# NERC DATA GRID (NDG)
# MEDIUM SIZED INITIATIVE (MSI)

## Final Report

Bryan Lawrence (STFC CEDA), Roy Lowry (BODC) , James Doughty (DIASS Ltd)

and the NDG Team*

01/10/09

* Comments and improvements to this document from Jon Blower, Steve Donegan, Phil Kershaw, Dom Lowe, Stephen Pascoe, Matt Pritchard and Andrew Woolf.

# 1. Executive Summary

The NERC DataGrid was established to provide integrating technology to allow data to be discovered, manipulated, understood and visualised from the managed archives of the NERC community (including but not limited to the archives of the NERC designated data centres). As an integrating activity it requires a community of data providers who share common standards of data documentation and data interfaces and who can deploy tools which expose and conform to those standards.

The NERC DataGrid project was originally launched as part of the UK e-Science programme (NDG), with continuation funding (NDG2) under that programme and via two subsequent direct NERC contracts: the "portals project" (NPP) and the "NDG Medium Sized Initiative" (NDG-MSI also known as NDG3).

NDG1 and NDG2 concentrated on the issues associated with integrating data and information in the atmospheric and oceanographic sciences while keeping abreast of the issues and technologies to be encountered when expanding efforts into the wider environmental sciences. The NPP both extended that work within atmospheric and oceanographic sciences, and extended the discovery work along with some visualisation into the wider NERC community in the process of developing a demonstrator "Environmental Data Portal". The goals of NDG3 were to further develop the key technologies, influence and educate relevant communities, put some key elements of the work on a stronger footing for deployment "operationally", and maintain expertise in anticipation of implementing a new NERC information strategy which would need it.

## *1.1 Purpose of this document*

This document summarises what the NDG3 project delivered, the key problems encountered along the way, identifies future work needed, and makes recommendations about how NERC should best exploit the work done in this and the previous projects. Some underlying assumptions about the information environment are included. This document is not intended to be a technical summary of the work done, but some technical expertise may be needed to interpret all the findings.

## *1.2 Strategic Context*

NERC currently has six designated data centres which are essentially discipline based along with a number of managed archives within other NERC owned and collaborative facilities. There are a number of cross-disciplinary areas (e.g. paleo) and projects (e.g. the RAPID climate change programme) which are not currently well catered for with this structure. The NERC science strategy requires integrated and multi-disciplinary data access for nearly all the themes identified, again, such multidisciplinary data access is not necessarily deliverable by the current structure without integrating activities.

It is also a commonly held view that the sort of integration and interoperability of data and information envisaged will provide major savings in time and money for both those doing science and those exploiting science. Indeed, recent calculations (see section 3.4 ) suggest that the effective financial benefit of data management and integration within the NERC community could reach the order of 35 million pounds per annum,

NERC has recently promulgated a new Science Information Strategy (July 2009), and is in the process of developing a concomitant implementation plan. Some of the recommendations in this report are intended to be fed into that process.

## *1.3 Key Project Achievements*

1. Technology Development: all the major NDG technology components (the discovery service tools, vocabulary service tools, information schema and tools, security infrastructure, data delivery and visualisation services) have been enhanced either within the scope of the MSI funding, with co-funding, or under the auspices of complementary activities.

2. Education and Knowledge Transfer (here we are discussing education and knowledge transfer about NDG concepts, not the education and knowledge transfer which arise from the use of NDG): a significant number of workshops have been held, and more are planned. Defra uses NDG solutions as INSPIRE exemplars, other Defra groups are interested in exploiting NDG, the Environmental Research Funder's Forum is considering using NDG technologies to maintain information about environmental research in the UK, and the Marine Environmental Data Information Network (MEDIN) is currently in the process of contracting NDG to deliver their discovery service to enable discovery of their data. A number of projects delivered for NERC and other stakeholders by the NDG partners have benefited from NDG technology (including but not limited to SeaDataNet, the UKCP09 User Interface, and the QUEST[1] Earth System Data Initiative. QESDI). A number of NERC centres not originally part of the NDG consortium are investigating deploying, developing or extending NDG technologies. A monthly newsletter has appeared, the NDG mailing lists have become public, and the NDG website has had been redeveloped.

3. Operational Services: both the NERC data discovery service (based at the NCAS/BADC and NCEO/NEODC) and the NERC vocabulary service (based at BODC) have had improvements aimed at easier usage, more reliability and clearer indications of service level. A report outlining realistic operational service levels and mechanisms for the development and evolution of these services will be prepared in the context of the implementation of the NERC science and information strategy.

4. Personnel: Key individuals from the original NDG team have been kept working on relevant problems, and are or will be feeding into the ongoing development of the implementation plan of the NERC science information strategy.

## 1.4 Project Finances

The original budget for NDG3 was £220K, but this was augmented by a further £96K from associated budgets. Of the £316K managed under the auspices of NDG3, £240K was spent as of the end of July, and the remaining £76K is committed to be spent by the end of the 2009/2010 financial year. Of that 76K, 52K has been allocated to run and develop the operational NERC data discovery service, which would otherwise be unfunded.

NDG3 was delivered by the UK Science and Facilities Research Council (STFC) with major contracts to the UK NERC British Oceanographic Data Centre (BODC), the Australian Commonwealth and Scientific Industrial Research Organisation (CSIRO), and minor contracts to DIASS Ltd for specialist project management and Orchard Professional Ltd for specialist website development and Tessala Ltd for programming support late in the project.

## 1.5 Summary of Recommendations

Further details can be found in section 3.

### Research & Operations

1. The operational "NDG" services need a transparent governance body (or bodies) which respects the intellectual property of the originators, the requirements of their stakeholders, and the fact that both already exist in environment where the main stakeholder may be NERC, but that stakeholders are not limited to NERC alone.

2. The implementation plan for the NERC science and information strategy needs to recognise the need for corporate funding for the "NDG" operational services.

[1] QUEST: Quantifying and Understanding the Earth System programme: http://quest.brist.ac.uk

3. NERC needs to understand that the operational services cannot stand-still, they exist in a fast changing technological world, and both will need to evolve to remain relevant. Such evolution will continue to require targeted informatics research and development funding.

4. There needs to be a clear procedure which controls the evolution of the operational Data Discovery Service and the Vocabulary Services so that the level of service is stable and predictable.

## Spatial Data Infrastructure Architecture.

5. The 'operational' Data Discovery Service (DDS) should become the primary *generic* entry point to NERC data.

6. The team responsible for running and developing the DDS should have responsibility for a NERC interface to generic UK, European and International data discovery activities.

7. NERC centres and surveys should deliver standards compliant metadata to the DDS using standards consistent with the requirements of the DDS governing body, and expect those requirements to evolve.

8. Not withstanding the existence of generic entry points, discipline specific entry points to NERC data should continue to evolve consistent with discipline specific norms and requirements.

9. Interdisciplinary data access and visualisation depends on tools which have common interfaces regardless of the underlying data formats and metadata standards. NERC needs to continue to invest both in these tools, and the underlying standards, for the foreseeable future.

10. The NERC IT community should support and encourage appropriate "federation level" access control to data over and above existing internal access control mechanisms. In doing so, they must recognise they cannot control or mandate the systems used by all parties, and they should exploit the experience already gained in the NDG team.

## Data

11. Whatever integrating technologies exist, NERC needs to continue to invest in documenting and formatting data to common standards (which are likely to differ between communities) so as to ensure the maximum benefit from the investment in software tools to deliver and visualise such data.

## Community Data Infrastructures

12. NERC should make a clear statement about the future and finances of the discovery and vocabulary services (and any other services developed in the future), so that the NERC community, and potential data providers and users can invest in the use of those services with confidence.

13. NERC should provide funds to kick-start the development of community data infrastructures that integrate into the corporate data infrastructure as recommended by the Data Portals Project Board in June 2008.

14. NERC should expect and encourage (and where possible, allow the commercial sale of ) access to NERC operational services (and recognize the governance implications, which is implicit in our first recommendation under research and operations).

## Marketing, Branding & Communication

15. The 'look and feel' of the NERC single entry point to its data should be more closely aligned to the NERC corporate identity.

16. NDG should be renamed to reflect that it isn't a data grid but a set of technologies and methodologies suitable for wide environmental applicability.

17. NERC should require and fund marketing and communications activity as an integral part of developing and expanding the (NDG) services.

18. NERC should continue to invest in having a showcase prototype portal demonstrating what can be achieved beyond the current operational activity (which might include demonstrating new service capabilities as well).

## Knowledge Exchange

19. NERC should refine the organisation, delivery and marketing of the informatics courses created by this project.

# 2. Summary Review of Technical Progress

In this section of the report we present a summary review of the major components of work carried out in the NDG-MSI. There were three sorts of work carried out in the NDG-MSI: service development, software development, and communication activities. In the sections that follow each section has a brief summary of the objective of the work, and where appropriate, a brief description of the software or service component itself, followed by a description of what was delivered by the NDG-MSI, and what the future for that component of work should be.

## *2.1 Data Discovery Service*

The discovery service provides both a web portal and an underlying database of "discovery"[2] metadata harvested from data providers subscribed to the data discovery service. The interface between the portal and the underlying database is provided by a publicly accessible web service which currently conforms to NDG bespoke specifications. The intention is that other communities can (and do) exploit this interface to marshal and present data discovery searches in their own websites and portals. The NDG portal itself, while aimed at being "the" NERC data discovery portal, is not intended to provide data manipulation and visualisation.

