Distributed Data Management of Scientific Data within the eMinerals Project

R. Tyer[1], L. Blanshard[1], K. Kleese[1], R.J. Allan[1], M.T. Dove[2], M.Calleja[2,], J.Wakelin[2]

[1]*CCLRC – Daresbury Laboratory, Daresbury, Warrington, Cheshire, WA4 4AD, UK.*
[2] *Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, UK*

1. INTRODUCTION

The 'Environment from the molecular level' (eMinerals) project[1] is a NERC funded eScience pilot project focused on fundamental science problems associated with key environmental issues such as nuclear waste storage and pollution.

Aside from the scientific issues, this research is challenging both in terms of the computational power required to tackle realistic system sizes with the required accuracy and the data management issues related to handling large amounts of data over a distributed virtual organisation. Hence the use of Grid computing together with associated data management technology provides enticing opportunities to facilitate and enhance this work.

The project involves the collaboration of environmental scientists, scientific code developers and computer scientists from Bath University, Cambridge University, CCLRC Daresbury Laboratory, Reading University, the Royal Institute and University College London.

From a data centric perspective, scientific research can be viewed by the sequence of events shown in Figure 1.

Figure 1 – Data centric view of scientific workflow.

In general terms, research begins with some data, for example a crystal structure. Some analysis is performed on this data which in turn generates more data, i.e. results. These results are then stored and annotated in some fashion after which a subset may or may not be selected for publication or distribution to a wider community.

One of the principle eScience challenges of the eMinerals project and the eScience programme in general is providing an effective and integrated infrastructure which facilitates each step of the data lifecycle for a distributed (both geographically and political) community. To this end a number of pieces of middleware have been developed and deployed for the project. This paper will elaborate on the data lifecycle shown in Figure 1, describing the pieces of middleware used, the issues motivating their need, and the benefits their use brings to the project.

2. DATA DISCOVERY

The process of data discovery is currently somewhat haphazard. The majority of scientific data is distributed through peer review journals. Although there do exist mechanisms to search this literature, the data held in these publications only represents a small fraction of the total data generated in order to write these papers. Also the data is not available in a machine readable format so manual manipulation is necessary in order to make use of any data harvested from a particular publication.

In addition to scientific publications, there exist a number of databases holding scientific data for specific user communities. In general these databases each have their own access mechanisms, search mechanisms and database schema. This heterogeneity results in it being non-trivial to search across this different data holdings in a coordinated fashion.

The CCLRC Data Portal[2] is a web portal which can perform parallel searches across multiple distributed databases. Its use facilitates the discovery of scientific data by allowing the scientist to search multiple data holdings simultaneously. In addition, the user is abstracted from both the access controls of the remote databases and their internal schema.

