Enterprise specification of the NERC DataGrid

Andrew Woolf[1], Ray Cramer[3], Marta Gutierrez[2], Kerstin Kleese van Dam[1], Siva Kondapalli[3], Susan Latham[2], Bryan Lawrence[2], Row Lowry[3], Kevin O'Neill[1].

[1]CCLRC e-Science Centre
[2]British Atmospheric Data Centre
[3]British Oceanographic Data Centre

Abstract

NERC DataGrid (NDG) will provide discovery of and access to a range of environmental data. Here we describe the deployment of the Reference Model for Open Distributed Processing (RM-ODP) to architect NDG development. RM-ODP was adopted as a formal architecture methodology because of the close match between Grids and ODP system concepts. The process provides a number of views of the system being designed – in this paper we concentrate on the Enterprise view. The RM-ODP Enterprise language specifies the purpose, scope and policies of a system using readily-understood concepts. It provides a structured approach to requirements capture and analysis. An ODP *community* corresponds to a classical Grid virtual organisation (VO). *Roles* are identified for VO participants, together with *activities* they engage in, and governing *policies*.

1. Introduction

The complexity of Grid infrastructures demands a structured approach to software engineering. This is a prominent distinguishing feature of many "e-science" projects over other scientific computing efforts which generally favour a more ad-hoc development path. Indeed, an undoubted benefit of the program is that practising scientists are exposed to state-of-the-art information technologies and methodologies. (Of equal benefit is for computer scientists to learn that FORTRAN remains the language of choice for many complex research codes.) There is as yet, however, no consensus on the respective merits of different Grid design techniques. The NERC DataGrid project [1] has adopted the Reference Model of Open Distributed Processing [2,3,4,5,6] (RM-ODP) as a formal architectural framework. We describe elsewhere the considerable synergy that exists between Grid and RM-ODP concepts, and concentrate here on the requirements capture and analysis aspects of the design problem.

The remainder of this paper is organised as follows: section 2 introduces RM-ODP and its application to requirements capture and analysis, section 3 describes its use with NERC DataGrid, section 4 discusses some practical lessons from the process, and section 5 provides some conclusions.

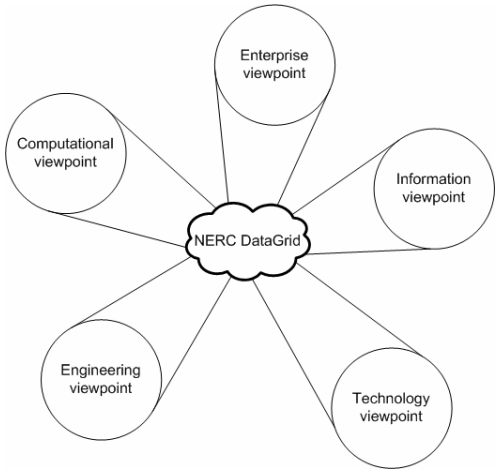

Figure 1: RM-ODP viewpoints approach

2. RM-ODP background

RM-ODP uses a viewpoint approach to specifying the architecture of a distributed system. A viewpoint on a system is an abstraction of the system specification focussing on a particular set of concerns (Figure 1). The following five viewpoints are defined in RM-ODP:

1. **Enterprise viewpoint**: concerned with the purpose, scope and policies governing the activities of the system

2. **Information viewpoint**: concerned with the semantics of information and information processing in the system
3. **Computational viewpoint**: a functional decomposition of the system in terms of computational objects and their interfaces
4. **Engineering viewpoint**: concerned with the infrastructure required to support distribution. Whereas the computational viewpoint is concerned with when and why objects interact, the engineering viewpoint is concerned with how they interact.
5. **Technology viewpoint**: specifies particular technology choices for the system

While the RM-ODP standards [3,4,5,6] do not specify how the viewpoints should be used in practice, typically the Enterprise viewpoint is applied first to produce an overall description of the system. It provides a structured framework for requirements capture and analysis.