The key objectives within NDG3 were to modify the service so that management by data providers along with usage and logging information is more complete, to enhance the interface so that a wider range of discovery requests could be supported, to use the logging information to help order result sets, and to upgrade the metadata harvest format. Some bug-fixes and improvements to the portal were also commissioned.

Management: The project has improved the logging capability of the framework of the Data Discovery Service (DDS) by keeping a record of what search text has been entered and which data links have been followed. This will provide more details to management on how extensively the DDS is being used and which links in which datasets are most often followed. This information may help prioritise further investment.

Discovery Service interface: The DDS Application Program Interface (API) (the service that interacts with requests from portals) was upgraded to allow requests to order results and to store information on searches into the logging database mentioned above. The metadata ingestion functions now also capture and record management information

Harvesting & Security: Metadata harvesting from the data providers was previously a daily scheduled process or done on an *ad hoc* basis. A new interface to allow data providers to have better control over what is provided to (or deleted from) the DDS has been developed and the beta test version can be found at http://ndg3beta.badc.rl.ac.uk/oai-info-editor/. The metadata harvester only allows authorised users from the data providers to invoke harvesting and now reports ingestion failures back to the data providers. A key problem for operational use of the DDS is that the metadata provided by data providers cannot be automatically validated for correctness, and errors cause ingestion failures which have to be fixed manually. Moving to a validating profile (ISO – see section below) is expected to eliminate the need for a major component of manual intervention with the running service. The new ingestion system now provides information directly to users which metadata files are causing problems.

ISO Metadata Profile: Upgrading the underlying discovery syntax to a profile of ISO19139 was largely dependent on NERC agreeing to a NERC specific ISO profile. That work was not under the control of the NDG team, being carried out by another NERC team; at the time of writing, no profile has been formally agreed. However changes to the ingestion routines have been made so that when a profile does become available, the update is not complex. Alternative ISO metadata profiles have been investigated and the MEDIN profile will be adopted as an interim profile. This work needs to be completed as a matter of urgency as it is a key requirement to comply with INSPIRE.

[2] For a discussion of metadata categories, and a definition of "discovery" metadata, see section 4.1

Completion of this work is further complicated by the fact that some of NERC's Data Centres are not in a position to provide ISO metadata records and thus the current format (DIF) will need to continue to run for a while yet. It is also probable that a DIF to ISO translation routine will need to be created to automate the transition from DIF to ISO.

Data Discovery Service Portal. Changes to what the users see were restricted to functional changes rather than look and feel changes. Changes include the ability to sort search results by Start Date, End Date, Data Centre, Popularity of Data Centre (by used links), Number of Times a Dataset has appeared in a Result-set, Proximity, Proximity Near Miss, Dataset Update Order, Dataset Name and Text Relevance (default setting). Other improvements included: refinements to the Vocabulary Service (run from BODC) to provide users with relevant alternative search suggestions, the implementation of a news feed to provide the community with news updates, and the implementation of a new date box to assist with definition of a date range.

The Environmental Data Portal, which was an alternative portal developed during the NPP to showcase some other attributes of portal services, was migrated to point at the operational discovery service rather than at a development service. The EDP is currently being operated by BGS, and continues to attract interest.

### Problems Encountered

The tools to deploy and configure the NDG software are still relatively immature, having had little effort expended on them in the previous projects. This led to problems getting the beta version of the DDS running as the underlying infrastructure changed, and delays to the timing of some DDS deliverables and suggested a number of improvements to the way the DDS is configured and managed are necessary.

While NDG3 has significantly improved the configuration management, with more modularity, and more sophisticated deployment techniques, there is still room for improvement, particularly to help deployment into groups which don't include NDG developers. Further work has been initiated within the existing funding envelope to investigate where improvements are necessary and, if appropriate, to implement them.

### Remaining Issues and Future Work

The data discovery service needs to be sustained as an operational service within the UK location strategy. The implications of this are that

1. Funding needs to be longer term than year by year,

2. The underlying interfaces need to expose INSPIRE syntax discovery metadata (possibly in addition to other more extensive NERC discovery metadata), and

3. The interfaces to those metadata need to include INSPIRE compliant interfaces, not just a NERC bespoke interface.

Neither of the last two activiites are difficult, and should be achievable within the near future, and possibly within the current funding envelope. In the longer term it is clear that NERC needs to also engage with GEOS and other UK, European and global discipline independent data discovery services through one interface, and this is the logical place to do it, not only from efficiency grounds, but also because it is likely that any ranking system such as that used by Google is likely to highly rate metadata exposed through one common interface as well as through others, thus achieving a higher profile for NERC data than can be achieved by the institutions operating alone.

It is also clear that the investment in the EDP, showcasing other aspects of data discovery and integration continues to yield interest in wider communities, suggesting that the EDP should continue to be maintained at least into the short-term. Given that the EDP also exposes data visualisation services as well as data discovery services, it is important to note that the existing data providers should continue to ensure these data are exposed appropriately. To that end NERC should mandate that their data providers maintain these interfaces within their core funding through to at least mid 2010.

## 2.2 Archive Metadata: CSML

The Climate Science Modelling Language (CSML) is an application schema of the Geographic Markup Language (ISO19136) which was developed for describing the key data types prevalent in atmospheric and oceanographic sciences. Along with the schema itself, NDG developed tools to read and manipulate data described in CSML for use by higher level services (such as those described in section 2.3 ).

Relatively little effort was expended within the NDG-MSI on CSML and directly associated tools, as that work is currently funded with a NERC knowledge transfer grant (C-SEKT). However, during the project it became apparent that there was a gap in funded work that could and should be filled. The COWS software stack (section 2.3 ) is implemented in Python, and while that has certain advantages, there are communities for whom a Java platform is more suitable, and being able to exploit CSML in Java environments would open up CSML exploitation to a wider audience.

Accordingly, Dr Jon Blower of the Reading University managed a component of work to develop a Java library for handling environmental data using the Climate Science Modelling Language (CSML).

This activity delivered:

1. Java interfaces defining the key elements of the CSML data model;

2. Software routines to read gridded data adapted from existing Reading code and to read other types of data from files and databases and express them using CSML interfaces.

3. Some focus on vertical profile data (CSML ProfileFeatures, e.g. Argo floats) and timeseries data (CSML PointSeriesFeatures, e.g. tide gauges) (Much of this sort of data had yet to be exposed within other NDG activities.)

4. Initial software routines to express these data using the CSML XML encoding.

**Problems Encountered**

Many key CSML concepts, such as units, geophysical quantities, coordinate reference systems and calendar systems, are “soft-typed” in CSML, meaning that there are many possible ways to encode (define) this information. In order to permit practical data exchange, it is necessary to agree upon conventions for encoding these concepts.

These agreements need to be made regardless of the programming language used, and their necessity here was exposed by having multiple implementations of CSML tooling, which demonstrates in passing that where “standards” are being proposed, multiple groups need to be involved in building implementations to ensure that all relevant issues/conventions are documented in the standard.

**Future Work**

The CSML conceptual issues need finalising. Once complete, the necessary software changes can be made and tested before the technology can be released publicly. The final project board approved funding a workshop to address this issue.

The current version of CSML has some gaps associated with describing various types of climate means, and it needs to be updated to be compliant with the emerging ISO standard for observations and measurements. This work is underway with other funding.

