OGSA-DAI Status Report and Future Directions

Mario Antonioletti[1], Malcolm Atkinson[2], Rob Baxter[1], Andrew Borley[3], **Neil P. Chue Hong[1]**, Brian Collins[3], Jonathan Davies[3], Desmond Fitzgerald[4], Neil Hardman[3], Alastair C. Hume[1], Mike Jackson[1], Amrey Krause[1], Simon Laws[3], Norman W. Paton[4], Tom Sugden[1], Paul Watson[5], Martin Westhead[1] and David Vyvyan[3]

1. EPCC, University of Edinburgh, JCMB, The King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK.
2. National e-Science Centre, Universities of Edinburgh & Glasgow, Edinburgh EH8 9AA, UK.
3. IBM United Kingdom Ltd, Hursley Park, Winchester S021 2JN, UK.
4. Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK.
5. School of Computing Science, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK.

Abstract

The OGSA-DAI middleware has been publicly available for over two years. OGSA-DAI facilitates Data Access and Integration (DAI) of data resources, such as relational and XML databases, within a Grid context. Project members also participate in the development of DAI standards through the GGF DAIS WG. The standards that emerge through this effort will be adopted by OGSA-DAI once they have stabilised. The OGSA-DAI developers are also engaging with a growing user community to gather their data and functionality requirements. Several large projects are already using OGSA-DAI to provide their DAI capabilities. This paper presents a status report on OGSA-DAI activities since the last AHM and announces future directions. The OGSA-DAI software distribution and more information about the project is available from the project website at http://www.ogsadai.org.uk/.

1 Present Status

1.1 Project Background

The Open Grid Services Architecture – Data Access and Integration (OGSA-DAI) project started in February 2002. It received £3.3 million funding for two years from the UK Core e-Science funding programme to develop Grid enabled middleware to facilitate data access and integration capabilities for UK based e-Science projects. The project was tasked with producing software based on the Globus Toolkit 3 which, in turn, was based on the then emerging Global Grid Forum's (GGF) Open Grid Services Infrastructure (OGSI) specification [1].

Over this first funding period three major and four minor releases of the OGSA-DAI distribution have been produced. These had increasing levels of sophistication and functionality as well as being able to interoperate with other Grid middleware available at the time. OGSA-DAI was regarded as being a key Grid technology to the extent that it became a contributed component to the Globus Toolkit (GT). OGSA-DAI is now also distributed with the Globus Toolkit starting with the GT3.2 release.

Part of the project's remit has also been to try to standardise data access interfaces for Grids. This led to the formation of the GGF Data Access and Integration Services (DAIS) Working Group, which had its first working group meeting at GGF5. An important part of the DAIS effort comes from the OGSA-DAI team. Since the GGF5 meeting DAIS has been attempting to standardise data access interfaces through the GGF. A series of draft specifications have been produced for every GGF meeting. The convergence of these standards has been slow but the slowness merely reflects the overall fluid state of Grid standards and infrastructures.

The original funding for OGSA-DAI came to an end in September 2003. Additional funding, £1.3 million, for a further two years of development was provided by the UK e-Science Core Programme II, administered by the EPSRC. Funding was also made available by the UK Open Middleware Infrastructure Institute. The continuation project, called DAIT (DAI Two), commenced in October 2003. The established development team, an academic-industrial partnership consisting of EPCC and IBM teams, was largely preserved albeit slightly reduced in numbers. The software product name, OGSA-DAI, has been maintained.

The first OGSA-DAI release under DAIT, release 4.0, was made available in mid-May 2004. A release roadmap has been produced and provisional functionality and capabilities for the releases have been made. DAIT plans to produce four major releases over the two-year funding period. The OGSA-DAI contribution to the GGF DAIS standardisation effort is also planned to continue.

