OGSA-DAI Usage Scenarios and Behaviour: Determining good practice

Mario Antonioletti[1], Malcolm Atkinson[2], Andrew Borley[3], Neil P. Chue Hong[1], Brian Collins[3], Jonathan Davies[3], Neil Hardman[3], Alastair Hume[1], Mike Jackson[1], Amy Krause[1], Simon Laws[3], Norman Paton[4], Keke Qi[1], Tom Sugden[1], David Vyvyan[3], Paul Watson[5] and Martin Westhead[1]

[1]EPCC, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ.
[2]National e-Science Centre, 15 South College Street, Edinburgh EH8 9AA.
[3]IBM United Kingdom Ltd, Hursley Park, Winchester S021 2JN.
[4]Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL.
[5]School of Computing Science, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU.

Abstract

OGSA-DAI has been developing Grid middleware for over two years now. A high profile project within the Grid community OGSA-DAI is increasingly being used by Grid based projects to provide their Data Access and Integration (DAI) requirements. From a simple set of services relatively sophisticated usage scenarios may be realised. This presentation examines a number of DAI scenarios identified by OGSA-DAI and the GGF DAIS working group and demonstrates how the existing OGSA-DAI framework satisfies them. A number of real use-cases from the projects that are using OGSA-DAI are outlined and gaps within the existing OGSA-DAI framework are identified. The OGSA-DAI software distribution and further information about the project is available from the project website at http://www.ogsadai.org.uk/.

1 Data Access and Integration Scenarios

To date, most of the effort in the *Open Grid Services Architecture – Database Access and Integration* (OGSA-DAI) project has concentrated on developing an extensible document-based framework that provides uniform data access mechanisms for a number of different data resources, i.e. mainly relational and XML based databases although there is also some support for accessing files through OGSA-DAI. Currently the framework used by OGSA-DAI is described in the draft *Grid Database Service Specification*, [1], of the Global Grid Forum's (GGF) Data Access and Integration Services (DAIS) Working Group as presented at GGF7. Although this viewpoint no longer aligns with the current direction of the DAIS specifications there are no real inconsistencies between the two approaches – differences arise mainly in the granularity of the requests going to the services. At some future point OGSA-DAI will adopt the presently defined DAIS interfaces although the document based access method will still be actively supported.

Basic data integration capabilities are currently not as well supported through the existing framework although it is possible to do this using the existing core OGSA-DAI distribution. However, this relies on an OGSA-DAI client explicitly chaining services together – clearly this is neither a scalable solution nor necessarily an easy one to implement. Work is currently underway at EPCC investigating simple data integration scenarios with a view to understanding how well the existing OGSA-DAI distribution serves this and to try and identify what additional functionality might be added to facilitate data integration capabilities. It is worth noting though that OGSA-DAI does not have to provide the whole of the data integration layer. Instead, it can exploit data integration capabilities provided by the data resource themselves or through some other third party means and wrap these data resources, already constituting a federated resource, which may then be further federated with additional data resources of different types by an OGSA-DAI layer. This model of operation is similar to one used by the eDiaMoND project (see below).

However, as things stand, we find that the data access scenarios (1-6) and data integration scenarios (7,8) identified by Greg Riccardi and Simon Laws, shown in Figure 1 below, can all be supported by the current OGSA-DAI distribution. In addition, many of the data access and integration requirements of existing projects employing OGSA-DAI may be classified under these scenarios.

More sophisticated data integration capabilities are realisable via Distributed Query

Processing (DQP), currently available as a separate package that may be downloaded from the OGSA-DAI web site. DQP is layered over OGSA-DAI providing an additional abstraction layer that allows queries to be applied to various relational data resources as though they were a single logical resource.

Key to Symbols			
Data Flow			
→	Call	⇢	Response
Actors			
○	Non-OGSI process	□	OGSI processes
A	Analyst	G	GDS
C	Consumer	P	Producer
Data			
Q	Query definition	D	Delivery definition
S	Status	R	Result
U	Update data	I	Data id (URI)

Table 1: Key to Symbols in Figure 1

This is done through an additional set of Grid services that extend the scope of OGSA-DAI: one of these services acts as the point of contact for a client and orchestrates other services behind the scenes, including services that evaluate queries on each data resource. Data integration scenarios can be managed at either the client or service end; DQP illustrates an extension to OGSA-DAI at the service end, enabling data integration. Future work within OGSA-DAI will investigate to see whether it is possible to integrate this type of functionality directly on to the core OGSA-DAI distribution.