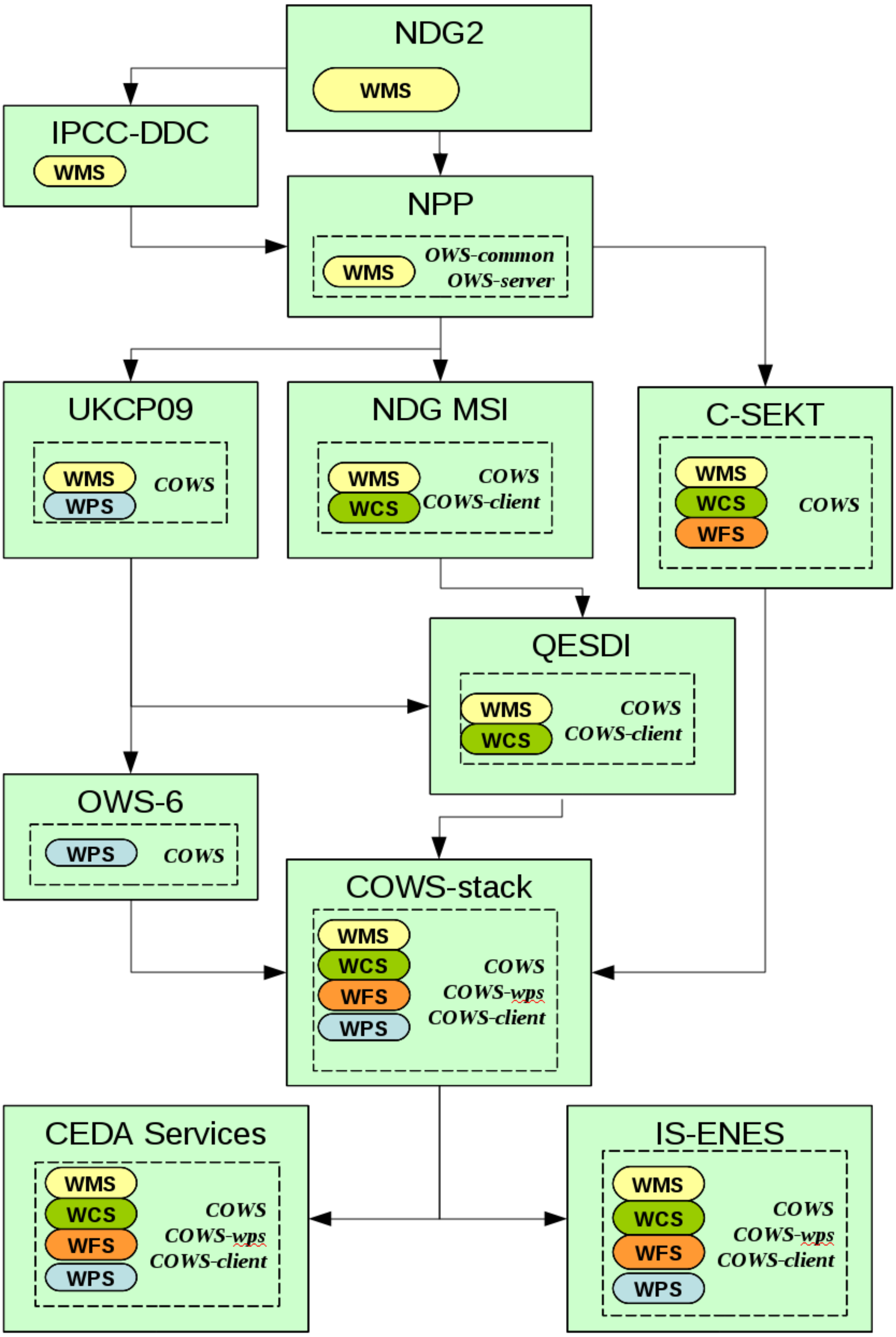

*Figure 1: Evolution of OGC web service components in NDG and other associated projects. Original components from NDG2 have been extended and deployed in support of the IPCC Data Distribution Centre, the UK Climate Projections 2009 (both funded by Defra), the C-SEKT knowledge transfer project, the QUEST Earth System Data Initiative (QESDI), even as they have been improved within the NPP and NDG-MSI. The WPS component was used in an international OWS testbed experiment (OWS6) which in turn is informing development. The new COWS software stack is now starting to be deployed operationally in CEDA, and will form a crucial part of the EU funded European Network for Earth Simulation (IS-ENES).*

## 2.3 Data Service Software

Early in the development of the NDG it became apparent that there was no suitable software stack to expose atmospheric and oceanographic data through common standards compliant service interfaces. At the same time it became apparent that while there were technical issues with the service interface specifications of the Open Geospatial Consortium (OGC), and despite discipline specific success stories such as OpenDAP, the OGC services provided the best roadmap to wide integration of environmental data in visualisations applications and for download. Eventually it became apparent that this was also the INSPIRE roadmap.

The key OGC services of immediate interest were the Web Map Service (WMS), the Web Coverage Service (WCS), the Web Feature Service (WFS) and the Web Processing Service (WPS). These provide maps, data download, sophisticated data and metadata exploration and sub-setting, and remote processing respectively. These are complex web services, and a gradual approach of building up capability has been followed, beginning in the original NDG and proceeding through subsequent projects. Capability has been added according to the requirements of the various projects. The evolution of the resulting "COWS – CEDA OGC Web Service" stack is depicted in Figure 1. A key facet of the COWS stack is that it can exploit CSML (section 2.2 ) descriptions of data in the configuration, allowing CSML semantics to be exposed, which improves the usability and configurabilty of the services.

Within the NDG-MSI, the key aims of the work were to improve some aspects of the basic COWS services (WMS, WCS), and to separate out a software package which provides web clients to the COWS services. All code was to be made open source. The resulting packages were then to be deployed with real data within CEDA, as a live test of how these services could be deployed. Longer term CEDA funding will be used to operationalise this software once the resulting clients and services could be secured so that access to specific data via these services could be controlled.

All this work was completed as planned: The services developed under the auspices of this work were deployed in a test environment[3] which served data from the NERC RAPID Programme[4]. They were secured using NDG Security (see section 2.4 ). Both the COWS server and the Client software are open source and available for other scientists and data curators to install and use with their own data holdings. See http://proj.badc.rl.ac.uk/ndg for more information. Many of the tools have already been exploited in the NERC QUEST Earth System Data initiative project (QESDI).

**Problems Encountered & Future Work**

A key part of the metadata spectrum is linking between the getCapabilities documents within the various OGC services, the links at which they are deployed, and how they are represented in discovery metadata. There is no robust paradigm for how to do this. An ad hoc solution was used for the work done in the NDG-MSI, but even within CEDA this mechanism doesn't scale for operational use. Realistically the connection between the service deployments and discovery metadata has to be done in intermediate metadata (such as MOLES, section 2.6 ), but the live version of MOLES deployed at BADC (which is much earlier version than the one worked on within the NDG-MSI) did not yet support this linking appropriately. Even with those handled properly, users need to be able to load "service contexts" discovered in catalogues into visualisation clients, this work too is in a very initial stage (with some progress in the context of the QESDI project).

Development of the OGC services is continuing with NERC funding via the core data centre lines, the C-SEKT knowledge transfer activity, and with European Commission funding via the EU. Deployment continues to be funded via CEDA, and the next step will be to migrate from development environments to operational environments.

[3] At the time of writing these services are deployed at http://ndg3beta.badc.rl.ac.uk

[4] http://www.nerc.ac.uk/research/programmes/rapid/

## 2.4 Data Access Control (aka Security)

A key aspect of delivering any distributed access to data is controlling that access so as to protect either any licensing restrictions associated with the data, or the service itself (from overuse resulting in inadvertent, or deliberate, denial of service attacks).

“NDG security” has been progressively developed through the NDG projects, with a major improvement funded by the EPSRC via a “software hardening” grant from the Open Middleware Interoperability Institute (OMII/UK). The key concepts of NDG security are described elsewhere[5], but the main concept is that software gatekeepers use web service interfaces to query user attributes and compare them with resource access criteria before making access control decisions.

Security related issues remain one of the major barriers to cross-disciplinary environmental research. One such problem scientists and researchers encounter is the need to hold multiple user accounts across different sites hosting the data and services they wish to access. NDG Security uses OpenID, a technology for single sign-on which allows users to hold a single login account which is then recognised across multiple institutions.

For Data Centres, the security system provides a flexible, decentralised and sustainable mechanism to allow (or restrict) access to data and services in a manner that encourages cross-institution agreement. The system uses established standards and specifications to make it easier for organisations to participate in such a federation. The system will be deployed as part of the Earth System Grid with partner organisations in the USA, Germany and elsewhere.

The key objective of the NDG-MSI was to provide a new NDG security interface as a Web Service Gateway Interface[6] middleware layer, and to apply that to secure the COWS server and clients. In the context of the latter, a key part of the requirement was to ensure that the open source javascript OpenLayers library which a key cows-client component could provide access control tokens to a COWS server, itself secured with NDG security. An additional requirement was to experiment with providing secure access to data exposed by OPeNDAP using NDG security.

All the developments planned under the NDG-MSI were completed. The new middleware layer allows underlying service applications to be secured without any modification, although the clients need to be modified to provide user tokens. Test datasets were also secured using OPenDAP. This work is an important practical demonstration of how OGC and OPeNDAP based services can be secured and will provide valuable input into defining future international security standards in these areas. In the case of OGC based services, the work will in turn inform OGC interoperability experiments and INSPIRE security developments and is likely to impact upon the implementation of INSPIRE throughout the EU.

### Problems Encountered

NERC’s security requirements are diverse and have had little management focus outside of CEDA since the early days of the NDG initiative. This document recommends that the Strategy Implementation Team addresses security as a specific set of activities since access control is often cited as a reason for organisations not to participate in federated architectures. It is important that NERC recognise that no single proprietary solution can satisfy NERCs distributed access control requirements, since they need to be inclusive of applications and data deployed in Collaborative Centres and Universities.

### Future Work

[5] Lawrence, B.N., P. Kershaw and J. Blower, 2007: Practical access control with NDG-security. http://www.allhands.org.uk/2007/proceedings/papers/788.pdf

[6] WSGI, see http://www.wsgi.org/wsgi/

The promotion and adoption of access control standards is vital to the establishment of a system to secure interoperable data services between organisations. Resourcing for individual data centres is critical to creating a sustainable operational infrastructure across NERC. A federated security infrastructure must be recognised (and therefore funded) as a distinct entity, separate from but linked to, the existing site based access control services.

Although work thus far has concentrated on OpenID (based on analysis carried out in the EU Metafor project), further development work to enable interoperability with Shibboleth (a federated security infrastructure) should also be pursued in order to open up interoperation to a broader range of organisations as represented in the UK Access Management Federation.

## *2.5 Vocabulary Services*

Interoperable metadata requires fields populated using terms from controlled vocabularies under responsible content governance managing the list of terms in the vocabulary and ensuring their meanings are clearly understood. Such governance has long been established in the oceanographic and atmospheric science domains through bodies such as IOC GETADE (Intergovernmental Oceanographic Commission Group of Experts on Technical Aspects of Data Exchange) and CF (Climate and Forecast) Standard Names Committee. However, a sticking point has always been technical governance: accessibility of the vocabularies from a centrally maintained source in a form readily usable by software agents without the need for multiple local copies (that evolve like Galapagos Finches and destroy interoperability).