1.2 Dissemination and Requirements Capture

Documentation is regarded to be an integral and important part of the OGSA-DAI distribution. The OGSA-DAI documentation is extensive and comprehensive. For release 4.0 of the OGSA-DAI distribution the format of the bundled documentation shifted from using HTML and PDF formats to XHTML only. This was mainly done to curtail the onerous task of keeping various versions in different formats synchronised. Also, the XHTML format can, in principle, easily be transformed into other formats. Tutorial material has also been added covering the client toolkit API for application developers.

Courses provide an important way of disseminating information about OGSA-DAI. These not only provide users with an opportunity to learn about OGSA-DAI but also enable them to provide feedback about their usage of the current OGSA-DAI distribution and to express their data and functionality requirements directly back to the development team. A number of OGSA-DAI courses have been given at NeSC as well as other high visibility events such as the 2003 UK All Hands Meeting (AHM), GGF7, GGF11 and the first and second Grid Summer Schools held in Naples, Italy. An OGSA-DAI course is also planned for this AHM meeting as well as an OGSA-DAI mini-workshop where users and developers will recount their experiences and use of OGSA-DAI.

A more formal attempt to gather user requirements occurs through user group meetings. At these users can engage directly with OGSA-DAI developers. The first OGSA-DAI user group meeting was held at NeSC in early April 2004. Thirty-two attendees were present representing at least twelve projects that have used, are using or intend to use OGSA-DAI. Representatives from some of these projects gave presentations where they recounted their OGSA-DAI experiences. A couple of breakout sessions were organised to capture user requirements for future versions of OGSA-DAI. One of these concentrated on functional requirements, the other on deployment infrastructures.

The functional requirements were enumerated under various categories, and then prioritised. The top three short-term requirements for future releases of OGSA-DAI were deemed to be:

- *Reliability*: OGSA-DAI should be stable and terminate gracefully without causing disruption to the service container after a critical failure. For instance, memory errors are not handled well by Java and can cause random threads to silently die in an Apache Tomcat container leading to unpredictable behaviour.
- *File access*: various groups would like OGSA-DAI to provide data access to files. These requirements came predominantly from the biological sciences groups but were also shared by other groups.
- *Large result sets*: OGSA-DAI should be able to handle requests that return large data sets. The production of large result sets should not lead to memory problems in OGSA-DAI.

These requirements have largely been taken into account in Release 4.0 OGSA-DAI distribution. Some protection has been added against `java.lang.OutOfMemoryError` errors and the OGSA-DAI implementation will try to return an exception if the JVM heap is almost full.

There are more file access prototypes available within this release and the handling of large result sets has been much improved. The main constraining factor is no longer OGSA-DAI but rather the quality of the underlying database driver.

The second parallel break out session discussed the infrastructures that OGSA-DAI should deploy to. From this it became clear that various groups would like a web services implementation of OGSA-DAI that deploys to *vanilla* installations of Tomcat and Apache Axis, a "WS-I version". There was also a requirement for a WS-Resource Framework (WS-RF) compatible implementation of OGSA-DAI, i.e. one that would deploy to GT4.0 (expected at some point in the third quarter of this year). OGSA-DAI is endeavouring to see whether both these requirements can be met.

More details about the first OGSA-DAI Users' Group meeting are available from [2]. It is intended that future user group meetings will

be held at approximately six month intervals close to major OGSA-DAI releases, driven by a chair and panel nominated by users. The next OGSA-DAI user group meeting has been provisionally scheduled for October 2004.

Additional feedback was gained from a user survey. This received an initial twenty-nine responses primarily from UK e-Science projects. All respondents signalled an eventual need for OGSA-DAI although only eleven are currently using OGSA-DAI. Most were interested in a "WS-I version" of OGSA-DAI in the near future, many were also interested in a "WS-RF version" of OGSA-DAI later "when WS-RF is settled". Primarily all were interested in the functionality provided by OGSA-DAI rather than being concerned about the underlying specifications as long as it was robust and interoperable. It should also be noted that this survey represented projects with a combined total of over 15000 users.

