The *e*Minerals project: developing the concept of the virtual organisation to support collaborative work on molecular-scale environmental simulations

MT Dove[1,2], E Artacho[1], TO White[1], RP Bruin[1], MG Tucker[1,3], P Murray-Rust[4], RJ Allan[5], K Kleese van Dam[5], W Smith[5], RP Tyer[5], I Todorov[1,5], W Emmerich[6], C Chapman[6], SC Parker[7], A Marmier[7], V Alexandrov[8], GJ Lewis[8], SM Hasan[8], A Thandavan[8], K Wright[9], CRA Catlow[9], M Blanchard[9], NH de Leeuw[10], Z Du[10], GD Price[11], J Brodholt[11], M Alfredsson[12]

[1] *Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ*
[2] *National Institute for Environmental eScience, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA*
[3] *Present address: ISIS Facility, Rutherford Appleton Laboratory, Chilton, Didcot OX11 0QX*
[4] *Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW*
[5] *Daresbury Laboratory, Daresbury, Warrington, Cheshire WA4 4AD*
[6] *Department of Computer Science, University College London, Gower Street, London WC1E 6BT*
[7] *Department of Chemistry, University of Bath, Bath BA2 7AY*
[8] *Department of Computer Science, The University of Reading, Whiteknights, Reading RG6 6AY*
[9] *Davy Faraday Research Laboratory, Royal Institution, 21 Albemarle Street, London W1S 4BS*
[10] *School of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX*
[11] *Department of Earth Sciences, University College London, Gower Street, London WC1E 6BT*

Abstract

The *e*Minerals project has established an integrated compute and data minigrid infrastructure together with a set of collaborative tools,. The infrastructure is designed to support molecular simulation scientists working together as a virtual organisation aiming to understand a number of strategic processes in environmental science. The *e*Minerals virtual organisation is now working towards applying this infrastructure to tackle a new generation of scientific problems. This paper describes the achievements of the *e*Minerals virtual organisation to date, and describes ongoing applications of the virtual organisation infrastructure.

Introduction: The vision of the *e*Minerals project

The long-term vision of the *e*Minerals project [1] is to create a new paradigm for the way that project teams within the computational sciences will be able to carry out studies of increased complexity. In many disciplines, it is typical for studies to be carried out by individuals who manage their own computations, their own workflow, their access to resources, and their data files. As computational power increases, it becomes possible to run simulations of ever greater complexity and with a higher degree of realism. We have anticipated that this will lead to the point where large simulation studies will be performed by collaborations of scientists working as members of a virtual organisation (VO), rather than by individual scientists, in the same way that large-scale experiments

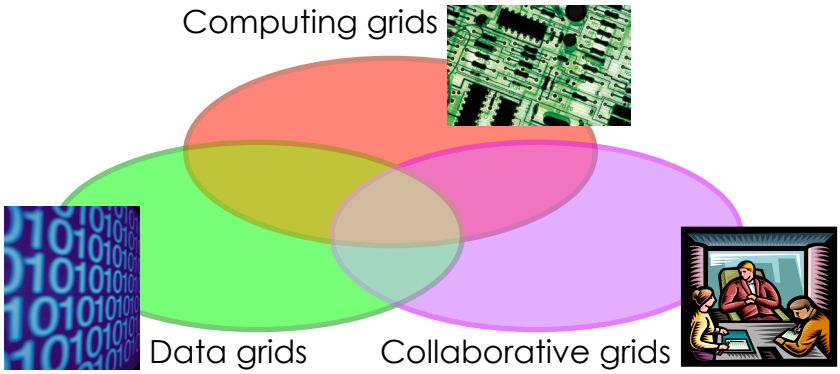

Figure 1: The grid infrastructure of the eMinerals project encompasses three components of grid computing: compute, data and collaborative grids. The compute and data components are met using the eMinerals integrated compute/data minigrid infrastructure, as described in the text. The collaborative grid infrastructure is met within the eMinerals project using the desktop Access Grid and a new application sharing tool developed within the project.

now require teams of scientists.

The challenge we set ourselves in the first stage of the *e*Minerals project was to establish an infrastructure to support collaborations between simulation scientists [2,3]. Our approach, shown schematically in Figure 1, has been to develop an infrastructure that encompasses three components: a compute grid, a data grid and a collaborative grid. This stage has mostly been completed, as described below and elsewhere [2–5].

The second stage of the *e*Minerals project is to expand the scale of what simulation scientists can achieve when working together as a VO, and progress to this end will be reviewed in this paper. The point is that a VO will include much more expertise and experience than can be found within a single research group, and the *e*Minerals project is attempting to exploit this fact in an attempt to tackle some large-scale environmental problems.