Technical governance in the form of the Vocabulary Server was one of the first services to be established by the NDG. The system is based on a list of lists, each with its own URI, containing terms, again each having a URI, that are linked together using a set of standard semantic relationships taken from version 1 of SKOS (Simple Knowledge Organisation System). The URIs are resolved as SKOS documents providing term labels, definitions and the mappings between terms thereby providing a simple pseudo-RESTful Web Service interface. A more extensive Web Service API with both pseudo-RESTful and SOAP interfaces was also developed.

The potential of the Vocabulary Server was quickly realised by many communities beyond the NDG. It now forms the semantic framework for SeaDataNet[7] and provides a version-controlled serving mechanism for the CF Standard Names. It is clear from usage statistics (the server took almost 50,000 hits in September 2009 with crawler access disabled) that these are not the only communities using the system operationally.

Vocabulary Server development in NDG-MSI had two objectives. The first was to generate a new release of the Version 1.1 of the API (V1.1.4) that implemented versioned list serving and fixed a number of known bugs. Although previous releases of the server allowed selected lists to be labelled with version numbers, a well-documented caveat was that the latest version was always served. The V1.1.4 release now delivers list content conforming to the version number specified. This allows metadata document managers more control in an active scientific environment where vocabulary content inevitably changes with time.

Secondly, a secured vocabulary editor API and client were developed to empower external content governance authorities to maintain vocabularies held in the server without the need for them to contact BODC to make changes. The API allows both batch and single changes to be made to one or more authorised controlled vocabularies by authenticated users.

### Issues and Future Work

The Vocabulary Server has two restrictions that limit its future growth. First, the system back end only supports internal mappings. Consequently, if a new semantic resource needs to be added then it must

[7] Seadatanet: is an EU funded project aiming to create and operate a pan-European, marine data management infrastructure, see http://www.seadatanet.org/.

be physically imported into the database that supports the server. Whilst this may be feasible for small lists, there are significant semantic resources such as SWEET (NASA ontology), the Marine Metadata Interoperability Ontology Repository and the EU Environmental Thesaurus (GEMET) available. The ability to cross-link Vocabulary Server concepts into these would significantly increase its value as a Semantic Web resource.

Secondly, it needs to be able to support a richer set of semantic relations between concepts than those allowed by SKOS 1 and therefore moving the server payload documents to another standard such as SKOS2 or OWL becomes necessary. The API and payload was informally examined by a computer scientist who provided pointers to resources with interesting functionality that could provide guidance for future development. These are being investigated and their potential for basis of V2 of the Vocabulary Server API is being documented as part of NDG-MSI. Resources to develop V2 of the API have been written into an EU FP7 proposal currently going through negotiation with the Commission.

The utility of vocabulary services is enhanced through the use of tools which exploit the underlying vocabularies. Within the NERC community it would be advantageous to continue developments (such as the CEDA vocabulary editor) which include vocabulary maintenance via the vocabulary web service.

## *2.6 "Browse" Metadata - MOLES*

The Metadata Objects for Linking Environmental Sciences (MOLES) were originally developed within NDG to fill a missing part of the "metadata spectrum", that is, a framework within which to encode the relationships between the tools used to obtain data, the activities which funded their use, and the datasets produced. MOLES would be primarily of use to consumers of data, especially in an interdisciplinary context, to allow them to be able to establish some details of provenance, and to compare and contrast, without recourse to discipline specific metadata or private communications with the original investigators. MOLES might also be of use to the custodians of data, providing an organising paradigm for the data and metadata.

The original (NDG1 and NDG2) MOLES were developed in a vacuum, with little to work with: existing discovery standards were thought to be too high level, and other schema to be too discipline specific, or describing the data itself, rather than the context. Subsequent work, funded within the Centre for Environmental Data Archival (CEDA) exposed many issues with the practicality of using the first version (V1) of MOLES, and resulted in the development of a simpler version (V2) which was deployed during 2009. Both of these versions of MOLES were however poorly documented, and had many other issues, which meant they were not really suitable for deployment in the wider community.

The aims of the NDG-MSI component of work were threefold: (1) to take the lessons learned in early MOLES versions (V1, V2) and develop a more standards compliant information model, (2) to exploit knowledge held in the Australian Commonwealth Science and Industrial Research Organisation (CSIRO) to build tooling which would allow the evolution of the MOLES information model to be more easily managed, and (3) develop a wider NERC community feeding into MOLES model development and as potential users of MOLES (whether internally or as an interface to their internal metadata holdings). All three aims have been met.

All three components of the work have been completed. A fully documented application schema of the ISO standard Geographic Markup Language (GML) has been developed and is available at http://proj.badc.rl.ac.uk/moles. This new "data model" for browse metadata is fully compliant with the relevant standards (not just GML), and XML schema can be automatically generated by the open source HollowWorld UML2XML tool. A small workshop was held bringing together a new community of potential MOLES users, and another workshop is planned imminently. MOLES3 was presented to the European Geophysical Union. Many instances of MOLES are now available as UML examples in the repository. It is likely that MOLES will have application in supporting the Environmental Research Funder's Forum (ERFF) as well as other venues outside of CEDA.

**Problems Encountered**

No significant problems have been encountered during this work, although a number of technical issues remain unsolved. One of the sub goals of the original work was to produce an Atom format serialization of MOLES, and this was postponed in favour of putting more effort into the fundamental data modelling.

**Further Work**

MOLES needs to be confronted with actual applications before it can be successful, and to that end, an important missing component of the available software infrastructure needs to be constructed: a tool for taking a UML domain model and automatically constructing an object to relational database mapping using a high-level object-relational-mapper. With that in place a rapid development cycle could be put in place to confront MOLES with real user experiences: both in creating and exploiting MOLES.

The work on atom serialisation needs to be completed.

There are already obvious lines of improvement to the existing data model, both to support new requirements (such as those of ERFF) and to better improve the usability and understandability of the current model.

## 2.7 Training

In order to share data management expertise across the NERC data management community the project funded, designed and delivered a series of workshops listed below:

- Data and Information Modelling (run by CSIRO & STFC) 12 attendees.
- Structuring Scientific Metadata (run by STFC) 7 attendees.
- Scientific Data Services (run by BAS & STFC) 9 attendees.
- Vocabulary Service (run by BODC & BAS) 12 attendees.

These were fully attended (40 places) and will be run again in the autumn of 2009 with the intention of inviting participation beyond NERC.

A workshop planned to discuss future technical and operational strategy was not delivered because of the co-existence of the NERC strategy implementation activity - the assumption being that the data and information community was already being widely consulted for their input into the new strategy.

## 2.8 Collaboration & Wider Communications

During the project several collaborative opportunities were identified and initial meetings were held with representatives from the following organisations:

- Defra’s Geographical Information Community
- FERA’s Information Management Community
- Oceans2025 Research Programme
- Environmental Researchers Funder's Forum
- Marine Environmental Data and Information Network
- The Centre for Ecology and Hydrology

All of these organisations or communities can benefit from the experience of NDG and several were keen to invest in, use or collaborate with NDG. At least one is currently investigating procuring a service level agreement for STFC to deliver an operational service.

**Problems Encountered.**

In general, the NDG has limited operational capability, as it has not been not yet been funded or resourced to delivery an operational service that has some guaranteed longevity beyond existing short funding cycles. Some sort of longer term funding and future is necessary for both third parties and NERC research centres alike to invest in conforming to NDG requirements.

Some of the front-end interfaces to the NDG, and especially the project website, look dated, despite exploiting cutting-edge technology in the background. A new NDG website is currently being configured to become the first point of call for those seeking more information about what NDG does and has achieved.

# 3. Detailed Recommendations

In this section we make some recommendations based on experience in the NDG projects thus far, and in particular, the NDG-MSI. These are aimed at two audiences: some general recommendations, aimed at the NERC community, and some specific technical recommendations, aimed at implementers of spatial data infrastructures (SDIs, see sections 3.3 and 4.3 ) where the requirements of the SDIs exceed those of currently available commercial products (pretty much any research organisation with complex metadata and/or data which doesn't fit within "normal" GIS systems).

## *3.1 Key principles and assumptions*

Underlying the recommendations to be presented there are some principles and assumptions, which we list here.

1. NERC should look like one entity to outside organisations.
2. Individual NERC centres (whether collaborative or otherwise) are engaged in scientific activities which are likely to push over the boundaries of any a priori standards, whether for data or metadata. Any system cannot record *all* the information, expose *all* the data using "standards compliant" formats, and provide *all* the relevant data services (whether "interoperable" or not). These "boundary activities" are likely to be different in different organisations.
3. Interoperability (see definition in section 4.2 ) is facilitated by using common vocabularies, services and data formats, but such common vocabularies, services and formats cannot be the only vocabularies, services and formats used within NERC centres.
4. Efficiency of work is facilitated by providing, and using where possible, common discovery systems, and common vocabularies, services and data formats.
5. Whatever metadata and data standards are in use within (or to deliver across the boundaries outside of) NERC, are likely to have to evolve rapidly to follow both usage and requirements from within NERC and from external scientific and user communities.
6. The way that scientific users interact with data is likely to be discipline specific, although there is a growing demand for scientific users from one discipline to exploit data from another.

