

The ^{my}Grid Information Model

Nick Sharman¹, Nedim Alpdemir¹, Justin Ferris², Mark Greenwood¹, Peter Li³, Chris Wroe¹

¹ Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL

² IT Innovation Centre, University of Southampton, Southampton SO16 7NP

³ School of Computer Science, University of Newcastle-upon-Tyne, Newcastle-upon-Tyne NE1 7RU

Abstract

^{my}Grid aims to develop high-level middleware to support an e-scientist in conducting *in silico* experiments in biology. An important part of this is to define a conceptual space – the ^{my}Grid information model – where both biological data and experimental processes are modelled, with emphasis on the latter. As such, it defines the basic concepts through which different aspects of the e-science process are represented and linked. The information model provides shared data abstractions that underpin important service interactions and so promote synergy between ^{my}Grid components. This paper describes how the information model captures the e-science process, including the collection of data, experiments, provenance and annotation.

1. Introduction

The scientific method has long been established as an effective means of building an understanding of the natural world. However, since the discovery of the central rôle of DNA in the chemistry of life, an increasing amount of data on genes and proteins has become available to biologists and biochemists. With the advent of cost effective computer power and network bandwidth, these data are increasingly available to the scientific community regardless of location, and powerful analytical programs have been developed to exploit them. As a result, the computer science community has become involved in research that aims at harnessing these data and programs with distributed systems technology to ease their exploitation, leading to a new discipline termed as e-science. E-Science, taken together with the developments in the Grid technology, introduces many opportunities as well as challenges [1]. Notably, typical e-science practice requires the integration of many distributed data sources and orchestration of diverse analysis services in a semantically rich, collaborative environment [2]. The implication is that without a relatively disciplined approach to the generation, transmission and interpretation of data and knowledge, e-science may introduce more problems than it solves.

To help address these challenges, the ^{my}Grid project seeks to understand, describe and model the e-science processes to support the e-scientists in conducting their day to day practices; in particular by creating an information model that captures the key concepts of the e-science method. Although ^{my}Grid as a whole is targeted at biologists and bioinformaticians, most of its information model is of interest to the e-science programme as a whole.

This raises the questions:

- What is the relationship between e-science and science?
- Can we adapt the scientific method to create an analogous e-science method?

The ^{my}Grid project seeks to answer these questions by describing and supporting such a method, and in particular by creating an information model that captures the key concepts of the e-science method. An aim of this paper is therefore to demonstrate that it is possible to model the fundamental e-Science processes in reasonable clarity, using the established methodologies of the scientific community.

The paper is organised as follows. Section 2 outlines the established scientific method, and how it relates to e-science. Section 3 then describes the key aspects of the ^{my}Grid information model, including modelling the scientific process, collecting e-science observations, personalisation and annotation. Sections 4 and 5 describe current implementation and discuss possible future work. Finally, Section 6 draws conclusions.

2. Scientific Method

The established scientific method can be summarised as follows:

1. Observe and describe of phenomena – *in vivo* and *in vitro* – and study existing knowledge – *in libris*
2. Formulate a hypothesis to explain the phenomena
3. From the hypothesis, predict other phenomena
4. Develop and perform *in vivo* and *in vitro* experiments that test these predictions. These experiments must be repeatable – and successfully repeated – by independent ex-

perimenters before the hypothesis can be generally accepted

The application of each of these aspects in an e-science context is now discussed.

2.1 Observation and study

Increasingly, information that has traditionally been disseminated in print is available electronically. This includes the full text of papers and journals and the catalogues and indices that describe them. Thus *in silico* methods supplement and can replace *in libris* study – an important part of the first step in the scientific method – and the ^{my}Grid project is developing tools that contribute to this.

We believe that *in silico* methods can also supplement the *in vivo* and *in vitro* observations and experiments of this step. We can apply increasingly sophisticated analyses to mine the published raw data, for example for correlations with similar phenomena in other organisms. Such analyses typically involve many data and program resources. We capture the orchestration of program resources using *workflows*, and the integrated orchestration of data and analysis resources using *distributed query processing*.

2.2 Hypothesis formulation and prediction

The second and third steps of the method are principally mental exercises in both science and e-science. In an e-Science context, we can use *ontologies* to define a language (at least a common vocabulary) to describe the domain of the study, and *statements* (e.g. as RDF statements [6]) in the language of the ontology to represent claims and hypotheses.

