DATA DISCOVERY IN THE NERC DATAGRID

Marta Gutierrez[2], Bryan Lawrence[2], Michael Burek[4], Ray Cramer[3], Kerstin Kleese van Dam[1,]
Siva Kondapalli[3], Susan Latham[2], Roy Lowry[3], Don Middleton[4], Kevin O'Neill[1], Ag Stephens[2],
Andrew Woolf[1]

[1]CCLRC e-Science Centre
[2]British Atmospheric Data Centre
[3]British Oceanographic Data Centre
[4](U.S.) National Center for Atmospheric Research

Abstract

The NDG discovery service provides access to all the data held in the British Atmospheric Data Centre, and substantial datasets held in the British Oceanographic Data Centre and the U.S. National Center for Atmospheric Research. A software package is available to aid data providers in making their data available. This paper introduces the architecture of the Discovery Service, based in part on OAI/PMH, and outlines a number of technical issues associated with implementing it.

1. INTRODUCTION

The advent of whole earth system science means that better tools are needed to discover, share and utilise data with the maximum of format transparency. The NERC DataGrid (NDG) has been designed and built in response for the UK academic environmental science community.

In this short paper we describe the discovery component of the NDG, covering the architecture, implementation, and practicalities of supporting millions of data entities and terabytes of data. The scalability of the design is demonstrated by implementing interoperability with the U.S. National Center for Atmospheric Research (NCAR).

The main components of the NDG architecture are discussed in [1]. The broad concepts of the discovery service were discussed in [2]. Here we concentrate on the architecture of the discovery service itself and the issues associated with metadata transformations, and populating the metadata archives, and ensuring international connectivity.

2. ARCHITECTURE

The main design policies underlying the NDG discovery implementation were that:

- It should be scalable across hundreds, if not thousands, of data providers.
- Discovery records should be completely public
- It should be possible for third parties to build their own discovery portals based on NDG discovery data.
- It should be interoperable with the library community and the electronic publishing industry, to ensure that datasets could eventually be formally published.
- It should be widely understood and capable of use on an international scale.

Existing connection-orientated services such as Z39.50 were rejected as unsuitable at the time given these constraints. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, [3]) developed by the digital library community met the requirements and has been adopted. This protocol allows single or multiple metadata records (or record headers) to be retrieved from a Data Provider, and introspection on the Data Provider's repository-level metadata and supported metadata harvest formats. Selective harvesting is supported on the basis of time/date stamps or set membership.

Any metadata format is supported with OAI-PMH (provided it can be cast in terms of a schema compliant XML document) although a harvest request for Dublin Core must always be honoured. NDG has chosen to use the Directory Interchange Format (DIF) of the NASA Global Change Master Directory (GCMD) as the initial discovery

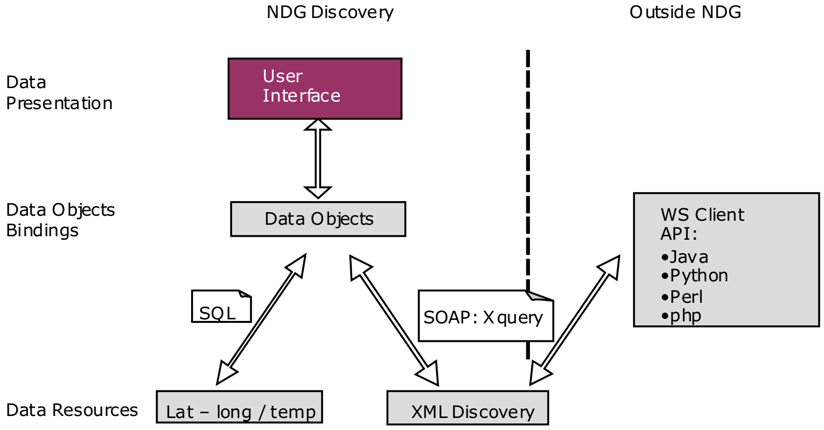

Figure 1: Web Services interaction with Discovery Metadata (at version one, later versions will provide the web service interface at the data object level providing access to both the geographical search interface and the xml documents).

format, although migration to ISO19115 is planned as soon as the underlying schema is stable enough for interoperable implementation at the international level.

The overall architecture is then simply described as followed:

- Individual data providers must produce OAI compliant repositories of discovery documents.
- The initial format is NASA GCMD DIF, but is planned to evolve to ISO19115.
- A discovery service harvests documents from the data provider discovery repositories and produces a new aggregated discovery repository.
- These documents are inserted in an XML database, and a web-service interface is provided.
- A discovery portal interacts with the web-service interface (as can any 3rd party service).
- The discovery documents must include appropriate identifiers that provide: (1) links to data delivery services that can directly access the data described in the discovery documents; and (2): links to more sophisticated metadata objects. For the NDG, the latter are the more semantically rich NDG-B metadata[1] which include further descriptions and relationships with other data objects across the NDG data centres.

