FirstDIG: Data Investigations using OGSA-DAI

P. J. Graham[1], **T. M. Sloan**[1], A. C. Carter[1], I. Gregory[2]

[1]EPCC, The University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh, EH9 3JZ, UK

[2]First South Yorkshire, Midland Road, Rotherham, S61 1TF, UK

Abstract

The FirstDIG (First Data Investigation on the Grid) project deployed OGSA-DAI (Open Grid Services Architecture: Data Access and Integration) within the First South Yorkshire bus operational environment. A generic Grid Data Service Browser has been built and used to interrogate and combine data from OGSA-DAI enabled data sources to answer business questions posed by First South Yorkshire. This paper describes the project and its achievements.

1. Introduction

First plc [1] are the UK's largest public transport operator. Within their UK bus operations they have a huge range of data sources - vehicle mileage, fuel consumption, maintenance records, revenue, reliability, etc. Data are collected via various mechanisms, for example, manually at a depot or on the buses via ticket machines, and are typically stored in disparate databases. There are many reasons for this heterogeneity of data storage - company acquisitions, incremental system construction, geography and differing operating systems are just a few. This diversity is not an issue for day-to-day bus operations but does introduce challenges for data analysis, particularly when it involves data from more than one source.

For example within First South Yorkshire, a bus operating division of First plc, data is spread across diverse computer systems with no common interface and so no common reporting process. Statistics can be produced manually but this is so labour-intensive that it is rarely performed. The development of a common interface to rectify this situation is expensive and may involve significant disruption to existing operations.

This challenge of analysing data from diverse sources is NOT unique to First plc and the bus industry. It occurs in many companies in many different business sectors.

1.1 OGSA-DAI

OGSA-DAI (Open Grid Services Architecture: Data Access and Integration) is grid middleware software designed to assist with the access and integration of data from separate sources via Grid Services [2]. As such it therefore offers First plc a possible means to meet this data analysis challenge.

1.2 The FirstDIG project

The First Data Investigation on the Grid or FirstDIG [3] [4] project was funded through the UK e-Science Grid Core Programme[5] to investigate the use of OGSA-DAI within First South Yorkshire's bus operational environment. The project was a collaboration between First plc and the UK National e-Science Centre [6] as represented by EPCC [7]. The project had two aims. The first was to demonstrate the deployment of OGSA-DAI services in a commercial environment, and learn from this process. The second was to answer specific business questions posed by First South Yorkshire through a short data mining analysis using the OGSA-DAI service enabled data sources. In addressing these aims the project tried to determine if OGSA-DAI is appropriate for analysis of diverse data sources, is straightforward to deploy and is easy to use. The project started in May 2003 and finished in January 2004. This paper describes the project and some of its achievements.

2. Grid Data Services

OGSA-DAI Release 3.1 has been successfully deployed at First South Yorkshire and grid data services created for the bus mileage and customer contacts systems. Answering business questions posed by First South Yorkshire requires the extraction and combination of data from both these systems. The customer contact data are held within Microsoft Access databases, whilst the mileage data are held within dBASE IV databases. Using OGSA-DAI these databases can be represented as Grid

Services, which are then accessible from other machines in a secure and seamless manner.

During the project lifetime, OGSA-DAI Release 3.1 only supported access to a select number of database management systems (DBMS). Unfortunately those did not include Microsoft Access or dBASE IV. Fortunately, OGSA-DAI could in principle connect to any database that was accessible via a Java Database Connectivity (JDBC) driver. Microsoft Access and dBASE IV are accessible via Open Database Connectivity (ODBC) drivers. These drivers comply with standard ways of connecting to disparate DBMS. Further there are methods for JDBC drivers to communicate with ODBC drivers, known as JDBC-ODBC bridges. Using the appropriate bridges the FirstDIG project was able to connect OGSA-DAI to the First databases.

Further grid data services based on SQL Server databases and comma separated values file have also been investigated.

3. OGSA-DAI Release 3.1 Limitations

The deployment of OGSA-DAI and the construction of grid data services by the project were instrumental in revealing a number of issues with Release 3.1 of OGSA-DAI. Some of these are outlined below. These are relevant to any grid data access and integration project as they reflect issues when handling real data in an operational environment.

3.1 Large Results

OGSA-DAI could not cope with queries that resulted in large numbers of records being returned. This has since been addressed in OGSA-DAI V4.0.

