

Experiences with Supporting Mass Storage for Multiple Communities

P Berrisford, **D Corney**, T Folkes, J Jensen, J Mencák, D Ross, B Strong, O Synge
CCLRC Rutherford Appleton Laboratory

All Hands Meeting, Sep 2005

Abstract

Access to reliable high capacity and high performance storage is becoming increasingly important to many scientific projects.

Rutherford Appleton Laboratory (RAL) runs a Petabyte-scale mass storage facility (ADS) which serves a wide range of scientific communities, from within the research council, from other research councils, and other communities. In this paper we describe our experiences with supporting a diverse scientific community, and to which extent our facility can enable data integration between the communities, both with the storage system itself, and with added Grid interfaces and other higher-level Grid tools.

1 Introduction

This paper is a contribution from the CCLRC e-Science Data Management Group, Petabyte Storage Group and the Tier 1 centre for the UK LHC community. These groups are collectively responsible for providing Petabyte scale data archive, storage and data management services to:

- The UK Particle Physics community, GridPP. The GridPP2 project contains UK collaborators in the Large Hadron Collider and in other international High Energy Physics experiments.
- The facilities, departments and projects within CCLRC, including ISIS — the world's brightest pulsed neutron and muon source [6], Diamond Light Source [3] — the largest UK funded scientific facility to be built in the UK for over 30 years, and many other departments and projects.
- A growing number of groups within UK academia and research, including PPARC astronomy community [10], the National Crystallography Service at Southampton University [9], the Integrative Biology Project [5], and BBSRC [1] and its geographically distributed constituent institutes.

The facilities, departments and projects within CCLRC have been using the ADS for preserving their data securely (storage, integrity, migration) for over twenty years. The ADS tape-store has a current capacity of one Petabyte, expected to grow to ten Petabytes within five years. It is managed with software developed at CCLRC, which is very strong on efficiency and reliability, but has only a very basic user interface. ADS provides an efficient tool for scientific archiving, but neither for data management (DM) nor for data integration (DI). Thus, communities require higher level interfaces and tools for DM and DI. Many interfaces and tools are provided and managed by us, but some communities manage their own higher level tools.

One such interface is the Storage Resource Manager (SRM) interface used by the particle physics community. This interface usually forms a component of a computational and data management Grid where it enables DM and DI via higher level services.

The San Diego Supercomputer Center (SDSC) have produced a DM/DI software package that provides a uniform interface for connecting heterogeneous data resources over a network, the Storage Resource Broker (SRB).

1.1 The Tier 1 and ADS

The particle physics community [4] consists mainly of experiments participating in the Large

Hadron Collider Computing Grid (LCG, [7]), and other international experiments with collaborators in the UK. RAL operates an LCG Tier 1 centre, one of about 12 in the world (with the LHC at CERN being Tier 0). Data is distributed from Tier 0 to Tier 1, and from there to the Tier 2 centres, of which there are four in the UK, run by UK universities.

The LHC is due to start operation in CERN in 2007. It is expected to produce 12-14 Petabytes each year. CERN will use the Grid to transfer data into the Tier 1 centre's disks at RAL at a sustained rate of of several Gigabits/sec while simultaneously serving Tier 1s in other countries, national Tier 2s and the CPU cluster at the Tier 1. Part of this data will have to be written to the ADS at significant sustained rates, and, in some cases, be read back and analysed while it is being written.

2 Data Integration and the Grid

Storage and data access interfaces are generally optimised to prevent different VOs, users, jobs, etc, from overwriting each other's files (using separate user id mappings, access controls, separate name spaces, separate partitions and storage pools on disks, etc). The challenge for DI is to bind the data back across the divides, to enable data discovery and sharing. Discovery is largely encouraged via metadata catalogues, whereas the sharing is enabled via common protocols and interfaces.

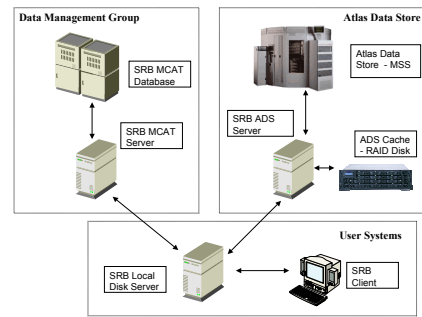
As mentioned in the Introduction, we provide several interfaces to the ADS. For the purposes of this paper, a *Grid interface* is loosely speaking one which interfaces to a Grid; in practice the distinguishing feature is often whether the protocol is carried over Globus GSI sockets.

