The myGrid project: services, architecture and demonstrator

Carole Goble, Chris Wroe, Robert Stevens
and the myGrid consortium
EPSRC e-Science Pilot Project myGrid
http://www.mygrid.org.uk

Abstract

myGrid is an e-Science research project developing open source high-level middleware to support *in silico* experiments in biology. *In silico* experiments use databases and computational analysis rather than laboratory investigations to test hypothesis. This paper provides an overview of services the myGrid project is developing, and the architecture in which they fit. Registries provide information about available data and computational services, while remote legacy bioinformatics applications are wrapped using a consistent distributed analysis framework Soaplab. As in conventional science, experimental method is as important as final results. myGrid formalises these methods as workflow or query specifications and provides service based middleware components to enact them. e-Science for the individual often has a narrow focus and so personalisation forms a key theme in myGrid service design. Information repositories, service registries and change notification systems are all being developed to provide personalised views of resources. myGrid components make extensive use of metadata to support this need for personalisation and the project is pioneering the use of semantic web technology, to manage annotation, ontologies and semantic discovery. The ultimate goal of myGrid is to supply this collection of services as a toolkit to build end applications. To demonstrate this concept the project is building its own application (the myGrid workBench).

1 Introduction

myGrid aims to develop open source high-level service-based middleware to support *in silico* experiments in biology. *In silico* experiments are procedures using computer based information repositories and computational analysis adopted for testing hypothesis or to demonstrate known facts. In our case the emphasis is on data intensive experiments that combine use of applications and database queries. The user is helped to create workflows (a.k.a. experiments), sharing and discovering others' workflows and interacting with the workflows as they run. Rather than thinking in terms of data grids or computational grids we think in terms of Service Grids, where the primary services support routine *in silico* experiments. The project's goal is to provide middleware services as a toolkit to be adopted and used in a "pick and mix" way by bioinformaticians, tool builders and service providers who in turn produce the end applications for biologists. The target environment is open, by which we mean that services and their users are decoupled. Services are not just used solely by their publishers but by users unknown to the service provider, who may use them in unexpected ways.

myGrid focuses on speculative explorations by a scientist to form discovery experiments. These evolve with the scientist's thinking, and are composed incrementally as the scientist designs and prototypes the experiment. Intermediate versions and intermediate data are kept, notes and thoughts are recorded, and parts of the experiment and other experiments are linked together to from a network of evidence, as we see in bench laboratory books. We aim to collect, share and reuse:

- *Experimental design components*: workflow specifications; query specifications; notes describing objectives; applications; databases; relevant papers; the web pages of important workers, and so on.
- *Experimental instances* that are records of enacted experiments: data results; a history of services invoked by a workflow engine; instances of services invoked; parameters set for an application; notes commenting on the results and so on.
- *Experimental glue* that groups and links design and instance components: a query and its results; a workflow linked with its outcome; links between a workflow and its previous and subsequent versions; a group of all these things linked to a document discussing the conclusions of the biologist and so on.

Discovery experiments by their nature presume that the e-biologist is actively interacting with and steer-

ing the experimentation process, as well as interacting with colleagues (in the simplest case by email). [6] gives a detailed motivation for the project.

The project produced a requirements gathering prototype based on use cases for the functional analysis of clusters of proteins. Data was based on microarray studies, which showed the level of activity of genes associated with circadian rhythms in the fruit fly, Drosophila melanogaster. Studies in this model organism can provide insights into how the human brain's internal circadian clock regulates sleep, body temperature, blood pressure, and hormone levels. We then developed a more detailed set of scenarios for the examination of the genetics of Graves' disease, an immune disorder causing hyperthyroidism [15]. This latter case study is our test bed application for our initial full prototype and the rest of the project. We have built an electronic laboratory workbench demonstrator application as a vehicle to crystallise our architecture and experiment with our services: their functionality, their deployment and their interactions [17]. In addition, Talisman is a third party application that is prototyping the use of our workflow components [12]. The paper is organised as follows. Section 2 gives an overview of the services in myGrid, its chief components and its architecture. Section 3 runs through an example of the workbench demonstrator. We conclude in section 4 with a statement on status and an outlook for the remainder of the project.

