

# Data Integration in Bioinformatics Using OGSA-DAI

Shirley Crompton<sup>1</sup>, Brian Matthews<sup>2</sup>, Alex Gray<sup>3</sup>, Andrew Jones<sup>3</sup>, Richard White<sup>3</sup>

<sup>1</sup>CCLRC, Daresbury Laboratory, Warrington WA4 4AD

<sup>2</sup>CCLRC, Rutherford-Appleton Laboratory, Didcot, Oxfordshire OX11 0AX,

<sup>3</sup>Cardiff School of Computer Science, Cardiff University, Cardiff CF24 3AA

## Abstract

The BioDA project is investigating how Bioinformatics GRIDs, a data and compute intensive domain, could benefit from using a standard framework, such as OGSA-DAI, to manage access to, and integration of, distributed heterogeneous data resources. In this paper, we outline the common data access and integration requirements from the bioinformatics community. We then highlight some specific issues encountered while designing an OGSA-DAI exemplar application for BiodiversityWorld, a Biodiversity GRID that specialises in extracting knowledge from correlating a plethora of distributed heterogeneous data sources in the study of biodiversity patterns.

## 1 Introduction

Within the diverse field of bioinformatics, there are many types of data analysis, both inter- and intra-disciplinary, which generate as many types of data and databases. The increased computational capacity of the GRID makes it possible for scientists to correlate and identify patterns arising from combining large numbers of datasets to formulate hypotheses, which can be tested using further datasets and turned into useful knowledge [1]. Such research activities, in turn, generate yet more datasets that will need to be integrated into further analyses.

The bioinformatics projects supported within the UK eScience programme recognised the need for data access and integration to new and legacy data sources. As a consequence, solutions to the data management problems have been implemented individually by each project according to their needs.

For example, the Biodiversity World (BDWorld [2, 3, 4]) project is creating a problem-solving environment targeted at providing support for biodiversity researchers to use common software tools in a GRID environment, and to use them to analyse data held in a variety of databases and data stores. BDWorld's middleware for data access and communications has been developed so that it can cope with changes caused by the evolving GRID middleware. Because the BDWorld sites have collaborated in two previous projects, they are interfacing to software they developed in the SPICE project [5] to co-ordinate access to some of the databases. In SPICE a CAS (Common Access System) hub was created to allow heterogeneous databases (often managed by legacy database systems) to be wrapped, accessed and linked to form a Catalogue of Life for the international Species 2000 project [6]. In the BDWorld project this prototype Catalogue of Life is used to provide taxonomic data about species which is then linked with data from biotic and abiotic

datastores. These are wrapped in a somewhat different way from the SPICE databases, as will be explained later, since a fixed common data model is not appropriate for the more diverse range of data used in BDWorld. This data includes geographical data about the distribution of species, climate data, genetic structure and sequence data. Existing analytic tools, such as tools for modelling a species' climatic niche, are also wrapped for inclusion in the PSE. The data is linked by the problem solving environment's tools to enable bioinformatics users to investigate scientific questions such as the biodiversity richness in regions of interest; the effect of climate change on the biodiversity of a region, and the usefulness of geographical data in refining phylogenetic hypotheses.

In parallel with the application projects in bioinformatics, the Open Grid Services Architecture - Data Access and Integration (OGSA-DAI) project [7] has been concerned with constructing generic middleware to assist with access and integration of data from separate data sources via the GRID, for use in a wide variety of e-Science projects. OGSA-DAI has produced a software package which is integrated with the Globus toolkit [8] to support accessing, querying and processing data stored in relational/XML databases and flat files. Further extensions consider using OGSA-DAI directly for distributed query processing (OGSA-DQP [9]).

However, it became clear that OGSA-DAI was not being used in bioinformatics projects to any great extent. In discussions with investigators from various projects we found that this was mostly due to reluctance on their part to use emerging prototype software part-way through their projects. These projects were high risk developments because of the large scale collaborations involved and the utilisation of immature GRID software. Thus the staff of these projects did not want to add another unknown factor into their development strategies. It should also be remembered that the start of the OGSA-DAI project

coincided approximately with the start of these projects.