2.3 Experiment design and execution

As in the observation and study step, we can supplement *in vivo* and *in vitro* experiments by *in silico* methods. Again, we use workflows to represent experiment designs. In addition, we extend the workflows with the ability to link inputs and outputs with statements that (when created) make claims in support of an experiment's hypothesis.

2.4 Relating the scientific and e-scientific methods

The above suggest that we can devise a method that mimics the structure of the scientific method. We can use this e-science method in various ways, including:

- As an extension to the observation and study step of the scientific method. The entire e-science method, including its experimental step, is used to suggest interesting phenom-

ena for an otherwise conventional exercise of the scientific method

- To complement the scientific method. Increasingly, *in vivo* and *in vitro* experimental data are captured digitally and e-science workflows can perform the post-processing that will generate the indirect observations that will confirm or refute the hypothesis
- As an alternative to the scientific method. As more data are captured and curated for public access, we can use the e-science method to 'mine' the data to test hypotheses unrelated to the research that drove their original creation. It is however moot whether at this stage such e-experiments would be regarded as conclusive: instead we would expect the scientific community to demand some more direct demonstration.

In the next Section we describe how the ^{my}Grid information model has been designed to support these approaches to e-science.

3. A Summary of the ^{my}Grid Information Model

The ^{my}Grid Information Model attempts to support the scientific method emerging from work on e-science. It has been derived by taking a view of the established science practices and seeking to adapt the scientific method to create an analogous e-science method. It breaks down the conceptual modelling space into a set of aspects which are summarised in the rest of this section.

3.1 Modelling the e-Science process

^{my}Grid borrows from the CLRC Scientific Metadata Model [6] to represent the e-Science process (Figure 2), though there are some differences in the two models, as will be described.

The CLRC Scientific Metadata Model proposes a schema for collecting and managing experimental data. ^{my}Grid adapts this schema for e-science data. In particular, the CLRC model defines investigations as including (*in vivo* or *in vitro*) experiments, measurements and (*in silico*) simulations. We extend the model with a fourth class of e-experiments distinct from these. Many of the properties in both models correspond to terms defined by the Dublin Core Metadata Initiative [2]; in the ^{my}Grid model this is made explicit.

Following the CLRC metamodel, we define a *study* as either a *programme* or an *experiment instance*. Each study can have a research focus, which we record as a collection of semantic

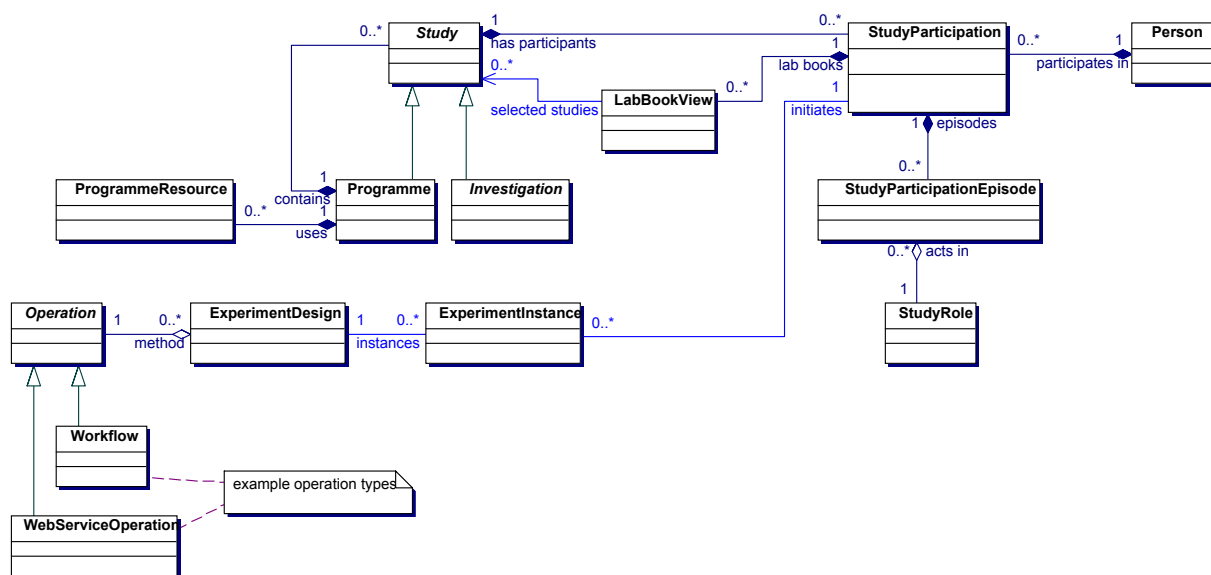


Figure 1: The e-Science Process

concepts in ontologies relevant to their domains. A programme is a structuring device for grouping other studies and can be used to represent e.g. a project or sub-project. *myGrid* extends the CLRC Scientific Metadata Model by allowing any e-Science resource to be associated with a programme under a local name (via a set of *programme resources*).