2 myGrid Services and Architecture

The myGrid middleware framework employs a service-based architecture, firstly prototyped with Web Services but with an anticipated migration path to the Open Grid Services Architecture (OGSA) [18]; [13] gives an account of the conversion of two myGrid services to OGSI services. The middleware services are intended to be collectively or selectively adopted by bioinformaticians, tool builders and service providers who in turn produce the end applications for biologists. Figure 1 shows the layered middleware stack of services. The primary services to support routine *in silico* experiments fall into four categories:

- services that are the **tools** that will constitute the experiments, that is: specialised services such as AMBIT text extraction [5], and external third party services such databases, computational analysis, simulations etc, wrapped as web services by Soaplab [14] if required;
- services for **forming and executing experiments**, that is: workflow management services [3], information management services, and distributed database query processing [4];
- **semantic services** for discovering services and workflows, and managing metadata, such as: third party service registries and federated personalised views over those registries [9], ontologies and ontology management [20];
- services for supporting the **e-Science scientific method** and best practice found at the bench but often neglected at the workstation, specifically: provenance management [16] and change notification [11].

The final layer (e) constitutes the applications and application services that use some or all of the services described above.

2.1 Services that form the experiments

myGrid middleware must go hand in hand with corresponding development of domain specific scientific services that can deliver data and computation analysis. Therefore bioinformaticians within the project have been developing service based access to bioinformatics tools and data.

Bioinformatics services: Services such as database retrieval and analysis tools need to be wrapped and offered in a form that accommodates their distribution and variety of data formats. myGrid has acquired or wrapped a range of bioinformatics Web Services including: the complete EMBOSS application suite of over eighty analysis tools, MEDLINE, SRS, OMIM and NCBI & WU BLAST sequence alignment tools. The project has developed Soaplab[1], a universal connector for legacy command line based systems. The majority of services that we want to be able to make use of are shell scripts, PERL fragments or compiled architecture specific binaries rather than web services; Soaplab provides a fairly universal glue to bind these into web services. Soaplab is freely available; see [14] for further details.

Text extraction services: AMBIT is a system for Acquiring Medical and Biological Information from Text developed under the auspices of this and the CLEF e-Science project. The majority of biomedical knowledge still persists as free text in the published literature. More automated assistance in the delivery of this knowledge to the scientist requires at least some of the information is extracted into a more structured machine interpretable form. AMBIT provides an information extraction service based on natu-

