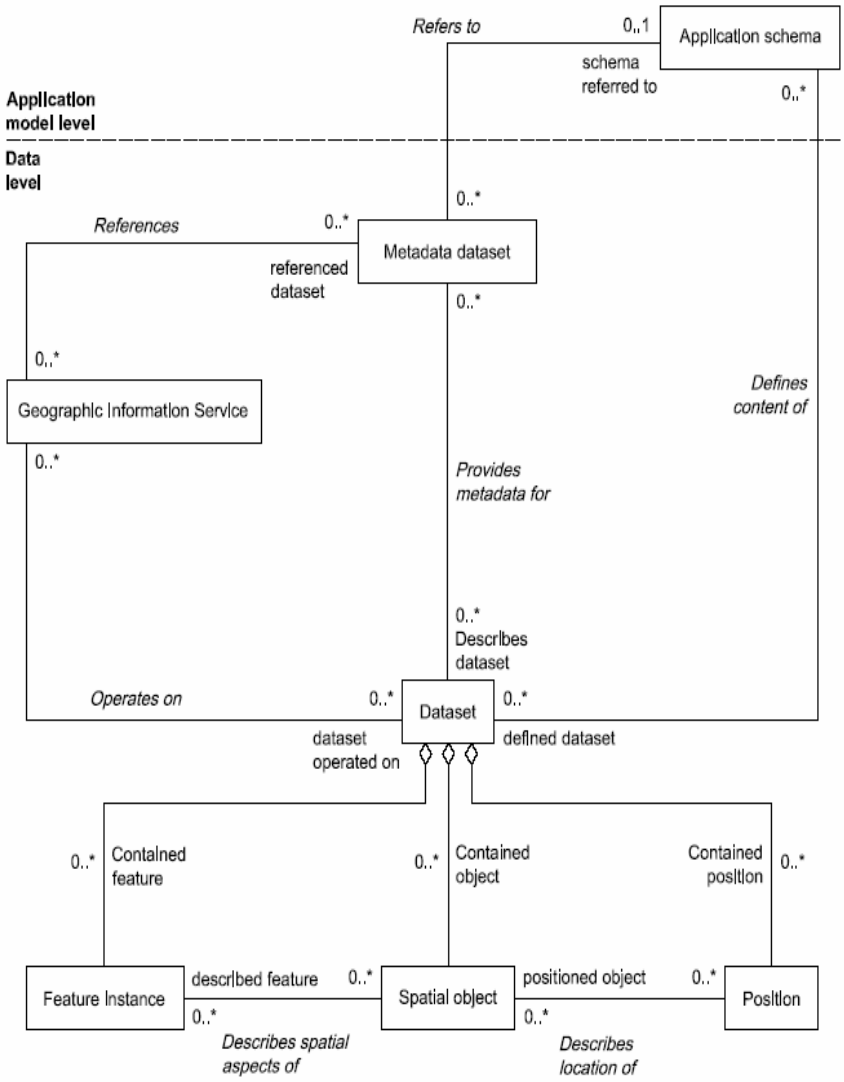

Figure 2: ISO TC211 Domain Reference Model for Geographic Information

Core to the Domain Reference Model is a geospatial *Dataset*. Logical datasets in the NDG are configured by individual Data Providers. They may be based around defined activities (e.g. the 'ECMWF ERA-40' reanalysis or the 'ACSOE' campaign measurements), or particular instruments (e.g. measurements from the 'Chilbolton radar'), or in any other manner deemed appropriate by the Data Provider. The question of what constitutes a dataset is an important one, and usually non-trivial. A dataset contains *Feature instances* and related objects (a feature is a semantically meaningful abstraction of a specific data type – for instance a snapshot of a simulated model field, or an atmospheric temperature sounding).

The logical structure and semantic content of a dataset is described by an *Application schema* [7] that defines the feature types that may appear. These are defined for the NDG in a data model named the Climate Science Modelling Language (CSML) [8]. The current version of CSML can be thought of as three components:

- Extensions to the Geography Markup Language (GML), to support MetOcean data.
- The application schema itself, which describes seven basic feature types: trajectories, points, profiles, grids, pointseries, profileseries and gridseries.

- Software which interfaces to the actual data in storage, whether in files or databases. The earliest versions of CSML support NetCDF and NasaAmes, but later versions are expected to support a wider variety (Grib, Met Office PP, HDF) via a further markup language abstraction layer exploiting CDAT/CDML [9] and perhaps NCML [10].