The remainder of this paper describes usage scenarios provided by projects using OGSA-DAI for real data-centric Grid applications and compares them with the abstract usage scenarios that have been identified. Perceived gaps in the existing OGSA-DAI framework are also outlined and some areas for future investigation are summarised.

2 Project Scenarios

Figure 2 shows a number of projects that use OGSA-DAI and the following sections provide a brief summary of how the software is used.

2.1 AstroGrid

The AstroGrid project (www.astrogrid.org) is trying to build a data Grid for UK astronomy, which will form the UK contribution to a global virtual observatory. OGSA-DAI is being used to build a prototype of a Grid data warehouse in which large data extracts from many data

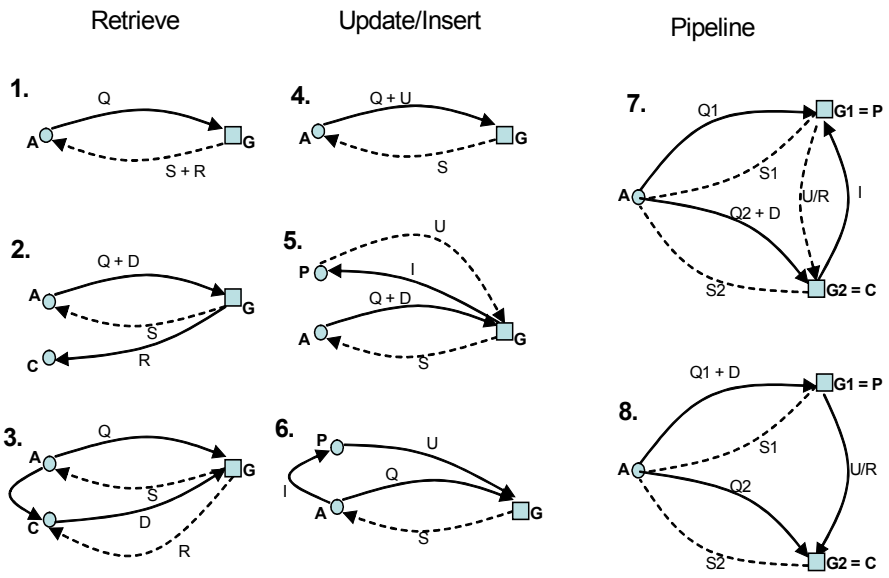

Figure 1: Data Access and Integration Scenarios

centres can be combined for easier analysis. The extraction and combination of large data sets means the emphasis must be on efficient streaming and processing of data.

2.2 myGrid

The myGrid project (www.mygrid.org.uk) uses OGSA-DAI and the DQP package to provide uniform access to multiple MySQL data resources via a single query interface. The DQP software is being developed within the context of the myGrid project.

2.3 GeneGrid

The scenario realised in the GeneGrid project (www.qub.ac.uk/escience/projects/genegrid) has OGSA-DAI providing integrated access to GeneGrid databases, public biological databases and proprietary databases. These databases are in differing formats: flat files, XML databases and relational databases; OGSA-DAI wraps these resources with varying degrees of success, and supplies a unified mechanism for accessing the data contained therein.

2.4 BioGrid

The BioGrid project (www.biogrid.jp) has developed a system for the federation of bio-related databases with OGSA-DAI, aiming at producing an application for Drug Discovery. The system has currently bridged 11 databases in heterogeneous communities, such as biology, medical science and pharmaceutics.

2.5 BioSimGrid

The BioSimGrid project (www.biosimgrid.org) is building an open software framework system based on OGSA and OGSA-DAI to deliver data analysis and data mining services to the biomolecular simulation and structural biology communities. Associated applications can be developed independently as distributed services and integrated into the system as well as using off-the-shelf middleware components.

2.6 eDiaMoND

Another scenario is realised in the eDiaMoND project (www.ediamond.ox.ac.uk). OGSA-DAI does not intrinsically support data federation, so in eDiaMoND the data resources, IBM® DB2® databases, are federated at the product level and in turn these are wrapped using OGSA-DAI. Additional data resources, which are not constrained to just DB2®, may then be integrated at the Grid service level using OGSA-DAI.