[1] http://industryhttp://industry.ebi.ac.uk/soaplab/

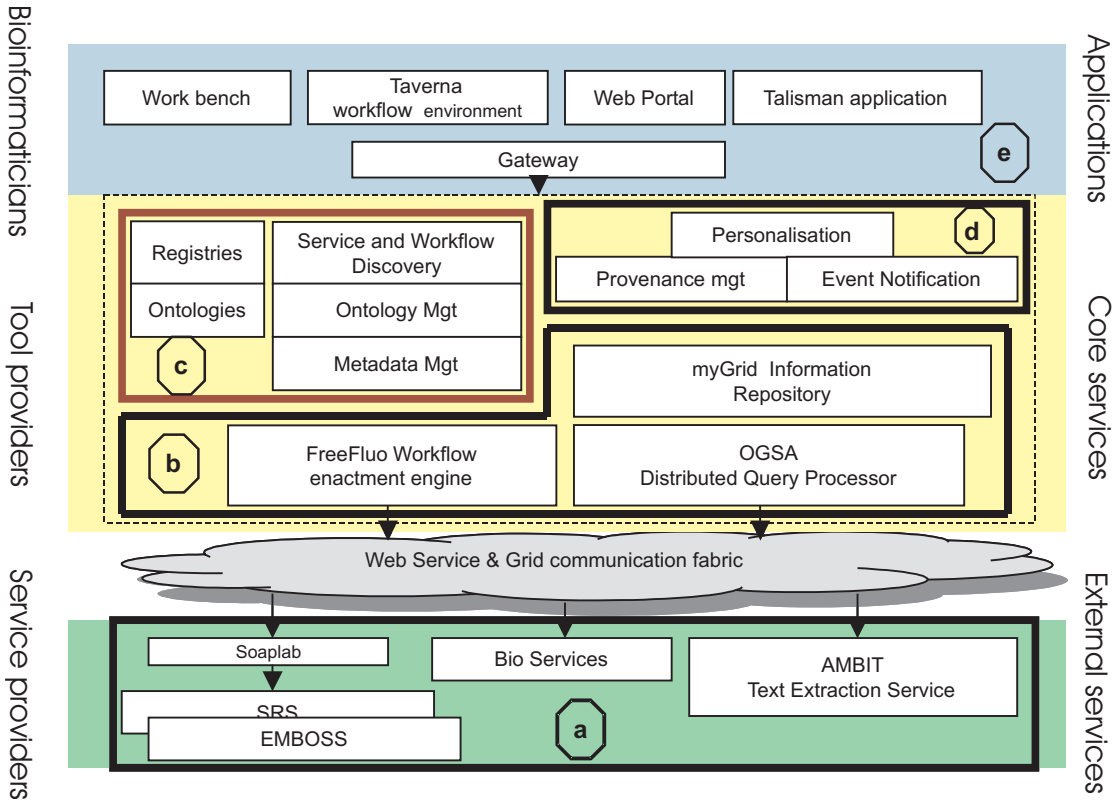

Figure 1: The myGrid services and middleware stack

ral language processing. Biological abstracts are processed and terms of various classes are recognised and isolated such as genes, proteins, protein structures and biological species. The information extracted from the texts is held in a relational database and viewed via dynamically generated web pages or a web service. See [5] for further details.

2.2 Services for forming experiments

myGrid regards *in silico* experiments as distributed queries and workflows. Data and parameters are taken as input to an analysis or database service; then output is taken from these, perhaps after interaction with the user, as input to further tools or database queries.

Workflow enactment, creation and management: Once discovered or built, a workflow is run by our powerful FreeFluo[2] workflow enactment engine, which can handle WSDL based web service invocation. FreeFluo supports two XML workflow languages, one based on IBM's Web Service Flow Language (which we used early on) and our own, XScufl, developed as part of the Taverna project, in collaboration with the Human Genome Mapping Project [1]. The FreeFluo engine and the Taverna workflow development environment are open source and downloadable. See [3] for further details.

Distributed database query processing: The OGSA-DAI project [3] and myGrid are together building a distributed query processing system that will enable a user to specify queries across a set of Grid-enabled information repositories in a high level language (initially OQL). Complex queries on large data repositories may result in potentially high response times, but the system can address this through parallelisation. The initial prototype is to be released in August 2003. See [4] for further details.

The **myGrid Information Repository** (mIR) acts as a personalised store of all information relevant to a scientist performing an *in silico* experiment. It implements an information model tailored to e-Science. Experimental data is stored together with provenance records of its origin. It is used to store workflow specifications ready to be submitted to the enactor together with records of running or completed workflows. These workflow records form a major basis for internally generated data provenance and are discussed in section 2.4. The mIR has also been designed to store information about people and projects both

[2] http://freefluo.sourceforge.net

[3] http://www.ogsadai.org/

directly linked to the investigation and from the wider scientific community to aid collaboration.