The Bioinformatics and DAIT (BioDA [10]) project is a one-year study to investigate the benefits of using OGSA-DAI in bioinformatics GRIDs, by establishing communication between bioinformatics projects and the OGSA-DAI team, by eliciting requirements from bioinformatics projects, and through case studies involving existing bioinformatics projects. For example, BDWorld's database handling is characterised by the diversity of the types of database used, the heterogeneity of the data with respect to its representation, and the variety of data being held and used in the analysis environment. This makes it an ideal test bed for OGSA-DAI as it will present many of the problems that such database middleware should be able to overcome more easily than traditional approaches to interoperability can.

In this paper we will highlight the generic data integration requirements gathered from the bioinformatics community. Then we examine some specific data integration issues arising from introducing OGSA-DAI to the BDWorld GRID, and discuss other OGSA-DAI integration tools that we are not using at present.

## 2 Generic Bioinformatics Data Access and Integration Requirements

The BioDA project organised a one-day workshop to bring together architects and infrastructure developers from the bioinformatics domain and the DAIT project to examine the community's needs for data access and integration in GRIDs with particular reference to OGSA-DAI.

At the workshop, 17 key requirements were gathered and these were refined through a survey of 8 bioinformatics projects at various stages of development (see [11]). Our findings indicate that these projects are particularly keen to see OGSA-DAI offering more features and support for:

- a. schema integration
- b. schema mapping
- c. mixed language query
- d. complex join across databases
- e. provenance data
- f. flexible resource discovery facilitated by a richer metadata registry
- g. RDF database access

The first four requirements map directly to data integration functionalities. The remaining three items reflect implicit needs for better metadata which will facilitate the selection and the location of distributed data resources via a metadata-driven two-step access to data [12].

The DAIT team may see many of the listed items as outside its original remit, and, therefore, as features which could be provided elsewhere. For instance, schema mapping has been implemented by others as part of a higher-level mediation service layered over

OGSA-DAI (see Section 4). We appreciate this argument, but these requirements are highly desirable to bioinformatics project practitioners, and their implementation would greatly enhance OGSA-DAI's appeal to potential users in this domain. If the DAIT team is unable to develop such features itself, we recommend that any team that does take on this responsibility co-operate closely with DAIT.

Apart from the ubiquitous call for more functionality, our findings show that bioinformatics projects with commercial users/partners are very anxious about the security of their data. They have sought reassurance over the security of the data delivery mechanisms and even the latency of the subsequent footprint that the data leaves on the server. The issue is further complicated by the lack of coherent security models with the evolving WS-RF [13] and WS-I [14] specifications which OGSA-DAI now supports. This issue needs to be resolved, if bioinformatics projects with commercial users/partners are not to be deterred from adopting the product despite its utility.

OGSA-DAI's recent migration to Globus WS-RF and OMII WS-I platforms has also affected users' confidence in the product. Infra-structural changes are disruptive and perceived as risky to project development. Our respondents have highlighted the need for OGSA-DAI to provide backward compatibility and to minimise the effects of new developments on current client users. On the issue of support, we suggest that an official policy relating to the establishment of a medium to long-term support service, i.e. beyond the current funding lifetime of the OGSA-DAI/DAIT project, would help reassure potential users that the product is not going to become unusable through the lack of continued maintenance.

We have highlighted the principal data access and integration requirements gathered via the BioDA workshop and survey. Further details may be found in the BioDA Interim Report [11]. The requirements gathered are useful to BioDA in the specification of OGSA-DAI case studies for bioinformatics projects. The information has also been fed back to the OGSA-DAI/DAIT team to assist with the development of their product.

## 3 BDWorld Data Integration Issues

In this section, we outline the initial design and some data access and integration issues encountered while developing an OGSA-DAI (R5) exemplar for BDWorld.

First we review the distinctive characteristics of the BDWorld GRID that have influenced the design of this exemplar. A particular feature of BDWorld is its usage of heterogeneous and "legacy" data resources with diverse structures and data standards. Many of these are internet information resources only accessible via HTTP/XML protocol, or even as HTTP/HTML, in which case 'screen scraping' techniques are required. In contrast to the diversity of its data resources, a

limited range of operations on these resources is typically required. For instance, one operation is to create a study data set by aggregating data from iterative searches of remote data collections using the same taxon object (representing a species or other group) as the search parameter.