Now consider the following basic cases for DI:

- I. A scientific community uses more than one interface, and needs to enable data access across more than one interface;
- II. Two scientific communities share data across the same interface;

Case (I) can be roughly subdivided into the following:

- I.1. Recovery: Data is written via the Grid interface, but the experiment wishes to be able to recover it via a non-Grid interface;



CCLRC Data Archival Management Architecture

Figure 1: SRB and ADS structure

- I.2. Migration: Data is written via the non-Grid interface, but the experiment wishes to analyse it using a Grid; or it is written via one Grid interface, but the interface is later replaced with another;
- I.3. Partial dual access: Data is written and accessed via one interface, but read-only access is granted via another interface;
- I.4. Full dual access: Data is written and read simultaneously via two distinct interfaces.

In contrast, DI on the same interface (case II) is much less common, even in the case where more than one experiment share the same interface. Nevertheless, the tools developed to support case I also enable DI for case II.

In the following sections, we look at this in more detail for the two main Grid interfaces, SRB and SRM.

2.1 Data Integration via SRB

SRB has provided a useful tool for managing geographically distributed data across different computational platforms, and has provided much-requested desktop accessible tools for data management. But many users also need the capability to archive large quantities of data managed by SRB.

We have found that the integration of the data management using SRB services and archiving capabilities using the ADS provides a facility much more powerful than either facility alone, and one that has proven to be of major interest to scientific users with large-scale data archival needs. Figure 2.1 shows the typical architecture used for this managed archive facility.

ISIS was interested in developing an archival data management solution using SRB. The ISIS facility at CCLRC manages 20 instruments that

provide pulsed neutron and muon beams for visiting teams of scientists to use in their own individual research programmes. While experiments are running they produce data 24 hours a day, which must be archived and made available to the scientists, both immediately during the running of their experiments and longer term after they return to their home institutions. Depending on the science of the experiment, a “run” of data is produced from the instrument anywhere from every 2 minutes to every 2 days. A run will produce one large file of data (about 100 MB) and possibly ten or more small files of descriptive information. Recent developments in SRB will allow ISIS and Diamond SRB resources to be federated, thereby allowing DI between the two facilities.

CCLRC has recently undertaken development for BBSRC to manage an SRB system for their data archival needs, including a special-purpose GUI to manage the end-to-end transfer and tracking of archive packages. Their data transfer must take place in 2 steps: from BBSRC local sites to a central site over slow network connections, and then nightly during specified hours from the central site into the ADS over a higher speed network. Users then require email notification when 2 copies of their data are resident on tape. This infrastructure will provide an archive service to approximately 5000 scientists scattered across the UK among the many component institutes of the BBSRC. This is the first time DI across the entire organisation of the BBSRC has been attempted.

2.1.1 Enabling DI via SRB Metadata

The ISIS and BBSRC solutions are enabled through the use of metadata at a number of levels. Each SRB “zone” consists of a number of storage servers, each with one or more storage vaults (resources), and a metadata catalog (MCAT) that manages the data held at each location. It is the system metadata within the MCAT that enables efficient data retrieval from, and movement of data between, these resources.

System metadata describes resources, files and logical collections of files. Each of these areas is now considered in more detail.

Resources: a key aspect of a Grid infrastructure is the ability of a user to view and manipulate data without any awareness as to the physical location of that data. Any type of storage device with a suitable SRB driver can be defined as an SRB resource. Each SRB server effectively communicates with virtual storage de-

vices. Given that files may be replicated across multiple resources (to enhance performance or data security), sufficient information needs to be recorded about each resource to enable sensible decisions to be made as to which replica to use for a given operation, e.g. if a replica of a file exists in an ADS tape resource, but another is also held on a disk cache resource, the latter should be used for retrieval purposes.

Files: System metadata is maintained for all files ingested into SRB. This includes information about the physical location of the file, file ownership, file attributes such as size, and access control information.