Metadata storage is a central feature of the mIR, with annotation possible for all internally stored objects in addition to objects stored in disparate remote repositories. Annotations are currently stored in an RDF triple like manner [4] and the project is considering the use of "off the shelf" RDF triple stores such as the Jena Semantic Web toolkit[5]. Several types of annotation are used from free-text notes of the object's significance with respect to the investigation, to more structured DAML+OIL based ontology annotations of what the object represents [8]. Annotation is a key tool used to link related objects and so answer wide-ranging queries such as 'What workflows have been recently run by members of my project?' and 'What other data is available on this topic?'

An organisation would typically have a single mIR, which would be shared by many users, each using it to store their own provenance, data and metadata. Different users can be provided with different views of the information it contains; in addition views will be based on criteria such as experiment, project and subject topic. These types of views can be built by exploiting the rich metadata associated with each object.

The mIR is an early adopter of the OGSA-DAI service, using it to make the repository accessible to local and remote components over a Grid. The OGSA-DAI distributed query processing service allows data from the mIR and one or more remote data repositories to be federated, producing unified information views to the biologist. The first version of the mIR has been built over the relational database product DB2[6] primarily because of its extensions to support query of stored XML documents. The second version is likely to take on a federated architecture, using a mediator and extensive use of annotation and shared identifiers such as Life Science Identifiers (LSIDs)[7] to interconnect data objects in distributed heterogeneous repositories.

2.3 Services for discovery and metadata management

Much of e-Science depends on discovering and pooling resources especially services but also experimental designs, data, people and projects. myGrid has developed several components to facilitate this discovery process.

Registries and registry views. These are a key feature of web services infrastructure in which service descriptions are centrally published. myGrid extends the idea of a registry in three ways:

- *Personalised views over distributed registries*. It has become clear that multiple distributed registries will exist, some community wide, some specific to an organisation. To accommodate this registry views are been developed that aggregate distributed information based on a personal profile [10].

- *Extensible metadata storage*. Originally designed to support the web services standard UDDI, the registry has now been underpinned with a flexible RDF storage component which enables it to support additional metadata standards such as DAML-S and BioMOBY.

- *Additional semantic descriptions* to allow more precise searching by both people and machines. These DAML+OIL semantic descriptions build on the work of the DAML-S coalition[8] and have been used to guide the construction of workflows by constraining the choice to those services, which have semantically compatible inputs and outputs. Similarly semantic description of workflows has been used within the myGrid workbench to discover relevant workflows given an item of data selected from the mIR.

We work with the BioMOBY project [19] and the Interoperable Informatics Infrastructure Consortium (I3C) registry group [9], on the architecture and development of these semantically-enabled registries.

Discovery components: These components take advantage of the richer metadata within the registry view to enable more sophisticated semantic discovery. Indexing and searches over DAML+OIL based descriptions of services and workflow specifications are supported by a "find service" which is underpinned by description logic based reasoning. A service browser module within the workbench provides hierarchical categorisation based on this reasoning. See [9] for more details.

Annotation components: A rich metadata framework only becomes useful when it is practical for users to add sufficient metadata. myGrid is using semantic web annotation tools such as COHSE[10] both to capture annotation and dynamically link resources based on those annotations. The first use of this is described in section 2.4 on provenance management.

Ontology services: myGrid services can be described as semantically aware. DAML+OIL concepts

[4] http://www.w3c.org/RDF/
[5] http://www.hpl.hp.com/semweb/jena.htm
[6] http://www.ibm.com/software/data/
[7] http://www.i3c.org/wgr/ta/resources/lsid/docs/index.htm
[8] http://www.daml.org/services
[9] http://www.i3c.org
[10] http://cohse.semanticweb.org

are used throughout the myGrid components. An ontology service has been developed to provide a single point of reference for these concepts and to support description logic reasoning of concept expressions. The use of semantic web technology such as ontology services and semantic annotation tools makes myGrid an early example of a "Semantic Grid" [7].

