

Bridges: Security Focused Integration of Distributed Biomedical Data

**Dr Richard Sinnott, Prof David Gilbert, Dr David Berry,
Dr Ela Hunt, Prof Malcolm Atkinson
National e-Science Centre
ros@dcs.gla.ac.uk**

The BRIDGES project will incrementally develop and explore database integration over six geographically distributed research sites with the framework of a Wellcome Trust biomedical research project (the Cardiovascular Functional Genomics project) to provide a sophisticated infrastructure for bioinformaticians. One of the key issues to be investigated in Bridges is data security. Different classes of data can be defined: public data sources; data sources for usage of project members only, and private data sources, e.g. patient records. The project will seamlessly handle the federation of these databases, incorporating features to transparently address security concerns. The project will provide valuable insights into the application of OGSA-DAI and the IBM DiscoveryLink technologies for this purpose, and propose and implement needed extensions/wrappers. This paper highlights the issues that we expect to address in the project and initial ideas we have to overcome them.

1. Introduction

The Wellcome Trust has funded a large (£4.34M) collaborative project (Cardiovascular Functional Genomics - 'CFG' [1]) over 5 years and involving 5 UK and 1 Dutch site to investigate hypertension. This disease affects 25% of adults in westernised societies and is the major cause of cardiovascular morbidity and mortality. Hypertension is caused by a combination of genetic and environmental factors. The CFG project is pursuing a translational strategy, which combines studies on rodent models of disease with parallel studies of patients and of large family and population DNA collections. As such, the project exemplifies large-scale computational problems of modern biology with requirements to combine information about three species, human, mouse and rat.

Typical activities that the CFG scientists will perform include: large-scale sequence comparisons, integration of sequence analysis with other data (phenotyping and genotyping results, genetic and radiation hybrid maps, micro-array gene expression profiling and protein function prediction), generation of cross-species maps of genes and markers and statistical analysis of large data sets needed in the context of gene discovery. Currently many of these activities are performed in a time consuming and largely non-automated manner often requiring navigation to many different data resources and following multiple links to related information.

The 23 month BRIDGES project [2] has recently been funded by the DTI to directly address the needs of the CFG scientists and provide a thorough investigation of relevant technologies for this purpose. Specifically, BRIDGES will investigate the application of OGSA_DAI [3] and IBM's DiscoveryLink product [4] to deal with federation of distributed biomedical data. In addition security is extremely important for the scientists. The scientific data itself may have different characteristics:

- **Public data:** data from public sources, such as SwissProt and EMBL, copies of which may be held locally for performance reasons or shared throughout the consortium.
- **Processed public data:** public data that has additional annotation or indexing to support the analyses needed by CFG. These must be held within the consortium, but one copy can serve the entire consortium.

- Sensitive data: the data about individuals in the cohorts of patients and the data derived from animal experiments. These require careful enforcement of privacy and may be restricted to one site, or even part of a site.
- Special experimental data: this may fall into a particular category, e.g. micro-array data, which has special arrangements for its storage and access already agreed.
- Personal research data: data specific to a researcher, as a result of experiments or analyses that that researcher is performing. This is not shared even among the local team. It may later become team research data.
- Team research data: data that is shared by the team members at a site or within a group at a site. It may later become consortium research data, e.g. when the researchers are confident of its value or have written about its creation and implications.
- Consortium research data: data produced by one site or a combination of sites that is now available for the whole consortium.
- Personalisation data: metadata collected and used by the bioinformatics tools pertinent to individual users. This data is normally only needed to support the specific user to which it pertains, but it may need to move sites when bioinformaticians visit sites or work together.

The distribution of CFG partners and the data security needs are depicted in Figure 1.

