



Introduction

The Natural Environment Research Council's environmental data holdings are one of its most significant resources. The maximum benefit may only be obtained from these data if they are readily available to potential users. NERC DataGrid (NDG) is a programme funded jointly by NERC and the UK Research Councils e-Science Core Programme to address data accessibility issues using e-Science technology.

Project Objectives

The ultimate objective is to combine NERC's data holdings into a seamless distributed data centre, which is available as a resource across the organisation. Whilst the project has to be concerned with metadata to address issues of data discovery, the fundamental concern is access to the data themselves. Initially the work will focus on oceanographic and atmospheric data, but the technology is being designed to facilitate support of data from the wider environmental science community.

Project Participants

The project partnership involves the British Atmospheric Data Centre (BADC), the British Oceanographic Data Centre (BODC), the Council for the Central Laboratory of the Research Councils (CCLRC), the National Oceanography Centre, Southampton (NOCS) and Plymouth Marine Laboratory's Remote Sensing Data Analysis Centre (RSDAS).

e-Science and Grids

e-Science is the term that has been coined for large scale scientific activities undertaken through global collaborations enabled by the Internet. It is envisaged that such science will need to be underpinned by access to high-powered computing facilities and information held in dedicated databases that is as easy to access as information held in web pages.

The architecture proposed to deliver this is termed 'the Grid', defined by Foster and Kesselman as '*An infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources.*' Their vision for the Grid has been developed into a middleware implementation called Globus that is currently being used for many Grid development projects.

Grid projects fall into two categories, Computational Grids and Data Grids. Computational Grids provide transparent access to processing on remote systems. This may be used to run computations on spare capacity in computers in many locations or to enable large data sets to be utilised by taking the application to the data. Data Grids provide access to information held in dedicated, distributed databases without the multiple logins and knowledge of several data systems currently required.

A central concept of Grid architecture is user authentication by means of a digital certificate. These are electronic messages that accompany requests for services and unambiguously identify the requester. They allow systems to automatically ascertain whether the request for data or computing power is authorised. This will replace the authorisation mechanism of user identifiers and passwords that is currently in use. Human networks and procedures are currently being implemented to assure the secure distribution of digital certificates.



User Interface

The first port of call for users is the Discovery Service, a web portal for humans (<http://ndg.nerc.ac.uk/discovery/>) or a web service Application Program Interface for software. The service delivers the discovery metadata records that match the search criteria. These records may be viewed or used as links to three types of additional service:

- NDG metadata service, exposing the MOLES records and their linkages leading to further discovery of related datasets.
- NDG data service, accessing the data from one or more datasets through the CSML metadata and returning a graphical presentation for visualisation, a harmonised data file or the source files.
- Data host local services, which can be anything accessible through a URL that the data host considers to be relevant to the dataset.



Metadata

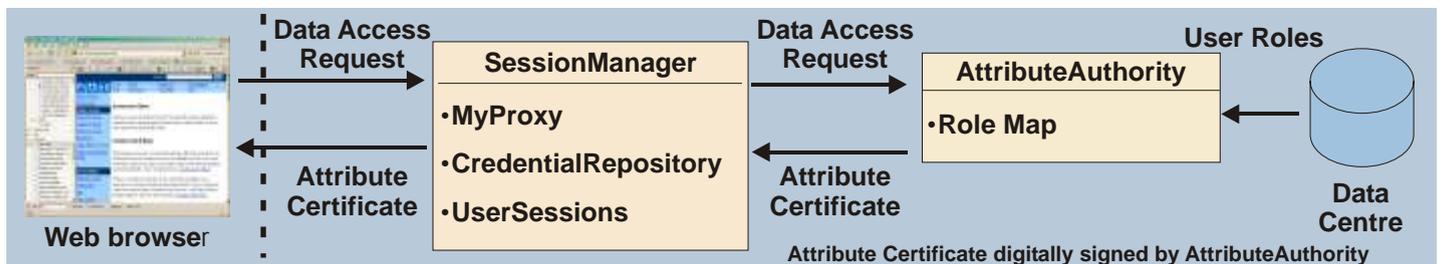
Metadata are the life blood of data systems designed for automated usage. They are necessary to describe the datasets so that interested parties can find them (discovery metadata) and to tell applications what to do when they have been found (usage metadata). Some metadata repositories fulfil both functions, but separate schemas have been employed for each function by NERC DataGrid.

Discovery metadata are held in an intermediate schema known as MOLES (Metadata Objects for Links in Environmental Science). Document reformatting technology (XQuery and XSLT) is used to create metadata records for the discovery service in standard formats such as Directory Interchange Format (DIF).

Usage metadata are expressed in Climate System Modelling Language (CSML <http://ndg.nerc.ac.uk/csml/>), which is based on the Open Geospatial Consortium's Geography Markup Language (GML). This enables the data to be handled as virtual objects known as features.

The NDG architecture is aware of other types of metadata, such as annotation metadata (user comments on datasets) and data documentation (e.g. data reports) that will be incorporated into the schemas as the system matures.

Security



One of the characteristic features of a data grid is that a user's access to data is delivered by single-point login. Access is then based on trust agreements between the data suppliers on the grid rather than user registration with every data host.

NERC DataGrid has addressed this by establishing an attribute authority, providing a proxy certificate when the user has been authenticated on a host where they are registered by user id and password.

The certificate also includes role information (e.g. 'academic' or 'NERC employee') assigned by the NDG host on which the user is registered. Other hosts hold a mapping between these roles and their equivalent local roles and use this to determine the data access rights for that user on the host's system.