Logical Collections: An important aspect of successful DI across distributed storage locations is the ability to view data according to a purely logical structure, independently of physical storage location. While files are stored physically within resources attached to SRB storage servers, they are organised logically according to a hierarchical directory structure that exists purely within the MCAT. These directories are referred to as ‘collections’. Thoughtful naming and organisation of data to reflect semantic relationships between files helps to promote effective use of the data, providing a level of documentation that aids DI and long-term understanding of the data in combination with user-defined metadata as described below. It also assists the efficient retrieval of data from the ADS, helping to ensure that related data is stored on a single tape. In some cases, data already has an appropriate structure having been generated automatically from an experiment, but additional higher-level organisation and annotation is still beneficial. Key benefits of the logical collection approach are the ability to physically move data between resources without affecting the logical structure, and the enforcement of access control across a logical collection of data, irrespective of the underlying location or storage mechanism.

As well as system metadata, SRB also supports the association of user-defined metadata with SRB objects (i.e. files or collections). Prior to the release of SRB version 3.3 this has taken the form of attribute / value pairs. The BBSRC archival system has used this to good effect, associating descriptive metadata with the top-level collection for each “archive package”. It is possible to perform searches on user-defined metadata, effectively creating a new logical collection of data representing the search results, e.g. all archives originally submitted by a specific user, or archives associated with a specific project.

Where the simple attribute / value pair meta-

data has been insufficient to meet the more complex metadata requirements of some projects, an external metadata database is used for discovery purposes, with links into SRB space. An example of this approach is the e-Minerals project. With SRB version 3.3 comes a new extended metadata schema facility, allowing far more complex metadata schemas to be associated directly with the MCAT. Additional indexing can be performed to improve performance when searching metadata.

2.2 SRM

Introduction

Particle physics needs to develop DM software to manage large amounts of data across a multinational Grid. The challenges are predominately data volume and performance across a highly distributed environment: large data volumes and highly concurrent access the norm within the physics community. These challenges will also affect bio-genetics, as numerical methods become increasingly data driven.

Most development work within LCG, EGEE and, previously, EU DataGrid, has been based upon a service-oriented architecture, with specified and decomposed services. The great benefit of well-specified interfacing through SOAP is that slow or less reliable prototype services can be replaced by more robust or performant ones without changing services dependent on it.

High Capacity Data Grid architecture

DM services within the LCG community are typically broken down into three services:

- A Grid interface to storage using the SRM protocol, the Storage Element (SE),
- One or more replica catalogues, and
- Data movement services.

These basic DM services are complemented by domain-specific metadata catalogues which are used to facilitate and enable data discovery.

Although multiple implementations exist for each of these three core service components, their function is consistent across implementations due to the service oriented architecture. Interoperability depends on how well the protocol is specified. The API of a Web Services protocol is defined via the Web Services Description Language (WSDL), but further information is required to define the semantics of the API. We have found with SRM that extensive testing is required to achieve interoperability due to parts

of the semantic specifications that were open to interpretation.

The service oriented nature of the SRM protocol has enabled multiple higher level services to provide similar DM and/or DI functionalities. As yet, common interface specifications and standardisations have not become established within these DM/DI communities.

Storage Elements

The functionality in the SRM protocol consists of an agreed common core, and further components which are optional. The common core allows services to interoperate in a site neutral manner irrespective of the underlying storage infrastructure. Storage elements also publish metadata containing status indicators such as access time (latency) and free space to higher level services. This is published using a work in progress standard called the Glue Schema. SRM-based SEs are intended to run virtualisation of storage resources at all scales of storage from the tape silos at CERN to the smallest university cluster.

The SRM API works with an asynchronous model simulated over a synchronous protocol. Within the grid community, the requirement for a consistent use of SOAP means that SOAP is used synchronously, as Remote Procedure Calls. The asynchronous API model provides distinct advantages, as SEs must cope with the difference in access latency. Latencies for online and offline storage vary greatly between, say, a RAID array and a human being asked to retrieve an offline tape to honor a request. Since the time taken to retrieve data can often outlast a typical HTTP time-out, requests are queued and return request identifiers that can be queried as to the current state of an operation. The SRM protocol also provides built in redirection, which facilitates clustering of data servers.

Replica catalogues

The purpose of the catalogue is to provide data discovery across multiple SEs. Replica catalogues are databases of file references, allowing discovery of all instances of a file within the Grid. The query retrieves the host SE, the local storage name. Higher level services locate the optimal copy based on “cost” of access, which is based on the latency of getting the file ready for transfer and the network bandwidth/availability. The catalogue is populated at the time of data transfer/creation, requiring files to be registered and used through higher-level services, which wrap and invoke the replica catalogue.

