Integrating Grid Computation and Data Services for the Simulation of Complex Materials

Lisa Blanshard l.j.blanshard@dl.ac.uk

Rik Tyer r.p.tyer@dl.ac.uk

Kerstin Kleese van Dam k.kleese-van-dam@dl.ac.uk

@ CCLRC Daresbury Laboratory, Warrington, UK

Key words to describe the work

grid, metadata, database, data, portal, materials science, workflow

Abstract

CCLRC is involved in the development of grid and data management tools for the Simulation of Complex Materials e-Science project [1] otherwise known as *e-materials*. The aim of the project is to bring together computer and computational scientists to exploit new and existing technologies for key current areas of materials science relating to:

- the development of *combinatorial materials chemistry*, with specific applications to catalysis and ceramics
- the *prediction of polymorphs* of organic-pharmaceutical compounds and their properties

Currently scientific data is distributed across a multitude of sites and systems. Scientists have only very limited support in accessing, managing and transferring their data or indeed in identifying new data resources. In a grid environment that spans numerous sites and organisations, it is essential to ease many of these processes. Therefore the aim of the project is to help with automating many of these tasks. Our first step is to manage the data the materials scientists produce in the areas of polymorph prediction and combinatorial methods. Currently they have no formal methods for storing and accessing the numerous files created from the applications they use.

During the first year of the project we have developed applications or modified existing applications for managing workflow and data.

The *workflow application* allows the user to string together a sequence of calculations. The calculations are performed by a number of computational chemistry programs. Each program requires a set of input files containing some data, such as the molecule name or structure, and additional parameters that control the flow of execution. Program execution may take a number of hours or days. Output files are generated during execution. Information contained within the output is used as input to the next calculation in the workflow sequence.

Data applications are required to manage the input and output files and provide a number of web interfaces so that the scientists can catalogue their data using *metadata.* Metadata includes details of the person who created the data, programs and facilities used and the location of the associated *data files* i.e. input and output files from each calculation.

Paper outline:

- *Background* and *Project requirements* - background to the science and working practices at the beginning of the project; main issues
- *Progress so far* - the tools and databases that have been provided for computation and data storage as part of the project

- *Uploading files automatically from computation to storage* - how files generated by the simulations during the computation workflow will be moved into a distributed file system; how access to the files will be provided
- *Uploading molecular data I* - how temporary data parsed from the simulation output files will be exported into a relational database of molecules, conformations and crystals; how access to the data will be provided
- *Automatic metadata generation* - how metadata will be automatically generated where possible and moved to a relational database of metadata
- *Uploading molecular data II* - finally how data parsed from the output files will be moved into a relational database of molecules, conformations and crystals; how access will be provided

Background

The project covers the following research areas:

- Combinatorial Materials Science

Combinatorial materials chemistry has specific applications to catalysis and ceramics. The main scientific focus is on two areas: acid sites in zeolites and metallocene characterisation. Computer simulations are used to find the properties of various materials such as the total energy and binding energy of the molecular structure.

- Polymorph Prediction [2]

Organic crystalline materials are prevalent in many industries, including pharmaceuticals, agrochemicals, pigments, dyes, explosives, and specialty chemicals. Many compounds can crystallize into polymorphs, multiple crystalline forms of the same molecule. Polymorphs may differ in key properties such as shelf-life, bioavailability, solubility, morphology, vapour pressure, density, colour, and shock sensitivity. Therefore, when working in the solid state, it is important to know how many polymorphs are possible as well as how their properties might differ. Once a particular form is chosen for its desired properties, researchers need to control the crystallization and formulation conditions so that unwanted polymorphs do not appear. In order to do so, they need to fully understand the structural aspects of each polymorph. This knowledge is also important for patenting and registration purposes.

The most common method for determining a crystal structure is to grow quality crystals for single-crystal X-ray diffraction. However, sometimes such crystals are difficult to crystallize. Furthermore, one can not be certain that all possible polymorphs have been discovered experimentally. Thus, methods that predict potential polymorphic structures, starting with just the contents of the asymmetric unit, would be extremely valuable.

Project requirements

Computational requirements

During the last decade, Professor Sally Price’s group at UCL Chemistry department has developed a computational approach for predicting the crystal structures of small, rigid, organic molecules. [3] However each search involves running large iterations of computationally expensive calculations and currently takes a few months to perform. Studies on larger molecules, which are more typical of manufactured organic materials, are not feasible using the existing computing infrastructure.

