

Environment from the Molecular Level e-Science project and its use of CCLRC's Web Services based Data Portal

L. Blanshard
e-Science Centre
Council for the
Central Laboratory of
the Research Councils
Daresbury Laboratory
Daresbury,
Warrington, UK

K. Kleese van Dam
e-Science Centre
Council for the
Central Laboratory of
the Research Councils
Daresbury Laboratory
Daresbury,
Warrington, UK

M. Dove
Department of Earth
Sciences,
University of
Cambridge
Downing Street
Cambridge, CB2 3EQ,
UK

Abstract - CCLRC's web services based multidisciplinary data portal, uses a metadata model of scientific data [6] to explore and access the content of data resources within CCLRC's main laboratories in the UK and other facilities in Europe. CCLRC are currently adapting the data portal for a number of projects, such as the Environment from the Molecular Level, so that earth scientists may store and access their own metadata and datasets, and simultaneously access related metadata and datasets from other facilities around the world. To achieve this the data portal has been redeveloped using Web Service technology so that the internal services could be accessed via any user interface or system specific to the e-science project community. Previously, access was provided only via a standard web browser.

Key Words - Web Services, Environmental Science, Portals, Data Management, Grid.

1. Introduction

The Council for the Central Laboratory of the Research Councils (CCLRC) <http://www.clrc.ac.uk/> is one of Europe's largest multidisciplinary research support organizations. From its three sites, CCLRC operates several large scale scientific facilities for the UK research and industrial community including accelerators, lasers, telescopes, satellites and supercomputers which all create copious quantities of data. Currently CCLRC is holding data in excess of 50 TB, deployed through a number of data centres including one World Data Centre, three National Data Centres and a range of facilities for particular communities or instruments. However, it is expected that much more data will be collected in the future with the advent of new instruments and facilities which will lead to data volumes in excess of several PB within the next 5-6 years. The data held at CCLRC

covers most major science areas e.g. Astronomy, Biology, Chemistry, Environmental Science, Physics.

CCLRC is directly involved in the e-minerals project *Environment from the Molecular Level* <http://eminerals.org/> funded by the UK's Natural Environment Research Council (NERC). Many environmental problems, such as transport of pollutants, development of remediation strategies, weathering, and containment of high-level radioactive waste, require an understanding of fundamental mechanisms and processes at a molecular level. Computer simulations at a molecular level can give considerable progress in our understanding of these processes. The aim is to link current developments in atomistic simulation tools with Grid technologies in order to facilitate simulation studies that can be performed with realistic conditions, and that scan a wide range of physical and chemical parameters. The project brings together simulation scientists, applications developers and computer scientists to develop UK Grid capabilities for molecular simulations of environmental issues. A common set of simulation tools will be developed for a wide range of applications, and the Grid environment will be established which will result in a giant leap in the capabilities of these powerful scientific tools.

CCLRC will be adapting their three portals [2] for use on the e-minerals project, HPCGridPortal <http://esc.dl.ac.uk/HPCPortal/>, DataPortal <http://esc.dl.ac.uk:9000/index.html>, and Advanced Visualization Portal. The HPC Portal [4] allows submission of jobs in a grid environment, resource discovery and file transfer for input and output of the jobs. The

Data Portal [3][9] provides a single point of access to search for related datasets across a number of institutions and scientific areas. These portals are currently being integrated to form a generic CCLRC portal architecture. CCLRC are also developing a database for the storage of metadata by the e-minerals project. The data will be stored and accessed by the Data Portal collection of web services or via the web interface. Once published the data may then be downloaded via the web to the scientific community at large. In order to store data it will be necessary to enter metadata such as details of the study, input parameters for the molecular simulations, data extracted from output files etc. along with links the datasets themselves.

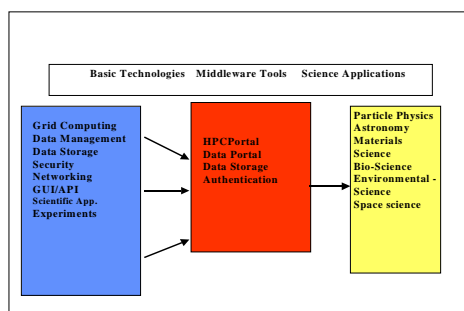


Figure 1: Building Blocks for an Integrated Solution

2. E-minerals Project Background

Many problems in environmental geochemistry cannot be solved without an understanding of processes at an atomistic level, and many of these problems are central to a complete earth systems science. We have reached the stage where we can approach such problems from large-scale atomistic simulations and computational quantum mechanics, using grid-enabled technologies. The e-science challenge is to scale up the length and time scales of the simulation techniques from the molecular level, through the mesoscopic scale, and towards the length and time scales of human experience. This will build upon terascale high-performance and distributed computing with new grid-specific scientific algorithms and middleware tools.

