The Grid: An Application of the Semantic Web

Carole Goble
Department of Computer Science
University of Manchester
Oxford Road
Manchester, M13 9PL, UK
carole@cs.man.ac.uk

David De Roure
Dept of Electronics & Computer Science
University of Southampton
Southampton,
SO17 1BJ, UK
dder@ecs.soton.ac.uk

Abstract

The Grid is an emerging platform to support on-demand "virtual organisations" for coordinated resource sharing and problem solving on a global scale. The application thrust is large-scale scientific endeavour, and the scale and complexity of scientific data presents challenges for databases. The Grid is beginning to exploit technologies developed for Web Services and to realise its potential it also stands to benefit from Semantic Web technologies; conversely, the Grid and its scientific users provide application pull which will benefit the Semantic Web.

What is the Grid?

The Grid is "flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources—what we refer to as virtual organizations." Foster et al, 2001[1].

Large-scale science and engineering are undertaken through the interaction of people, heterogeneous computing resources, information systems, and instruments, all of which are geographically and organizationally dispersed. "The Grid" is an emerging platform to support coordinated resource sharing and problem solving on a global scale for data-intensive and compute-intensive applications [1]. The name arose from an analogy with an electricity power grid: computing and data resources would be delivered over the Internet seamlessly, transparently and dynamically as and when needed, just like electricity. Thus, the overall motivation for Grids is to facilitate the routine interactions of resources in order to support large-scale science and engineering.

The Grid was originally focused on sharing computational power and resources for advanced science and engineering. The 'metacomputing' projects of the early 1990s set out to build virtual supercomputers using networked computer systems. The target applications were, and primarily continue to be, large-scale science. For example, trans-national experiments, such as the particle physicist's quest to find the Higgs boson by building a Large Hadron Collider. This device generates petabytes of data in a few seconds and the complex analyses can take months of computational processing [2].

Increasing the computational power by combining large numbers of geographically diverse systems raises the issues of scalability and heterogeneity. Scalability brings a number of challenges: the inevitability of failure of components, the need for automation, the need to exploit the locality of resources due to network latency, and the increasing number of organisational boundaries, emphasising authentication and trust issues. Larger scale applications may also result from the composition of other applications, which increases the complexity of systems. Heterogeneity is addressed by middleware, such as the Globus Toolkit [3], to provide uniformity through a standard set of interfaces to the underlying resources.

Early Grid middleware exploits a range of protocols such as LDAP for directory services and file store queries [4], GridFTP for large-scale reliable data transfer and SSL for security. Higher level functionality, such as tolerant scalable data replication [5], exploit these. Some attention has been paid to data intensive rather than compute intensive Grid use; for example, the Storage Request Broker provides applications with uniform access to distributed file storage [6]. However, research and development activities relating to the Grid have generally focused on applications where data was stored in files, and there is little support for transactions, relational database access or distributed query processing [7].

The Grid community is now actively developing fundamental mechanisms for the interaction of any kind of resource including documents, databases, instruments, archives and people. Support for data interaction is focused on consistent access to databases from Grid applications and coordinated access to databases from Grid applications [8].

This is partly in response to the adoption of the Grid by scientific disciplines other than particle physics (e.g. biology, earth science, chemistry, astronomy), that are less concerned with the size of data than the need for data integration (*see Buttler et al in this issue*).

Grid Services

Grid middleware should enable new capabilities to be constructed dynamically and transparently from distributed services. In order to engineer new Grid applications it is desirable to be able to reuse existing components and information resources, and to assemble and co-ordinate these components in a flexible manner. Partly for this reason the Grid is moving away from a collection of protocols to a service-oriented approach: the Open Grid Services Architecture (OGSA) [9]. This unites Web Services with Grid requirements and techniques.

The Grid's requirements mean that Grid Services extend Web Services considerably. Grid service configurations are:

- *dynamic and volatile*. A consortium of services (databases, sensors, compute servers) participating in a complex analysis may be switched in and out as they become available or cease to be available;
- *ad-hoc.* Service consortia have no central location, no central control, and no existing trust relationships;
- *large.* Hundreds of services could be orchestrated at any time;
- *long-lived*. A simulation could take weeks.

These requirements make strenuous demands on fault tolerance, reliability, performance and security [9]. Whereas Web Services are presumed to be available and stateless, Grid services are presumed to be transient and stateful.

Grid services are broadly organised into four tiers:

1. Fabric (security, data transport, certification, remote access, network monitoring, ownership and digital watermarking, authentication);
2. Base (resource scheduling, data access, event notification, metadata management, provenance, versioning);
3. High Level (workflow, database management, personalisation);
4. Application (a gene sequence alignment, a Swiss-Prot database, a gene finding algorithm).