The project scientists use a number of simulation programs. These are a combination of commercial codes (Gaussian 98, Cerius2) and open-source (Molden, Molpak, Dmarel) plus others. For example, one of the steps to predict polymorphs from a molecular formula is to use Gaussian 98 to calculate the molecular properties such as density and population.

The applications can take a number of days to run and many runs are completed in sequence to ascertain scientific results.

Data management requirements

Typically scientists are forced to manually relate between experimental, data, computing and analysis facilities that are available world wide, with little infrastructure support. In the future it is hoped that the grid and associated middleware will provide these functions, enabling the scientists to choose much more easily from a wide range of services, connecting and combining desired services for an optimal working environment. Much of the access to the grid is envisaged to take place through customisable, community oriented *portals*. The e-materials project is one of a range of projects within CCLRC have been chosen to provide the initial building blocks of an integrated solution for users of experimental, computing and data facilities, demonstrating on a few selected examples how basic technologies can be used to build middleware components that support high level scientific grid applications.

Data will play a pivotal role in the success of Grid or e-Science developments. Virtually all envisaged applications will need to be able to draw from and deliver to the distributed heterogeneous information/data sources with a variety of contents. Hence three major challenges are posed: data accessibility, data transfer and management of personal data. Data accessibility implies the capability to locate information/data without prior knowledge of its physical location or format. Furthermore scientists, as well as applications, need to combine results from different sources.

In terms of the e-materials project, the scientists generate a large amount of data while running simulations and typically they have been stored on individual's machines or on the machines that are used for simulations. This has made access to the data within and outside the group very difficult. In addition, securing access to the data is complex due to the number of machines involved. There is a high risk of data loss as few backups are taken and this is disorganized.

To improve the situation a number of requirements have been established as a result of consultation with the scientists.

The three main aspects for data management concern management of *files*, *metadata* and *data.*

File Management

- *file storage* – input and output files that are created/generated from the many simulation runs. Initially a single file store will be sufficient, however later in the project, as the number of files increase, it will be necessary to increase the number of storage devices and locate them at different sites;
- *interface(s)* for the scientists to manage their own files i.e. upload files from their machines to the storage device; download/view files; organize in a directory structure
- *file transfer* – facility for auto-upload of files to storage by the workflow application
- *file sharing* - functionality to share their files with scientists in their working group at different locations
- *publishing* – ability to publish files on an internet site so that other interested parties may download/view them;

Metadata Management

- *storage of metadata* – a database to store the metadata catalogue – note that this is may be on a separate device to the one where the data files themselves are archived
- *interface* - allow scientists to create and edit metadata to catalogue the files including the location of the files
- *automatic metadata generation* – facility to generate metadata by the workflow

application where possible e.g. when the workflow application uploads files, the final location of files on the storage device could be written to the metadata database

Data Management

- *storage of data* – a database to store domain specific information from the files such as molecular structure of a compound and other properties, possible conformations or arrangement of atoms and possible crystal structures. This information should link to metadata so there is a record of how it was generated.
- *automatic data extraction* – tools to retrieve data from the output files. Note that the output files contain more or less free text and are different for each computational chemistry program so some conversion to XML may be necessary
- *automatic data insertion* – tools to insert the data into the database by the workflow application
- *interface* – for scientists to search for and view/download data
- *data sharing* – allow other scientists within the group to view/download data
- *publishing* – facility to publish data on the web for other interested parties to view/download; requires search facilities

In each case appropriate user interfaces will be necessary especially access via the web for project scientists and external parties once the data is published.

Progress so far

The project has been running for a year and we have made significant progress in the areas of computation and data storage. These tools are currently in use by the project scientists.

Computational chemistry on the grid

Making use of early implementations of the OGSA specification the UCL team have wrapped the Fortran binaries into OGSI-compliant service interfaces to expose the existing scientific application as a set of loosely coupled web services. The OGSA implementation facilitates the distribution of such applications across a large network, radically improving performance of the system through parallel CPU capacity, coordinated resource management and automation of the computational process. A computational workflow service enables users to distribute and manage parts of the computational process across different clusters and administrative domains. The web service coordination language, Business Process Execution Language (BPEL) makes such workflow services easily configurable. The aim is to provide services for running applications across a grid that scientists can configure themselves using relatively user-friendly languages such as XSLT and BPEL.