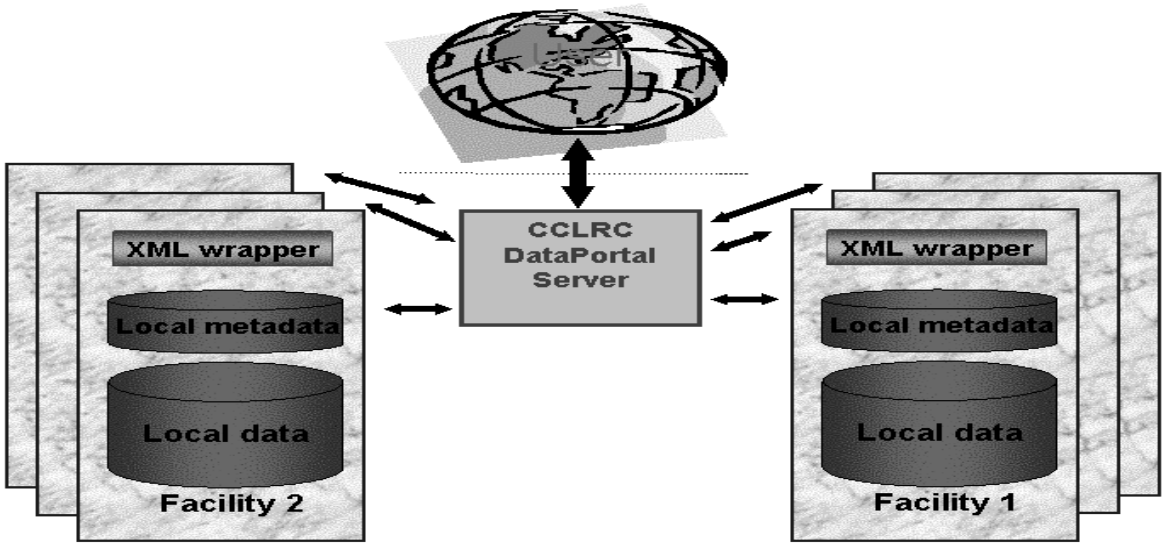

Figure 2 - CCLRC Data Portal Architecture

Essentially the Data Portal performs a brokering service between the user and the individual data holding facilities. A XML wrapper resides at each of the remote facilities as shown in Figure 2. These XML wrappers translate the search requests from the Data Portal into the SQL appropriate to the local database. Once these queries have been run, the XML wrapper then translates the results back into the CCLRC Metadata Format[3], which is a metadata schema developed to be a superset encompassing most scientific metadata. The highest level of the CCLRC Metadata schema is shown in Figure 3 and further details can be found in Reference 3.

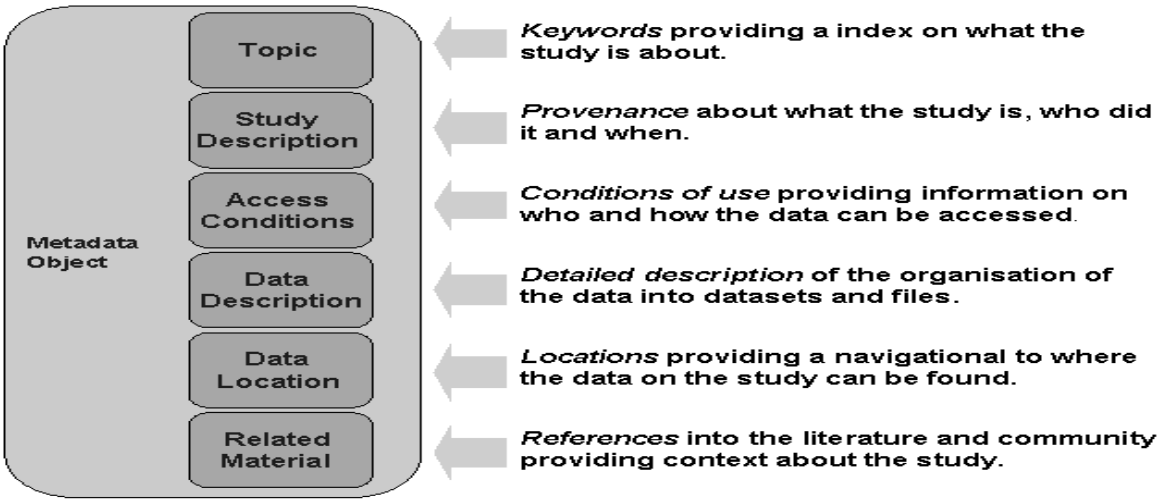

Figure 3 - CCLRC Metadata Schema

The use of the CCLRC Metadata schema allows search results to be returned in a common format. This provides the scientist with a single view of the search results and provides further abstraction from any database heterogeneity.

Search queries are specified by 'drilling down' within non-unique topic hierarchies. An example hierarchy could be:

Physics → Condensed Matter → Magnetism → Oxides → Manganites

These topic hierarchies form the basis of the search queries which are sent to the XML Wrappers. Essentially the topic e.g. "Manganites" is sent to all the wrappers in parallel and in turn they search through the metadata databases for studies relating to that topic. The returned studies are presented to the user and there is the option to download associated data files for further analysis. The files are typically stored on separate file stores.