A *Metadata dataset* provides metadata related to a dataset. It includes necessary information to support access to, and transfer of, the dataset. A standardised schema [11] provides metadata element definitions for dataset description, topic, point of contact, quality, etc., and may include references to an application schema. Metadata in the NDG is generated as a by-product of a sophisticated conceptual meta-model (or ontology) for datasets and the relationships between them.

Discovery metadata is generated from the more complete metadata dataset, and conforms (or will) to international standards (such as ISO19115).

The relationship between CSML, the metadata dataset, and the discovery metadata is described in the NERC Metadata taxonomy ([12, 13], Figure 3), where they can be thought of as A,B and D metadata respectively. More generically, these taxonomic classes are:

- [**A**]rchive: Usage metadata generated from (or about) the data. Normally generated directly from internal metadata. This is equivalent to the 'Application schema' of the ISO Domain Reference Model discussed above
- [**B**]rowse: Generic complete metadata, semantic, including summary of syntactic (S), not including discipline-specific (E).
- [**C**]omment: Metadata generated to describe both documentations and annotations (as opposed to binary data).
- [**D**]iscovery: Metadata suitable for harvesting; the 'Metadata dataset' of the ISO Domain Reference Model.
- [**E**]xtra: Additional metadata, discipline-specific.
- [**S**]ummary: Summary metadata (overlap between A and D).

Finally, the ISO Domain Reference Model includes *Geographic Information Services* that operate on the dataset [14]. A broad taxonomy of services is imagined – for human interaction (e.g. viewing data), information and model management, workflow, general geo-processing, annotation, transfer etc. The NDG is being built fundamentally as a service-oriented architecture to provide a basis for service chaining. Initial delivery services do not conform to any standard, de facto or otherwise, but future services are envisaged based on the Open Geospatial Consortium web services as well as the Web Services Resource Framework (WSRF, [15]).

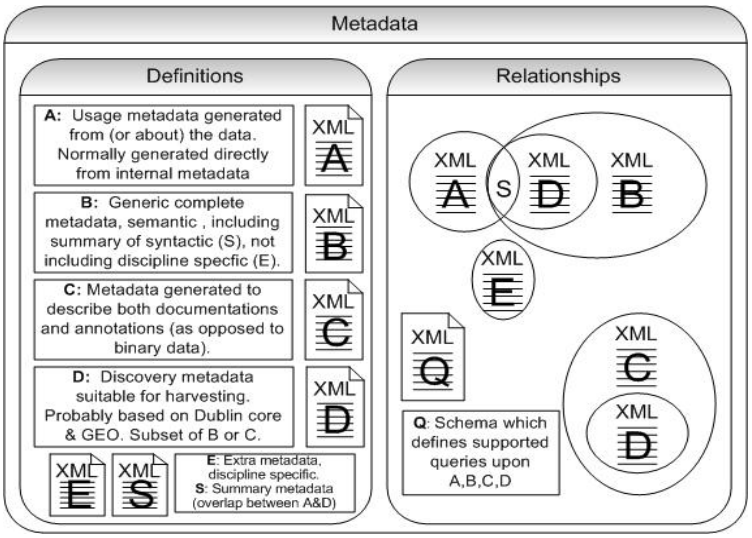

Figure 3. NDG Metadata taxonomy

4. Interoperability

As outlined above, interoperability can be achieved on a range of scales. Even within the NDG, we expect two classes of data providers; those who initially can only provide discovery metadata coupled with archives of data that are not accessible via the NDG delivery and security paradigms, and those that are "fully NDG-enabled". The NDG itself expects to interoperate with other groups, as witnessed by the international collaborations.

The first level of interoperation is discovery, and discovery interoperability has been established between the NCAR Community Data Portal and the NDG. Although the A and B metadata structures are different at each site, both can deliver D in NASA GCMD DIF format, and both are working to ISO19115. At the time of writing, demonstration of metadata harvesting from NCAR to NDG and vice-versa has been completed, and live discovery to respective archives is expected soon.

At NCAR, the main B catalogues are Thredds [16] catalogues, generated from A metadata which is NCML. In the NDG, B catalogues conforming to the NDG metadata schema are generated in different ways at the different data providers (at the BODC, the B metadata is generated from more complete metadata holdings in an Oracle database, at the BADC, the B documents are currently manually created, with input from the BADC metadata catalogue). Migration to the additional use of CSML is underway, but CSML itself was only fully

defined in January 2005, so there is much work to do to exploit the flexibility it will allow.