The reason for using BPEL for workflow management is that we do not necessarily know how the computational process will develop in the future and we are trying to give the scientists freedom to alter their methodology. Our solution prevents software from dictating their agenda.

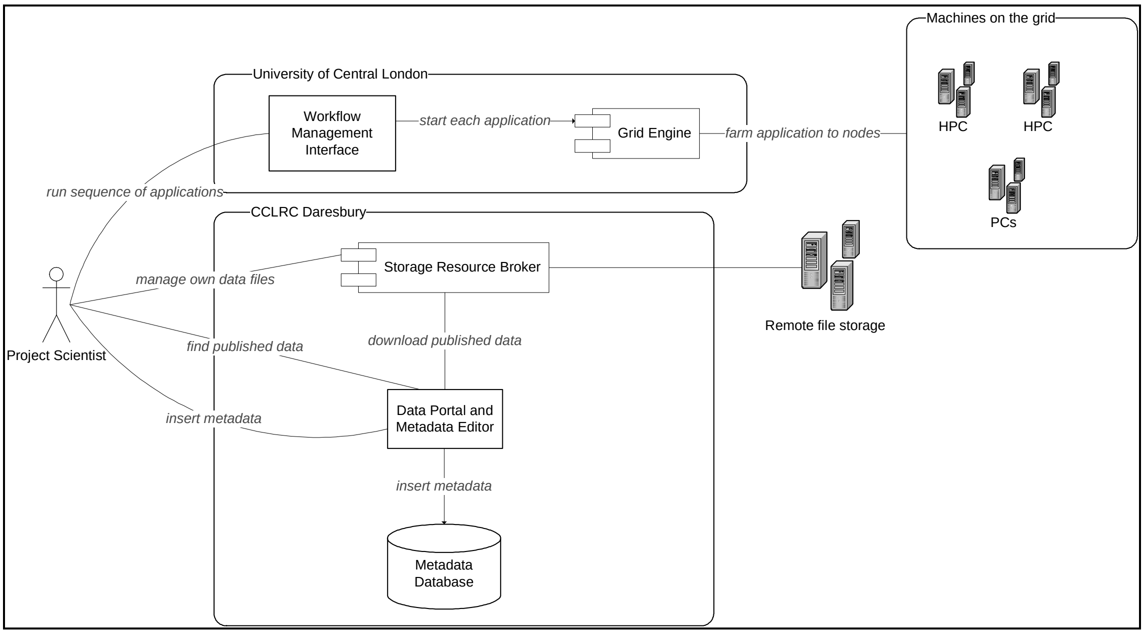

Figure 1 Component diagram of the current interaction between simulations and data storage

Managing scientific data

A number of middleware tools have been developed or installed in the areas of file management and metadata management. These are the Storage Resource Broker (SRB) [7] for file management developed by SDSC and the Data Portal [5] developed by CCLRC. A relational database [6] housed at CCLRC is used to store metadata.

Storage Resource Broker

SRB is client-server based middleware initially developed by SDSC in the mid-Nineties to provide uniform access interface to different types of storage devices. SRB provides an uniform API that can be used to connect to heterogeneous resources that may be distributed and access data sets that may be replicated.

SRB is a means of allowing users to manage data storage and replication across the wide range of physical storage system types and locations available within UK e-Science, while still allowing having a single, stable, access point to the data.

The SRB Client is an end user tool that provides a user interface to send requests to the SRB server. There are three main implementations of this: command line *S-commands*, MS Windows GUI *InQ* or Web based *MySRB.* A recent addition is the MySRB server component. This allows all access to storage via a thin client. The MySRB server is actually an application server middleware component that acts as a client to service multiple thin client sessions.

Metadata database

CCLRC has provided a database to store scientific metadata i.e. information about a particular area of study, who was involved, where and how it was it carried out. The metadata stores the location of the data files. Figure 2 shows the schema of the database. This has been designed to serve as a multi-disciplinary metadata catalogue and is used on a number of projects.

The same physical metadata database will be used for both scientific areas of the project.

The schema illustrates the axioms for our domain. The important axioms follow:

A study may be (dotted line) classified by a number of key words – a key word can then be used later to find a particular study using SQL (Structured Query Language – the database query language), or through an application such as the Data Portal.