2.7 ODD-Genes

The ODD-Genes project (www.epcc.ed.ac.uk/oddgenes) employs OGSA-DAI to perform tightly linked queries on gene identifiers against remote, independently managed databases. Researchers use OGSA-DAI to remotely browse the results of data analysis batch jobs stored in a relational database. Based on the results of their own analyses, researchers search an OGSA-DAI registry of genetics databases. The registry

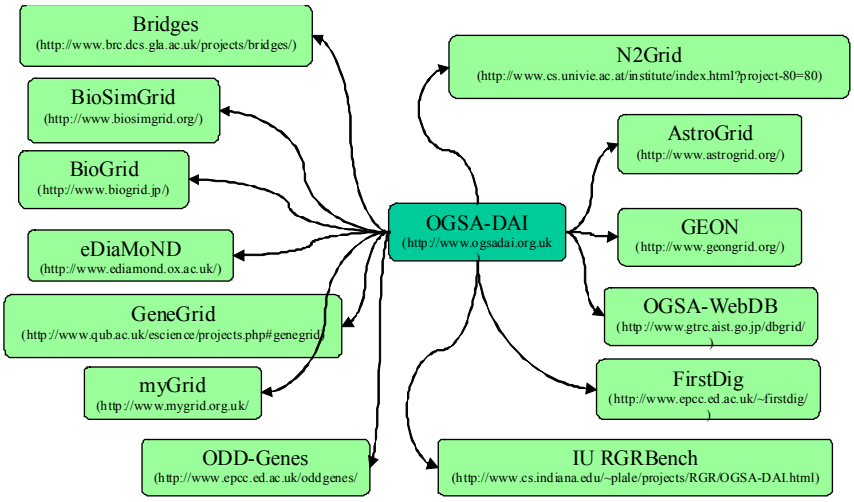

Figure 2 - Projects that employ OGSA-DAI

provides Grid locations of other databases with potentially relevant data that can then be queried for further analysis.

2.8 BRIDGES

The BRIDGES project (www.brc.dcs.gla.ac.uk/projects/bridges) is investigating the application of OGSA-DAI and IBM's Information Integrator product to deal with federation of distributed biomedical data. The data is stored in multiple locations and formats and each location has differing security policies governing access to the data.

2.9 N2Grid

N2Grid (www.cs.univie.ac.at/institute/index.html?project-80=80), a neural network environment based on the Grid, is an evolution of the existing NeuroWeb system. The system realises all components of an artificial neural network as data objects in a Grid enabled database.

2.10 GEON

The GEON project (www.geongrid.org) uses OGSA-DAI as part of the GEON systems software stack deployed on all GEON nodes. OGSA-DAI is currently being evaluated to provide access to data from databases running on remote GEON nodes.

2.11 OGSA-WebDB

OGSA-WebDB (www.gtrc.aist.go.jp/dbgrid) has been developed to bring existing web database resources into the OGSA environment. The project has created an extension to the OGSA-DAI relational resource implementation allowing users to query web database resources via OGSA-DAI.

2.12 FirstDIG

The FirstDIG project (www.epcc.ed.ac.uk/~firstdig) used the client toolkit API introduced in Release 3.1 of OGSA-DAI to build a more sophisticated client that allows queries to be run against a virtual table representing the distributed join of data from geographically separated, heterogeneous relational databases. This was achieved without needing to modify the core OGSA-DAI components. Instead, the browser, utilising the client toolkit, managed the creation of temporary tables to achieve this pattern. The project finished in February 2004. Further information on the project and its outcomes is available from the project web site.

2.13 IU RGRBench

IU RGRBench (www.cs.indiana.edu/~plale/projects/RGR/) is a Grid information services benchmark/workload used to better understand resource information management in Grid information servers. The benchmark suite comprises a set of queries and scenarios that are defined over a data model of Grid resources and was run against OGSA-DAI wrapping a relational database. The project also created a portlet client interface for OGSA-DAI, enabling data access via a portal.

3 Scenario Evaluation

Comparing the project scenarios described in section 1 against the data access scenarios and data integration scenarios shown in Figure 1, it is clear that the majority of projects are using OGSA-DAI for data access only and are mostly following the retrieval scenarios (1-3). As OGSA-DAI stabilises and users gain confidence and experience with the software, projects are starting to look at exploiting some of the more complex scenarios that can be realised with OGSA-DAI. For example, the AstroGrid project plans to employ a data integration scenario where data is transferred from the original source into dynamically created user tables for subsequent analysis.

