PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 1015–1019
doi:10.1098/rsta.2008.0238
Published online 16 December 2008

The NERC DataGrid services

By S. E. Latham[1], R. Cramer[2], M. Grant[3], P. Kershaw[1], B. N. Lawrence[1,*], R. Lowry[2], D. Lowe[1], K. O'Neill[1], P. Miller[3], S. Pascoe[1], M. Pritchard[1], H. Snaith[4] and A. Woolf[1]

[1]*STFC Rutherford Appleton Laboratory, Didcot OX11 0QX, UK*
[2]*British Oceanographic Data Centre, Proudman Oceanographic Laboratory, Liverpool L3 5DA, UK*
[3]*Plymouth Marine Laboratory, Plymouth PL1 3DH, UK*
[4]*National Oceanography Centre, Southampton SO14 3ZH, UK*

This short paper outlines the key components of the NERC DataGrid: a discovery service, a vocabulary service and a software stack deployed both centrally to provide a data discovery portal, and at data providers to provide local portals and data and metadata services.

Keywords: Web Map Service; Web Coverage Services; geospatial; DataGrid; Open Archives Initiative Protocol for Metadata Harvesting

1. Introduction

The information environment required to support an environmental data grid is described in Lawrence *et al.* (2009, hereafter L08). In this paper, we describe the services that have been deployed in the NERC DataGrid (NDG) to support data discovery and reuse in the environmental sciences. The main components are laid out in figure 1. It is important to note that individual data providers will have existing (legacy) data and metadata systems, and it is exceedingly unlikely that there is any benefit to trying to homogenize their approaches to data management. Hence, the NDG approach was to build tools to support the deployment of coupling systems between the existing data and (i) NDG metadata middleware and (ii) Open Geospatial Consortium (OGC) Web service middleware.

In this paper, we briefly discuss the details of the coupling systems, the services and the way the system is joined together. The services built for the NDG also include support for access control based on distributed authentication and authorization, but these are discussed elsewhere (Lawrence *et al.* 2007).

2. Discovery

The metadata middleware is connected to the NDG discovery service, which provides a centralized harvested database of D-type (see L08) metadata. Metadata is harvested from all the NERC data centres, other key NERC groups,

* Author for correspondence (bryan.lawrence@stfc.ac.uk).

One contribution of 24 to a Discussion Meeting Issue 'The environmental eScience revolution'.

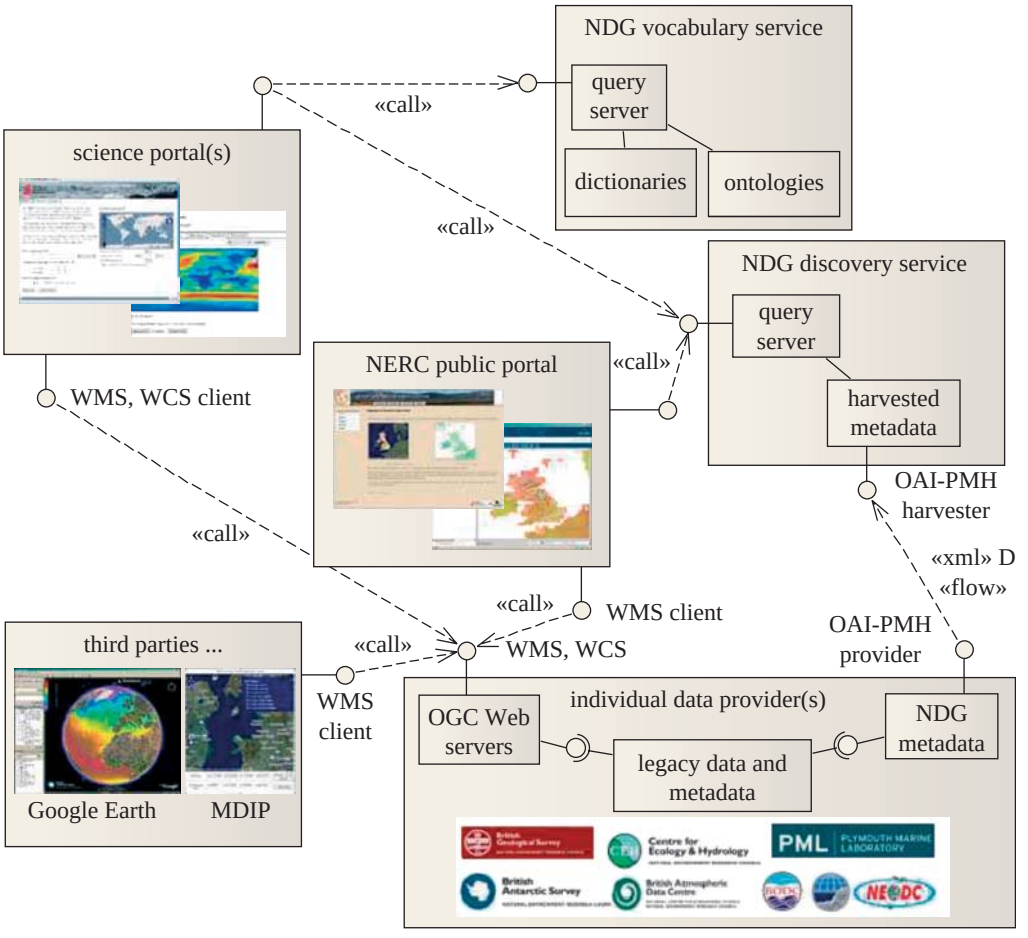

Figure 1. Key components of the NDG (neglecting security and access control components).