The e-science challenges fall into two main areas of activity, namely to be able to simulate surfaces and interfaces within realistic environmental contexts, and to

simulate bulk properties. Both types of simulations require both an empirical and quantum mechanical representation of the forces between atoms. There are challenges in developing new algorithms across the range of applications, particularly to work with large samples. In both types of work we are looking for increases in the sample sizes that span 2–3 orders of magnitude. A range of applications are used within the project. Examples of these are DL_POLY developed at CCLRC which is a large-scale molecular dynamics (MD) simulation package and SIESTA that uses first-principles Quantum Mechanics (QM) techniques based on density-functional theory (DFT) that scale linearly with system size (as opposed to the usual cube scaling).

The grid challenges also span the range of new initiatives, including development of computational and data grids.

The outcomes will be a genuine mixture of new science insights and new science possibilities for the application of molecular simulations to environmental problems. These will also include new grid environments, accessible through an integrated software system, for remote computation, data archival, data management and metadata creation.

2.1 Data Access & Download

The Data Portal and additional software will facilitate automatic recording of the data preparation, modeling and analysis cycles and enable the creation of metadata linked to the datasets produced. Results of runs with different parameters, temperature, pressure or other external condition, can then be later retrieved and used in post-processing analysis with, for instance, data mining techniques. This can provide the input for new runs if new areas of the parameter space must be explored. Results may also be combined in more sophisticated inter-disciplinary models.

2.2 Data Archiving

A new database will allow the archiving of massive simulation configurations with appropriate metadata definitions. The Data Portal will provide access to the data and that of partner researchers. A critical category of exploitation of the grid for large-scale simulations is data management. Both MD and QM calculations generate large volumes of data, which are amenable to further analysis or

visualization. Computational data may also be linked with experimental data in external mineralogical or geochemical databases for fluid-flow and thermodynamic geochemical modeling tools.

3. Data Portal Redevelopment using Web Services

The Data Portal (version 2) has recently been redeveloped to take advantage of web services technology (version 3).

Originally the portal consisted of two modules: the web interface that provided searching capability via a browser, by allowing the user to *drill down* to select a topic, and an XML Wrapper that translated the query into the local data querying language such as SQL. A separate XML wrapper existed for each metadata archive at a facility or institution. After translating the query, the XML wrapper would send the query to the database and wait for the resulting metadata to be returned. This was then converted into CCLRC's scientific metadata format in XML and transferred back to the web interface. The web interface would then display the results on the browser in various levels of detail according to user selection. The metadata contained links to the associated datasets that could then be added to a *shopping cart* i.e. the Data Portal's local database. The user could then return at a later date to view the metadata again and possibly download some or all of the associated datasets.

Some of the issues associated with this version of the data portal were:

- over-reliance on the web interface. Since the data portal functionality was to be used on a number of e-science projects and application areas it was considered likely that a more specific user interface would be required for each project or work area.
- difficulty in interchanging or removing parts of the functionality if they were not required or desired, or even if they were developed in other languages and platforms. For example, should the e-minerals project have their own method of authentication then it would be difficult to swap it with the Data Portal's own method without redevelopment
- inability to spread the load over a number of servers should traffic become high

The solution was to create a new modular design where each module represented an area of functionality and used web service technology. This meant that each module could run on separate servers (or all on one) as required. In addition services could be swapped in and out without redeveloping the code.

4. Data Portal Web Services used by E-minerals Project

The use of web services in the Data Portal is paramount when designing user interfaces or portals specific to a scientific area such as Environmental Science. It allows the most suitable functionality to be selected for a tailored solution. Some of the web services that we plan to use are as follows:

The Query & Reply web service allows a user to search one or more facilities for studies relating to a chosen topic such as surface dynamics. Other web services in the Data Portal handle the request and pass the query on to the chosen facilities. The results are collated and returned to the e-minerals interface in XML. The interface may be tailored to display the results as appropriate. Contained within the XML are links to the actual datasets which may be added to the user's *shopping cart*. The advanced query allows further search parameters to be specified such as a range of dates when the study was performed.