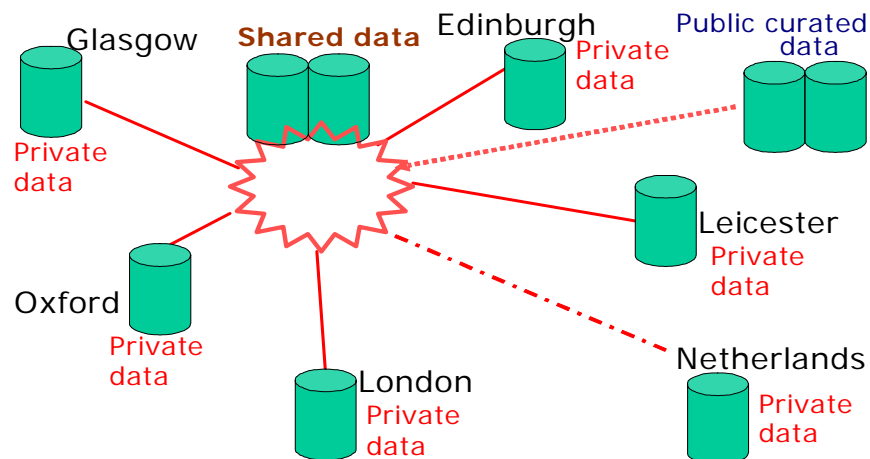


Figure 1: Data Distribution and Security of CFG Partners

2. Data Access and Integration

Data access and integration within Bridges will be investigated through results from the OGSA_DAI and the recently funded follow up project, Data Access and Integration 2 (DAIT), and IBM DiscoveryLink. OGSA_DAI/DAIT is a collaborative programme of work involving the Universities of Edinburgh, Manchester and Newcastle, the National e-Science Centre, with industrial participation by IBM and Oracle. Its principal objective is to produce open source database access and integration middleware which meets the needs of the UK e-Science community for developing Grid and Grid related applications. Its scope includes the definition and development of generic Grid data services providing access to and integration of data held in relational database management systems, as well as semi-structured data held in XML repositories. OGSA_DAI is one of the key driving forces behind the Grid Data Access and Integration Services (DAIS) standardisation activities at GGF.

IBM's DiscoveryLink product – which will soon be renamed as IBM Information Integrator – has been developed to meet the challenge of integrating and analyzing large quantities of diverse scientific data from a variety of life sciences domains. IBM DiscoveryLink offers single-query access to existing databases, applications and search engines. The DiscoveryLink solution includes the combined resources of DiscoveryLink middle ware and IBM Life Sciences services. Using this software, IBM Life Sciences services can create new components that allow specialized databases—for proteomics, genomics, combinatorial chemistry, or high-throughput screening—to be accessed and integrated quickly and easily. This is depicted in Figure 2.

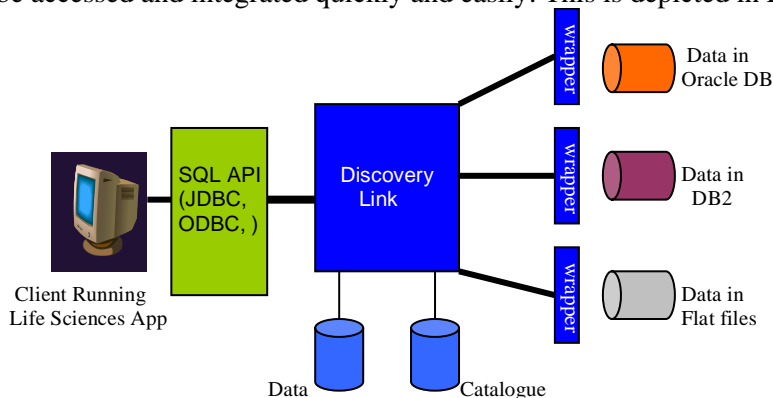


Figure 2: IBM DiscoveryLink Data Access and Integration

At the far right of Figure 2 are the data sources. DiscoveryLink looks to these sources like an application - they are not changed or modified in any way. DiscoveryLink talks to the sources using wrappers, which use the data source's own client-server mechanism to interact with the sources in their native dialect. DiscoveryLink has a local catalogue in which it stores information (metadata) about the data accessible (both local data, if any, and data at the backend data sources). Applications of DiscoveryLink manipulate data using any supported SQL API, for example, ODBC or JDBC are supported, as well as embedded SQL. Thus a DiscoveryLink application looks like any normal database application.

The focus of OGSA_DAI and DiscoveryLink is thus upon access and integration of data and not specifically upon security concerns. Security in the context of the Grid is an area that is currently receiving much attention since it is a crucial factor in the wider uptake of the Grid.