Other factors also arise when considering scenarios using OGSA-DAI, such as scalability and performance. The OGSA-DAI development team have focussed on these issues in recent releases, improving both memory and processing performance. OGSA-DAI release 4.0 also scales far better than previous releases – enabling integration scenarios where millions of database rows are transferred between a data source and a data sink.

Data integration scenarios explicitly realised within OGSA-DAI include the DQP package described in section 1 and the functionality provided by the OGSA-DAI Data Browser. The Data Browser is a client that allows queries to be run against a distributed join of data from multiple relational databases. The Data Browser was contributed to OGSA-DAI by the FirstDIG project (see section 2.12). EPCC is currently investigating some simple data integration scenarios using OGSA-DAI with a view to creating some more 'off-the-shelf' data integration tools and extending the OGSA-DAI activities framework to facilitate the easy implementation of bespoke data integration solutions.

4 Further Investigations

Clearly there are many additional usage scenarios which are of interest. The OGSA-DAI project is starting to consider these, particularly in the area of data integration. As well as producing software which can be used to tackle specific data integration problems, this work also provides valuable feedback on the completeness and efficiency of the underlying OGSA-DAI data access components which it is based on, in a similar manner to the ongoing OGSA-DQP work.

Some features under consideration which may aid in the provision of data integration tools include:

- Reuse of activity and process definition allowing more rapid development of data integration patterns, akin to using stored procedures.
- A higher level workflow framework which would describe the relationships between GDSs and data consumers, how they cooperate with each other, and how data transportation is defined. This would also allow better interoperability with other services, e.g. job submission services.
- Transactional capabilities within the perform document would allow more complex data scenarios to be expressed. This would be dependent on existing web services transactions standards being implementable within the data access context. Some preliminary work has already taken place within the project.
- Embeddable code within perform documents could help reduce data transfer and increase the scalability of OGSA-DAI. An approach similar to the Java code snippets described in BPELJ [2] could be taken, with a code sandbox within the GDS allowing access to activities running on the service.
- User-definable exception handling, like the ability to define transactions, would allow more complex workflow patterns to be expressed within a perform document.

Obviously there is insufficient available effort within the current project to cover all of the possible areas in the available time. We are interacting closely with our users to ensure that the scenarios and features that are most desirable are given priority. Regular twice-yearly user group meetings have been set up to pool experience and gauge opinion. More details about the first OGSA-DAI user group meeting are available from [3].

Other fields that OGSA-DAI has not currently explored include much of the data virtualisation space, such as a common namespace, a common schema, a common query language and functionality such as replication, caching, distribution and federation. Again, OGSA-DAI plans to investigate these areas in the future.

5 Conclusions

The OGSA-DAI middleware has now been in existence for over two years. It is increasingly being used by real projects, 15 to date[1], to facilitate the provision of Data Access and Integration capabilities within a Grid context. This paper outlines some of the usage patterns being employed in order to document and better understand how the Grid is already being successfully employed by data-centric applications. Through this process, as well as a series of user group meetings and surveys, OGSA-DAI hope to be able to elicit additional usage scenarios and identify areas where OGSA-DAI may be developed to contribute more effectively to the deployment of future data Grid applications.

6 Acknowledgements:

This work is supported by the UK e-Science Grid Core Programme, whose support we are pleased to acknowledge. We also gratefully acknowledge the input of our past and present partners and contributors to the OGSA-DAI project including: IBM UK, University of Manchester, University of Newcastle and Oracle UK.

7 Copyright

[1] For current information on usage of OGSA-DAI see www.ogsadai.org.uk/projects.

8 References

[1] Chue Hong, N., Krause, A., Malaika, S., McCance, G., Laws, S., Magowan, J., Paton, N.W., Riccardi. G. *Grid Database Service Specification*, 16th February 2003.

[2] BPELJ: BPEL for Java technology whitepaper. Available from: http://www-106.ibm.com/developerworks/webservices/library/ws-bpelj/

[3] Presentations and notes from the first OGSA-DAI users' meeting are available from: http://www.ogsadai.org.uk/docs/docs.php#ug1.