The use of the Data Portal can vastly simplify and automate scientific data discovery, which can in turn deliver significant benefits. For example, it can facilitate collaboration, identify particular knowledge gaps and prevent unnecessary repetition of work. Also a systematic and more automated mechanism of data discovery enables larger scale studies to be undertaken which incorporate or extend previous work.

3. DATA ANALYSIS

The next stage in the data lifecycle is Data Analysis which represents performing some manipulation on the data. In the context of the eMinerals project, this would typically involve running one or more scientific codes on the data. The computational requirements of the applications vary significantly. However they can be broadly classed as either High Throughput Computing (HTC) or High Performance Computing (HPC) applications.

The project makes use of a number of HPC facilities within the UK including the HPCx and the CSAR machines. In addition the project has a number of small (16 node) Beowulf clusters.

For HTC applications, the project has set up a number of Condor Pools, the most notable of which is the UCL pool which currently comprises around 800 machines.

In order to integrate these different resources, the Globus Toolkit 2 has been used as a layer above the individual queuing systems, e.g. PBS or Condor.

4. DATA STORAGE

The output of the scientific applications is currently all stored as flat files. These files are often distributed over a number of machines and /or different types of media. Hence there are two data management problems associated with this scenario, firstly the data is organised physically rather than logically and secondly the data is all stored in different file formats.

In order to access the first concern, the Storage Resource Broker (SRB)[4] has been deployed on the nodes of the project. The SRB is a tool developed by San Diego Supercomputing (SDSC) which facilitates data management across a number of distributed heterogeneous file systems. Essentially the system abstracts the user from the physical location, media and protocols of the underlying storage systems. This allows data to be organised logically into a single virtual file system. In addition SRB simplifies the sharing of data across a distributed virtual organisation such as the eMinerals. In particular it is straightforward to configure the access permissions on individual files to allow other members of the project access. There is also a facility to allow colleagues outside the project access using a ticketing scheme.

The architecture of a SRB domain is shown schematically in Figure 4. Each data storage resource runs a SRB server. The SRB domain needs one master SRB server which is attached to a local database containing the MCAT (Metadata Catalogue) which maps from locations within the virtual file system to physical locations on individual resources.

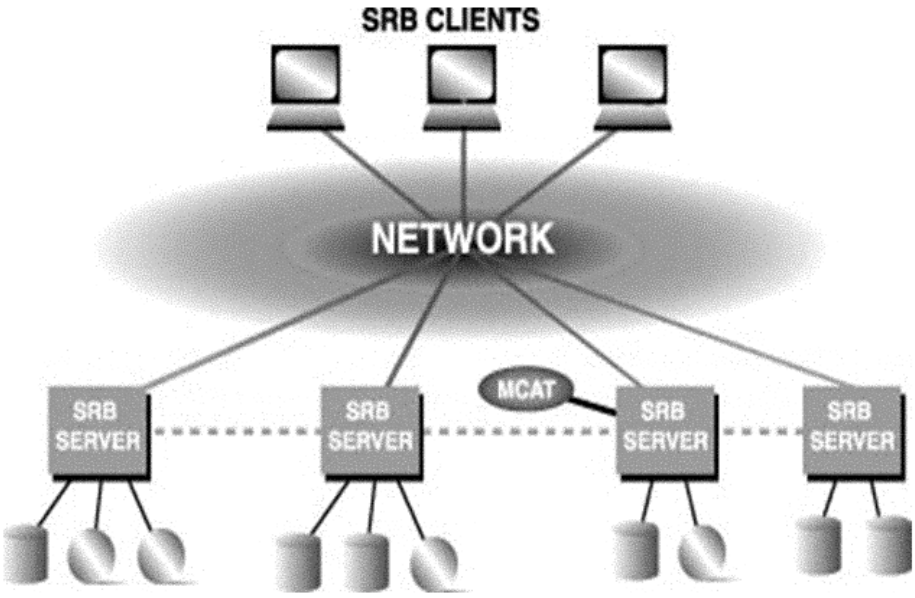

Figure 4 - SRB Architecture

There are a range of client tools which allow the scientists access to the data in SRB. These include a Windows client, InQ, a web interface, MySRB and Linux /Unix style shell commands known as the Scommands. This flexibility in choice of interfaces has proved very useful in encouraging the adoption and use of SRB by the scientists.