3. Security

In order to control access to resources, authentication and authorization are needed [5]. Authentication is the establishment and propagation of a user's identity in a given system. Authentication in the UK e-Science community has been based upon ITU-T X.509 digital certificates [6] using Public Key Infrastructures (PKIs) [7,8,9]. Authorisation is concerned with controlling access to services based upon specific policies. Authorisation typically requires tools to specify and manage policies; mechanisms to distribute and obtain policies; services that use policies to make an access decision; and mechanisms that request and enforce access decisions. Security and especially authorisation are currently the focus of much work within the Grid community [10]. Examples of works currently investigating authorisation in the context of the Grid include Community Authorisation Service [11], Akenti [12], VOMS [13] and VOM [14].

In addition to these, the PERMIS project [15,16] built and validated the world's first X.509 attribute certificate based authorisation infrastructure. The PERMIS team are working closely with the Globus team to design a standard Security Assertion Markup Language (SAML) [17] interface to any authorisation infrastructure. This will allow Grid applications to plug and play

any authorisation infrastructure. The Bridges project has agreed to work with the PERMIS team and provide a rigorous investigation of security authorisation in a Grid context.

In addition to authentication and authorisation security aspects, a key requirement of the CFG scientists is related to privacy. Privacy relates to the use of data, in the context of consent established by the data owner. There is little prior art in privacy grid science, although there is useful UK background in privacy including hospital systems [18]. Web based standards such as P3P [19] may contribute to only a small fraction of the necessary security mechanisms.

4. Other Information

At the time of writing the BRIDGES project has yet to formally begin. We have recently completed our recruitment activities and it is expected that the full team will be on board and the project shall fully commence on 1st October.

5. References

- [1] Cardiovascular Functional Genomics project, <http://www.brc.dcs.gla.ac.uk/projects/cfg/>
- [2] BioMedical Research Informatics Delivered by Grid Enabled Services (BRIDGES), www.brc.dcs.gla.ac.uk/projects/bridges
- [3] Open Grid Service Architecture – Data Access and Integration project (OGSA-DAI), www.ogsadai.org.uk
- [4] IBM DiscoveryLink, <http://www3.ibm.com/solutions/lifesciences/solutions/discoverylink.html>
- [5] E-Science Security Roadmap: Technical Recommendations v0.5, UK e-Science Security Task Form, draft executive summary v0.51
- [6] ITU-T Rec. X.509 (2000) | ISO/IEC 9594-8 The Directory: Authentication Framework
- [7] C Adams and S Lloyd (1999), Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations, Macmillan Technical Publishing.
- [8] Adams, C., Lloyd, S. (1999). “Understanding Public-Key Infrastructure: Concepts, Standards, and Deployment Considerations”, Macmillan Technical Publishing, 1999
- [9] Austin, T. “PKI, A Wiley Tech Brief”, John Wiley and Son, ISBN: 0-471-35380-9, 2000
- [10] Grid Security, <https://forge.gridforum.org/projects/sec>
- [11] L Pearlman, et al., A Community Authorisation Service for Group Collaboration, in Proceedings of the IEEE 3rd International Workshop on Policies for Distributed Systems and Networks. 2002.
- [12] M Thompson, et al., Certificate-Based Access Control for Widely Distributed Resources, in Proc 8th Usenix Security Symposium. 1999: Washington, D.C.
- [13] VOMS Architecture, European Datagrid Authorization Working group, 5 September 2002.
- [14] Steven Newhouse, Virtual Organisation Management, The London E-Science centre, <http://www.lesc.ic.ac.uk/projects/oscar-g.html>
- [15] D. Chadwick and A. Otenko. The PERMIS X.509 role based privilege management infrastructure, in Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies, Monterey, California, USA. 2002.
- [16] Privilege and Role Management Infrastructure Standards Validation project www.permis.org
- [17] P Hallem-Baker and E Maler, Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML), OASIS, SAML 1.0 Specification. 31 May 2002. <http://www.oasis-open.org/committees/security/#documents>
- [18] I. Denley and S.W. Smith, Privacy in clinical information systems in secondary care. British Medical Journal, 1999. 318: p. 1328-1331.
- [19] Platform for Privacy Preferences (P3P) Project, W3C, <http://www.w3.org/P3P/>