3.2 Certain characters cause problems

XML does not allow certain character codes (eg &, >) in its text. This is a limitation of XML. These are typically rarely used characters, however some have been found in fields in the Mileage database. Although these were erroneous, their presence caused any queries on that field to fail. The solution was to remove the characters by hand. This is not ideal, and this issue has since been fixed in OGSA-DAI Release 4.0.

3.3 Bit data type

This data type is used to represent Yes/No fields in Microsoft Access databases. This has since been addressed in OGSA-DAI V4.0.

3.4 CSV data sources

Comma separated value (CSV) data sources are common but current CSV JDBC-ODBC drivers do not have sufficient functionality to support the querying necessary to answer First's business questions. This is not an OGSA-DAI issue per se.

3.5 Date Formats

Queries on Date fields from Access and dBASE IV databases returned a DBMS internal identifier for the date rather than the date itself. The OGSA-DAI team are investigating how to transform the identifier into a human readable date format in the result set.

3.6 Usability

Within the resulting OGSA-DAI set up, initial sessions were executed using the command line client tool that came with the OGSA-DAI distribution. There was also a graphical client available with the distribution, but both these tools were meant as demonstrators. Using these tools meant the creation and utilisation of several XML files. These are not particularly straightforward to use and can be confusing to those unfamiliar with them.

3.7 Data Integration

OGSA-DAI allows access to heterogeneous data sources in a uniform manner but it does not provide support for integrating data from such sources. A user has to perform this integration manually. For example, to answer one of the business questions posed by First South Yorkshire a table from the customer contacts system needs to be joined with a table from the bus mileage system. To perform this join, OGSA-DAI was used to retrieve the tables from the distributed databases into a local database management system where the tables were joined. OGSA-DQP (Distrbuted Query Processor) [8] extends OGSA-DAI to provide this type of functionality. However this functionality is only available on Linux and not on the platforms available at First South Yorkshire at the time of the project. Moreover OGSA-DQP uses OQL as the query language rather than the similar but more common SQL.