other UK data providers (including participants in the UK Marine Data Information Partnership (MDIP) and the Defra Central Science Laboratory) and some overseas data providers (the US National Centre for Atmospheric Research and the German World Data Centre for Climate). Harvesting is accomplished using the Open Archives Initiative Protocol for Metadata Harvesting (Lagoze *et al.* 2002).

The centralized database can be queried programmatically from remote locations via a bespoke Web service interface and supports a variety of D-type metadata formats (both as input and output). Those remote locations currently include the NERC data discovery service (aimed at the scientific community), a public-orientated NERC data portal, an MDIP portal and each data provider that instantiates the NDG service stack (discussed below). Future versions will be linked into the Global Earth Observation System of Systems discovery portal and other major international initiatives.

The discovery service supports general textural queries, geospatial queries, temporal queries, authorship queries and supports limited semantically assisted searching via the vocabulary service discussed in §3. The latter provides

alternative suggested search topics based on user input: for example, a search for rain currently provides 15 hits, but the portal also suggests precipitation as an alternative (which yields 92 hits).

3. Vocabulary service

Key capabilities of the semantic interoperability deployed in the NDG are the ability to define objects accurately with terms from controlled vocabularies, and to map between terms in different vocabularies. The NDG vocabulary server was developed to provide both governance (version control, reliability, etc.) and access to such lists and mappings from anywhere on the Web (not just the NDG).

The vocabularies served are aggregations of concepts, each described by a semantically neutral identifier, full and abbreviated human-readable labels and a definition. This information is held in an Oracle database with automatically maintained versioning, time stamping and audit trails, with the content heavily protected by triggers and constraints.

The system is populated by over 120 000 concepts aggregated into 105 lists. Each concept can be uniquely addressed by a Universal Resource Identifier (URI), which allows mappings between related concepts to be held in a relational Resource Description Framework triple store. This has over 75 000 triples mapped using the relationships specified by the simple knowledge organization system (SKOS). Future development will support a richer relationship set. There is considerable scope for the expansion of this work to exploit other semantic Web technologies.

Each concept URI may be expressed as a Universal Resource Locator (URL), which delivers a SKOS document containing the concept labels and its mappings. Complete vocabularies may also be addressed as URLs. Web service interfaces to the server containing a range of access methods are also provided.

The system is fully operational and attracts several thousand unique concepts and vocabulary accesses per month.

4. Metadata and services

Prior to the advent of NDG, pre-existing commercial and open-source OGC service software was primarily orientated towards supporting maps and simple features.[1] Supporting data objects of interest to the environmental sciences requires support for more complex data types, and, particularly for the oceanographic and atmospheric sciences, requires support for CSML (see L08). Supporting the ability for users to understand the datasets held by the providers requires support for serving and browsing MOLES (see also L08) metadata.

These extensible markup language (XML) metadata requirements are significant, but not onerous, and their acquisition is summarized in figure 2. Data that are created with appropriate internal metadata (e.g. CF[2] compliant netCDF) can be ingested using a scanning tool, which is part of the software library developed to manipulate CSML. (If the original data include little or no

[1] A simple feature can essentially be fully portrayed by a point, line or polygon on a map.
[2] Climate forecast conventions, see http://www.cfconventions.org.