## *3.2 Research & Operations*

It is obvious that the sorts of spatial data infrastructure envisaged and required by the research community are rather more sophisticated than can be delivered with current technology – commercial or otherwise. What's perhaps less obvious at the management level is that it will always be so; the expectations of research communities will always exceed the tools available! Hence, in designing an SDI, one needs to design from the beginning support for evolution but couple that with robust delivery systems, and sensible mechanisms for promoting research & development products into operation.

In previous NDG projects, the line between development products and operational products has been blurred, primarily because previously there hasn't been a community requiring any sustained operational service: the most that anyone has really required is that NDG services are stable for "demonstrations" of functionality. In the NDG-MSI we have both improved the software infrastructure for the discovery service, and migrated it into a more "production" environment. There are now external (non-NERC) communities in the process of negotiating service level agreements for the use of the discovery service. The Vocabulary service, because it became integral to a number of other projects outside of NDG earlier, was already in a more stable state (albeit with some communities desperately wishing it was less stable and evolving faster[8]).

[8] There is always a tension between stable services, and those that meet the requirements of the user community. Where service upgrades are needed because of bugs or lack of crucial functionality, the balance may tip towards requiring less stability and more responsiveness.

The Discovery service is currently run by the Centre for Environmental Archival, CEDA, and the Vocabulary Service by the British Oceanographic Data Centre within NERC. Both groups have invested considerably in these services, and not only with funds which originated via current or previous NDG projects (nor even with NERC funding alone). It is clear then that the evolution of these services needs to be governed not only by what is possible, but by what is needed by the stakeholder community (and such needs are a balance between enhanced functionality and adequate stability).

We recommend that:

1. The operational “NDG” services need a transparent governance body (or bodies) which respects the intellectual property of the originators, the requirements of their stakeholders, and the fact that both already exist in environment where the main stakeholder may be NERC, but that stakeholders are not limited to NERC alone.

Such a body would ensure that the service was governed independently of the requirements of any one stakeholder, as well as provide some assurance to those stakeholders that the services would meet their requirements over long enough periods for them to invest in complying with interface standards of those services.

We would anticipate that individual stakeholder communities would have service level agreements with the service providers. Clearly NERC is one of those stakeholder communities, and there are two possible funding modes that NERC could envisage: (1) individual NERC bodies could have individual service level agreements, or (2) NERC could corporately have one service level agreement with each service provider. Only the second would provide NERC with the ability to evolve the NERC SDI consistently across disciplines, and only the second is consistent with the current NERC science information strategy. Clearly, if NERC wishes to mandate an operational service level, it needs to fund it, therefore we recommend that:

2. The implementation plan for the NERC science and information strategy needs to recognize the need for corporate funding for the “NDG” operational services.

As discussed above, the operational services will need to evolve, both in response to stakeholder push, and external “legislation” (here we include both “real” legislation, such as the implementing rules for INSPIRE, and “de facto” legislation, such as the requirements of global collaborations such as the Global Earth Observation System of Systems, GEOSS). More operational services can be expected. Accordingly:

3. NERC needs to understand that the operational services cannot stand still, they exist in a fast changing technological world, and both will need to evolve to remain relevant. Such evolution will continue to require targeted informatics research and development funding.

The exact mechanisms by which such funding is delivered are out of scope for this report, but for the same reasons that we have suggested centralised corporate funding of the operational systems, we would recommend the same for targeted development for the NERC SDI. Whether the work is done by any one team is different matter, but however it is done NERC should maintain both the capacity and intellectual leadership in SDI development that it already has, since it provides a competitive advantage for NERC science.

Clearly, stability of the operational systems is important, so

4. There needs to be a clear procedure which controls the evolution of the operational Data Discovery Service and the Vocabulary Services so that the level of service is stable and predictable.

And these procedure will need to respect the various service level agreements reached by the governing body.

## 3.3 Spatial Data Infrastructure Architecture.

In this section we discuss specific recommendations about the requirements for spatial data infrastructure, both for NERC, and as foreshadowed in the section introduction, in general.

The basic assumption of a distributed spatial data infrastructure is that it provides integrating views of capability distributed between multiple organisations, that is, an SDI provides discoverable services exposing discoverable products, be they raw data, metadata, or visualisations. Those distributed organisations can (and in the case of NERC data centres, should) have internal data management structures which are more extensive and sophisticated than those exposed into any given SDI, because to an extent, an SDI provides a form of "lowest common denominator" integration. This point is often overlooked by proponents of GIS based SDIs for whom the lowest common denominator is :"what you can put on a map", but of course there is potentially much more geospatial information in common across communities, than can be rendered on a map, and much of that information is research and policy important. Two prominent examples of information products which fall into the latter category are GeoSciML[9] and CSML (see section 2.2 ).

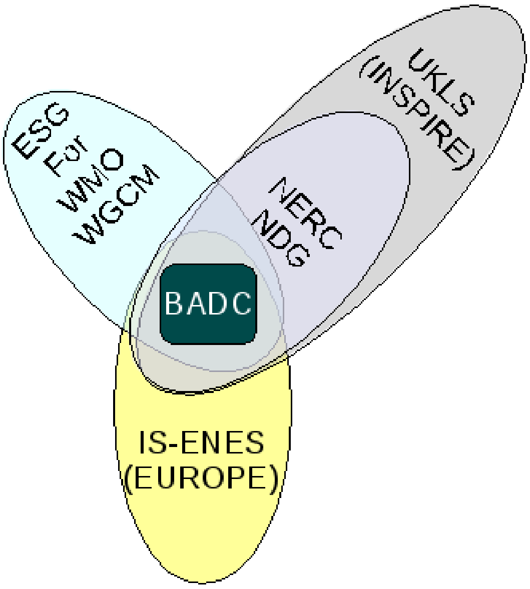

*Figure 2: The BADC in multiple SDIs: part of the Infrastructure for a European Network for Earth Simulation (IS-ENES), part of the Earth System Grid (ESG) supporting the World Meteorological Organisation's Working Group on Global Climate Modelling, and part of the NERC DataGrid, itself to fit in the UK SDI for INSPIRE.*

In practice we might even expect NERC data centres to form part of multiple spatial data infrastructures, with many, but not all, services common between the different SDIs. For example, Figure 2 depicts the BADC in forming spatial data infrastructures with, amongst others, the US Earth System Grid to provide access to petascale climate model archives. The problems and solutions for that activity are different from those required for a NERC spatial data infrastructure (itself, to form part of the UK contribution to a European INSPIRE compliant SDI). Even at the European level we then find a European Network for Earth Simulation, that will be INSPIRE compliant, but much, much more.

### 3.3.1. Generic spatial data infrastructures for scientific communities.

There are three irreducible minimum components of an SDI, content, content exposure, and discovery. That is, we need:

- Repositories of commonly constructed information (content), with tools to create and manage that information;
- some web service to expose that content to prospective users, and;
- some way of finding (discovering) that content.

Other services can (and should, for maximum benefit) exist:

- vocabulary services providing common (and managed) definitions and relationships between entities,
- visualisation services providing a range of graphical views of data,

[9] Geoscience Markup Language, see http://www.geosciml.org/

- data and information manipulation services (ranging from sub-setting to the on demand production of new products from multiple distributed inputs), and;
- portals which orchestrate the services to satisfy actual and potential user requirements.

In most cases, organisations already have content,and so to establish an SDI, one simply (!) needs “common content” paradigms (and methods to migrate the existing content into that common paradigm), exposure systems, and discovery systems which provide search and browse of that exposed content.

Where organisations don't already have suitable managed content, then they can choose to establish managed archives using the “common” format, or they can establish more sophisticated information systems, and follow the same path described in the previous paragraph. The latter approach is almost certainly the approach to take if the organisation has aspirations to use SDI techniques to further scientific goals **within** the community it supports (as opposed to the community the SDI might be aimed at supporting), particularly where one anticipates intersecting in multiple SDIs (as one might expect for any given discipline based activity, where national and international collaborations are discipline based, as well as institutionally based, again, for example, see the BADC role in figure 2).

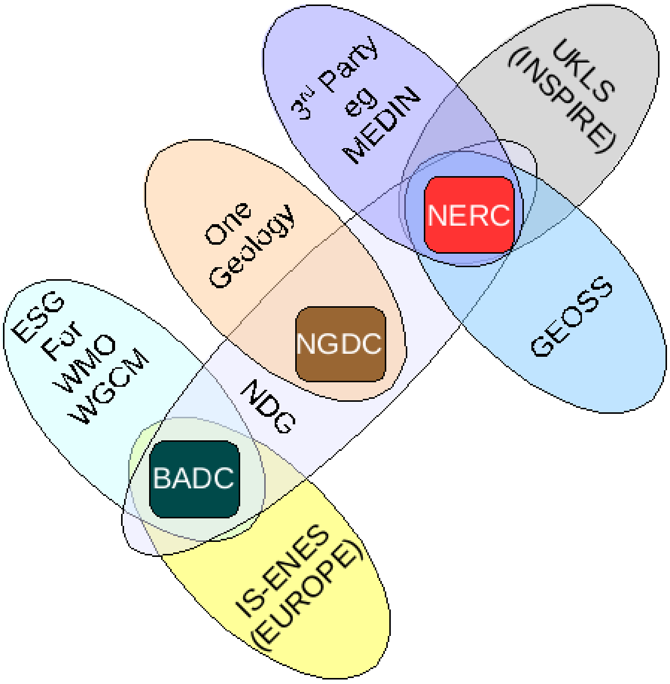

*Figure 3: A NERC SDI based on the NERC DataGrid could allow the discipline specific data centres such as BADC and the BGS National Geoscience Data Centre to concentrate on discipline specific SDI issues, while the NDG SDI team looked outwards at the UK Location Strategy, INSPIRE and GEOSS, and under contract, provides services for 3rd parties such as the Marine Environment Data and Information Network, MEDIN. (In reality of course there would be a number of other NERC data centres involved, each intersecting with their own communities.)*

In both cases, there needs to be at the very minimum, SDI level discovery. While there are proponents of distributed queries, nearly all successful web-based indexing systems which scale to any size use a three step process: (1) harvest information, (2) index the harvested information, (3) provide search interfaces to the harvested indexes.