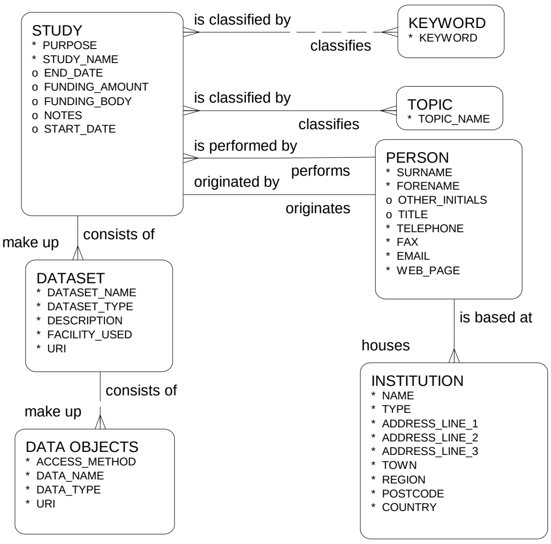

Figure 2 Metadata database schema

A study is classified by a number of topics e.g. */computational chemistry /polymorph prediction /aspirin.* The Data Portal allows searches using a *drill-down* list. The user selects *chemistry* from a list and then *computational chemistry* etc. to find the topic they are interested in. Each study must link to all relevant topics.

A study is performed by a number of people i.e. investigators. Also *a person originates a study*. So there are two relationships between person and study.

A study consists of a number of data sets – a dataset entity represents information about a directory or folder of files created during a particular simulation run, such as the remote location (URI) and type e.g. binary or ASCII. In the case of the e-materials project, the URI translates to a specific directory in SRB.

A data set consists of a number of data objects – a data object holds the location of a particular file from a simulation.

Metadata Editor

Using a web interface the scientists in the group can create new studies and datasets in the database and edit existing ones. It is here where they enter the location of the directory of files in SRB.

Data Portal

This provides high-level access to multidisciplinary data via the web, linking to existing or new data catalogue systems. These catalogues include metadata as well as links to the data itself. The data may be held in various storage resources from local disks, over databases to multi terabyte tertiary tape systems. At the moment all the data for the ematerials project is held in file storage managed by the Storage Resource Broker.

The DataPortal provides common search capability via a scientific metadata format in XML [4] developed by CCLRC. Information from the metadata database is transferred in this format. The common format also allows a cache to be held in memory of metadata from a number of metadata databases to be combined so that searches across all of them is possible if desired.

The Data Portal is used to share data with interested parties and amongst the group. SRB can also be used to share files amongst the group.

Uploading files automatically from computation to storage

As you can see from Figure 1 the scientist must manually collate the results of computation and then upload them manually to the SRB. Obviously this is not ideal. The next step is to integrate this process so that the data files are stored automatically in some predefined directory structure in the users home directory in the SRB. The following requirements have arisen from discussions:

- the workflow manager will upload files at the end of each calculation
- new directories will be created in the relevant user's home directory on SRB to store the files
- a security mechanism must be in place so that only designated programs/users may upload files
- the above functionality must be accessible from a number of different

heterogeneous machines in the various clusters on the grid as this is where the applications would run.