1.3 OGSA-DAI Releases and Usage

OGSA-DAI has gained a high profile within Grid communities. By the end of June almost 2500 downloads have been recorded from the OGSA-DAI web site alone. In addition to this, OGSA-DAI is also independently distributed with the Globus Toolkit 3.2. We do not have information about these downloads.

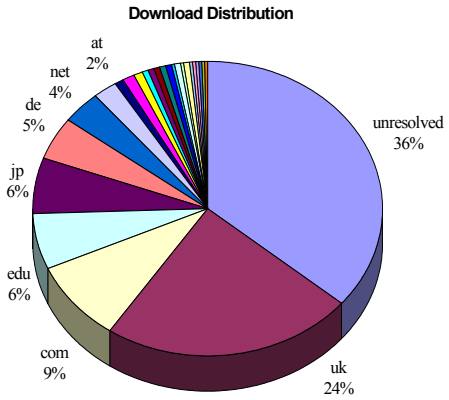

Figure 1 - OGSA-DAI downloads by geographical distribution

The pie chart in Figure 1 classifies the total number of downloads, for all releases, according to the top level DNS domain of the machines that people have used to download the OGSA-DAI software. Only domains with over 2% of the total number of downloads are labelled. As can be seen downloads primarily originate in the UK. A sizeable number of the *unresolved* domains (IP addresses only) originate in China.

As far as we know at least fifteen projects are evaluating, using, or have used OGSA-DAI. These range from the bioinformatics community (GeneGrid), through metacatalog services, to the transport industry (FirstDIG). The ones that have volunteered information about their usage are enumerated with some information about their usage at the project website (http://www.ogsadai.org.uk/projects). More details of how these projects are using OGSA-DAI are described in a companion paper submitted to this AHM [3].

Release	Date of Release	Number of Downloads (at 6/04)
4.0	May 2004	348
3.1	February 2004	573
3.0	July 2003	792
2 interim	June 2003	291
2.0	April 2003	249
1 interim	February 2003	108
1.0	January 2003	104

Table 1 - OGSA-DAI downloads by release

1.4 Current Release

The current release of OGSA-DAI is 4.0. Some of the main highlights of this release are:

- Works with the GT3.2 release which is the current stable version of the Globus Toolkit;
- The client toolkit library is now a fully supported OGSA-DAI component. This should make it easier to develop OGSA-DAI client applications and should provide future OGSA-DAI applications that use the toolkit with some protection against inevitable changes to OGSA-DAI service interfaces;
- Memory and performance issues have been addressed. OGSA-DAI can now return result sets with millions of rows. The main delivery activities, e.g. GDT (Grid Data Transport portType), HTTP,