Thus, we have the following generic SDI requirements:

- Establish a discovery service which can carry out the function described above, and possibly build a portal to that service.
- For each organisation in the SDI, develop or acquire tools which expose internal information through SDI compatible interfaces.
  - Note the assumption is that internal tools exist to collect and organise that information!
- Develop services which expose more than just discovery information
- Develop portals which orchestrate/exploit those extra services in the SDI.

### 3.3.2. The NERC SDI

If NERC were to implement an architecture based on Figure 2, each and every contributing organisation would have to have their own interface with the UKLS, with INSPIRE, with GEOSS and all the other major activities, even as they concentrate on their discipline specific requirements.

A much more effective option is shown in Figure 3, where we see how a NERC SDI (called NDG in the figure) could be used to minimise the overhead of taking part in some classes of SDI, even as all the benefits are accrued.

Accordingly, for NERC, we recommend that:

5. The 'operational' Data Discovery Service (DDS) should become the primary *generic* entry point to NERC data.

6. The team responsible for running and developing the DDS should have responsibility for a NERC interface to generic UK, European and International data discovery activities.

Initially the "NDG" component need not be substantially different from the INSPIRE component .NDG is currently nothing more than a requirement to expose common format metadata and some limited OGC services which can be orchestrated for visualisation in some common portals. The next step in NDG is to move to ISO19115 INSPIRE compliant metadata, at which NDG functionality could be indistinguishable from INSPIRE functionality. However, one would expect the NERC SDI to move past the INSPIRE requirements alone, and look to facilitating NERC interdisciplinary science requirements (by for example, requiring the exposure of MOLES metadata, or something with similar functionality). The details of the future roadmap are unimportant, the point here is simply that one expects a future for a NERC SDI that goes beyond the requirements of INSPIRE alone.

This sort of infrastructure would rely on the individual NERC information holding entities conforming to the common standards of the NERC SDI. Hence one of the main aims of the project was to move NDG discovery metadata to the ISO metadata standards that will enable NERC to fulfil INSPIRE obligations. In order to do this NDG needed an agreed ISO metadata schema for NERC which has only just been agreed (after the main body of work described here). Even now there are some details which have yet to be resolved, and in any case, some evolution can be anticipated.

7. NERC centres and surveys should deliver standards compliant metadata to the DDS using standards consistent with the requirements of the DDS governing body, and expect those requirements to evolve.

8. Not withstanding the existence of generic entry points, discipline specific entry points to NERC data should continue to evolve consistent with discipline specific norms and requirements.

One of the obvious consequences of this architecture, and the necessity for discipline specific centres to engage in their discipline specific SDIs is that NERC will need more than just a central team of experts in this technology – although a central team could and should help coordinate and share information between the discipline specific activities.

NERC has invested considerably in the development of CSML (section 2.2 ) and MOLES (section2.6 ), which address two specific parts of a proposed NERC spatial data infrastructure. However, there are still some significant steps to be taken to be able to fully exploit these information schema. Although both CSML and MOLES are being deployed in mission critical activities now, both need more work before they can be exploited more widely. In the former case, CSML needs to be upgraded to conform to newer ISO standards, and the two reference implementations of CSML tools (in java and python) need more development. In the latter case, the new version of MOLES developed within NDG3 is much superior to previous versions, but there is currently no tooling which can exploit it, and there are cross-NERC applications which need further evaluation before it can be used in such a role.

While this NERC SDI vision is compelling and achievable now, moving beyond discovery and vocabularies (which, although not discussed explicitly above, fit within the same conceptual service pattern) is not yet so easily achieved. While the OGC Web Map Service provides a limited class of maps, even within WMS there are profiles targeted at specific communities, and the wider use of download and more sophisticated visualisations across communities is in its infancy. Similarly, the use of the Web Feature Service to expose complex metadata structures and methods to manipulate them is still not prevalent, not least because developing, populating and exploiting complex metadata structures is not yet easy. Accordingly:

9. Interdisciplinary data access and visualisation depends on tools which have common interfaces regardless of the underlying data formats and metadata standards. NERC needs to continue to invest both in these tools, and the underlying standards, for the foreseeable future.

### 3.3.3. Access Control Infrastructure

The integration of data access across different organisations requires an integrated security infrastructure, at least from the user perspective: in practice, integrating the infrastructure within and between institutions is not possible, but the user experience can be. This is achieved by separating authentication and authorisation in such a way that users can authenticate at one institution, and let that institution assert their identify to others (within a federation using agreed techniques). Authorisation can then be controlled by the institution protecting the resource (data, service, information etc).

Such techniques do not preclude institutions having resources they do not make available to federated users, nor can they guarantee they cannot be hijacked should someone be set upon unauthorised access, but they do allow a wide range of activities to be much easier to carry out with an appropriate (but not foolproof) level of access control.

The NDG has developed tools which are being deployed now, in international federations, and which utilise standards (OpenID and SAML) which are suitable for deployment within NERC for protecting most data resources. These standards do not preclude the use of internal NERC “single-sign-on” technologies, but are designed for between institution use. Many of the known security flaws in the “out-of-box” configuration have been resolved and there are no practical reasons why NERC should not move to deploying this sort of technology now. While there are other possible technologies, the important thing is to get some “federation-level” experience now, rather than wait for a perfect solution.

We recommend that

10. The NERC IT community should support and encourage appropriate “federation level” access control to data over and above existing internal access control mechanisms. In doing so, they must recognise they cannot control or mandate the systems used by all parties, and they should exploit the experience already gained in the NDG team.

## *3.4 Data*

As indicated in section 3.3.1., content is a crucial component of a spatial data infrastructure. Clearly the NERC community has a lot of data, and much of it is in the managed archives of the NERC designated data centres, but even in those data centres the data is not necessarily in a fit condition for exposure to a wider community. Issues which make such data less fit range from documentation issues (it might not be adequate for a non-expert user, it might not be in a format that can be easily parsed), to data format issues (it might not be possible to expose that format data through the services available, or the semantics might not be precise enough for service configuration).

Within all the NDG projects, specific datasets have been “conditioned” for usage in NDG, but in the longer term, it should be the expectation that the majority of NERC data should be conditioned fit for download and visualisation within a NERC SDI (perhaps in addition to whatever conditioning they need for other community activities).

The NERC data centres are already charged with “curating” the data they hold, and it fits well within the definition of curation[10], to expect to add/modify documentation and to reformat (or make copies in alternative formats) in response to the “designated community” requirements. Clearly NERC is part of the designated community, and the requirements of a NERC SDI are therefore within those of the designated community of data centres. Also just as clearly, the requirements of their discipline specific communities remain crucial – without serving their disciplines such data centres would become irrelevant.

Accordingly:

11. Whatever integrating technologies exist, NERC needs to continue to invest in documenting and formatting data to common standards (which are likely to differ between communities) so as to ensure the maximum benefit from the investment in software tools to deliver and visualise such data.

Exactly how that investment is delivered is within the scope of the Science Information Strategy, but we would expect that the NERC designated data centres would be expected to do this for existing data from within their existing funding lines (albeit with a prioritised schedule, and recognition that some data may just be too expensive to migrate into this sort of infrastructure - although even for those data adequate documentation metadata should be available via SDI metadata services). We would expect that the costs of conditioning new data form part of the ingestion budgets associated with those data.

**Value for money?**

It is instructive to estimate the costs of addressing the legacy data and putting it in a fit state for use in a NERC SDI. During the NPP 11 datasets (from BAS, BGS, CEH and the Defra CSL) were “conditioned” and exposed for visualisation in the portal, at an average cost of £4K each. Given 1318 datasets currently listed in the NERC data discovery service, we have a cost of approximately £5M to deliver interoperable visualisation services for the NERC legacy data. To provide additional download services might cost a little more, but we can be reasonably confident that it would be marginal on top of the visualisation costs, so we might anticipate total potential costs of order £6M. (Although as alluded to above, in practice not all data would be suitable for this sort of activity, so the actual costs would be a lot less, and if done within core budgets, delivered over a long period).

The NERC Science Information strategy included an example from Shell International of the cost savings of investing in data management. Their estimates suggest that for their exploration geoscientists, appropriate investment in data management resulted in moving time spent from data discovery and selection from 53% to 30% resulting in an increase in time spent using, interpreting and adding value from 23% to 46%. If we one used that example and extrapolated it to NERC, where the science budget in 2008/09 was £153M[11], and we assumed the same efficiency savings, we'd see approximately £35M per annum more science for an investment of order £6M. Of course these are just back of the envelope calculations, we might expect for example, that the existing investment in data management by NERC would have already made much of these savings, but they are indicative that there are major savings to be made by investing in making it easy for scientists to find and manipulate (easily) the right data.