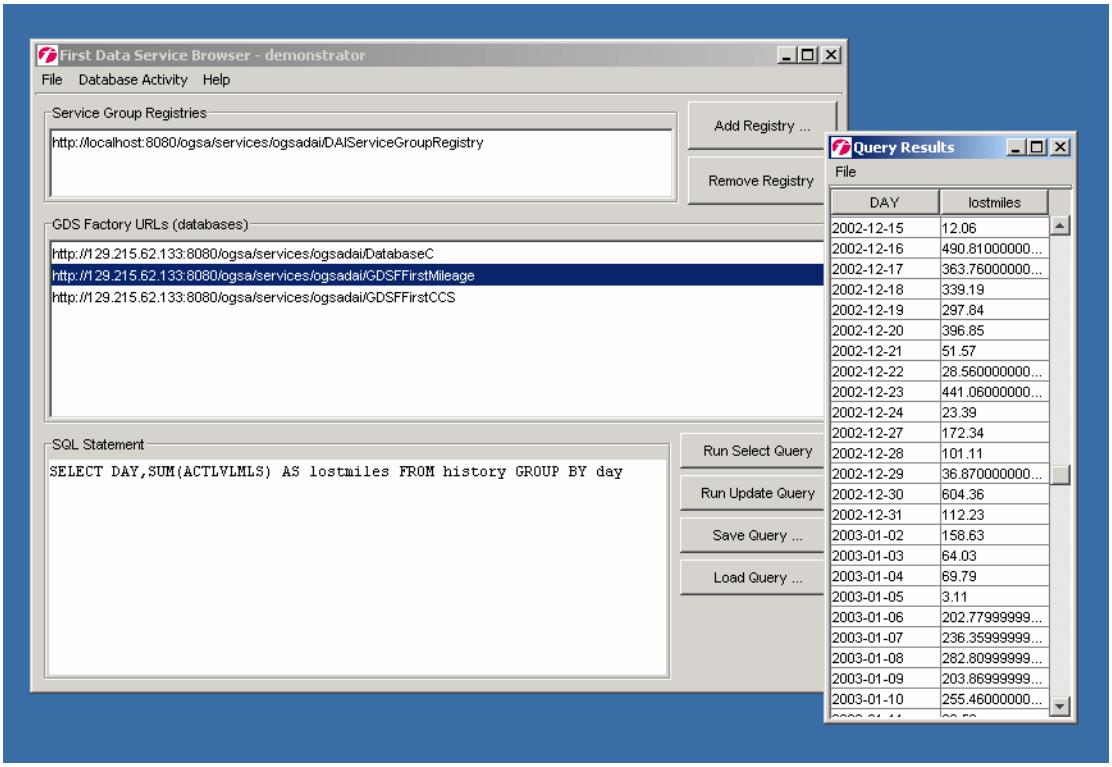

Figure 1: The Grid Data Service Browser

4. Grid Data Service Browser

To aid ease of use of OGSA-DAI and address the usability issues raised in section 3.6, the FirstDIG project built a generic Grid Data Service Browser [9] and deployed this at First South Yorkshire. This browser provides a user with a simple GUI (see Figure 1) thus removing the need to use complicated XML to interact with grid data services. Via the browser, a user can interact with the data services and hence the underlying databases through SQL queries and updates. Further, to help address the data integration issues raised in section 3.7, this browser provides a simple interface, the Join dialog, to enable SQL joins to be performed across any OGSA-DAI grid data service enabled databases. This browser has since been included in Release 4.0 of OGSA-DAI and used and extended in the INWA project [10].

4.1 System Outline

The browser is written in Java and interacts with OGSA-DAI sources via the OGSA-DAI Client Toolkit [11]. The OGSA-DAI Client Toolkit is a Java API that provides the basic building blocks for OGSA-DAI client development. It is intended to minimise the specialist knowledge required to interact with OGSA-DAI services and to shield a developer from future changes to the internals of OGSA-DAI.

The target platform for the browser was initially a Microsoft Windows 2000 PC however it has also been deployed successfully under Microsoft Windows XP. The browser is written in Java and so should work on any Java-enabled platform.

The design of the browser is based on the Model View Controller (MVC) design pattern. This pattern works in the following manner.

The Model is the core of the application and maintains the state and data that the application represents, for example the database details. When significant changes occur in the model it updates all of its Views.
The Controller is the user interface presented to the user to manipulate the application. When the user interacts with the controller this usually instigates some activity within the model.
The View is the user interface that displays information about the model to the user. Any object that needs information about the model needs to be a registered view

The main advantage of using the pattern is improved modularity (the model does not need to know about any particular view for example,

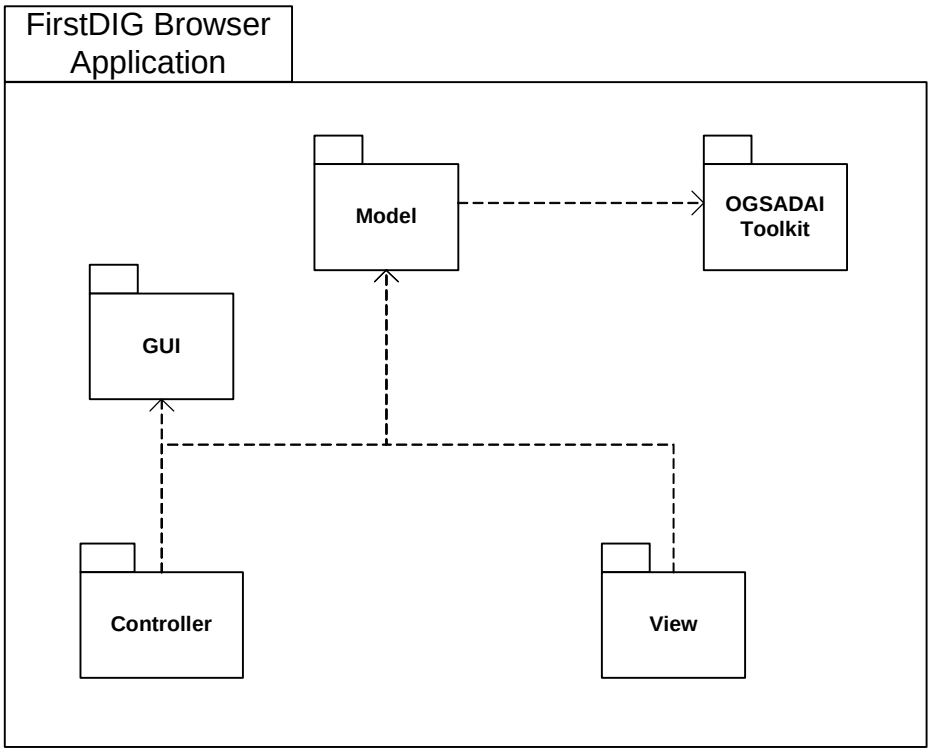

Figure 2: FirstDIG Browser Package Overview

and similarly the view does not need to know about the controller). This improves clarity and extensibility, and has made the software easier to implement and to maintain. Also it protects the View and Controller classes from any changes in the OGSA-DAI toolkit, as it is the model alone that interacts with this.