The Enterprise viewpoint [7] represents a system and its environment as a *community* of *enterprise objects* which is formed to meet some objective. An RM-ODP community corresponds to the classical Grid concept of a virtual organisation (VO) [8]. The Enterprise specification defines the VO participants, *roles* they play, *activities* undertaken, and *policies* that apply.

The RM-ODP does not mandate any particular specification language for an architecture. For NERC DataGrid, the Unified Modelling Language (UML) was chosen due to its maturity and the ready availability of tools. For the Enterprise specification, RM-ODP activities were factored into themed UML *packages*, with a UML *Use Case* detailing each activity, and UML *Collaborations* providing summary overviews. RM-ODP roles were factored on the basis of logical aggregation of RM-ODP *behaviour*, and represented as UML *Classifier roles* in the activity collaborations. The use of UML for an RM-ODP architecture specification is the subject of a current standardisation activity[1].

RM-ODP seems well suited as a Grid architecture framework. The publish-find-bind pattern of service-oriented architectures, for instance, is directly captured in Enterprise Viewpoint activities; web service portTypes correspond to Computational Viewpoint interfaces; and choices over WSRF or OGSI may be relegated to the Engineering Viewpoint. RM-ODP also defines a set of utility functions supporting common patterns of distributed architectures – security, repositories, coordination (e.g. notification, replication). These are important elements of many Grids. The depth of the RM-ODP – Grid synergy requires further research to establish, only the basic principles of the approach have been utilised here.

A number of viewpoint correspondences in the RM-ODP meta-model ensure consistency of an architecture. For example state changes in the Information specification occur at Computational interfaces, which support activities in the Enterprise specification.

3. NERC DataGrid: Enterprise specification

We describe here key elements of the NDG Enterprise specification – a summary of the requirements capture and analysis. NERC DataGrid (NDG) aims to enable search and discovery, and provide seamless access to a range of environmental data held across a number of NERC designated data centres[2] and research groups. The transition from data discovery to access will be transparent, with details of data location and storage format encapsulated. A wide range of data types will be supported including both observational (ship- and airborne instruments, radar, mooring, etc.) and model (e.g. numerical weather forecast, climate simulations).

3.1. NDG virtual organisation

The NDG virtual organisation (RM-ODP 'community') includes researchers, students, data centres and university research departments, and has the objective of exploiting UK environmental science data. The objective of NDG within the VO is to facilitate discovery of and access to this data.

3.2. NDG roles

The following roles are identified for NDG participants:

- **User**: the main role in NDG; participates in a large range of NDG activities including search and discovery, data browsing and delivery.

[1] The ISO/IEC JTC1/SC7/WG19 is developing ISO/IEC 19793 "Information technology – Open distributed processing – Use of UML for ODP system specifications".

[2] The initial focus of NDG is on data hosted by the British Atmospheric Data Centre, and the British Oceanographic Data Centre – two of the designated data centres of the UK Natural Environment Research Council (NERC).