3. IMPLEMENTATION ISSUES

Implementation issues fall into a number of categories; described here mainly for the NDG:

- There are a number of implementations of the OAI-PMH. At the NDG we have chosen the DLESE package [4].
- Implementing the metadata pipeline: In NDG we have developed XSLT codes which create discovery documents from NDG-B metadata [1]. By contrast, at NCAR, Thredds [5] catalogues are used as the base metadata and XSLT codes produce the discovery documents.
- Harvesting non-Dublin-core documents involves ensuring that there is an XML-XSD available for the discovery format. Both NCAR and NDG developed their own, as GCMD does not publish one.
- Installing an appropriate database technology for searching. NDG has used an eXist database, which we do not expect to scale to millions of documents, but which has allowed us to investigate the use of XQuery technologies. Because geographical searching based on bounding boxes is also required, records are also installed in a postgres/pgsphere database.
- Data Delivery services differ at different locations, and the Discovery service needs to provide some level of

introspection as to what service is available at a data URI. We have had to make slight modifications to the discovery documents to indicate what NDG services are available (if any). In the longer term one would hope that the data URIs themselves would respond to well-formed introspective queries.

4. CURRENT STATUS

The NDG discovery portal, at http://ndg.nerc.ac.uk/discovery now provides access to all the holdings of the British Atmospheric Data Centre and substantial holdings from the British Oceanographic Data Centre and the U.S. National Center for Atmospheric Research (NCAR).

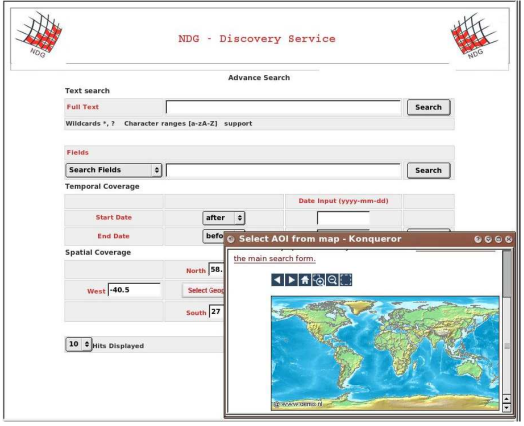

Figure 2: NDG Advanced Search Interface with popup map coordinate entry.

(http://cdp.ucar.edu) provides access to all the NDG holdings as well as their own. Hundreds of datasets, representing TBs of data in millions of files are now easily discoverable. Additionally, interested library communities can harvest Dublin Core versions of the discovery documents.

Appropriate software, including appropriate code and documentation, has been bundled so that new data providers can easily make their data discoverable via both interfaces.

5. DISCUSSION

While OAI has has become de facto interoperability standard for the Digital Libraries and institutional respositories (e.g. [6]), it is being widely adopted by scientific communities as a mean of publishing and archiving scientific metadata to access datasets, satellites images etc. For example, The Publishing Network for Geoscientific & Environmental Data PANGAEA [7] provides publicly available geo-referenced metadata under a similar framework.

Although OAI introduces a lightweight protocol for metadata harvesting, there is still no such approach to deal with a standard query and retrieval mechanism for federated parties. The Open Geospatial Consortium (OGC) is working under this line with the Catalogue Service implementation specification. CAT2.0 [8]. The specification covers scenarios such as: publishing , harvesting, and discovery of metadata. It proposes protocols bindings to SOAP, Z39. 50. An implementation of the CAT 2.0 is being developed by ESRI, and the GIS Portal Toolit includes an earlier version of this specification.

With approaches built on OAI, it will still remain important to identify common community profiles of metadata via crosswalks or standard subsets of ISO19115 (we expect differing scientific communities to have quite different profiles of ISO19115). Although OAI mandates Dublin Core (DC) format as a minimum denominator for exchange, DC does not provide enough useful metadata to establish generic data discovery services.

REFERENCES

[1] Integrating distributed climate data resources: The NERC DataGrid. A. Woolf et.al., A. Proceedings of the ELEVENTH ECMWF WORKSHOP on the Use of High Performance Computing in Meteorology (2004), *To Appear.*

[2] The NERC DataGrid: “Googling” Secure Data. B. Lawrence et.al., Proceedings of the UK e-Science All Hands Meeting 2004. (http://www.allhands.org.uk/2004/proceedings)

[3] The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[4] http://www.dlese.org/oai/index.jsp

[5] http://my.unidata.ucar.edu/content/projects/THREDDS

[6] http://epubs.cclrc.ac.uk/

[7] http://www.pangaea.de

[8] http://www.opengeospatial.org/specs/?page=specs

[9] http://www.esri.com/software/arcgis/gisportal-toolkit/index.html