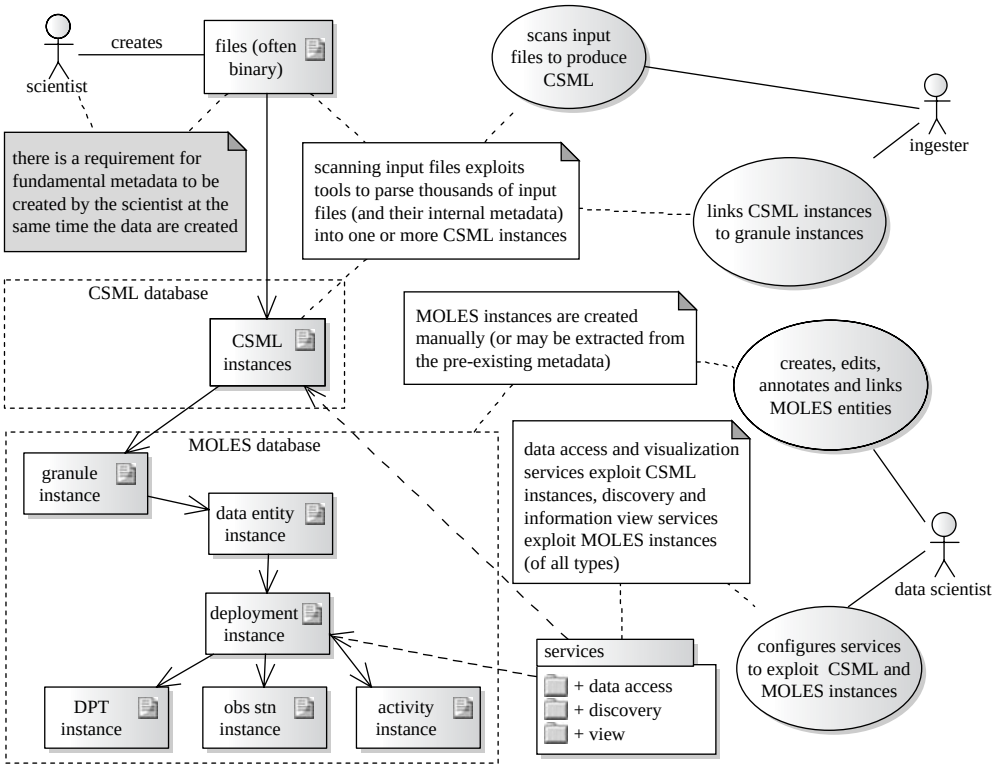

Figure 2. Key actions in configuring metadata.

internal metadata, then it is nearly impossible to automate and ingest the data into any sort of environment, grid or otherwise.) This process results in relatively few CSML instances per dataset, with potentially thousands of files (with millions of records) reduced to a handful of CSML instances. These CSML instances should be associated with MOLES granules, with one feature type per MOLES granule. The process of scanning and associating CSML files, termed ingestion, currently requires human interaction, but at a relatively low level. This procedure results in data granule entities that, without accompanying MOLES information, would be difficult to exploit. Experience suggests that these 'MOLES class' metadata are crucial to the successful exploitation of data outside the original data collection team. Constructing the required MOLES information is generally manually intensive and time consuming (unless it can be harvested from the previously collected metadata), requiring skilled 'data scientists', but it does result in scientific contextual information, which makes it possible to compare, contrast and select data with confidence about the data provenance and quality.

With a MOLES metadata repository in place, it is possible to automatically generate discovery metadata, and this is now operational in the British Atmospheric Data Centre (BADC), where, as of August 2008, discovery records are generated automatically for 160 data entities associated with 229 activities, 863 data production tools and 400 observation stations. Those 160 discovery records appear in the NERC data discovery service that currently advertises 1360 datasets.

The data discovery and vocabulary services are currently operational at http://ndg.nerc.ac.uk/discovery and http://vocab.ndg.nerc.ac.uk/.

NDG currently recommends storing MOLES and CSML documents in an eXist (see http://exist.sourceforge.net/) XML database. When data providers do so, an NDG software stack can be deployed, which provides (i) CSML aware OGC Web Map Service (WMS) and OGC Web Coverage Services (WCS), (ii) a discovery service client, (iii) eXist clients to support export to discovery metadata, (iv) a MOLES browse server, (v) a MOLES browse client, and (vi) a WMS and WCS client (visualization and download service). As well as being deployed at data providers to serve up data and metadata, an instance of this software stack which provides only the discovery service client component is deployed as the operational NERC data discovery service portal. As indicated above, there are also stand-alone scripts provided with the CSML stack to scan some types of input data to produce CSML. Currently, the CSML software stack is only configured to a few prototype datasets, as the operational connection between CSML production and MOLES granule instances is expected first at the BADC, which is currently reconfiguring older metadata handling software to exploit the new paradigm.

5. Future work

In the development of the NDG thus far, two key lessons have been learned: (i) Web services based on the representational state transfer model are superior to those based on the simple object access protocol both in terms of interoperability and in terms of the benefits of mapping URIs to resources, and (ii) domain modelling in the unified modelling language coupled with serialization into XML schema is superior to direct modelling in XML because it is much easier to build and populate distributed systems that are clearly articulated and well understood by all participants. These two lessons are prompting the development of new versions of MOLES, which should have wider applicability and easier usage, as well as easier to deploy and maintain services.

References

Lagoze, C., Van de Sompel, H., Nelson, M. & Warner, S. 2002 The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. See http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

Lawrence, B. N., Kershaw, P. & Blower, J. 2007 Practical access control with NDG-security. In *Proc. UK e-Science All Hands Meeting, 2007* (ed. S. Cox).

Lawrence, B. N., Lowry, R., Miller, P., Snaith, H. & Woolf, A. 2009 Information in environmental data grids. *Phil. Trans. R. Soc. A* **367**, 1003–1014. (doi:10.1098/rsta.2008.0237)