2.4 Services for supporting e-Science

myGrid aids users in finding appropriate resources, offering alternatives to busy resources and guiding users in the composition of resources into workflows. In addition, myGrid offers:

Notification services: A workflow may need to be re-run when new or updated data and analytical software become available. myGrid has a notification service to mediate an asynchronous interaction between services. Servers may register the type of notification events they produce and clients may register their interest in receiving updates. For example, users can register their interest in an object in the mIR, and be notified of any relevant new information. Notifications may also be used to automatically trigger workflows to analyse new data. The type and granularity of notification events will be defined with ontological descriptions in metadata exchanged with the notification service. See [11] for further details.

Provenance management: Biologists routinely record the provenance of their bench experiments in lab books and this should be true for computational experiments too. As well as being important for traceability, provenance information enables the utilisation of notification events generated by services to determine whether a workflow needs to be re-run, e.g. if a new version of a databank used by a workflow is released. When a workflow is executed, FreeFluo generates provenance logs in the form of XML files, recording the start time, end time and service instances operated in this workflow. Data, and metadata about the workflow and the provenance logs are stored in the mIR. All mIR objects carry provenance attributes: hence the provenance log has who created it, when, in what context, and so on. In addition, a set of metadata is associated with this workflow invocation instance: the input and output relationships between the workflow instance and data items, the 'is defined by' relationship between the workflow instance and its associated definition documents. Other annotations regarding the hypothesis of the experiment, thoughts and opinions by the scientist and quality of results are also stored as XML in the mIR or as regular web documents. This provenance information is extracted to answer questions such as "what recent workflows were run by Dr. Pearce using BLAST". As we make liberal use of ontologies, by annotating provenance logs with concepts drawn from the myGrid ontology, we can experiment with building a dynamically generated hypertext of provenance documents, data, services and workflows based on their associated concepts and reasoning over the ontology [21]. See [16] for further remarks on provenance and our plans.

Personalisation opportunities: We intend to make all services personalised to the scientist in their own appropriate ways. For example: different users can be provided with appropriate views of the mIR; the registry view gives a user perspective over the services they can use, and the opportunity to attribute their metadata to third party services they do not own, as well as publish their own workflows and services; and the event notification system allows users to define their own choice of events. The user is represented by a User Proxy service.

2.5 Applications and Application services

The intention is that services can use some or all of the myGrid services on offer. Applications can interact with services directly or via a Gateway. For example, the Talisman rapid prototyping application [12] directly interacts with FreeFluo. The Gateway provides an optional unified single point of programmatic access to the whole system, for ease of use, isolating client software from the detailed operation and interactions of the core architecture and adds value in respect of support for collaboration, provenance and personalisation by overlaying provenance metadata and semantics relationships on non-myGrid services (such as legacy Web Services).

3 myGrid demonstrator workBench

In order to experiment with the interaction of services, and to garner concrete requirements from stakeholders, we have built a demonstrator workbench application. The myGrid workBench demonstrator has been developed using the NetBeans Platform[11], a JAVA based infrastructure backplane to allow more rapid development of complex desktop applications. The application is not intended to be a fully-fledged end product but rather a vehicle for the project.

The workbench has been seeded by experiments in genetic studies of Grave's Disease. Graves Disease is caused by an autoimmune response against the thyroid, causing hyperthyroidism. It is one of the commonest autoimmune diseases and has a strong genetic component. By gathering information about genes,