An *experiment instance* represents some executing or completed bioinformatics task, typically by enacting a workflow. The record of the execution (for example, input and output parameters; the service(s) invoked) are known as the Experiment Instance's provenance (Section 3.3). The CLRC Scientific Metadata Model has no direct equivalent to a *myGrid* experiment instance, though simulation, since it involves computation, has some similarities. Its use of Experiment refers to *in vitro* or *in vivo* rather than *in silico* activities. *myGrid* does not currently model CLRC Experiment, Simulation or Measurement entities.

An experiment instance is in turn an application of an *experiment design*, which represents the method to be used (typically as a workflow script). Finally, the *lab book view* allows a study participant to view Studies in a way relevant to their current concerns. For example, they may select all the experiments they initiated in a particular project, ordered by date, in one view, while viewing all subprojects in which they participate in another. The association with selected Studies is thus derived from the selection rule(s).

The participation of investigators in a study is captured by a *study participation*, which

qualifies a *person's* relationship to the study by a set of *study roles*. (e.g. Chief Investigator, Investigator, Researcher; or Team Leader, Researcher, Tester). Since a person's relationship with a study will change over time, the history of their relationship is captured as a collection of *study participation episodes* each with a definite start date and a (possibly undefined) end date.

Each study role represents a set of rights and responsibilities within the Study (and related Studies); thus a Person's overall rights and responsibilities within a Study at a particular time are the unions of those of the currently associated Study Roles. The Study Participation is the primary means of authorizing a Person to access or modify resources associated with a Study and of identifying provenance (e.g. creator of a document, instigator of a workflow enactment).

3.2 Collecting e-science observations

Scientific data are held in many distinct types and formats, and it is necessary to identify the type and format of each datum of interest so that it can (only) be input to type-compatible viewers, services and workflows. The set of such types cannot be fixed once and for all, even in the subdomains of biology and bioinformatics and so the *myGrid* information model contains a metamodel for types and formats that can be populated by its users. This is shown in Figure 2.

Types are represented explicitly as instances of the class *transfer type*. The scientific community in general and the bioinformatics community in particular represent data in a variety