BDWorld has also taken the position that any computationally intensive task will be carried out within a single resource [2]. This has influenced their design to focus on achieving resource inter-operability rather than maximising performance. The BDWorld architecture has an abstraction layer (the BDWorld-Grid Interface (BGI)), which provides a syntactically uniform interface to all BDWorld resources — both databases and analytic tools — including an invocation mechanism. Resources are wrapped to conform to this interface; wrapped resources are then able to interact with various Grid or Grid-like implementations via an adaptor specific to the Grid infrastructure currently in use. Other BDW components are designed to use the same mechanism. In particular, the user interface to BDW is provided through the Triana [15] workflow management system, which has been extended to act as a BGI client.

But it is still necessary to be able to discover and use these resources: knowledge is needed about resource types, operations supported, data types, etc., and about conversions between data types, etc. To this end, BDWorld is building a metadata repository connected to an ontology in order to manage resource heterogeneity. This is designed to support semantic equivalence testing when locating and, in particular, integrating datasets from autonomous data providers which, for example, may employ non-standard species names to index their data or may use an unusual data representation.

To access and harvest data from the remote data resources, BDWorld resource wrappers must publish metadata on their capabilities and implement the BGI. This includes implementing a uniform method call, *invokeOperation(resourceHandle, operationHandle, XMLDataCollectionString)*, that allows all resources (be they data or analytic tools) to be invoked by the BGI in a uniform manner. An implication of this uniform resource invocation mechanism is that BGI data calls are not expressed in terms of standard SQL queries. Another feature is that data passing to and from the resources is communicated over the BGI as an XML document or a simple string. This generalisation permits the transmission of either the data or, if the volume is large, the handle for the data.

There are two main ways in which we could introduce OGSA-DAI into BDWorld. One possibility is to augment the BGI to make it possible for queries to be included in workflows and to be sent directly to OGSA-DAI enabled databases. Distributed query processing facilities could be developed to the point

where they could assist in planning the execution and distribution of data-orientated parts of a workflow. (For the current status of OGSA-DQP see Section 4.) But this would be a very major revision to the BDW protocols, and does not take account of the fact that many of the resources of interest are simply not exposed as databases. The other option is to provide facilities within individual wrappers that benefit from OGSA-DAI.

We have adopted the latter approach for our initial exemplar. Figure 1 shows the basic design of this case study. It maintains existing BGI design features such as the invocation method, accessing web databases via resource wrappers. This design is based on a virtual Grid Data Service (GDS) that is not mapped to a particular data resource. The BGI can invoke the OGSA-DAI client with a resourceHandle for any of the wrappers that the GDS represents (Step 1). The OGSA-DAI client then creates the GDS and composes the activity request to perform the operation (Step 2). The *BDWQueryActivity* calls the appropriate wrapper to search the target database (Steps 3, 4) which returns the download URL for the result file (Step 5). The data is retrieved by the *deliverFromURL* activity (Step 6) and input into the *XSLTransform* activity (Step 7) which also takes a XSL format file as input. On completion, the output is returned to the OGSA-DAI client (Step 8). This scenario shows how OGSA-DAI could be modified to access web databases but the data integration is still handled by the BDWorld system. This basic design may be adequate for small volumes of data. For large volumes of data, we could write the output to a cache and return a handle for the data to the BGI. Nevertheless, this is still not exploiting OGSA-DAI's capability for data integration.

Figure 2 illustrates how we could take advantage of OGSA-DAI's flexible activity framework and add a new activity to the basic design to support data integration. This scenario requires the BGI to pass a list of resourceHandles to the OGSA-DAI client and to provide a target location for the delivery (Step 9). (The BDWorld Workflow may need to be amended to achieve this.) We have added a custom *mergeOutput* activity (Step 8) to integrate data returned by the separate web databases. To maximum bandwidth, we could deliver the consolidated dataset via high performance gridFTP to the Workflow unit or to a downstream operation.

Our scenario represents a simple distributed union. We are confident that OGSA-DAI will be able to support this usage. However, it remains to be seen whether this unorthodox application of OGSA-DAI to screen-scrape web data is scalable or will be any more efficient than the current BDWorld methods of wrapping resources. This scenario will be tested in the next phase of the project over the next few months.