Distributed Data Services

Distributed data services move data from one SRM to another. The data distribution services are responsible for storing files, scheduling transportation, distributing and retrieving files while keeping the replica catalogue and the SRM's databases synchronised, while optimising bandwidth usage of the network and the SRM's nodes.

This class of service typically hide the complexities of the SRM service and the Replica Catalogue services and provide the end user with all the tools and commands required. This is to prevent the replica catalogue from synchronisation issues with the underlying SRM.

The algorithms and approaches used to distribute data appear to be domain specific and so many competing services have been developed. This reflects the distributed nature of resources and the user communities' need for high performance: competing data services have typically been developed to serve an individual user community. This diversity implies that considerable research is required before a single solution for DI between the experiments can be developed that matches all requirements.

User Tools

Data management within the Grid is a large stack of software and each layer of software has its own end user tools. Typically the distributed data services provide tools to facilitate reading and writing data as well as providing explicit control for moving data across Grid resources.

Most Grid applications are not developed specifically for the Grid so compatibility tools also needed to be developed allowing non-Grid software to work within a Grid environment. GFAL is currently the only application production grade tool that provides native POSIX file access to Grid Storage through LCG's distributed data services. GFAL uses the library overloading of system function calls to intercept native file operations and interoperate with the distributed data services, replica catalogue services and with the SRM functionality.

User Metadata

Currently most physics communities consider their experimental metadata search and retrieval patterns to be too specific for needs to use common tools. This has not stopped standardisation and requirements gathering efforts to attempt to establish commonality between the communities. We hope that this work will provide the basis upon which high performance flexible frameworks can be developed across user communities.

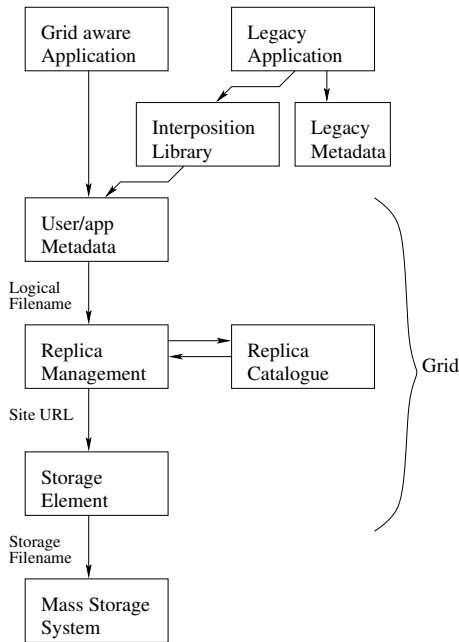


Figure 2: Simplified Grid structure

Figure 2 illustrates the simplified view of the Grid DM. DI is enabled by enabling data discovery at various levels. As the ADS and the SE are the lowest level services, shared between all experiments, DI can be achieved by enabling data discovery through higher level Grid services.

The majority of experiments using SRM require a single namespace, which enables DI by allowing filenames to be shared. Experiments have read access to the whole namespace but are not allowed to write into each others' parts of the space.

In the SE, path names in the namespace are independent of the interface used to access the service. This data written in through an SRM interface can be retrieved through a GridFTP interface or even a non-Grid interface in the event of a "Grid Failure".

2.3 Non-Grid access

We return now to the DI case I. The subcases I.1 and I.2 have one thing in common, that data is written into the storage system via one interface, but ownership is taken over by another.

In the more difficult subcases, I.3 and I.4, data is accessed simultaneously by two distinct interfaces, in our case a Grid interface and a non-Grid interface. By accessing data via the non-Grid interface, the Grid access control and auditing is bypassed, and the metadata held by the Grid interface may become out of date.

For all of these subcases, the essential component is that the ADS metadata is synchronised (in the case of migration) and kept synchronised with those of the higher level service. To resolve this issue, we need to look closer at the ADS DM metadata.

ADS metadata

The main application of the ADS metadata in the context of DI is to ensure *data integrity*, as a prerequisite for DI. However, we shall see that by publishing part of the ADS's metadata we can enable DI across higher level services.

The metadata stored in the ADS catalogue can be split into the following areas:

1. Authorisation. The ADS has the concept of a primary user associated with an ID that is registered on the system. This user can delegate access to other users from complete access and control down to access to an individual file.
2. Physical Tapes. The catalogue maintains a mapping of which user files are stored on which physical tape.
3. User files (virtual tapes). This information includes the owner and name of the file, which physical tape the file is on and the block addresses for start and finish of the file (for each tape the data is on) and file size (current and maximum), date and time information, and checksums.
4. Active files. Information about file that are actively being used, such as which data node the file is on, priority of request, etc.
5. Tape drives. The information includes the status of the drive (active, idle, dis/mounting).
6. Data servers. This list contains the location and type of data server node.