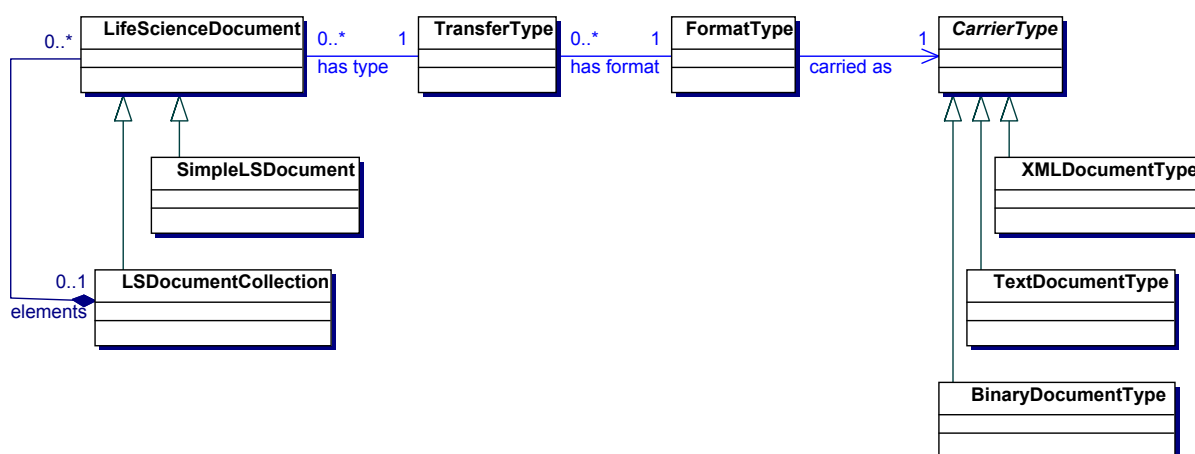


Figure 2: Modelling Scientific Data

of formats. Some of the more recent formats are defined by XML Schemas (or DTDs), but many existing formats are either binary formats or some form of non-XML structured text. Binary types can include formats specific to a particular domain, but also include general purpose formats such as JPEG (for images) or PDF (for documents). To capture format information, all kinds of transfer type have an associated *format type* which in turn has a *carrier type*. The carrier type identifies the coarse representation of the data necessary to pass it to an operation, i.e. Text, XML or Binary. We distinguish between transfer and format types because some syntactic types can represent several semantically-distinct concepts. For example, a FASTA-format entity can contain either a gene sequence or a protein sequence: FASTA is a shared format type while FASTA-gene and FASTA-protein are distinct transfer types that share a common format type.

Computational operations may take and produce both full records and simple record identifiers and so each must be modelled by distinct transfer types. Identifiers often have to be interpreted in the context of a particular database (the value '12345' may retrieve entirely different documents of different formats when used to access two distinct databases), but recent work has developed a URN scheme for a Life Science Identifier (LSID) [2]. We see considerable advantage in adopting LSIDs to describe e-science documents held and managed by ^{my}Grid components, and for annotating documents under the control of other authorities. However, the LSID proposal has not yet been universally adopted, and in the short to medium term there

is a need, reflected in the Information Model, to handle simple – non-LSID – identifiers.

3.3 Representing e-experiments and their provenance

As noted above, we typically represent experiment designs by workflows. These are reusable and repeatable (in the sense of re-executable, not necessarily giving identical results): e-experiments are instances of executions of these experiment designs.

An e-experiment comprises links to the experiment design (workflow), the precise resources accessed, its particular input and output data sets, annotations capturing the domain-specific meaning of the experiment plus properties identifying execution times and the initiating scientist.

These collections together provide the *provenance* of the e-experiment, and at the minimum provide an 'audit trail' that links raw input data through chains of e-experiments and intermediate results to the ultimate outputs of a study. The ^{my}Grid Information model divides provenance data into two linked categories: data provenance and experiment provenance (Figure 3).

A *data provenance* record, containing attributes corresponding to terms in the Dublin Core MetaData Initiative (DCMI) set, can be associated with any Document. For documents managed by ^{my}Grid, the provenance data will include a link to a *creation method*, which is either a *direct creation* (e.g. by file upload or edit) or an output of some e-science operation.