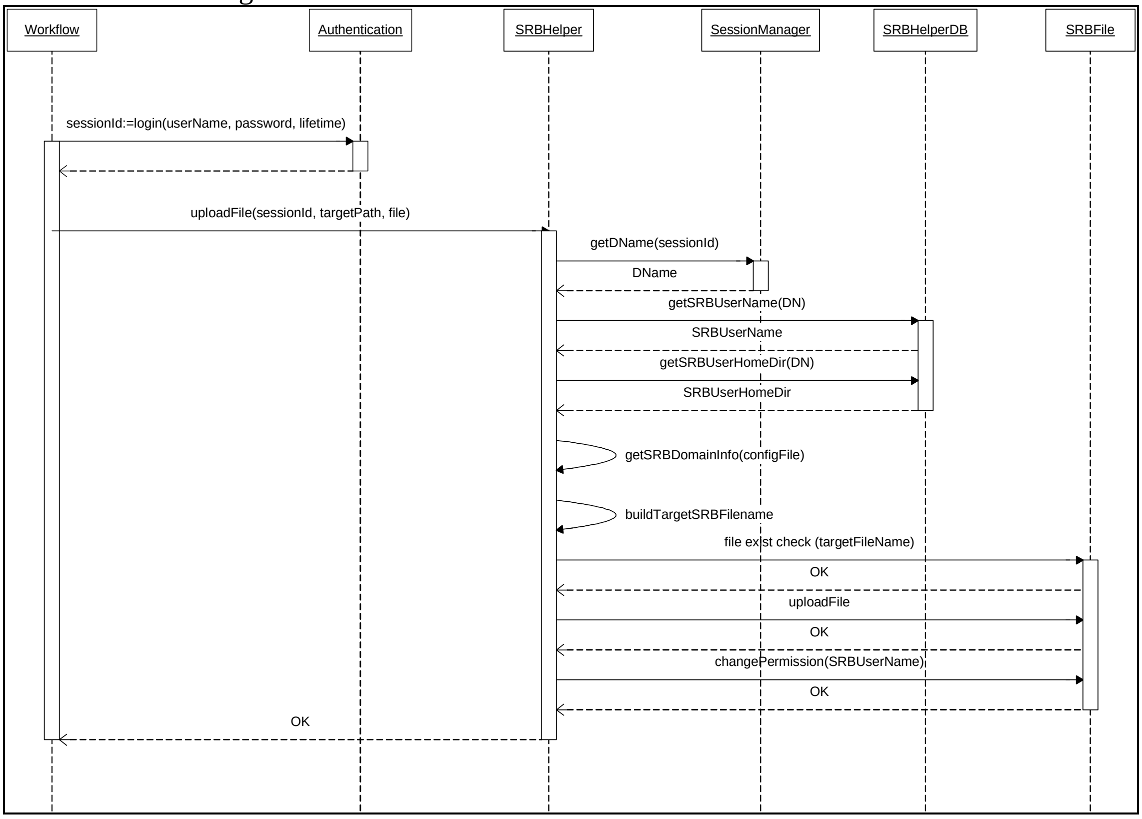

Figure 3 "Upload file to SRB" sequence diagram

To meet these requirements an SRB Helper web service will be provided by CCLRC:

- *SRB helper web service* - used to upload a single file to the SRB to a chosen directory in the user's home; also to create new directories in SRB
- *SRB Helper local database* – to store the name of the user's home directory in SRB

Since the compute services do not maintain conversational (session) state, the workflow would need to provide authentication details as part of each request to upload files or create a directory and hence degrade performance. To alleviate this we will use two other services that have been developed as part of the Data Portal:

- *authentication web service* – uses x.509 certificates issued by the UK e-Science Certificate Authority[1] to verify the identify of a user and returns a session identifier
- *session manager web service* – has access to a local database of sessions and associated *distinguished names* i.e. the user name and *lifetime left* i.e. the number of hours left on the certificate

The process of uploading files is as follows:

1. User logs on to the Workflow Manager and selects a sequence of applications to run

[1] in fact we require the user to upload their e-science certificate first into a *MyProxy* server that issues a *proxy certificate* in its place. The proxy has a limited lifetime so security risks are reduced. For more information see http://grid.ncsa.uiuc.edu/myproxy/

2. The Workflow Manager sends the jobs to the grid engine
3. When each job ends the Workflow Manager authenticates, creates appropriate directories in SRB and uploads the files.

Figure 3 shows the interaction between the Workflow Manager and the various web services during a file upload. Note that further files may be uploaded until the session has expired. The Workflow Manager then needs to re-authenticate with the user's details.

Uploading molecular data I

As part of the polymorph prediction computation workflow, the scientist manually selects a number of low energy structures with which to continue processing. The scientist is presented with a scatter plot of the various results and selects them via the user interface. The scatter-plot contains a number of properties that were generated during the simulations within text-based output files. As part of the workflow the output files are parsed and those properties put into a local intermediary database.

The next step on integration therefore will be to transfer this data into final storage as part of *data management* aim of the project. Also we need to extend the web interface in the Data Portal to access data as well as files.

To meet these requirements the following must be developed:

1. A database schema to hold information about molecules, conformations, crystals and their properties
2. SQL script(s) to select data from the intermediary database in to the above database schema
3. Database link or export facility to transfer the data into its final storage
4. Interface for users to view the data in a suitable format via the web that is secure. This will most likely use existing services such as authentication and session management that are used as part of the Data Portal. It is envisaged that the user will want to find the data by searching through the metadata catalogue as they currently do when looking for files.