While the user can access these values most of the information is of little use, and at the moment only the virtual tape information is used. The checksums do not match those calculated outside the system, and physical tapes are recycled if the usage becomes sparse so they will change over time. The fields for file names is limited as what the user creates is a virtual tape, and so it follows the tape standards for these values.

To over-come these limitations software has to be layered on-top of the ADS system to provide the user with a "logical" view of the data.

One such system is the pathtape server. This takes a path name (i.e. /home/gtf/testfile) and will return a unique virtual tape name that the user can use on the create option. The server allows lookups to map path names to virtual tape names. This is not ideal, and should be integrated closer into the system.

In summary, the ADS DM metadata can enable DI mainly by ensuring data consistency and storage robustness. As yet, the higher level services have not made use of the ADS metadata but it can be used to perform optimisation of tape access and version control.

The authorisation systems also has implications for DI in case I, because it needs to be managed across all interfaces. However, due to space limitations, we are unable to cover security aspects in this paper.

2.3.1 Tier 1

The Tier 1 has 60 disk servers with 40TB deployed as part of the main SE cluster providing SRM and GridFTP interfaces for 3 LHC experiments: CMS, LHCb and Atlas. There are also another 160TB accessible through NFS and other non-Grid methods to the Tier 1 batch worker farm. The size of the farm is partly to fulfill the requirements of LCG, that data can be buffered if the ADS is temporarily unreachable, but also to enable data to be analysed while data is still being written. While data will be ingested at a constant rate, some experiments will be reading it back to analyse it.

This multiple disk, multiple interfaces approach means that data can be accessed either via the SE, or directly in the SE caches or on the disk farm via GridFTP, or using non-Grid access via NFS. In practice, this allows case I.3 and I.4 access for the jobs running on the computing clusters. Clearly, the problem with case I.4 access is that if the job updates the file via NFS, to let the SE know that the file has been changed.

3 Current and future challenges

3.1 Planning

Running a centralised service for a large number of diverse communities clearly has advantages, but also presents many challenges. The storage cost modelling shows that different staffing thresholds exist for the size of both disk and tape archive services, below which they are un-economical to staff, and above which economies

of scale exist (up to other thresholds). Furthermore, the centralisation enables a dedicated and consequently more professional service including, tested disaster recovery services, off site storage, media circulation facility, routine media migration facilities, and so on. Planning for large scale operational services provide many challenges including:

- Constant cycle of hardware upgrades.
- Longer term development and replacement of underlying software tools and services.
- Automated migration of very large data volumes:
 - To new media
 - To new data formats
 - To new data management tools

Due to growing demand and rapidly changing technology the data storage aspects of the service are in a constant cycle of hardware upgrades. In the last twenty years the data archive has gone through five major upgrades and has used four different types of tape robot. It is essential that any underlying operation service is independent from vendor hardware to ensure that this process can be continued with minimum disruption both to the operational staff and to the user.

The growing demand for increased volumes also affects the underlying software system on which the archive services are built. Although the repeat cycle for this is much longer term than for the hardware — often in the order of tens of years rather than two or three years, when it does happen it can be far more significant, since it is far more closely linked to the user level than the hardware. The archive system software in use on the ADS was developed 15 years ago, and at that time it was never envisaged that it would eventually be required to scale to multiple Petabytes and billions of files. We are currently in a process of review to consider options between implementing and adapting other, recent systems such as CASTOR, or to continue to develop and maintain our existing underlying software system. The challenge is to migrate seamlessly and maintain the service level for the higher level services.

The archive service expects to change the underlying hardware frequently. Consequently the process of data migration from one media form to another is a necessary and more or less ongoing part of the service about which users need

not be aware or concerned. However, the problem of migration of data from old to new formats is a much more serious problem. Since the archive service is not aware of (or interested in) specific formats of different data sets (since it sees only a binary data stream), the problem of migration to new data formats is faced and resolved by the users, or by specific community data centres which are close to their own user community, and whose role is to support that community. In the worse case where data has been stored in local, specialised formats which are poorly documented and poorly understood due to long-term staff turn over this can lead to loss of data. This is one area which is being considered within Digital Curation Centre [2], whose development centre is based at CCLRC Rutherford Appleton Laboratory. One approach being considered is the use of specialised tools such as EAST [8] and DFDL to define the obsolete data formats and provide an automated conversion from the old to the new data format.