Each tier relies on metadata. To achieve the flexible assembly of Grid services requires information about the functionality, availability and interfaces of the various services. Service discovery and brokering uses metadata descriptions [10]. Service composition is controlled and supported by metadata descriptions [11]. Metadata is key to achieving the Grid Services vision.

What is the Semantic Web?

The Semantic Web is "…an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is the idea of having data on the Web defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications… data can be shared and processed by automated tools as well as by people." World Wide Web Consortium [12].

The ambition is of an environment where software agents are able to dynamically discover, interrogate and interoperate resources, building and disbanding virtual problem solving environments, discovering new facts, and performing sophisticated tasks on behalf of humans [13]. However, simple metadata and simple queries give a small but not insignificant improvement in information integration [14]. Simple or complex, automated processing of Web content requires explicit machine-processable semantics associated with Web resources to describe what it is *about* and what it is *for*.

The Semantic Web can be thought of as three tiers:

1. Fabric, made up of a unique global *identity* for a resource; *metadata* for asserting facts about resources and the claims on those assertions, and a common language for expressing metadata and knowledge embodied by *ontologies* for a shared understanding and a common vocabulary for the metadata and rules for *inferring* new metadata and knowledge.
2. Base Services, for example reasoning and querying over metadata and ontologies, explanation of those inferences, trust management, agents, search engines, ontology servers);
3. Application Services; e.g. a travel agent service.

Considerable efforts are in progress to develop the languages and technologies for the fabric and base services; notably RDF(S) [15], DAML+OIL [16] and OWL [17].

The Grid and the Semantic Web

Until very recently the Grid and the Semantic Web communities were separate, despite the

convergence of their respective visions. Both have a need for computationally accessible and sharable metadata to support automated information discovery, integration and aggregation. Both operate in a global, distributed and changeable environment.

The Semantic Web base services can be Grid Base Services. The Semantic Web fabric is the means by which the Grid could represent metadata: both for Grid *infrastructure*, driving the machinery of the Grid fabric, and its base and high level services, and for Grid *applications*, representing the knowledge and operational know-how of the application domain.

Semantic Web for Grid infrastructure

Semantic Grid Services
The description of a service is essential for automated discovery and search, selection, matching, composition and interoperation, invocation and execution monitoring. This choice depends on service metadata. Classification of services based on the functionality they provide has been widely adopted by diverse communities as an efficient way of finding suitable services, e.g. UDDI. Reasoning over service descriptions has a role to play when classifying and matching services. In Condor [11] a matching mechanism is used to choose computational resources. In an architecture where the services are highly volatile, and configurations of services are constantly being disbanded and re-organised, knowing if one service is safely substitutable by another is essential.

At the time of writing, the current state of describing Grid Services through semantics is by using the names assigned the portType and serviceType elements of a WSDL document, linked to a specification document [9]. Bringing together the Semantic Web and Web Services has already attracted attention (*see Bussler et al in this issue*). DAML+OIL has been explored in myGrid[1] [10]. The myGrid service ontology extends the DAML-S ontologies [18]. Service classifications are more expressive than UDDI's simple hierarchies and services are queried and matched by subsumption reasoning over the service descriptions. However, Grid Services dynamically create and destroy service instances, have soft state registration and form long-lived service configurations. How this affects the way Semantic Web technologies can describe and discover Grid services is a challenge yet to be adequately addressed.

Information integration
Complex questions posed by scientists require the fusion of evidence from different, independently developed and heterogeneous resources. In biology, for example, the hundreds of data repositories in active service have different formats, interfaces, structures, coverage, etc. (*see Buttler et al in this issue*). The Web and the Data Grid guarantee a certain level of interoperability in retrieving and accessing data. The next level of interoperability is not just making data available, but understanding what the data means so that it can be linked in appropriate and insightful ways, and providing automated support for this integration process [19].

Scientists typically link resources in two ways:
(a) *Workflow orchestration*: Process flows, or workflows coordinating and chaining services using a systematic plan, are the manifestation of *in silico* experiments, allowing us to represent the e-Scientist's experimental process explicitly;
(b) *Database integration*: dynamic distributed query processing, or the creation of integrated databases through virtual federations or warehouses [20].

Information mediation is not restricted to traditional scientific databases. Computational resources are discovered, allocated and disbanded dynamically and transparently to the user. The problem of mediation between different Grid compute resource brokering models, such as Unicore and Globus, closely resembles mediation between two database schemas.