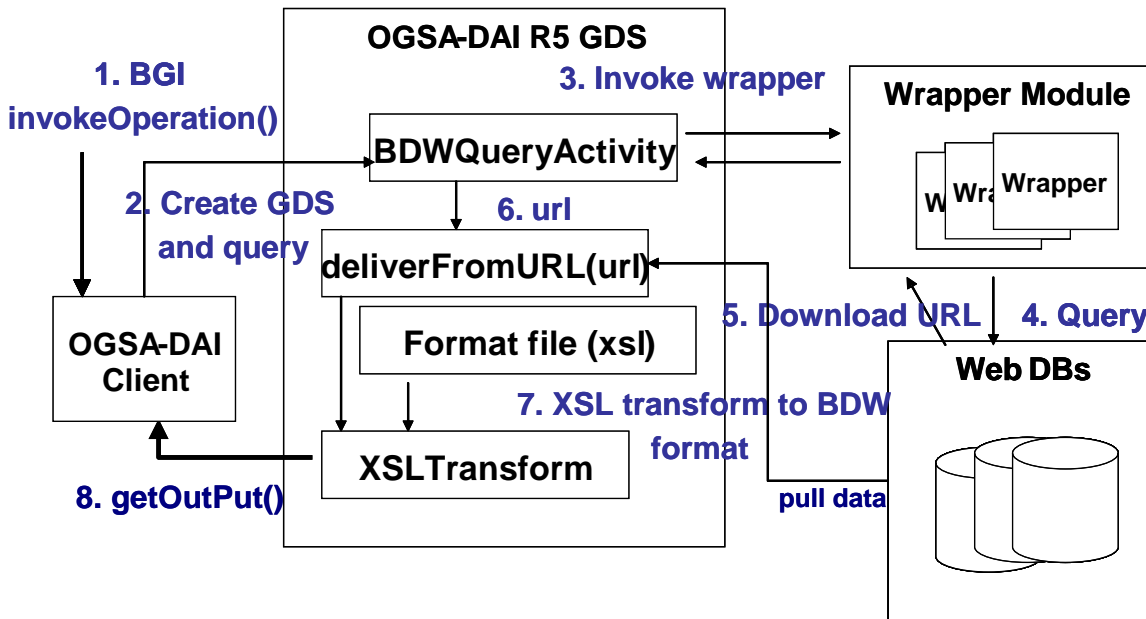


Figure 1. Basic Exemplar

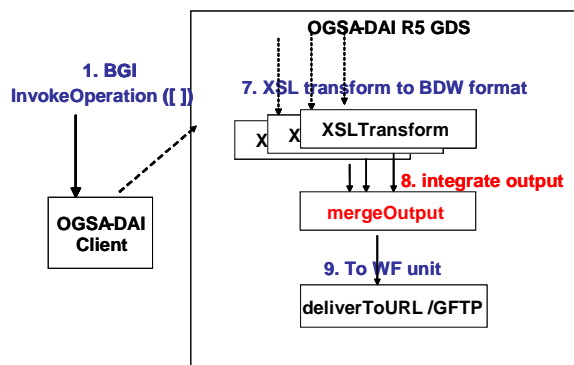


Figure 2. Data Integration

#### 4 Other OGSA-DAI Data Integration Tools

In this section, we consider three other GRID data access and integration solutions and discuss why these have not been used in the BDWorld exemplar. These systems are OGSA-WebDB [16], OGSA-DQP [9] and the Grid Data Mediation Service that has been developed by the GridMiner project team [17]. We have chosen to limit our discussion to solutions based on OGSA-DAI in line with the BioDA project scope (see [10]).

##### 4.1 OGSA-WebDB

OGSA-WebDB provides a GRID Data Service interface based on OGSA-DAI R5 to integrate existing web databases into the GRID environment. The system comprises a proxy database, mediator and database wrappers [16]. It allows users to perform distributed SQL queries on multiple web databases through a two phase processing scheme in which the

same SQL query is performed twice: once by the mediator and once by OGSA-DAI.

The mediator first converts the SQL query into site-specific Boolean search conditions for each target web database and invokes the site-specific wrappers to perform the search. Once all the results are loaded into the individual proxy tables representing each web database used in the query, OGSA-DAI performs the same SQL query on these proxy tables to integrate the data. The mapping between internal and external representation of the data is defined during the proxy configuration process.