The Shopping Cart web service offers a number of methods: *getCart*, *removeFromCart*, *addToCart*, *addNote*, *getNote*. A user may select datasets or files from the results that he is interested in and add them to his own shopping cart using the *addToCart* method. The result of this operation is that the link to the datasets are kept in the Data Portal's local database along with an identifier for the user. At a later point he can view the contents of the cart using the *getCart* method or use *removeFromCart* to remove the link. To download the data the user or the calling application simply uses the URL to retrieve the data manually. Alternatively the link may be passed to the HPC Portal or another grid computing portal for input to an application running on the Grid.

Data Insertion web service will be used to insert metadata about a study into one of the metadata archives, along with links to the data held elsewhere.

The User Admin web service provides a number of services to add a user to a virtual study, remove a user and amend a user's details.

The HPC Portal web service provides a link to allow submission of jobs on the Grid. The submitJob method takes parameters representing the name of an application, the version, a list of URL's pointing to the input files, and a URL where the output files should go. It then executes the job on an available and suitable resource transferring files as necessary using the Data Transfer web service (not shown).

5. Locating Web Services

We have an additional web service called the Lookup Module that is used to discover the web services available within the data portal. Basically it accepts two parameters that describe the web service such as "XMLW" (XMLWrapper) and "BADC" to select the facility and returns a WSDL (Web Services Description Language) file. WSDL is an XML grammar to describe web services and includes items such as the name of the web service, the number and type of parameters it takes and the type of message used to communicate with the service. The Lookup service allows the data portal modules to be interchanged easily without changing the interface to the data portal itself i.e. the e-minerals application would still access the lookup service in the same way and hence the data portal.

6. Further Work

An important function yet to be provided is the ability to insert metadata and possibly data into the archive. This web service will either be accessed via a form on a browser (for metadata that must be manually entered) or from another application particularly where metadata is gathered automatically.

Further research will be carried out by CCLRC into linking a number of web services together and there are a number of technologies available such as BPEL (Business Process Execution Language for Web Services; also known as BPEL4WS). This is especially relevant to the e-minerals project as a typical scenario would consist of:

- authenticate the user's certificate
- start a session
- send a query and return the results

- add datasets (referenced in the results) to the shopping cart
- start a grid application using the dataset as input
- use the output as input to another job etc

As can be seen, this is a fairly typical scenario and any existing technology will be evaluated as to its suitability for the e-minerals projects as well as other e-science projects the CCLRC are involved in.

7. References

- [1] P.Baumann, P.Furtado, R.Ritsch, N.Widmann: Geo/Environmental and Medical Data
- [2] Allan RJ, Hanlon DJ, Kleese-van Dam K, Sufi SA, Sastry S, Boyd DRS. An Integrated Grid Services Portal for HPC Computations, Data Management & Advanced Visualization. Technical Project Description 2001
<http://www.dl.ac.uk/TCSC/UKHEC/GridWG>
- [3] Ashby JV, Bicarregui JC, Boyd DRS, Kleese K, Lambert SC, Matthews BM, O'Neill KD. A Multidisciplinary Scientific Data Portal. High Performance Computing and Networking 2001; 2110: p.13-22.
- [4] Allan RJ. HPCGridPortal and Related Work. HPCGrid Magazine 2001: issue 26
- [5] Allan RJ. Developing a Web Portal for Computational Grids. CCLRC e-Science Centre 2001
http://esc.dl.ac.uk/TechReports/portal_guide/portal_guide.pdf
- [6] Matthews BM, Sufi SA. The CCLRC Scientific Metadata Model - Version 1.
<http://www.dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf>
- [7] McDonald M. The Book of VB .NET: Web Services, Part 1. 2002
- [8] Allan RJ, Wang SD, Chohan D, McKeown M, Colgrave J, Dovey M. UDDI and WSIL for e-Science. UK Grid Support Centre 2002
<http://esc.dl.ac.uk/WebServices>
- [9] J.Ashby, J.Bicarregui, D. Boyd, K.Kleese, S.Lambert, B.Matthews, K.O'Neill: The CCLRC Data Portal, ERCIM News 45, European Research Consortium for Informatics and Mathematics, 2001