[11] http://www.netbeans.org /products/platform/index.html

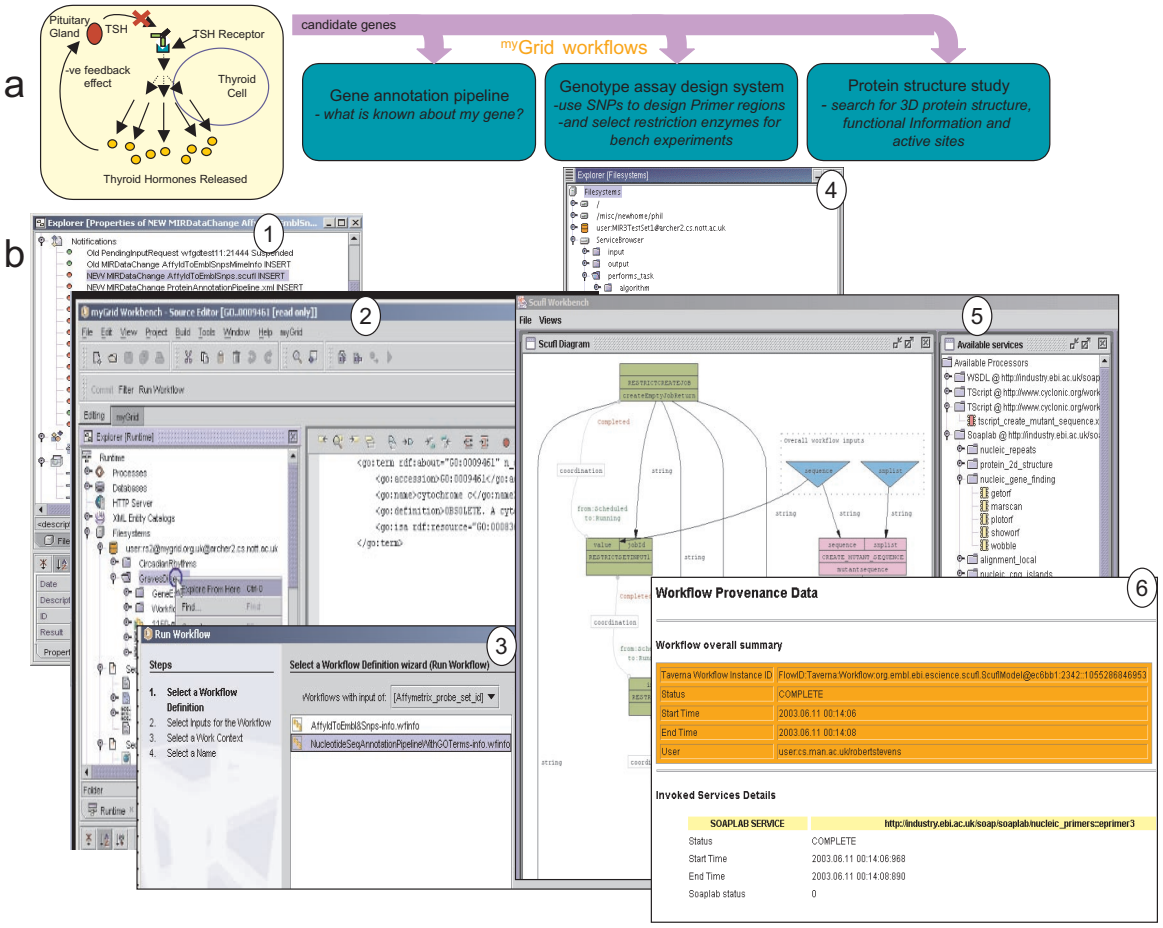

Figure 2: (a) Summary of Graves' Disease scenario and workflows involved. (b) Screen shots showing myGrid workBench and Taverna workflow editor during each stage of the scenario.

which have altered levels of activity in the cells responsible for the immune response, the researchers hope the gain new knowledge of the mechanisms of the disease and so ultimately inform the design of novel therapies. As soon as the identity of the relevant genes is known the myGrid workbench is used to run workflows that gather information about those genes, help design new molecular biology experiments to focus on the genes of interest, and to predict the 3D structure of the protein products of the genes.