The second issue associated with data storage is the problem of data formats being used by the different scientific applications. Currently the project uses around a dozen different scientific codes, most of which use their own bespoke formats for input and output data. The manual conversion between these formats can be both error prone and time consuming. Moreover the use of multiple bespoke data formats complicates more automated data management strategies. To overcome these issues, the project has adopted the use of a XML based language known as CML (Chemical Markup Language)[5].

The project is in the process of converting the codes they use to handle input and output in CML. The approach where possible has been to persuade the code developers to modify the applications themselves in order to import and export CML. Unfortunately this is not possible for all applications used by the project, for example some codes used are commercial and there is no access to the source tree. In this case XSLT stylesheets have been written which convert from the CML to the native input. For the output a regular expression matching engine written by the CML developers and known as Jumbo Parser is used to mark up the native output into CML.

There is also active collaboration with the CML developers to ensure that the language is rich enough to describe the broad spectrum of data relating to environmental science research.

Although the actual data itself is still held in flat files, albeit now in CML, the adoption of a XML based file format would make it relatively easy to move towards storing the data in a database which in turn would open a number of possibilities including data mining.

5. DATA ANNOTATION AND DATA PUBLISHING

The final stages of the data lifecycle are annotation and publishing. The data files themselves are published by uploading them to the SRB. Although the use of the SRB facilitates sharing and replication of data, the value of the data files would diminish rapidly if they were not annotated with relevant metadata describing the context and method of their generation. In addition this step is essential if the data files are to be retrievable at a later date in the Data Discovery phase.

The annotation is accomplished by separating the metadata (which is held in a relational database based on the CCLRC Metadata Schema) from the data (which is held in flat files in CML). The metadata contains links to the SRB which internally maps to the physical file locations. This use of the SRB also maintains referential integrity if the files are moved.

Currently the metadata relating to scientific work is entered using another web portal known as the CCLRC Metadata Editor. Within the metadata editor, individual computation simulations are grouped into studies which are labelled by various topic hierarchies that allow retrieval via the Data Portal at a later date.

This final step of publishing the metadata completes the data lifecycle loop shown in Figure 1.