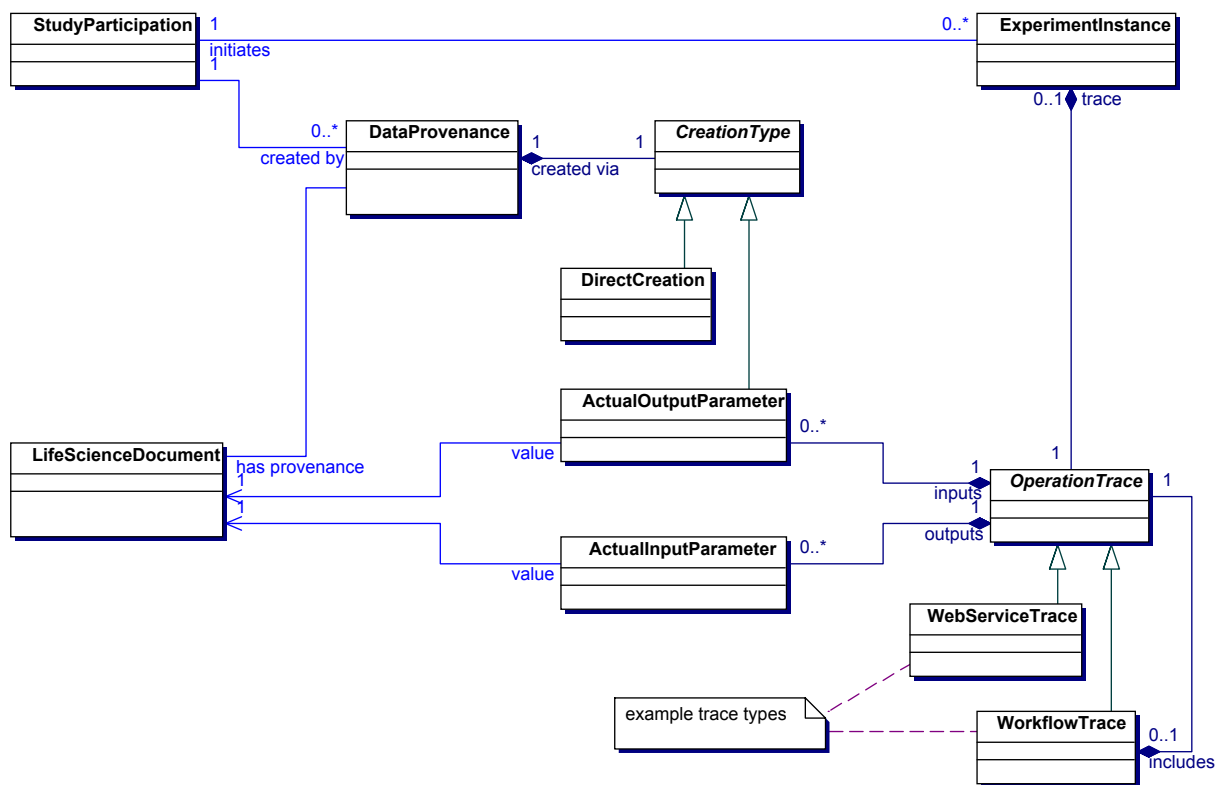


Figure 3: Provenance Metadata

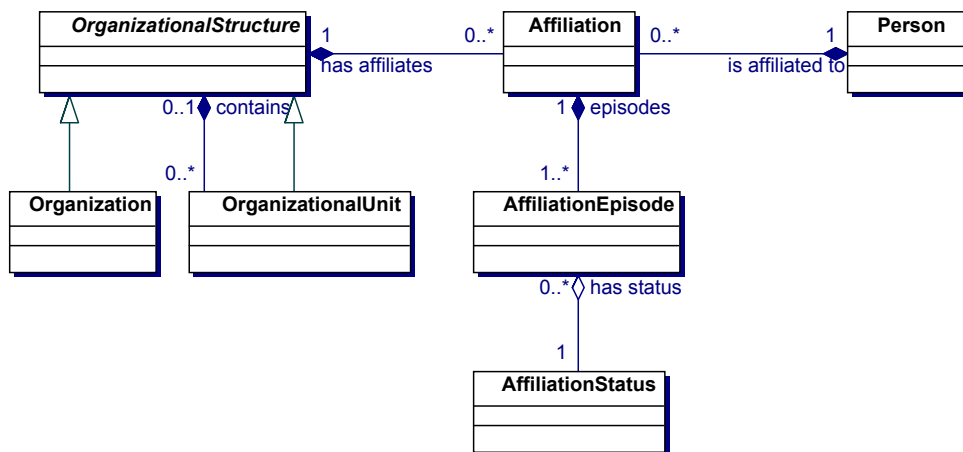


Figure 4: People and Organizations

E-science experiments, represented by experiment instances contain an *operation trace* that identifies exactly the inputs and output parameters. These actual parameters have types defined by the corresponding formal parameters of the operation’s interface. The exact trace information varies with the type of operation invoked. However, for a nested workflow operation, it includes traces of the sub-operations invoked.

3.4 Personalization

The myGrid information model follows the CLRC by representing individual scientists, their institutions, and their associations with particular e-science studies. This is shown in Figure 4.

Organizational Structures abstract Organizations— cooperative enterprises such as universities, corporations, national or international laboratories – and Organizational Units, the hierar-

chy of sub-units within them – such as faculties, departments and research groups.

A *person's* relationship with an *organizational structure* is captured by an affiliation, which identifies their status (e.g. Professor, Lecturer, Research Assistant, Doctoral Student; Director, Manager, Scientist, Technician). Since a person's relationship with an organization will change over time, the history of their relationship is captured as a collection of *affiliation episodes* each with a definite start date and a (possibly undefined) end date.

Persons participate in e-science studies as described in Section 3.1. They also have research interests which are recorded as a collection of semantic concepts in ontologies relevant to their domains. These interests can be used in the discovery process and for notification of relevant events.

Many of the properties in this group correspond to terms defined by the existing standards including the X.500/LDAP model [8] and VCARD [2].