Despite its relevance, we have chosen to develop our own OGSA-DAI solution rather than to make use of OGSA-WebDB in BDWorld. The decision is partly influenced by the ease with which OGSA-DAI may be customised: it provides multiple extension points for this purpose. OGSA-WebDB (version 1.0) is currently only available as a binary distribution. No API or Javadoc is included in the distribution and user interaction is restricted to a GUI client. Moreover, OGSA-WebDB poses each query in SQL which has to be run twice during an operation. The present version of the BDWorld BGI invokes the database wrappers using a universal invocation mechanism which passes the query parameters directly into the Java wrappers that it has already implemented (see [2, 4]). OGSA-WebDB also requires the wrappers to output an XML document that conforms to its own custom XML schema.

If the BDWorld wrappers are to be re-deployed in OGSA-WebDB, we would need to re-factor the BDWorld wrapper architecture and set the wrappers to extend OGSA-WebDB rather than the BDWorld-defined AbstractWrapper as well as to amend the output formatter. In view of these constraints, we concluded that the efforts involved in integrating

OGSA-WebDB in its current form outweigh its potential benefits to BDWorld.

## 4.2 OGSA-DQP

Another example of a GRID data integration tool is OGSA-DQP, a component of the myGrid Project [18] and an open-source product. It provides a service-based Distributed Query Processor based on OGSA and OGSA-DAI standards. OGSA-DQP can potentially offer declarative support for service orchestration and can be seen as a mediator component over OGSA-DAI R4 wrappers. It supports OQL (Object Query Language) queries over OGSA-DAI GDSs and other services available on the GRID. This design allows data analysis to be combined with data access and integration operations in one query statement. Its main focus, though, is efficient query execution rather than reconciling data source heterogeneity, which is a main issue with BDWorld data resources (see [2, 4]). OGSA-DQP currently only supports equi-join queries on standard GDSs interacting with RDBMS. Some of its operations use the native functionalities of a GDS to fetch database schemas and scan relational tables (Mukherjee, *pers. comm.*) This restriction makes it unsuitable for BDWorld as most of its data comes from autonomous web databases.

## 4.3 Grid Data Mediation Service

The Grid Data Mediation Service (GDMS) prototype provides a more general GRID data integration solution. Unlike OGSA-DQP, its main focus is on reconciling data resource heterogeneities. It supports inner-joins and unions. It can be used with relational, XML and CSV flat file data sources but not web databases. GDMS has a mediator-wrapper architecture which permits the federation of multiple data sources under a global schema and exposes them as a single virtual OGSA-DAI R3 GDS [19]. The relationship between the global schema and the participating data sources is defined in a statically configured mapping schema. Clients query the GDS in a standard manner using SQL expressed in terms of the global schema. The GDMS uses the mapping schema to reformulate and partition the query and then execute the partitions against the data sources via custom query wrappers.

The GDMS framework could potentially be extended to incorporate web databases, as demonstrated by tools such as XMF [20]. As in the case of OGSA-WebDB, GDMS is an attractive solution that could possibly be modified to provide a dedicated data management component for BDWorld if the BGI is extended to support SQL-style distributed queries.

## 5 Conclusion

We have briefly summarised the common data access and integration requirements raised by the

bioinformatics community and presented one of the proposed OGSA-DAI exemplars for aggregating thematic data for BDWorld applications. We have also touched on how BDWorld is designed to use metadata and ontological knowledge to ensure semantic integrity when integrating data from different providers. We have highlighted other key data integration solutions based on OGSA-DAI and commented on their suitability for BDWorld. We have noted a range of features which, if incorporated into OGSA-DAI, would be beneficial to the bioinformatics projects surveyed. But we have also observed that BDWorld has restricted access to some of the databases it uses and so OGSA-DAI is not applicable in a straightforward manner, although our exemplar illustrates that there are still scenarios in which it may be usefully employed.

In the next stage of our study, we will develop further exemplars to measure OGSA-DAI's performance against the requirements highlighted by users from the bioinformatics community.