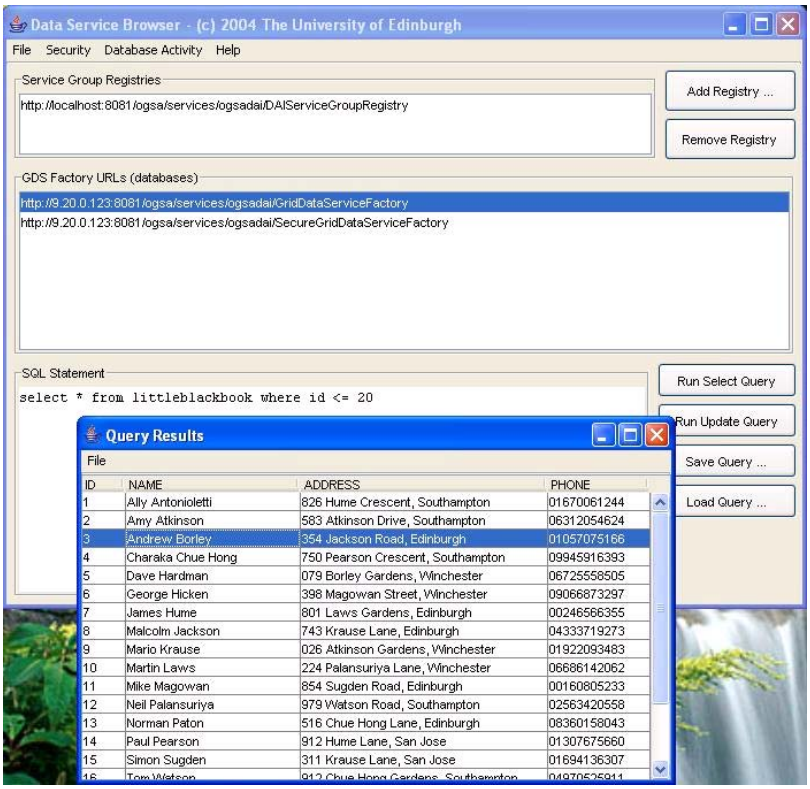

Figure 2 - The OGSA-DAI Data Browser

FTP, GridFTP, StreamToServlet, etc. are now fully streaming. The maximum size of a result set is now limited by the underlying database driver and not OGSA-DAI;

- The SQL Server and PostgreSQL relational DBMSs are now officially supported databases;
- A GUI Data Browser, see Figure 2 overleaf, is now available that allows interaction with OGSA-DAI services representing relational resources;
- Bulk load is now supported. This allows large data sets to be entered directly into a resource;
- OGSA-DAI now supports the final WebRowSet XML schema specified in JSR114 [4];
- Message level security has been added to the Grid Data Transport;
- OGSA-DAI documentation is now distributed in XHTML format rather that PDF. In part this was due to an internal rationalisation but also it also offers users greater flexibility;
- Static metadata defined in a GDSF configuration file is now registered with a DAISGR;
- There is some support for Stored Procedures (for DB2 only).

2 Future Plans

2.1 Compliance with Standards

The emerging Grid standards and middleware implementations that OGSA-DAI is dependent on are in almost as much flux two years on as they were at the beginning of the project. This has an obvious direct impact on the development and evolution of OGSA-DAI. For instance, it makes it very hard to produce a stable platform for Grid developers who seek a

reasonably long shelf life for their applications. OGSA-DAI faces these same problems but it hopes to try to protect developers who employ OGSA-DAI in their applications from many of these changes through the provision of a client toolkit. The OGSA-DAI client toolkit has been developed to make it easier to program OGSA-DAI applications and also shield developers from changes to the underlying interfaces. More information is available from [5].

OGSA-DAI is currently based on a GGF7 version of the DAIS specification [5]. This defines a single document-based operation called perform. Post GGF12 an attempt is planned to re-align with a subset of the current DAIS specifications to implement and evaluate some of the interfaces and functionality that have been defined in these new versions of the specifications. The existing document-based interface will be preserved as projects that have adopted OGSA-DAI are already using this interface and have found it to be useful.

OGSI has now been superseded by the WS-Resource Framework set of proposals for standards. The Globus Alliance is working hard to produce an early implementation of a version of these draft specifications with a release of the Globus Toolkit 4.0 expected to come out in the third quarter of this year. OGSA-DAI plans to produce a release of OGSA-DAI that will sit within the final release of GT4.

As has already been mentioned there is a strong call from some UK e-Science projects to implement a version of OGSA-DAI that operates using only Apache Tomcat and Apache Axis. This requirement is currently being investigated by OGSA-DAI and we should be in a position to report the details of our approach and subsequent plans for OGSA-DAI at the AHM.

In addition, the definition of a contribution policy covering legal, technical and product management requirements will allow other projects to develop and disseminate useful additional components to the rest of the community. The integration of the (internally produced) OGSA-DQP component is being used as a pilot study.

2.2 Development Strategy

Section 1.2 highlights how users will drive the future strategy for the development of OGSA-DAI in two basic ways; an infrastructure that is acceptable to the widest group of users and the desire for new function.