Figure 2 shows the basic package overview of the browser software. The software comprises five packages described below.

OGSA-DAI Client Toolkit: As previously mentioned, this is an API for developers wishing to write OGSA-DAI client applications.
Model: This package contains the methods for interacting with the OGSA Client toolkit and hence the underlying databases.
GUI: This represents the GUI container for the View and Controller packages and enables access to different views and controls.
Controller: This package contains the methods for allowing a user to interact with the OGSA-DAI service-enabled data sources.
View: This package contains the methods for displaying the results of interactions with the OGSA-DAI service-enabled data sources.

4.2 The Join Dialog

The browser includes a join dialog to help a user perform a join across two different databases (see Figure 3). Unfortunately this involves several steps to perform what would be a single line SQL statement if the tables were in the same database, but the join dialog assists the user in this.

Figure 4 illustrates the steps involved in a join using this dialog. Firstly the temporary tables to hold the data to be joined are created in database C. Then database A is queried, and the results sent to one of the temporary tables in database C. The same process occurs for database B. Then the join query is performed on the data at C, the results are displayed and the temporary tables destroyed. The dialog has text fields or drop-down selections for each of these stages. The drop-down selections are used to select which databases to use as A, B and C. The editable text boxes allow the user to enter the SQL statements required to create and destroy the temporary tables, get the intermediate results from databases A and B and perform the join.