## 3.5 Community Data Infrastructures

The uptake of NDG technologies and services by the NERC community (let alone the wider community) is predicated on knowing that NERC is committed to running and developing these services in the

[10] Here we define curation as performing the functions of an ISO 14721 compliant open archival information system (OAIS).

[11] 2009 Annual Report, page 41.

medium term (not just on year-by-year) project funding. Without such a commitment, both in aim (which now appears in the new science information strategy), and in funding (which we expect to appear in the implementation of that strategy), neither the NERC nor wider community, will seriously engage with NDG.

12. NERC should make a clear statement about the future and finances of the discovery and vocabulary services (and any other services developed in the future), so that the NERC community, and potential data providers and users can invest in the use of those services with confidence.

One of the recommendations from a previous project, the NPP, was that NERC should encourage the use of the prototype NDG based NERC SDI by targeted community website development, with pump-prime funding from NERC for a number of such activities. Besides the obvious benefit of delivering specific community requirements, use of funds in this way would inculcate an expectation (and actuality) that exploiting an existing SDI was both economic and technically efficient. In the longer term then we might expect future data delivery website developments would do this as a matter of course.

CEDA is following exactly this pathway with the development of the QUEST Earth System Data Initiative, where data will be delivered using OGC services, orchestrated by a portal orchestrating those services. However, while the advantages of doing it this way are obvious to the CEDA development team, they are less obvious to others in the wider NERC community, where the relevant expertise is less prevalent. The NPP project board argued, and the NDG-MSI team agree, that a pump priming programme would develop that expertise more widely. Accordingly:

13. NERC should provide funds to kick-start the development of community data infrastructures that integrate into the corporate data infrastructure as recommended by the Data Portals Project Board in June 2008.

There are communities outside of NERC which recognize the utility of the existing NDG SDI approach. Several parts of Defra have contacted CEDA, with an aim to exploiting NDG services, and the Marine Data & Information Network (MEDIN) are in the process of negotiating a commercial service level agreement for NDG service delivery. However, all of these communities are worried that NERC is not committed to them long term, and to an extent their own investment is predicated on NERC commitment. (The MEDIN team have taken the plunge based on the wording of the draft NERC information strategy, plus the commitments of the STFC team to migrate MEDIN support within whatever results within NERC.)

Clearly commercial service developments provide both cost recovery opportunities for NERC, and demonstrate NERC technology knowledge transfer. However, such commercial engagement means that these groups become stakeholders with an interest in service governance.

14. NERC should expect and encourage (and where possible, allow the commercial sale of ) access to NERC operational services (and recognize the governance implications, which is implicit in our first recommendation under research and operations).

## 3.6 Marketing, Branding & Communication

A key part of encouraging both metadata input to, and usage of, the NERC discovery and vocabulary services is that their portal interfaces should clearly belong to NERC, even though they are delivered by CEDA and BODC respectively. Accordingly:

15. The ‘look and feel’ of the NERC single entry point to its data should be more closely aligned to the NERC corporate identity.

As we have taken the NDG concepts into the wider community, it has become apparent that the name "NDG" carries two sets of unwelcome baggage, even as it grows positive brand recognition. The unwelcome baggage is that:

- Although envisaged as a "grid" in the sense understood by the Open Grid Forum (OGF), what has been delivered is a spatial data infrastructure as understood by the Open Geospatial Consortium. This leads to confusion.

- Because the original NDG was conceived by, and implemented within, the meteorology and oceanography communities (and sub-communities at that), the wider community imagine limitations of applicability that do not apply.

Accordingly:

16. NDG should be renamed to reflect that it isn't a data grid but a set of technologies and methodologies suitable for wide environmental applicability.

Clearly, renaming will imply some loss of brand recognition, but in practice NERC has not invested in marketing the NDG externally, nor has it invested in communicating it's benefits internally (beyond the NDG-MSI workshops).

17. NERC should require and fund marketing and communications activity as an integral part of developing and expanding the (NDG) services.

In doing this, it is perhaps worth recognising that to appeal to a non-academic audience, as one might wish to do, both for KT, and to attract commercial support, NERC could work in partnership with web-design companies, recognising the fact that website design is not a core skill of NERC scientists. (The importance of branding and appeal was recognised within the NDG-MSI as discussed in section 2.8 ; steps have been taken to address some of the fundamental problems within existing NDG websites.)

It is clear from the impact of all the NDG projects that interest from both potential data suppliers and potential data users is enhanced by the presence of both operational reliable services/portals and more edgy prototypes showing what can or will be achieved in the near future. With regard to the latter it is also clear that no web based service can remain static in the face of rapidly evolving user expectation. Accordingly:

18. NERC should continue to invest in having a showcase prototype portal demonstrating what can be achieved beyond the current operational activity (which might include demonstrating new service capabilities as well).

## *3.7 Knowledge Exchange*

The main emphasis of NERC knowledge exchange is on communicating environmental knowledge, and the tools developed and deployed within the NDG clearly contribute in a major way to those goals. However, the informatics tools developed within NDG are of interest in their own right (for example, NERC has funded the C-SEKT project to build on NDG tools and provide them to, amongst others, the Met Office). Many of the NDG tools and prototypes are being showcased by Defra in the context of the UK implementation of INSPIRE, not least because the NDG team are amongst the world leaders in developing and deploying the OGC compliant services which will form the heart of INSPIRE delivery.

The workshops held under the auspices of the NDG-MSI have been very successful, and it is obvious that a wider market exists, not only within the NERC community, but outside.

Accordingly:

19. NERC should refine the organisation, delivery and marketing of the informatics courses created by this project.

# 4. Useful Definitions

In this section we provide definitions for the concepts of metadata, interoperability, spatial data infrastructures, mashups, fusion and our usage of the term “standards”.

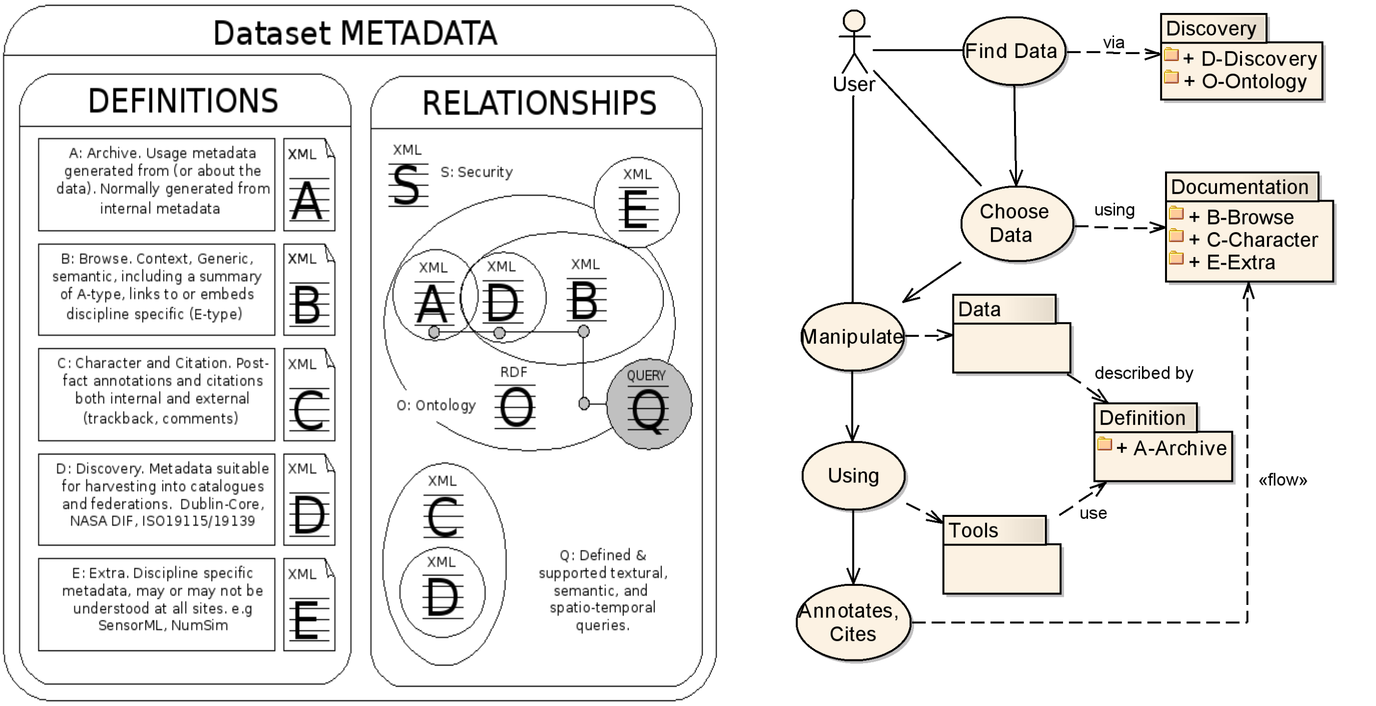
Dataset METADATA
DEFINITIONS
A: Archive. Usage metadata generated from (or about the data). Normally generated from internal metadata
XML
A
B: Browse. Context, Generic, semantic, including a summary of A-type, links to or embeds discipline specific (E-type)
XML
B
C: Character and Citation. Post-fact annotations and citations both internal and external (trackback, comments)
XML
C
D: Discovery. Metadata suitable for harvesting into catalogues and federations. Dublin-Core, NASA DIF, ISO19115/19139
XML
D
E: Extra. Discipline specific metadata, may or may not be understood at all sites. e.g SensorML, NumSim
XML
E
RELATIONSHIPS
XML
S
S: Security
XML
E
XML
A
XML
D
XML
B
RDF
O
O: Ontology
QUERY
Q
XML
C
XML
D
Q: Defined & supported textural, semantic, and spatio-temporal queries.
User
Find Data
via
Discovery
+ D-Discovery
+ O-Ontology
Choose Data
using
Documentation
+ B-Browse
+ C-Character
+ E-Extra
Manipulate
Data
described by
Definition
+ A-Archive
Using
Tools
use
«flow»
Annotates, Cites