Semantic Web and Database technologies offer great possibilities. A common data model for aggregating results drawn from different resources or instruments could use RDF. Domain ontologies for the semantic mediation between database schema [19], an application's inputs and outputs, and workflow work items could use DAML+OIL/RDF(S). Domain ontologies and rules can be used for constraining the parameters of machines or algorithms, and inferring allowed configurations. Execution plans, workflows and other combinations of services benefit from reasoning to ensure the semantic validity of the composition [21].

[1] http://www.mygrid.org.uk

So we can use Semantic Web services for:

- The classification of computational and data resources, performance metrics, job control; schema integration, workflow descriptions;
- Typing data and service inputs and outputs;
- Problem solving selection and intelligent portals;
- Infrastructure for authentication, accounting and access management.

Turning this around, we can envisage that the Base and Application services of the Semantic Web are implemented as Grid services.

Semantic Web for Grid Applications

The ultimate purpose of the Grid is to support knowledge discovery. The Semantic Web is often presented as a global knowledge base. Consider a scenario: A scientist posing the question "what ATPase superfamily proteins are found in mouse?" might get the answers (a) The protein accession number from the Swiss-Prot database she has permission to access; (b) InterPro is a pattern database but needs permission and payment. (c) Attwood's project is in nucleotide binding proteins (ATPase superfamily proteins are a kind of nucleotide binding protein); (d) Smith published a new paper on something similar in Nature Genetics two weeks ago; (e) Jones in your lab already asked this question last week.

A scientist may be advised of equipment or algorithm parameter settings, helped to choose and plan appropriate experiments and resources based on her aims and shared best practice, and ensure that conclusions are not drawn that are not fully justified by the techniques used. These are all applications of, or for, the Semantic Web, and include personalised agents or services, semantic portals onto services, recommender systems and a variety of other knowledge services [22].

The scientific community has embraced the Web. The result is commonly publication of information without accompanying accessibility. Many resources have simple call interfaces without APIs or query languages and only "point and click" visual interfaces. Scientific knowledge is often embodied in the literature and in free text "annotations" attached to raw data. The presumption that a scientist will read and interpret the texts makes automatic processing hard and is not sustainable given the huge amount of data becoming available. The Semantic Web is about making the computationally inaccessible accessible and to automate information discovery.

Provenance, quality, trust and proof.

Both the results and the way they are obtained are highly valued. Where data came from, who created it, when, why and how was it derived is as important as the data itself for user and service provider [23]. These are applications of the Proof, Trust and Digital Signatures of the Semantic Web (*see both Maximilien and Finin in this issue*). In molecular biology, data is repeatedly copied, corrected and transformed as it passes through numerous databases. Published data is actively curated automatically and by hand. Complex assemblies of programs create results from base data. Annotating results with commentaries, linking results with their sources, asserting which parameters were used when running an algorithm and why, are possible applications of Semantic Web and database technologies.

Assertions are also qualitative. Scientific knowledge is contextual and opinionated. Contexts change and opinions disagree. New information may support or contradict current orthodoxy leading to a revision of beliefs. Inferences on assertions can give new knowledge but inferences must be exposed or else the scientist will not use them. Dealing with multiple (diverging) assertions over resources, and inference engines capable of tolerating discrepancies, is a challenge of the Semantic Web.

So we can use the Semantic Web services for:

- annotating results, workflows, database entries and parameters of analyses with: personal notes, provenance data, derivation paths of information, explanations or claims;
- linking *in silico* and 'at the bench' experimental components: literature, notes, code, databases, intermediate results, sketches, images, workflows, the person doing the experiment, the lab they are in, the final paper;
- describing people, labs, literature, tools and scientific knowledge.

Scientific knowledge is replicated and archived for safe-keeping. It is essential to be able to recall a snapshot of the state of understanding at a point in time in order to justify a scientific view held at that time. This raises questions: What does it mean to garbage collect the 'Semantic Grid', and how do we recover a snapshot?

Grid Services come and go, which is why event notification is a Grid base service. As data collections and analytical applications evolve, keeping track of the impact of changes is difficult.

Scientists rerun their queries if base data changes, or new knowledge questions the underlying premise of an analysis. Mistakes or discredited information are propagated and difficult to eliminate. The ontologies and rules change. When an ontology changes in line with new beliefs, this does not wipe the old inferences that no longer hold (and how do we propagate those changes?). They must continue to co-exist and be accessible. Monitored events and items can be described using ontologies; database triggers can implement the notification mechanism.