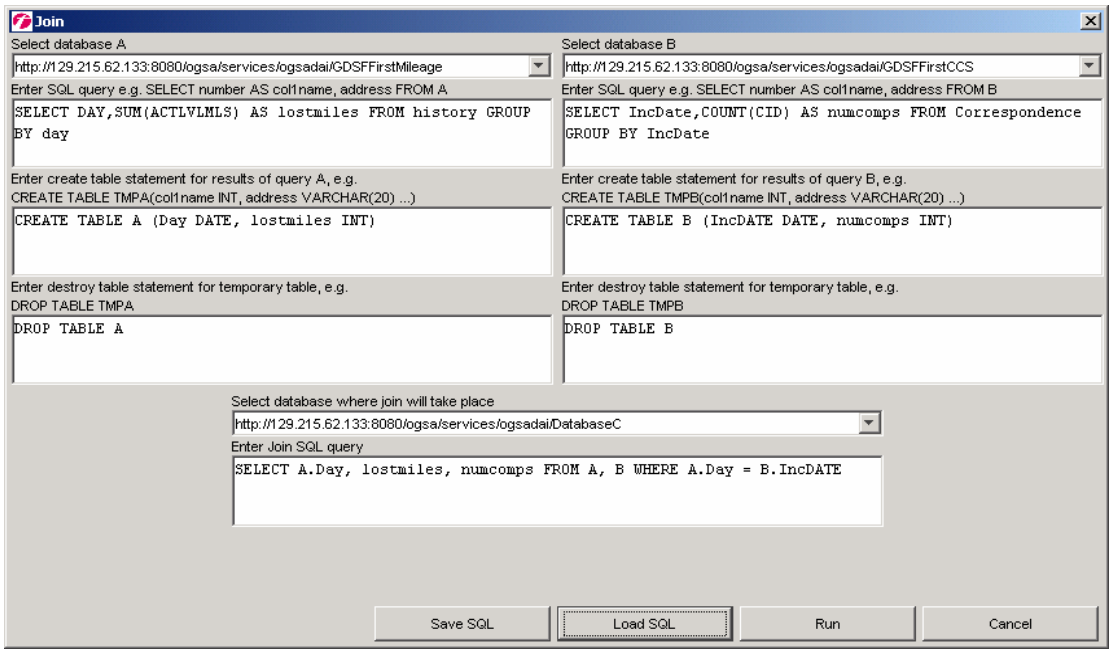

Figure 3: The Join Dialog

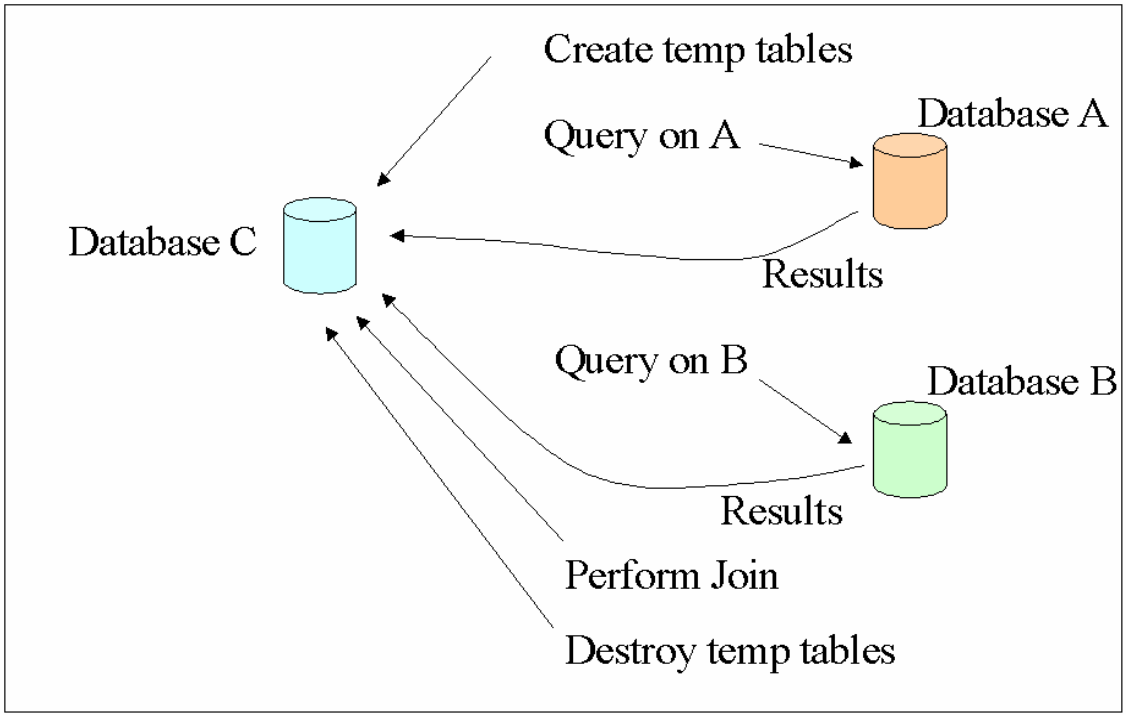

Figure 4: The steps in a join

5. Data Analysis

As reported in [4], the data sources used in the FirstDIG project were from the following systems within First South Yorkshire.

Customer Contact – this records correspondence with customers including commendations and complaints.

Vehicle Mileage – this records the daily vehicle mileage for bus services.

Ticket Revenue – this contains the daily tickets sold and the money taken for the bus services.

Schedule Adherence – a satellite tracking system that records whether a bus is arriving and departing on time from a bus stop.

These systems are located at various company sites, on differing platforms in different