6. EXAMPLE USE CASE

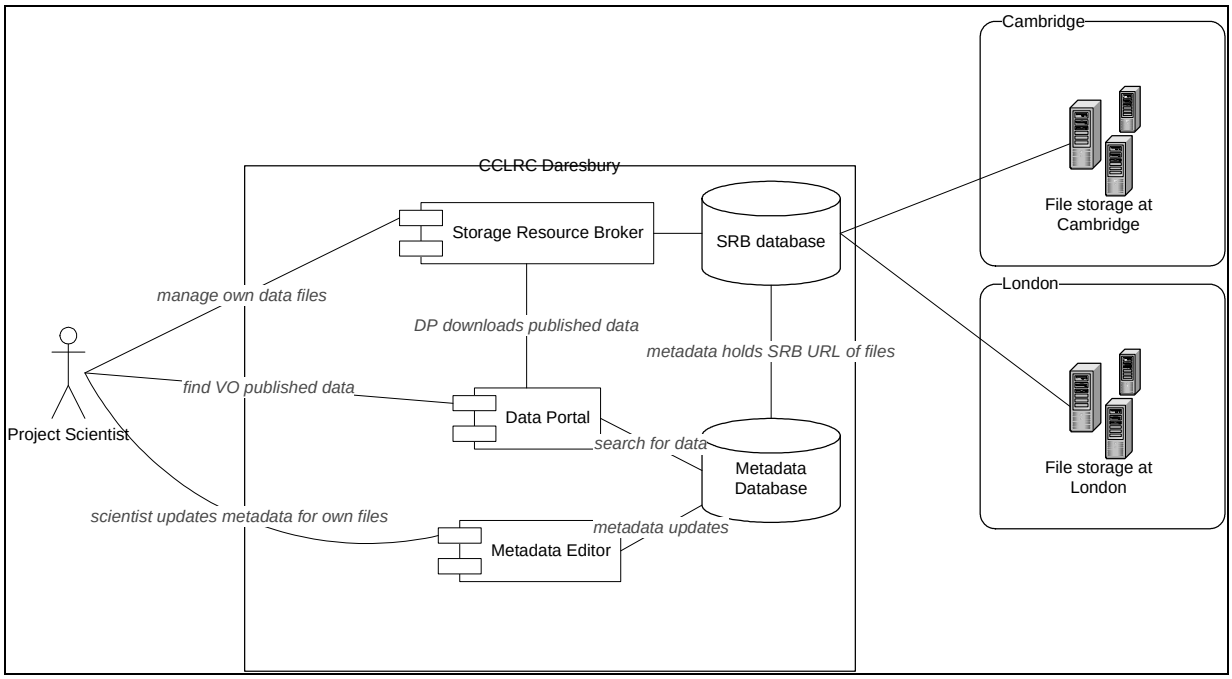

Figure 5 Distributed data management for the e-minerals virtual organisation (VO)

Figure 5 shows the current configuration that has been installed for the eMinerals project scientists. A typical scenario from the project scientist perspective would be:

1. Scientist generates some input files (possibly using data retrieved from a data holding using the Data Portal).

2. Using this data, a calculation is performed on a HPC resource which creates a set of output files.

3. Using the SCommands on the HPC facility, the output files are uploaded onto a file storage at Cambridge or London in their own home space within the SRB domain.

4. The output files can be viewed or downloaded using MySRB.

5. After the results have been analysed and checked the next step is share the data with the rest of the virtual organisation (VO). This is achieved by creating a new *study* using the metadata editor. This creates a record of the specific area of research, provenance information, etc. The study is associated with a particular topic or discipline e.g. sulphides.

6. Within this study, new *datasets* are created which link to the set of output files using the Metadata Editor. The dataset metadata includes the location of the data files as a SRB URL e.g. srb://eminerals.dl.ac.uk/home/joe.bloggs/... SRB then translates this location in the virtual file system to the physical files held in the SRB Vaults at, for example, Cambridge or London.

7. If another scientist wants to view the results, she would log on to the Data Portal and search using the drill down taxonomy e.g. Geology/Minerals/Sulphides. The user is then presented with a list of studies and would select the appropriate one. The user is then presented with the datasets belonging to that study. These data sets can then be selected and downloaded to the client machine. The Data Portal and SRB handle the transfer of files from the remote machine to the scientist's machine seamlessly.

7. SUMMARY

The project has deployed and integrated a number of pieces of middleware which were developed independently. In particular the entire cycle of a data centric view of scientific research has been addressed in terms of storing data, creating metadata and sharing information between members of the VO.

Future work will focus on further integration and simplification of the different tools to make their use more intuitive to the scientists. In the near future we aim to facilitate:

- more automatic generation of metadata where possible
- auto upload of files from the computation process to SRB
- searches across metadata catalogues from other facilities such as the ISIS facility at Rutherford Appleton Laboratory
- improvements to Data Portal, SRB, Metadata Editor resulting from project feedback

One of the strengths of the eMinerals projects is that it brings together computer scientists from the eScience programme and environmental scientists. This synergy will hopefully continue to drive both the science and eScience efforts within the project.

[1] e-minerals web site http://eminerals.org; http://www.e-science.clrc.ac.uk/web/projects/eminerals

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

[3] CCLRC Scientific Metadata Model http://www.e-science.clrc.ac.uk/documents/projects/dataportal/cclrcmetadatamodel.pdf

[4] Storage Resource Broker (San Diego Supercomputer Centre) http://www.npaci.edu/DICE/SRB/; http://www.e-science.clrc.ac.uk/web/projects/storage_resource_broker

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