# The NERC DataGrid Project

## Introduction

One of the most significant resources of the Natural Environment Research Council is the organisation's environmental data holdings. 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, currently in a series of quality data archives, into a seamless distributed data centre, available as a resource across the NERC Virtual 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 will be 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) and the Council for the Central Laboratory of the Research Councils (CCLRC) in collaboration with the U.S. Programme for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore Laboratory. There is also close cooperation with the US Earth Systems Grid.

## What are 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 as current access to 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 jobs using unused processor capacity 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 databases without the multiple authentication 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 that unambiguously identify the requestor and 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.

# Anatomy of the NERC DataGrid

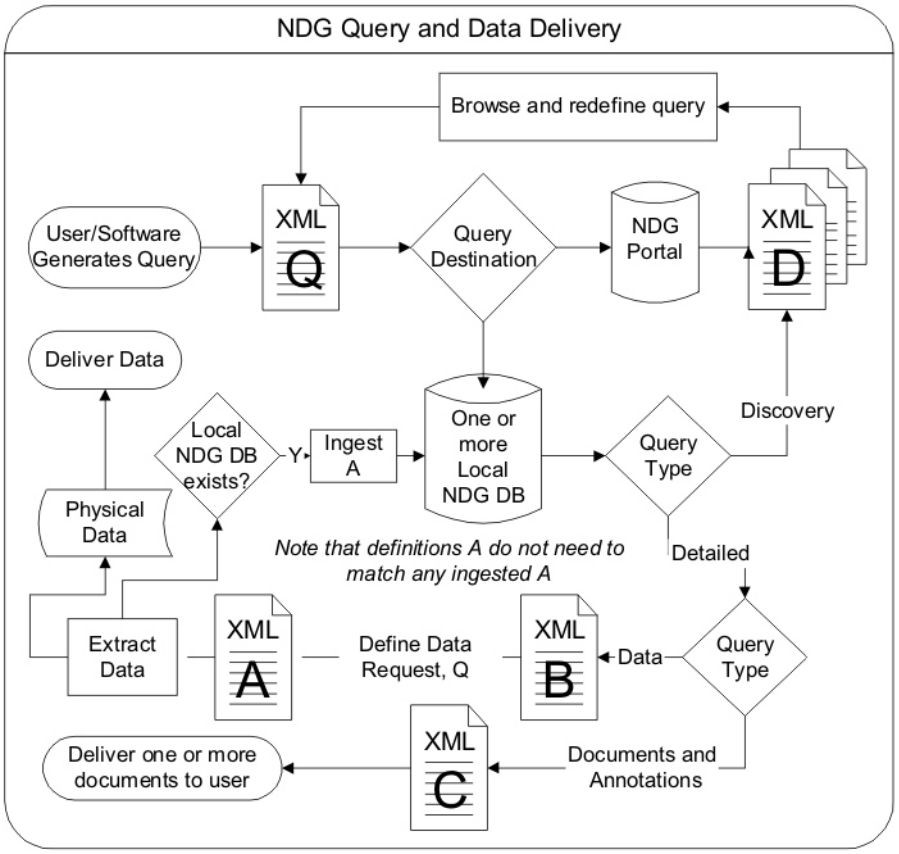
Users require data conforming to their specifications ingested into a tool that allows them to interpret, manipulate or visualise the data. Getting to this position requires several stages:

Discover what data sets might be suitable from a metadata catalogue.
Browse detailed metadata and formulate a data request.
Submit request and receive data.
Manipulate data to match the desired tool.
Load up data and do some science.

This presents the user with a lot of work, particularly if the data required are held by different data centres. Multiple authorisations will be involved and almost inevitably the data will be delivered in different formats.

The NDG will remove much of this burden. Discovery and browsing will still be required, but once the data request has been formulated the system will take over, merging individual elements of data to produce a 'designer dataset' that conforms to the user's specification and may be directly ingested into tools provided as part of the NDG interface.

Facilitating such a system requires extensive metadata to document the data to the level where automated systems can handle the data. Several types of metadata, to be implemented as XML files, have been identified. A metadata model has been developed to provide data set descriptions that form the basis of browsing (type 'B' in the adjacent diagram) and discovery (type 'D'). Automated data processing will be facilitated through data descriptions (type 'A' metadata) conforming to the NDG data model.




Some users will access the NDG through a portal interface. However, the system will also be accessible to software through an Application Program Interface (API). This means that applications will be able to automatically ascertain if appropriate data are available, request those data and ingest them. The potential of this aspect of the technology for data assimilation into models is enormous.