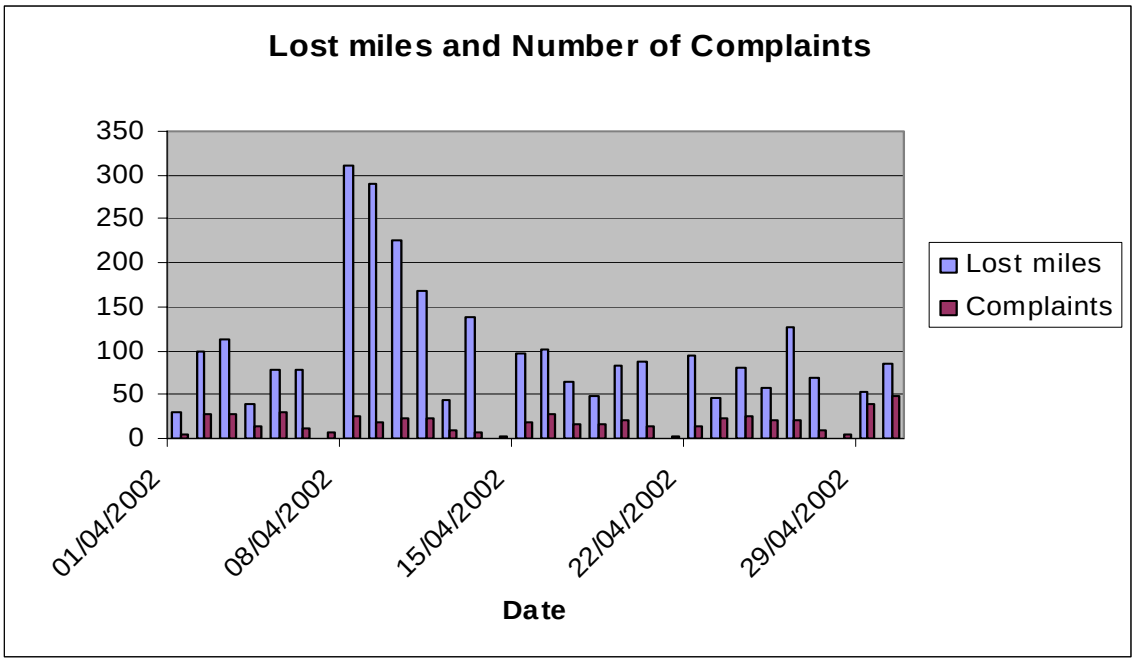

Figure 5: Lost miles and Number of Compliants per day

databases. The databases range from SQL sources to ODBC sources to COBOL files. It is precisely these issues that any technology must address in order to be applicable and useful to business.

First South Yorkshire formulated a set of business questions whose answers required consolidation of data from these various sources. The questions covered topics such as:

- the effect of lost mileage on revenue, where lost mileage is due to activities such as road works and breakdowns;
- the effect of lost mileage on the number of complaints;
- the effect of reliability on revenue;
- the effect of reliability on complaints received.

Using the grid data services and the Grid Data Service Browser, it has been possible to answer quickly and easily those business questions concerning customer contacts and mileage. Previously answering such questions would require much manual manipulation of the data and take upwards of an hour. This does not include the time for locating and requesting the necessary data. This is obviously an additional overhead. Using the grid data services and the FirstDIG browser answers to these questions can be retrieved in seconds direct if necessary from the operational data wherever it resides. Figure 5 shows the results of one particular join query across these grid data services using the FirstDIG browser.

The data in the revenue system is available in CSV files and a grid data service can be generated for a CSV file. Unfortunately, the available JDBC-ODBC drivers do not provide sufficient functionality (see Section 3.4) to answer First South Yorkshire's revenue-related business questions via the resulting grid data service. These questions therefore had to be answered via conventional means, ie. manually importing the data into a local DBMS and manipulating it there.

The Schedule Adherence system holds its data in SQL Server. Using JDBC-ODBC bridges, it was established that a grid data service can be provided for this data. However, due to a lack of project time no such service was deployed at First South Yorkshire.

Due to reasons of commercial confidentiality the answers to the business questions cannot be given here. However it is clear that these answers have provided important insights into bus operations, so much so that senior IT management in First South Yorkshire has stated the results will revolutionise the way they do business.

6. Conclusions

The project has shown that OGSA-DAI can provide a cost-effective solution that First South Yorkshire can utilise. It allows staff to access disparate data sources via a common interface

and straightforwardly analyse the data in a practical time scale. This means that with considerably reduced labour costs, First South Yorkshire can produce more accurate and more comprehensive information for business management.

Acknowledgements

FirstDIG was funded as part the UK e-Science Core Programme [5]. The project partners are the National e-Science Centre [6], represented in this project by EPCC [7], and First plc [1] as represented by First South Yorkshire.

References

[1] First plc, http://www.firstgroup.com/
[2] Open Grid Services Architecture: Data Access and Integration, http://www.ogsadai.org/
[3] First Data Investigation on the Grid, http://www.epcc.ed.ac.uk/firstdig/
[4] T.M.Sloan, A.Carter, P.J.Graham, D.Unwin, I.Gregory, "First Data Investigation on the Grid: FirstDIG", Proceedings of the 2nd UK e-Science All Hands Meeting, 2-4 September, 2003, Nottingham, UK
[5] UK e-Science Core Programe, http://www.escience-grid.org.uk/
[6] UK National e-Science Centre, http://www.nesc.ac.uk/
[7] EPCC, http://www.epcc.ed.ac.uk/
[8] 'OGSA-DQP: A service-based distributed query processor for the Grid', available from http://www.ogsadai.org/docs/OtherDocs/114.pdf
[9] First Data Service Browser User Guide, http://www.epcc.ed.ac.uk/~firstdig/DISSEMINATION/FirstDIGBrowserUserGuide.pdf
[10] INWA, http://www.epcc.ed.ac.uk/projects/inwa
[11] The OGSA-DAI Client Toolkit Tutorial, http://www.ogsadai.org/docs/current/tutorials/ClientToolKit/Client-Toolkit-Tutorial-Index.html