Research Challenges and Opportunities

The development of the Web was stimulated by Particle Physics. This community was a well-organised microcosm of the general community. It had definite and clearly articulated information dissemination needs, and it had a group of smart people prepared to co-operate with the means and desire to do so. The state of play of the Grid today is reminiscent of the Web some years ago. Currently, there is limited deployment, largely driven by enthusiasts within the scientific community (indeed, the High Energy Physics Community again), with emerging standards and a degree of commercial uptake. The same might also be said of the current state of the Semantic Web deployment, though it is not clear that the same drivers are in place as existed for Web and Grid.

Meanwhile, the Web itself has enjoyed massive deployment and continues to evolve; e.g. the shift from machine-to human communications (HTML) to machine-to-machine (XML), and the emergence of the Web Services paradigm. The requirements of one of the drivers, e-Commerce, are in line with those of e-Science. However, a typical Grid application is not a typical Web application. A Grid application might involve large numbers of processes interacting in a coordinated fashion, while a typical Web transaction today still only involves a small number of hosts (e.g. server, cache, browser). Moreover, Grid processes continually appear and disappear. Achieving the desired behaviour from a large scale distributed system involves technical challenges that the Web itself has not had to address, though Web Services take us towards a similar world.

The Grid relies on metadata-enabled services. The Semantic Web requires a metadata-enabled Web. In the same way as the components of information systems have moved to support HTML and XML in the last few years, we now need them to take on the support for creating and maintaining metadata. There are many obstacles. The manual creation of metadata is problematic. People are not always in the best position to create it and they might not produce accurate metadata, through circumstance or error; there are always alternative but equally valid descriptions. Grid applications have the requirement and the opportunity to automate the management of *quality* metadata. Addressing the problems of creating, managing and linking ontologies is paramount.

Grid and Semantic Web technologies appear symbiotic and their visions are related. Grid computing benefits from the Semantic Web fabric and services for the management of its semantics. The Semantic Web benefits from the application pull provided by the Grid and the Grid infrastructure itself. The base services of the Semantic Web – ontology servers, metadata generators, ontology alignment and so on – can be implemented as Grid Services (Figure 1).

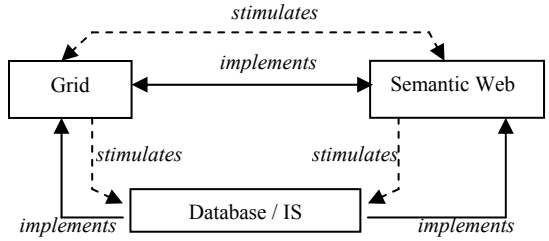

Figure 1: Grid, Semantic Web and DB/IS.

Database and Information Systems technologies are essential to build the kind of global and distributed infrastructure needed for both Grid and Semantic Web. To achieve these benefits requires that the Grid computing and applications community pay due attention to the Semantic Web. This applies to vertical projects, where the Semantic Web technologies can be applied within the application domain, and also to middleware developments, which can build in the Semantic Web infrastructure. There is a cost to taking on board new technologies, and here the benefits may not always be immediate. The Semantic Web community has a role to play in supporting the initial uptake, especially as many traditional Grid developers regard themselves as systems-oriented and the adoption of knowledge technologies seems irrelevant. One barrier to adoption is confusion over what can be achieved now and what is best treated as 'wait and see'.

Why should the Semantic Web and Database researchers be interested in the Grid?

1. It is a good example of the type of application envisaged for the Semantic Web. The essence of the Grid is the power provided *by large scale integration of resources*, and the scale and automation of the Grid necessitates the 'universally accessible platform that allows data to be shared and processed by automated tools as well as by people'.
2. It is a real application, with emphasis on deployment and performance, and is on a large scale and has established communities of users. Such applications are essential to the uptake of the Semantic Web and need Database technologies.
3. The Grid potentially greatly benefits from Semantic Web technologies. Even at the most basic level, Grid developers acknowledge that 'information islands' are being created and require an interoperability solution.
4. The Grid will stress Semantic Web solutions, and it raises some specific Grid-related issues, which will provide a useful challenge. Solutions to these issues are unlikely to be peculiar to grid computing – related issues will surely be evident in other Semantic Web applications.
5. It is self-contained, with a well-defined community who already work with common tools and standards.
6. Aspects of the Semantic Web could be applications of Grid computing, for example in search, data mining, translation and multimedia information retrieval.

The partnership between the Semantic Web and the Grid presents an exciting vision. Each partner has obstacles to its progress, but each stands to benefit from the other. Both need DB/IS input, and will stimulate DB/IS research.