Figure 2a provides a summary of this scenario and the types of workflow involved. Figure 2b shows screen shots of a typical walkthrough the scenario. (1) The notification service informs the user via a notification client in the workbench that new data has been added to the mIR which can be browsed in the workBench (2). In this case it is the identity of a new gene with changed expression in Graves' Disease.(3) The user can then discover which workflows have been published that can operate on data of this specific semantic type (an Affymetrix probe set identifier) via a wizard in the workbench. The wizard itself makes use of a semantic find service, which finds relevant services and workflows in the myGrid registry using description logic reasoning over associated semantic descriptions. A registry browser is also available in the workbench to allow the user to browse more freely for a workflow or service using a hierarchical categorisation based on each individual semantic description (4). If an appropriate workflow does not exist, a new one can be created in the Taverna editor (5). The workflow and associated data are submitted to the FreeFluo enactor. The enactor provides a detailed provenance record stored in the mIR describing what was done, with what services and when. This can also be viewed within the workbench (6), and the user can again be notified when the resulting output data from the workflow is deposited back in the mIR.

4 Status and Outlook

myGrid has its first demonstrator application in place. The project has 18 months left to run. All the components outlined in Section 2 have prototype implementations in various stages of maturity. A set of core components has reached a stage where they can effectively support *in silico* investigation and are their first releases are available for download. These include the workflow enactor FreeFluo, its sister development environment Taverna and the suite of bioinformatics services made available via Soaplab. The demonstrator together with this core set of components will enable us to both begin evaluation and gather more concrete requirements from molecular biology users. These will allow us to improve and refine the facilities of the myGrid services. Our industrial collaborators are keen to use myGrid to support their own e-Science activities, and partners such as GlaxoSmithKline have already taken snapshots of the development code for previewing.

Our current exploration of Graves' disease has been quite narrow and orientated around a single user group. To evaluate the myGrid components through the workbench thoroughly, we need to populate with data and associated metadata from multiple studies, investigations, projects, experiments and users.

Other components need more significant ongoing development following lessons learnt from early prototypes. Our initial investigations have revealed we need a more sophisticated model of provenance and other experimental data holdings [19]. This will allow us to store much more heavily linked metadata about provenance that will enable us to create views of the mIR along many axes.

The myGrid Information Repository has provided us with many useful insights into the types of data and metadata that need to be stored and the ways that data needs to be presented. The current mIR uses RDBMS technology and much of the information held therein is stored in a triple like manner. Much of the provenance information is stored as XML files; this makes it cumbersome to retrieve and process much of the metadata stored in the mIR. Consequently, we will be investigating more wide spread use of RDF technology in the future.

Currently, the notification service is coarse grained in the types of notifications it indicates. For instance, one topic is "data change", which is used for the arrival of new data, the update of data, etc. in the mIR. Much finer grained notifications need to be developed if users are to judge response to notification appropriately.

Semantic web technologies such as annotation, discovery and ontology services are still at an early stage of development. Current prototypes have been useful in crystallising requirements for semantics within e-Science and specifically how those semantics are integrated into myGrid components. The next phase of the project will aim to deliver more robust semantic components tailored to this environment.

At the end of the project once the components have been developed and evaluated we will be producing myGrid-in-a-box: a package of components that can be downloaded by a research organisation, from which an arbitrary selection can be installed and configured to support the local e-Science requirements of that institution.

Acknowledgements

This work is supported by the UK e-Science programme EPSRC GR/R67743, & DARPA DAML subcontract PY-1149, Stanford University. The authors would like to acknowledge the myGrid team (past and present): Matthew Addis, Nedim Alpdemir, Rich Cawley, Neil Davis, David De Roure, Vijay Dialani, Alvaro Fernandes, Justin Ferris, Robert Gaizauskas, Kevin Glover, Chris Greenhalgh, Mark Greenwood, Yikun Guo, Ananth Krishna, Peter Li, Xiaojian Liu, Phil Lord, Darren Marvin, Karon Mee, Simon Miles, Luc Moreau, Arijit Mukherjee, Tom Oinn, Juri Papay, Savas Parastiditis, Norman Paton, Steve Pettifer, Milena Radenkovic, Peter Rice, Angus Roberts, Alan Robinson, Tom Rodden, Martin Senger, Nick Sharman, Robert Stevens, Victor Tan, Paul Watson, and Anil Wipat.