Within the next calendar year or so, experiments addressing security interoperability will also be undertaken. We will explore interoperation between the ESG security paradigm underlying the IPCC data store, and the NDG security paradigm ([17], itself yet to be widely implemented beyond demonstrators). Both use RFC3820 proxy certificates for authentication, but NDG and ESG have very different authorisation systems. However, it will be necessary to find a way of addressing such issues, as without doing so, it will not be possible to attempt true interoperation of delivery services (currently one would have to stay completely within one or other paradigm for any given transaction).

5. Summary & Future Work

The NDG project has been running for nearly three years, and has undertaken a comprehensive analysis of the architectures required for interoperable data archives in the climate community. Not all components of the NDG architecture have yet been implemented, but the initial deployment of a discovery service within the NDG has been widened to interoperate with the NCAR community data portal, and further international interoperability experiments are planned.

NDG is also deploying more comprehensive architectures to support transparent secure data manipulation, but this is at an earlier stage. Further, although not discussed here, initial development of tools based on CDAT/CDML [9,18] have resulted in the deployment of an "NDG data extractor" to make TB of reanalysis and climate data available to the NDG. Future activities based on these NDG tools will include the deployment of the Big Data Analysis Network (BDAN) for the NERC Centres of Atmospheric Science. BDAN will consist of large "storage brick systems" in a number of sites, each consisting of tens of TB of storage, containing primarily simulation data. Within each storage brick, data access will exploit a parallel virtual file system (e.g. TerraGrid [19] or similar) to provide significant bandwidth to commodity storage. However, the same generic interoperability issues will apply between the BDAN bricks, the NDG, and the wider world.

Acknowledgements

The NERC DataGrid is funded under the UK e-Science program through grant NER/T/S/2002/00091 from the Natural Environment Research Council.

References

1. Intergovernmental Panel on Climate Change, http://www.ipcc.ch/
2. Stainforth, D.A., *et.al*,. 2005, "Uncertainty in predictions of the climate response to rising levels of greenhouse gases", *Nature*, **433**, 403-406, 27 Jan 2005. See also http://www.climateprediction.net.
3. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
4. ISO 23950:1998, Information and documentation – Information retrieval (Z39.50) – Application service definition and protocol specification
5. NetCDF Climate and Forecast (CF) Metadata Convention, http://www.cgd.ucar.edu/cms/eaton/cf-metadata/.
6. ISO 19101, *Geographic information – Reference model*.
7. ISO 19109, *Geographic information – Rules for application schema*.
8. Woolf, A.,*et.al.*, 2005, "Climate Science Modelling Language: standards-based markup for metocean data", *Proceedings of 85th meeting of American Meteorological Society* http://ams.confex.com/ams/Annual2005/techprogram/paper_86955.htm
9. The Climate Data Analysis Tools (CDAT) and Climate Data Markup Language (CDML), http://esg.llnl.gov/cdat/cdms_html/cdms-6.htm
10. The NetCDF Markup Language, http://www.unidata.ucar.edu/packages/netcdf/ncml/
11. O'Neill, K., *et.al*., 2003, "The Metadata Model of the NERC DataGrid", *Proceedings of UK e-Science All Hands Meeting 2003*, ISBN 1-904425-11-9.
12. Lawrence, B.N., *et.al*., 2003, "The NERC DataGrid Prototype", *Proceedings of UK e-Science All Hands Meeting 2003*, ISBN 1-904425-11-9.
13. O'Neill, K., *et.al.*, 2004, "A specialised metadata approach to discovery and use of data in the NERC DataGrid", *Proceedings of UK e-Science All Hands Meeting 2004*, ISBN 1-904425-21-6.
14. ISO 19119, *Geographic information – Services*.
15. Web Service Resource Framework, http://www.globus.org/wsrf/
16. Thematic Real time Environmental Distributed Data Services, THREDDS http://www.unidata.ucar.edu/projects/THREDDS
17. Lawrence, B.N., 2004, "The NERC DataGrid: 'Googling' Secure Data", *Proceedings of UK e-Science All Hands Meeting 2004*, ISBN 1-904425-21-6.
18. Stephens, A., Marsh, K.P., and Lawrence, B.N. Presenting a multi-terabyte dataset via the web. Proceedings of the ECMWF Ninth Workshop on Meteorological Operational Systems, 2003. http://ndg.badc.rl.ac.uk/public_docs/Ag_Stephens_ECMWF_Workshop_NOV2003.pdf
19. http://www.terrascale.com/prod_e.html