## References

1. Atkinson, M. (2004) Data Access and Integration. *Ercim News* 59, p 34-80.
2. Jones, A.C., White, R.J., Gray, W.A., Bisby, F.A., Caithness, N., Pittas, N., Xu, X., Sutton, S., Fiddian, N.J., Culham, A., Scoble, M., Williams, P., Bromley, O., Brewer, P., Yesson, C. and Bhagwat, S. (2005) Building a Biodiversity GRID. In Konagaya, A. and Satou, K., eds.: *Grid Computing in Life Science, (LNCS 3370)*, Springer-Verlag, p. 140-151. (Biodiversity World: <http://www.bdworld.org/>)
3. White, R., Bisby, F., Caithness, N., Sutton, T., Brewer, P., Williams, P., Culham, A., Scoble, M., Jones, A., Gray, W., Fiddian, N., Pittas, N., Xu, X., Bromley, O., Valdez, P. (2003) The Biodiversity World environment as an extensible virtual laboratory for analysing biodiversity patterns. In Cox, S., eds.: *Proc. UK e-Science All Hands Meeting*, Nottingham, UK, EPSRC, p. 341-344.
4. Jones, A.C., White, R.J., Pittas, N., Gray, W.A., Sutton, T., Xu, X., Bromley, O., Caithness, N., Bisby, F.A., Fiddian, N.J., Scoble, M., Culham, A. and P. Williams (2003) Biodiversity World: An Architecture for an Extensible Virtual Laboratory for Analysing Biodiversity Patterns. In Cox, S., eds.: *Proc. UK e-Science All Hands Meeting*, Nottingham, UK, EPSRC, p. 759-765.
5. Jones, A.C., Xu, X., Pittas, N., Gray, W.A., Fiddian, N.J., White, R.J., Robinson, J.S., Bisby, F.A., and Brandt, S.M. (2000) SPICE: a Flexible Architecture for Integrating Autonomous Databases to Comprise a Distributed Catalogue of Life. In *Proc. 11th International Conference on Database and Expert Systems Applications (LNCS 1873)*, Springer-Verlag, p. 981-992. (SPICE: <http://www.systematics.reading.ac.uk/spice/>)

6. Species 2000. (<http://www.sp2000.org>)
7. OGSA-DAI. (<http://www.ogsadai.org.uk/>)
8. Globus Toolkit. (<http://www-unix.globus.org/toolkit/>)
9. OGSA-DQP. (<http://www.ogsadai.org.uk/docs/dqp/>)
10. BioDA.  
(<http://isegserv.itd.rl.ac.uk/BioDA/pages/default.htm>)
11. Bioinformatics and OGSA-DAI (BioDA) Interim Report.  
([http://isegserv.itd.rl.ac.uk/BioDA/documents/BioDA\\_InterimRep5.pdf](http://isegserv.itd.rl.ac.uk/BioDA/documents/BioDA_InterimRep5.pdf))
12. Watson, P. (2003) Databases and the GRID. In *GRID Computing: Making the Global Infrastructure a Reality*, Wiley, p. 363-384.
13. WS-Resource Framework.  
(<http://www.globus.org/wsrf/>)
14. Web Services Interoperability.  
(<http://www.ws-i.org/>)
15. Triana. (<http://www.trianacode.org/>)
16. Kojima, I. and S.M. Pahlevi (2004) Design and Implementation of OGSA-WebDB – a service based system for making existing web databases grid-ready. The GGF10 Workshop, Berlin, Germany.  
(OGSA-WebDB:  
<http://www.gtrc.aist.go.jp/dbgrid/ogsa-webdb/>)
17. GridMiner. (<http://www.gridminer.org/>)
18. myGrid. (<http://www.mygrid.org.uk/>)
19. Wöhler, A. and P. Brezany (2004) *Mediators in the Architecture of Grid Information Systems*. Institute for Software Science, University of Vienna  
(<http://www.gridminer.org/publications/gridminer2004-01.pdf>)
20. Lee, Kangchan, Min, Jaehong, Park, Kishik and Kyuchul Lee (2002) A Design and Implementation of XML-based Mediation Framework (XMF) for Integration of Internet Information Resources. In: *Proceedings of the 35<sup>th</sup> Annual Hawaii International Conference on System Sciences*, vol. 07, no. 7, p. 202-210.

## Acknowledgements

The BioDA project is funded by the Biotechnology and Biological Sciences Research Council. The authors wish to thank all the participants who took part in the BioDA Workshop and those who participated in our requirements survey. We are also grateful to BDWorld, DAIT, OGSA-DQP and GridMiner colleagues for their helpful advice and help with designing the OGSA-DAI exemplars.