We also thank our industrial partners: IBM, Sun Microsystems, GlaxoSmithKline, AstraZeneca, Merck KgaA, geneticXchange, Epistemics Ltd, and Network Inference.

References

[1] Taverna workflow environment for bioinformatics: http://sourceforge.net/projects/taverna.

[2] *Proceedings UK OST e-Science 2nd All Hands Meeting*, September 2003.

[3] M Addis, T Oinn, M Greenwood, J Ferris, D Marvin, P Li, and A Wipat. Experiences with e-Science workflow specification and enactment in bioinformatics. In [2].

[4] MN Alpdemir, A Mukherjee, NW Paton, P Watson, AAA Fernandes, A Gounaris, and J Smith. OGSA-DQP: A Service-Based Distributed Query Processor for the Grid. In [2].

[5] R Gaizauskas, M Hepple, N Davis, Y Guo, H Harkema, A Roberts, and I Roberts. AMBIT:

Acquiring Medical and Biological Information from Text. In [2].

[6] CA Goble, S Pettifer, and R Stevens. *Knowledge Integration:* In silico *Experiments in Bioinformatics in The Grid: Blueprint for a New Computing*. Morgan Kaufman, 2003.

[7] CA Goble and D De Roure. An Application of the Semantic Web. *ACM SIGMOD Record*, 31(4), December 2002.

[8] I Horrocks. DAML+OIL: a Reason-able Web Ontology Language. In *Proc. of EDBT 2002*, number 2287 in Lecture Notes in Computer Science, pages 2–13. Springer, March 2002.

[9] P Lord, C Wroe, R Stevens, CA Goble, S Miles, L Moreau, K Decker, T Payne, and J Papay. Semantic and Personalised Service Discovery. In [2].

[10] S Miles, J Papay, V Dialani, M Luck, K Decker, T Payne, and Luc Moreau. Personalised Grid Service Discovery. Performance Engineering. In *Performance Engineering. 19th Annual UK Performance Engineering Workshop*, pages 131–140, 2003.

[11] L Moreau, X Liu, S Miles, A Krishna, V Tan, and R Lawley. myGridNotification Service. In [2].

[12] T Oinn. Talisman Rapid Application Development for the Grid. In *Proceedings of Intelligent Systems in Molecular Biology*, Brisbane, Australia, July 2003.

[13] S Parastatidis and P Watson. The NEReSC Core Grid Middleware. In [2].

[14] M Senger, P Rice, and T Oinn. Soaplab - a unified Sesame door to analysis tools. In [2].

[15] R Stevens, K Glover, C Greenhalgh, C Jennings, P Li, M Radenkovic, and A Wipat. Performing in silico Experiments on the Grid: A Users Perspective. In [2].

[16] R Stevens, M Greenwood, and CA Goble. Provenance of e-Science Experiments - experience from Bioinformatics. In [2].

[17] R Stevens, A Robinson, and CA Goble. myGrid: Personalised Bioinformatics on the Information Grid. *Bioinformatics*, 19:i302–i304, 2003.

[18] D Talia. The Open Grid Services Architecture - Where the Grid Meets the Web. *IEEE Internet Computing*, 6(6):67–71, 2002.

[19] MD Wilkinson and M Links. BioMOBY: an open source biological web services proposal. *Briefing in Bioinformatics*, 3:331–341, 2002.

[20] C Wroe, R Stevens, C Goble, A Roberts, and M Greenwood. A suite of DAML+OIL ontologies to describe bioinformatics web services and data. *International Journal of Cooperative Information Systems*, 12(2):197–224, 2003.

[21] J Zhao, CA Goble, M Greenwood, C Wroe, and R Stevens. Annotating, linking and browsing provenance logs for e-Science. In *2nd Intl Semantic Web Conference (ISWC2003) Workshop on Retrieval of Scientific Data*, Florida, USA, October 2003. submitted.