3.5 Annotation and argumentation

A crucial aspect of the ^{my}Grid model is the ability to annotate all resources – including data, services, workflows, people – by statements derived from a domain-specific ontology or vocabulary. These ontologies will typically be developed and understood by some community, from a small project team to all researchers in a discipline. Statements within the ontology can thus represent claims concerning the annotated resources and allow resources to be discovered by their meaning.

4. Current Status

The ^{my}Grid information model is captured in a conceptual model in UML [10]. From this model, we have developed a set of concrete XML schemas that are the basis of communication between the ^{my}Grid services, in terms of web services port and message types. These services include the ^{my}Grid information repository, metadata store and workflow enactment engine. The information repository and metadata store also support the LSID resolution interfaces, with metadata returned as XML-coded RDF statements. The LSID proposal defines port types for obtaining both raw data for a resource and metadata describing it.

5. Future Work

We will continue to refine and extend the information model and its implementation from experience and feedback from our bioinformat-

ics users. To date, we have concentrated on the more persistent e-science entities, which can be held in the ^{my}Grid information repository and metadata store. We foresee an extension of this work to cover the three (*in vitro* and *in vivo*) experiment types described by the CLRC Scientific Data Model; these can then form the starting points for our *in silico* e-science experiments.

More interestingly perhaps, we are beginning to model more transient entities – e-science events – in the information model. We believe that these will help us to explore and explicitly support common higher-level e-science patterns. For example, the workflow enactor will generate experiment instance life cycle events including workflow and operation start and end.

6. Conclusions

The ^{my}Grid information model is being used to define appropriate port types and message types to bind together ^{my}Grid services to form a coherent framework for e-science.

As noted above, the ^{my}Grid information model takes into account a number of existing standards and proposals. The CLRC Scientific Metadata Model has been particularly useful in developing the underlying 'spine' of the model.

By adopting the emerging Life Science Identifier (LSID) standard [2] we allow two-way integration between ^{my}Grid and other bioinformatics resources across the world. Similarly, by using RDF to represent metadata, biologists can make use of existing semantic web tools to 'mine' the knowledge in a ^{my}Grid installation. We believe this 'open world' model is important for the success of ^{my}Grid: no 'closed world' system can hope to replicate the tools its potential users already access. For example, we are already making use of third-party tools such as Haystack [9] to query ^{my}Grid metadata.

References

- [1] F. Berman and T. Hey. *The Scientific Imperative* in I. Foster and C. Kesselman, editors, *The Grid: Blueprint for a New Computing Infrastructure*, pp. 13–24. Morgan Kaufmann, 2004.
- [2] T. Clark, S. Martin & T. Liefeld: *Globally distributed object identification for biological knowledge bases*, Briefings in Bioinformatics Vol 5 No 1 pp 59-70, Henry Stewart Publications, March 2004
- [3] F. Dawson and T. Howes: *vCard MIME Directory Profile*, RFC2426, IETF (<http://www.ietf.org/rfc/rfc2426.txt>)

- [4] Dublin Core Metadata Initiative: *DCMI Metadata Terms*, November 2001 (<http://dublincore.org/documents/2003/11/19/dcmi-terms/>)
- [5] C. Goble, C. Greenhalgh, S. Pettifer, and R. Stevens. *Knowledge integration: In silico experiments in bioinformatics* in I. Foster and C. Kesselman, editors, *The Grid: Blueprint for a New Computing Infrastructure*, pages 121–134. Morgan Kaufmann, 2004.
- [6] G. Klyne, J.J. Carroll: *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C, February 2004 (<http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>)
- [7] B. Matthews and S. Sufi: *The CLRC Scientific Metadata Model, version 1*, DL TR 02001, CLRC, February 2001 (<http://www-dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf>)
- [8] OSI: *Information technology – Open Systems Interconnection – The Directory: Overview of concepts, models and services*, ISO/IEC 9594-2001
- [9] D. Quan, D. Huynh and D.R. Karger: *Haystack: A Platform for Authoring End User Semantic Web Applications*. In D. Fensel, K. Sycara and J. Mylopoulos (eds): *The Semantic Web – ISWC 2003*, pp. 738-753, LNCS 2870, Springer-Verlag 2003
- [10] J Rumbaugh, I. Jacobsen, G. Booch: *The Unified Modeling Language Reference Manual*, 2nd Edition, Addison-Wesley, December 1998

Nick Sharman, Nedim Alpdemir, Justin Ferris, Mark Greenwood, Peter Li, Chris Wroe June 25, 2004