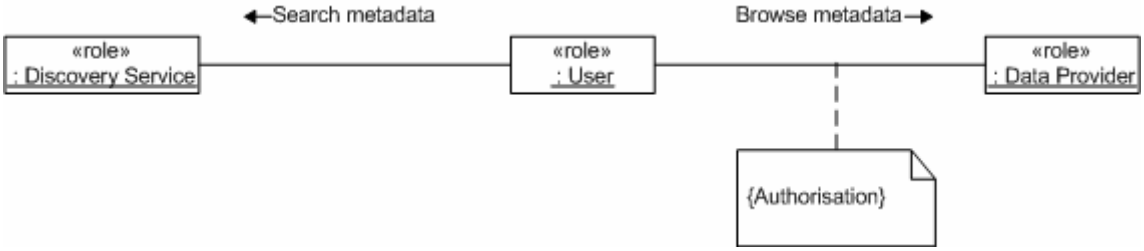

Figure 2: UML collaboration for NDG 'Search and discovery' Enterprise activities

- **Data Provider**: A Data Provider is a role concerned with data provision and delivery in NDG. It also answers requests for detailed metadata describing a dataset and provides discovery-level metadata for harvesting by a Discovery Service.
- **Discovery Service**: Provides the search and discovery facility for NDG. It harvests discovery metadata from Data Providers.
- **Delivery Broker**: Mediates requests for data across one or more Data Providers.
- **Attribute Authority**: Assigns security roles to Users for the purposes of data and metadata access control.
- **Workspace**: A persistent workspace exists for each registered user, providing storage, logging, and scripting facilities for the user.
- **Workspace Provider**: Supplies resources used for Workspaces.
- **NDG Manager**: Maintains registries of Data Providers, Discovery Services, etc.

3.3. NDG activities

NDG activities are factored around the following themes:

- **Search and discovery**: searching over discovery metadata (both free-text and guided); browsing of detailed metadata for datasets of interest. The metadata structures in NDG are described elsewhere [9,10,13], including a taxonomy of metadata types. Metadata 'browsing' here corresponds to interrogating the 'B' metadata of that taxonomy (describing relationships between data 'activities', 'observation stations', 'platforms' etc.)
- **Data browse and delivery**: browsing of dataset details, selection of data subsets, and delivery of data through one of several access methods. 'Data browsing' corresponds to interrogating details of the 'A' metadata of the abovementioned taxonomy. Dataset structure is defined in terms of the data model described by Woolf *et. al.* [11]. Subsets of data may be selected in time and space, and by parameter within the dataset. Selections may be made across multiple datasets. Delivery of a selected set of data is offered through any of a number of different mechanisms, including instantiation in a specified file format, or via a third-party service such as OPeNDAP or OGC web services. Since data products may be joined across more than one Data Provider, such requests are mediated by a Delivery Broker.
- **Workspace management**: provision of resources for Workspaces, and interaction of a User with Workspace facilities – including storage of metadata/dataset references, queries and results etc, a personal history log, and a workflow engine.
- **Metadata management**: updating of metadata and datasets, and harvesting of discovery metadata; registration of Data Providers and Discovery Services. The digital library protocols of the Open Archives Initiative [12] are used for metadata harvesting. This allows Data Providers and Discovery Services to federate as well with external OAI clients or servers.
- **User administration**: logging in and out of NDG, and assignment and retrieval of security credentials.

3.4. NDG policies

Policies in RM-ODP are specified as obligations, permissions, or prohibitions that apply to various activities. NDG has specified policies concerning security, resource usage and quality of service. NDG security policies are factored along the conventional 'AAA' lines:

- **Authentication**: All interactions (except discovery metadata searching) in NDG are required to be mutually authenticated using PKI. Authentication patterns like single sign-on and delegation are enabled through x.509 proxy certificates.
- **Authorisation**: Users are granted authorisation attributes (access control

Use Case Name:	Search metadata
Description:	A User defines and executes a search against discovery metadata.
Normal Course:	1. User specifies text for free-text metadata query. 2. Optionally, a delimiting geographical and/or temporal region may be specified. 3. The result set is configured ('full metadata record', 'summaries only', 'title and data links' etc), number of 'results-per-page' optionally specified. If search is interactive, specify whether results are to be 'displayed' or 'downloaded'. 4. The query target is specified: against the complete metadata store, the previous result set, or specific records. 5. The query is executed by the Discovery Service, and results returned to the User and displayed. 6. The result set may be explored further by the User by requesting subsequent pages of results or expanding the amount of detail displayed. 7. Individual records may be selected against which to run subsequent refining queries, or a single record selected to browse the detailed metadata (activity 'Browse metadata').
Alternate course:	*(Structured search query)* 1. A guided search may be constructed by selecting query parameters from lists of controlled vocabularies.
Notes:	

Figure 3: UML Use Case for 'Search metadata' activity

roles) by Attribute Authorities (typically individual Data Providers). Access to detailed metadata or data may be controlled on the basis of these authorisation attributes. Furthermore, mappings may be defined by Data Providers from attributes prescribed by other authorities to their own. For example, BADC may choose to map a BODC authorisation attribute to one of its own, on the basis of level of trust and the attribute definition published by BODC. Authorisation in NDG is discussed in more detail by Lawrence *et. al.* [13].