Another challenge facing the services in the years ahead is the need to automatically migrate the data and its associated metadata from the current to future generation of tools. Whilst SRB and SRM are highly suitable tools for DM services right now, it will one day be superseded by the next generation of tools — whatever they will be, and with it will come the need to migrate very large volumes of data from SRB into these new versions. It has already been noted that DI is expected to further exploit the separation of data from metadata. Maintaining the link between data and metadata in automated upgrades from known to unknown future tools is not a trivial problem. This is one of the issues being considered in the CLADDIER project.

3.2 Other issues

There are other DM/DI issues, which we have been unable to cover in this paper in sufficient detail to do them justice, or at all. These include:

- Security. Access control, Virtual Organisation management and user lifecycle management, single sign-on, confidentiality requirements;
- Data discovery in a full data and computational Grid;
- Garbage collection and “volatile” space;
- A full discussion of the future of web services in relation to mass storage interfaces;

- Thin client vs. thick client interfaces and portals;
- Language and portability issues;
- Metrics;
- Usability of interfaces and DI tools.

4 Conclusion

The growth of DI will be facilitated by the increased use of the Grid and its computational resources. As communities increasingly understand the benefits of these developing technologies, and start to realise the potential benefits to be gained from sharing their data, the levels of DI will increase further.

To service several diverse communities, we need to maintain multiple interfaces to the ADS. We find that the low-level ADS DM infrastructure is necessary but not sufficient to facilitate DI, but not enough use is being made of it to optimise the access.

We have discussed how the SRB and SRM interfaces take different approaches to DI. SRB enables DI for case II via an essentially self-contained DM Grid, allowing experiments to enable DI via SRB's own metadata services or higher level metadata databases.

SRM takes the "low level" storage service approach required to support a component-based, high performance, heterogeneous data and computing Grid infrastructure. In this Grid, data is often discovered at various levels of the hierarchy, and several metadata tools are available at the different levels. Unfortunately, these higher level services differ between experiments, even architecturally.

The heterogeneous Grid component approach, when not everybody uses the same interface, highlights the need for well-defined open standard protocols. We find that in the cases where such protocols are not available, DI is often hindered, because each site or each experiment will provide its own non-interoperable solution.

More generally, one important conclusion is that a carefully chosen DM architecture and infrastructure helps enable DI, provided the users are educated about what they can do and what the limitations are (cf. the complexities of supporting case I.4 vs. I.3).

There are advantages to *not* integrating data across interfaces, particularly when no experiment/facility uses more than one interface. It simplifies the service we need to provide, because

services were originally designed to provide different namespaces etc. for different interfaces.

Finally, we have found that DI to ADS is helped by integrating the work of different groups in CCLRC's e-Science centre, mainly the storage group and the data management group.

Acknowledgments

This work was funded by GridPP2, PPARC, BBSRC, and EGEE, and CCLRC.

We gratefully acknowledge interesting and useful discussions with Andrew Sansum and Steve Traylen from the RAL LCG Tier 1 centre.

This document was typeset with L^AT_EX.

References

- [1] Biotechnology and Biological Sciences Research Council. <http://www.bbsrc.ac.uk/>. (June 2005).
- [2] Digital Curation Centre. <http://www.dcc.ac.uk/>. (June 2005).
- [3] Diamond Light Source. <http://www.diamond.ac.uk/>. (June 2005).
- [4] GridPP2. <http://www.gridpp.ac.uk/>. (June 2005).
- [5] Integrative Biology. <http://www.integrativebiology.ox.ac.uk/>. (June 2005).
- [6] ISIS. <http://www.isis.rl.ac.uk/>. (June 2005).
- [7] Large Hadron Collider (LHC) Computing Grid. <http://lcg.web.cern.ch/LCG/>. (June 2005).
- [8] Denis Minguillon. EAST: A standard and its tools. <http://debat.c-s.fr/tour/documentation/conferences/estec19-02-03/cnes1.pdf>, February 2003. (June 2005).
- [9] National Crystallography Service. <http://www.ncs.chem.soton.ac.uk/>. (June 2005).
- [10] Particle Physics and Astronomy Research Council. <http://www.pparc.ac.uk/>. (June 2005).