Automatic metadata generation

As can be seen from Figure 2, metadata includes information about *studies* and *datasets* where a study concerns a particular area of research e.g. polymorph prediction of aspirin, and a dataset belonging to that study describes a directory of output files resulting from a particular simulation run. It also contains the location in SRB of the directory itself as a URL [7] e.g.

```
srb://emat.cclrc.ac.uk/home/li
sa/polymorph/aspirin/lowEnergi
es
```

At the moment the scientist uses a web form to enter the study and dataset information into the metadata database. However this was a interim solution until some of the metadata could be generated automatically. In this stage we aim to generate as much metadata as possible and insert it automatically as part of the workflow.

1. User uses a web form to create a new study in the database and enters details such as study name, description, notes.
2. Other details are auto generated such as the date, author, institution etc
3. User logs on to the Workflow Manager and selects a sequence of applications to run and also links the workflow to the study
4. The Workflow Manager sends the jobs to the grid engine
5. When each job ends the Workflow Manager authenticates, creates appropriate directories in SRB and uploads the files.
6. Then the dataset metadata is automatically generated from the workflow information. The location of the files in SRB is added to the dataset metadata

7. This is repeated at the end of each job until the workflow has finished
8. User can then use the data portal to open the study and view the new datasets and associated files.

We plan to develop the following to meet this aim:

- a web interface to create a new study for the user to enter limited information (this may be derived from the existing Metadata Editor tool)
- service to allow the Workflow Manager to insert dataset metadata e.g. insertDataset()
- changes to the Workflow Manager interface to allow the user to link the workflow to a study

Uploading molecular data II

One of the final stages is to replace the storage of output files with data parsed from the output files in a relational database. This will provide many benefits to the project scientists as they will be able to query the data in ways that are not currently possible, for example, is a particular property of a molecule more likely to result in the compound crystallizing into polymorphs.

It is envisaged that some of the output files will be kept for an interim period until the project scientist is happy that the data parsed from the files is as expected.

A significant challenge that we face is that the output files are in a legacy textual format that is different for each application. The output also depends on the parameters that were entered. To solve this problem there has been much effort into converting the applications where possible to output a standard XML format specific to the domain called Chemical Markup Language [9]. However there is further work to do in extending the language so that some of the information can be represented. Once the output is generated in CML for all cases then it will be relatively easy to extract the information and put it into the database. Oracle provides a number of tools to do this which we are investigating.

1. User uses a web form to create a new study in the database and enters details such as study name, description, notes.
2. Other details are auto generated such as the date, author, institution etc
3. User logs on to the Workflow Manager and selects a sequence of applications to run and also links the workflow to the study
4. The Workflow Manager sends the jobs to the grid engine
5. When each job ends the Workflow Manager authenticates, creates appropriate directories in SRB and uploads the files.
6. Then the dataset metadata is automatically generated from the workflow information. The location of the files in SRB is added to the dataset metadata. Metadata is uploaded to the database at CCLRC.
7. The output files (now in CML) are parsed and the relevant information is extracted. The WM calls the relevant services to insert the data into the database at CCLRC.
8. This is repeated at the end of each job until the workflow has finished.
9. User can then use the data portal to open the study and view the new datasets and associated files. Users can also view the data in a suitable format.

To meet these requirements the following will be developed/used

1. a number of web service methods to insert data about molecules and their properties, conformations and their properties and crystals and their properties. These web services will be called from the Workflow Manager:
   - insertMolecule()
   - insertConformation()
   - insertCrystal
   - updateMolecule()
   - updateConformation()

- updateCrystal()

2. all data must be output using CML or another XML format
3. functionality to extract the data from the CML/XML

References

[1] Simulation of Complex Materials e-science project http://www.e-science.clrc.ac.uk/web/projects/complexmaterials

[2] Overview of Polymorph Prediction http://www.accelrys.com/cerius2/polymorph.html

[3] Prof Sally Price (UCL) - Research Pages http://www.ucl.ac.uk/~ucca17p/home_all.html

[4] CCLRC Scientific Metadata Format http://www-dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf

[5] CCLRC Data Portal http://www.e-science.clrc.ac.uk/web/projects/dataportal

[6] CCLRC Database Services http://www.e-science.clrc.ac.uk/web/projects/database_service

[7] Storage Resource Broker http://www.npaci.edu/DICE/SRB/

[8] Implementing and using SRB, Proc UK e-Science All Hands Meeting 2003, © EPSRC Sept 2003, ISBN 1-904425-11-9 http://www.nesc.ac.uk/events/ahm2003/AHMCD/ahm_proceedings_2003.pdf

[9] Chemical Markup Language http://www.xml-cml.org/