- **Accounting**: All client-server interactions in NDG must be logged by the server.

As an example, Figure 2 is a UML collaboration showing the Enterprise viewpoint activities in the 'Search and discovery' package. User, Discovery Service and Data Provider roles are involved, and the authorisation policy constrains metadata browsing. The Use Case for the 'Search metadata' activity is detailed in Figure 3.

Note that while the Enterprise viewpoint specifies policies such as authorisation patterns, the implementation is the domain of the Engineering and Technology viewpoints. Thus, desired behaviour is specified separately from decisions about particular technology choices (PERMIS, VOMS etc).

4. Perspectives and lessons

The use of RM-ODP for capturing NDG requirements through the Enterprise specification proved valuable for a number of reasons.

First, it provided a structured approach to elaborating and analysing NDG requirements. The various stakeholders (primarily NERC designated data centres, a selection of prospective NDG users, and implementers) could arrange their conception of NDG in a structured way in terms of roles, activities and policies. (Or, at least, the analysis of requirements could be structured in this manner.) The elaboration of the requirements proceeded via a series of meetings, focussing on specific aspects of the Enterprise specification. Issues identified in the process of working up and analysing the discussions following a meeting provided a focus for a subsequent meeting. This staggered approach had the added benefit of educing assumptions and conflicts that may not be apparent at the level of a coarse requirement.

As a specific example, it was clear (in version 0.3 of the architecture) that a role-based access control mechanism would be used with security attributes assigned by Data Providers. However, it took some considerable discussion amongst stakeholders to agree for the next revision on the mechanism for, and granularity

of, federating authorisation within the virtual organisation. The principle eventually agreed was that Attribute Authorities (a role typically fulfilled by a Data Provider) would publish granting policies for any or all of the roles they assign, and that a resource owner (Data Provider) would be able both to define access in terms of direct authorisation requirements, and to define equivalence mappings between credentials for that purpose. A User denied a request for a resource is informed of those credentials required for success – a non-persistent 'attribute wallet' is populated with relevant attribute certificates during the course of an NDG session.

Similarly, details of what behaviour (i.e. Enterprise viewpoint activities) was intended by 'metadata searching', a 'user workspace', etc., came about only through focussed discussion.

Implementation details in RM-ODP are confined to the Engineering and Technology viewpoints. This clear separation of concerns was useful in the requirements capture in order to prevent system scoping from drifting towards implementation details. This was useful in NDG in determining the security requirements, for instance – stakeholders focussed their attention on defining Enterprise security policies and behaviour, rather than a technology-driven approach of considering how specific implementations might be employed.

It was very important, through the process, to involve individuals with a thorough understanding of the current operational technology base of NDG partners (NERC designated data centres). The myriad issues associated with the complexity of real-world environmental datasets, operational data management infrastructures, service-level agreements, data supply chains, deployment practicalities etc. provided important 'reality-checks' during the analysis process.

Finally, because of the implicit correspondence between viewpoints in RM-ODP, the Enterprise specification of NDG provided important guidance for its architecture based on the requirements analysis. For instance, a number of Enterprise viewpoint activities divide very naturally along lines of data and metadata – these become fundamental divides of the Information viewpoint. (The schema for metadata [9,10] and data models [11] were mentioned earlier, and are key components of the Information viewpoint.) Interfaces defined in the Computational viewpoint were motivated very strongly by the activities and behaviours specified in the Enterprise viewpoint. Thus, interfaces and operations associated with Data Delivery correspond directly to the Enterprise activities 'Get delivery options' and 'Deliver data'.