The few direct OGSA-DAI dependencies on GT3.2 will be removed. A version of OGSA-DAI will be produced depending only on basic web services technology, i.e. without reference to grid service extensions. This will have reduced function compared to the current version but the current GT3.2 based version will be retained in the short term. This provides the platform to further develop OGSA-DAI support for other infrastructures, such as WSRF, and other interfaces, such as DAIS.

Additional function will generally be added to OGSA-DAI as new activities. As DAIS support is added new WSDL will be defined to support new function. This new WSDL can either be contributed back to DAIS or maintained as an OGSA-DAI extension to DAIS.

2.3 Future Release Schedule

Provisional release dates for future versions of OGSA-DAI and possible are outlined below:

- **Release 5 (October 2004)**
  - The main release of OGSA-DAI R5 will continue to be based on GT3.2. Basic web services and WS-RF (using GT4-core) interface implementations of OGSA-DAI will be investigated in parallel with development on R5. These may be made available as technical previews of OGSA-DAI alongside the main release.
  - Possible alignment with the DAIS specifications if these are deemed to be stable. It is likely that the technical previews of basic web services and WS-RF versions of OGSA-DAI will also demonstrate a version of the DAIS interfaces.
  - Distributed Relational Query Processing may be integrated into the OGSA-DAI release. Currently distributed query processing is supported using a separate component.
  - Extended support for particular file formats (provisionally CSV and EMBL).
  - Improved dependability and security integration
  - Extended and integrated XML and relational facilities, in particular looking at database management functionality.

- Better support for stored procedures across more databases.
- Coordinated OGSA-DAI contributor community

- **Release 6 (April 2005)**
  - Functionality driven by user group
  - New facilities depend on user priorities, context and research
  - OGSA-DAI components from contributor community
  - Increased data integration tools
- **Release 7 (September 2005)**
  - Maintainable release for the user community

3 Conclusions

This paper provides a summary of the current status of the OGSA-DAI project and gives an indication to UK e-Science projects of the future direction and development plans of the OGSA-DAI software distribution.

It is clear that OGSA-DAI cannot pursue all avenues of interest and so requirements coming from users are valuable for providing direction to the project. We will be seeking further user requirements to help define the roadmap of OGSA-DAI through extended user interaction, as well as ongoing research efforts within the project. The AHM meeting affords OGSA-DAI the opportunity to report back on what has been achieved to the UK community and also communicates the future roadmap while soliciting feedback from UK users as to whether this meets their future data access and integration requirements in a Grid context.

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: EPCC, IBM UK, IBM US, NeSC, University of Manchester, University of Newcastle and Oracle UK.

4 Copyright

5 References

[1] S. Tuecke, K. Czajkowski, I. Foster, J. Frey, S. Graham, C. Kesselman, D. Snelling, P. Vanderpilt, Open Grid Services Infrastructure, Version 1.0, http://www.gridforum.org/ogsi-wg, March 13, 2003.

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

[3] Antonioletti, M., Atkinson, M., Borley, A., Chue Hong, N., Collins, B., Davies, J., Hardman, H., Hume, A., Jackson, M., Krause, A., Laws, S., Paton, N., Sugden, T., Vyvyan, D., Watson, P. and Westhead, M., *OGSA-DAI Usage Scenarios and Behaviour: Determining good practice*. AHM 2004.

[4] Bruce, J., *JSR 114: JDBC Rowset Implementations*. Final Release, 07 April 2004. See: http://jcp.org/en/jsr/detail?id=114.

[5] Hume, A., Sugden, T., Jackson, M., Antonioletti, M., Chue Hong, N., Krause, A. and Westhead, M. *Protecting Application Developers – A Client Toolkit for OGSA-DAI*. AHM2004.

[6] 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. Available from: http://forge.gridforum.org/projects/dais-wg/document/Grid_Data_Service_Specification-GGF7/en/1.