*Figure 4: Categories of metadata, and usage sequence. Users begin by finding data in discovery services, perhaps exploiting an ontology to locate things via a traversal from their vocabulary to that used to store the data, then they choose data by browsing between datasets and examining contextual and other detailed metadata. Having obtained datasets, they manipulate data using tools (or home-grown software) which are informed by the metadata describing the layout and meaning of the data objects themselves. Ideally, having analysed the data, they cite the data in publications and/or annotate the data collections directly.*

## 4.1 Metadata

At the scientific level, scientists need to document their data to varying levels of complexity, depending on the likely maturity of the users. The same is true of any metadata systems, and so it is helpful to categorise “metadata”, that is information about data[12].

1. At the most detailed level, scientific metadata may not be not useful to others, but still need to be recorded. It's a moot point as to whether an organisation wishes to capture all such metadata. Tools which manipulate data need to exploit metadata describing the layout and structure of the data, phenomenon names etc (this metadata appears as A-type metadata in Figure 4)
2. There is an intermediate level of detail which is of use to potential users of the data, providing enough detail and context for someone else to use the data without contacting the originator. This sort of metadata is both enough to enable the choice between similar datasets, and to provide adequate scientific provenance. Some of this metadata could be constructed using common cross-disciplinary standards (this metadata appears as B-type metadata in Figure 4), and some requires more specific detail encoded using discipline specific standards (termed as E-type metadata in Figure 4).

[12] A fuller discussion of this material appears in Lawrence et.al., 2009 in Phil. Trans. Roy. Soc. A, 367, 1003 - 1014. doi:10.1098/rsta.2008.0237.

3. Scientists often classify data, using annotations and citation, and capturing such "ranking" material to define the "character" and "fitness" of the data is an important activity. (This metadata appears as C-type metadata in Figure 4).
4. There is a high level "catalogue" level of detail, which is useful for organizing information and providing data "discovery" (usually via some sort of directed search exposing key parameter, services, or characteristics). This sort of metadata is usually only enough to advertise the presence and potential usefulness of data. (This metadata appears as D-type metadata in Figure 4).

The various types of metadata required to manage data are normally created by different individuals, using different tools. Key tools that need to exist include:

7. Graphical user interfaces (GUIs) which are designed so that humans can enter both selections from controlled vocabularies describing the data and free text.
8. Automatic metadata production tools which are either
   a) run over data products produced by instruments or software, run with or without human intervention (possibly, in the former case, including the addition of human generated metadata), or
   b) part of the software inherent in the instrument or software so that the original human operator has configured the instrument or software to produce the appropriate metadata. (Inevitably, with time, this class of metadata needs to be supplemented by metadata created by one of the previous methods, as metadata requirements generally evolve faster than the internal capability of instruments and production software).
9. Vocabulary services exposing controlled vocabularies which can be used by the GUIs to provide vocabulary selections, and by all metadata tools to validate metadata entry which is expected to be from controlled vocabularies.
10. Tools to create controlled vocabularies, which in practice means more GUIs, along with tools that can mine vocabularies from free text.
11. Tools which manage vocabularies and expose appropriate interfaces (generally web service interfaces unless these tools are to be in tightly coupled software systems, and not intended to expose the vocabularies to wider communities).
12. Tools which manage metadata (in various schema) and expose appropriate interfaces (generally web service interfaces unless these tools are to be in tightly coupled software systems, and not intended to expose the metadata to wider communities).

## *4.2 Interoperability*

The ability to exchange and use information requires:

1. Tools to exchange information (and the ability to use those tools). Where those tools consist of systems that integrate information from a variety of systems, *without special effort*, then we can talk about *service interoperability*.
2. The ability to interpret and use the information when it is acquired. When that use involves integrating information from disparate sources, we want *data interoperability* - that is, we want to be able to integrate the data/information to a common format/view *without special effort.* This last requires a range of activities:
   a) Standardised formats
   b) Standardised metadata structures, and
   c) Standardised (controlled) vocabularies, with
   d) Ontologies to mediate between community standard vocabularies.

e) Tools which can exploit vocabularies to mediate for humans (“reasoning” agents etc).

It's important to understand that interoperability is nothing more than the ability to do something quickly and easily without special effort by the user, it generally doesn't involve the doing of something that otherwise couldn't be done. What it does do is:

3. Enable the doing of something that wouldn't otherwise be done because it would take too long, or it's too much effort, (e.g. a wider spectrum of hypothesis testing), and/or
4. Provide efficiencies of scale, if a task needs to be done many times (by different communities or users), then it obviates the necessity for the development of specific solutions by each community/user (thus saving in time and cost).

Practical scientific uses of interoperable services include providing both “quick look” visualisation and sophisticated graphics, with relatively little effort by the consumer, data manipulation services, and a host of other useful tools for the scientific toolbox.

## 4.3 Spatial Data Infrastructure (SDI)

Spatial data infrastructure (SDI)[13]: A framework of spatial data, metadata, users and tools that are interactively connected in order to use spatial data in an efficient and flexible way. Another definition is *the technology, policies, standards, human resources, and related activities necessary to acquire, process, distribute, use, maintain, and preserve spatial data.*

Some of the main principles are that data and metadata should not be managed centrally, but by the data originator and/or owner, and that tools and services connect via computer networks to the various sources … To achieve these objectives, good coordination between all the actors is necessary and the definition of standards is very important.

## 4.4 Mashup

Mashup[14]: A derivative work consisting of two pieces of (generally digital) media conjoined together, such as a video clip with a different soundtrack applied for humorous effect, or a digital map overlaid with user-supplied data.

Mashup[15]: (web application hybrid), a web application that combines data and/or functionality from more than one source

In the web context, a mashup is generally a temporary construct, and in the context of this document, the place of mashups is in the construction of maps/views which show the spatial relationships between different data/information entities. Such mashups are generally done as part of hypothesis testing of the sort "is there a spatial relationship between these two quantities worth pursuing?" or "what data/information is available in the neighbourhood of this spatial feature?"

The delivery of mashups requires interoperable services capable of providing views of data in (or on) a common visualisation paradigm (e.g. a map) (Almost by definition, a mashup occurs because of the use of interoperable services, while one might achieve the same result - map or whatever - via a different technique, it wouldn't be a mashup without the underlying assumption that it was delivered via interoperable web services.)

## 4.5 Data Fusion

Data Fusion[16]: The use of techniques that combine data from multiple sources and gather that information in order to achieve inferences, which will be more efficient and potentially more

[13] From wikipedia: http://en.wikipedia.org/wiki/Spatial_data_infrastructure, accessed 30 September, 2009.

[14] From wiktionary: http://en.wiktionary.org/wiki/mashup, accessed 17 June, 2009.

[15] From wikipedia: http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid), accessed 17 June 2009.

> accurate than if they were achieved by means of a single source ... combines several sources of raw data to produce new raw data. The expectation is that fused data is more informative and synthetic than the original inputs.

It's clear that data fusion is a more mature activity than a mashup: the expectation is that data fusion results in a product, and that the product is "more informative" than the individual constituents. One might be tempted to suggest that a data fusion product is more than the sum of it's parts! In the NERC context data fusion is generally a scientific activity.

Another point of distinction between data fusion and mashups are that the latter are done "geospatially", data fusion can occur along any useful axis of the resulting data object (e.g. time, wavelength etc).

It's also clear that data fusion doesn't *require* service interoperability, nor does it *require* data interoperability, but it's obvious that delivering data fusion is easier with data in common formats, described in a common manner (data interoperability). It's also obvious that data fusion possibilities can be explored more quickly by using interoperable services.

## *4.6 Standards*

Throughout this document we have used the phrase “standards” and “standardised”. These phrases are meant to imply protocols and implementations which conform to something which has either been through a de facto or de jure process of community definition.

The issues for the applicability of standards resolve to “fitness for purpose” and “community acceptance”.

For some applications, the “community” will be NERC, for some it will be discipline specific, for others national or international federations (such as INSPIRE and GEOSS). Clearly then, the appropriate standards are those which are fit for the community use, and accepted within the community. It is quite likely that at times one NERC organisation may seek to push the same information through more than one “standard” because it is part of more than one “community”. That is to be expected!

[16] From wikipedia: http://en.wikipedia.org/wiki/Data_fusion, accessed 17 June, 2009.