5. Summary and conclusions

RM-ODP is a methodology for architecting distributed systems that structures a specification into five complementary viewpoints: the Enterprise, Information, Computational, Engineering and Technology viewpoints. Requirements capture and analysis corresponds to formulating the Enterprise specification. System requirements are specified as a series of activities involving a number of identified roles. These activities are constrained through stipulated policies.

RM-ODP was found to be useful as a framework for the formal architecture specification of NERC DataGrid. We have described key components of the Enterprise specification of NERC DataGrid, and how it was used to structure the requirements capture and analysis.

There is a good match between RM-ODP and Grid concepts, and the methodology may find general utility as a structured approach to specifying Grid architectures.

Finally, there has been some high-profile debate [14][3] and published research [15] on what exactly should be meant by a "Grid". This confusion undoubtedly contributes to the perception of 'hype' around Grid that has been identified [16]. The elaboration of an agreed semantics for the Grid would considerably advance its cause and carve out a clear space in the broader sphere of distributed computing. The use of RM-ODP as a structured approach to specifying Grids across a number of application domains and projects would contribute to the identification of common patterns and provide data to explore Grid semantics.

[1] Lawrence, B., *et. al.*, "The NERC DataGrid prototype", UK e-Science All-Hands Meeting, Nottingham, UK (2003). [Online: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/065.pdf, 23 June 2004]

[3] And see follow-up correspondence from Sun Microsystems (GridToday, vol. 1, n. 8), William Johnston (GridToday, vol. 1, no. 9) and others – referred to as 'the Grid identity crisis' in the editorial of GridToday vol. 1. no. 10 (http://www.gridtoday.com/02/0819/020819.html)

[2] Putman, J.R. (2001), "Architecting with RM-ODP", Prentice Hall PTR, New Jersey, 2001.
[3] ISO/IEC 10746-1:1998, "Information technology – Open Distributed Processing – Reference model: Overview"
[4] ISO/IEC 10746-2:1996, "Information technology – Open Distributed Processing – Reference Model: Foundations"
[5] ISO/IEC 10746-3:1996, "Information technology – Open Distributed Processing – Reference Model: Architecture"
[6] ISO/IEC 10746-4:1998, "Information technology – Open Distributed Processing – Reference Model: Architectural semantics"
[7] ISO/IEC 15414:2002, "Information technology – Open Distributed Processing – Reference model – Enterprise language"
[8] Foster, I. *et. al.*, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations", Int. J. HPC Apps., **15**(3), 200-222 (2001).
[9] O'Neill, K. *et. al.*, "The metadata model of the NERC DataGrid", UK e-Science All-Hands Meeting, Nottingham, UK (2003). [Online: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/129.pdf, 23 June 2004]
[10] O'Neill, K., *et. al.*, "A specialised metadata approach to discovery and use of data in the NERC DataGrid", UK e-Science All-Hands Meeting (2004).
[11] Woolf, A., *et. al.*, "Data virtualisation in the NERC DataGrid", UK e-Science All-Hands Meeting, Nottingham, UK (2003) [Online: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/094.pdf, 23 June 2004]
[12] OAI, "The Open Archives Initiative Protocol for Metadata Harvesting", Open Archives Initiative, Protocol Version 2.0 (2002) [Online: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, 23 June 2004]
[13] Lawrence, B., *et. al.*, "The NERC DataGrid: 'Googling' secure data", UK e-Science All-Hands Meeting (2004).
[14] Foster, I., "What is the Grid? A three point checklist", GridToday vol. 1 no. 6 (2002) [Online: http://www.gridtoday.com/02/0722/100136.html, 23 June 2004]
[15] Németh, Z. and V. Sunderam, "Characterizing Grids: Attributes, Definitions, and Formalisms", J. Grid Comp., **1**, 9-23 (2003).
[16] Barlow, R., "The Grid – Hype or Reality?", International Symposium on Modern Computing in honour of John Vincent Atanasoff, Iowa State University (2003).