Acknowledgements

This work is supported by the EPSRC and DTI through the UK e-Science programme, in particular myGrid (GR/R67743), Geodise (GR/R67705) and Comb-*e*-Chem (GR/R67729/01). The authors wish to thank all their colleagues on these projects. Special thanks to Nigel Shadbolt, Nick Jennings and Sean Bechhofer.

References

[1] I. Foster, C. Kesselman, S. Tuecke *The Anatomy of the Grid: Enabling Scalable Virtual Organizations*, International Journal of Supercomputer Applications, 15(3), 2001.

[2] S. Bethke, M. Calvetti, H.F. Hoffmann, D. Jacobs, M. Kasemann, D. Linglin *Report of the Steering Group of the LHC Computing Review*, CERN/LHCC/2001-004, February 2001.

[3] I. Foster and C. Kesselman, *Globus: A Metacomputing Infrastructure Toolkit*, Int. Journal of Supercomputer Applications, 11(2): 115-128, 1997.

[4] K. Czajkowski, S. Fitzgerald, I. Foster, C. Kesselman. *Grid Information Services for Distributed Resource Sharing.* Proceedings of the Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), IEEE Press, August 2001.

[5] A. Chervenak, E. Deelman, I. Foster et al *Giggle: A Framework for Constructing Scalable Replica Location Services*, SC2002, November 11-16, 2002, Baltimore, Maryland.

[6] A. Rajasekar, M. Wan and R. Moore, *MySRB & SRB - Components of a Data Grid*, The 11th International Symposium on High Performance Distributed Computing (HPDC-11) Edinburgh, Scotland, July 24-26, 2002.

[7] P. Watson, *Databases and The Grid*, UK e-Science Programme Technical Report Number UKeS-2002-01.

[8] N. Paton, M. Atkinson, V. Dialani, D. Pearson, T. Storey, P. Watson, *Database Access and Integration Services on the Grid*. UK e-Science Programme Technical Report Number UKeS-2002-03.

[9] I. Foster, C. Kesselman, J. Nick and S. Tuecke*, The Physiology of the Grid: Open Grid Services Architecture for Distributed Systems Integration*, GGF4, Feb. 2002. See http://www.globus.og/research/papers/ogsa.pdf

[10] C. Wroe, R. Stevens, C. Goble, A. Roberts, M. Greenwood, *A suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data*, International Journal of Cooperative Information Systems *in press*.

[11] R. Raman, M. Livny, and M. Solomon. *Matchmaking: An extensible framework for distributed resource management.* Cluster Computing: The Journal of Networks, Software Tools and Applications, 2:129-138, 1999.

[12] W3C Semantic Web Activity Statement, http://www.w3.org/2001/sw/Activity/

[13] J. Hendler, *Agents and the Semantic Web*, IEEE Intelligent Systems Journal, March/April 2001 (Vol. 16, No. 2), pp. 30-37.

[14] B. McBride, "Four Steps Towards the Widespread Adoption of a Semantic Web", in Proceedings of the First International Semantic Web Conference (ISWC 2002), Sardinia, Italy, June 9-12, 2002. LNCS 2342, pp 419-422.

[15] I. Horrocks, *DAML+OIL: a reason-able web ontology language*, in Proceedings of EDBT 2002, March 2002.

[16] *Resource Description Framework* http://www.w3.org/RDF

[17] *OWL Web Ontology Language 1.0 Reference* http://www.w3.org/2001/sw/WebOnt/

[18] A. Ankolekar, M Burstein, J Hobbs, Lassila, D Martin, S McIlraith, S Narayanan, M Paolucci, T Payne, K Sycara, H Zang *DAML-S: Semantic Markup for Web Services* in Proceedings of the International Semantic Web Working Symposium (SWWS), July 30-August 1, 2001.

[19] C.A. Goble, *Supporting Web-based Biology with Ontologies*, in Proceedings of the Third IEEE ITAB00 Arlington, VA (November 2000), pp. 384–390.

[20] I. Foster, J. Voeckler, M. Wilde, and Y. Zhao. *Chimera: A Virtual Data System for Representing, Querying and Automating Data Derivation.* Proceedings of the 14th Conference on Scientific and Statistical Database Management, Edinburgh, Scotland, July 2002.

[21] J. Cardoso and A. Sheth, *Semantic e-Workflow Composition*, Technical Report, LSDIS Lab, Computer Science, University of Georgia, July 2002.

[22] D. De Roure, N. Jennings, N. Shadbolt. *Research Agenda for the Semantic Grid: A Future e-Science Infrastructure*, UK e-Science Programme Technical Report Number UKeS-2002-02.

[23] P. Buneman, S. Khanna, K. Tajima, W-C Tan. *Archiving Scientific Data*, SIGMOD Conference 2002.