It should be remarked that there is a wide variation in how the concept of a VO is interpreted. The *e*Minerals VO is less dynamic, and somewhat smaller, than other VO's (such as VO's associated with particle physics experiments), and accordingly the tools we require are primarily concerned with collaborative working rather than tools for technical management of the VO (such as CAS and VOMS). The *e*Minerals model of a VO is one that will become increasingly common in many branches of computational science, and accordingly we are working towards the implementation of our vision in a manner that will enable the results to be easily transferred to other fields of computational science.

The *e*Minerals project

The *e*Minerals project has a focus on using molecular-scale simulations to help understand environmental issues [1]. The science issues include nuclear waste disposal, and adsorption of toxic atoms (such as arsenic) and molecules (such as dioxins) onto mineral surfaces [6,7]. The project uses a variety of molecular simulation methods, including large-scale molecular dynamics simulations using empirical representations of the forces between atoms, and lattice-energy relaxation methods based on a quantum mechanical description of the interatomic forces.

The use of a wide range of simulation methodologies means that no single computational resource is adequate for all simulations. The very largest simulations are firmly within the realm of high-

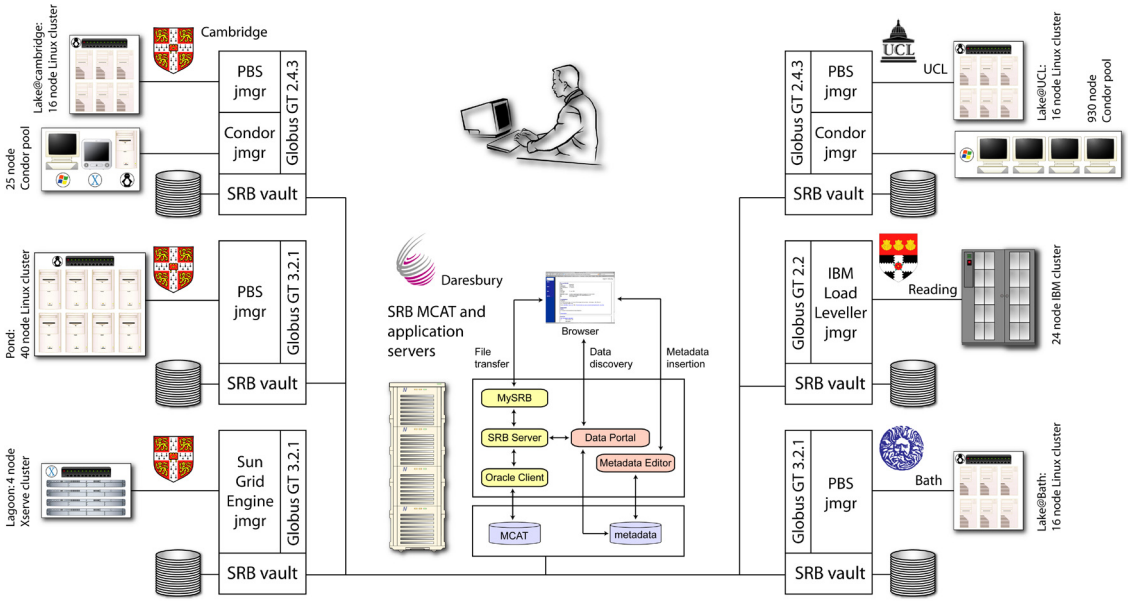

Figure 2: The eMinerals minigid, an integrated compute and data grid infrastructure designed to support a wide range of simulation requirements. The compute components include Condor pools, Linux and Apple Xserve clusters, and an IBM parallel computer. The data grid component is based on the Storage Resource Broker (SRB). The eMinerals minigrid provides an SRB data vault associated with each compute resource, with a central metadata catalogue (MCAT) server. The SRB is integrated with the eMinerals data portal to enable sharing of complete sets of data associated with individual studies.

performance computing, and in these cases the grid challenges are primarily concerned with data management (see below). However, many other simulations can be performed within a reasonable turn around time (several hours) on modern commodity computers, and in these cases the computational challenge is to provide a grid infrastructure that will support high-throughput operation. Different simulations will have different requirements for processor memory. Those simulations with low-memory requirements can often be met using Condor pools based on desktop computers, and those with higher-memory requirements can be performed on fit-for-purpose clusters.

The *e*Minerals project team [8] is distributed across a number of sites, and comprises a group of simulation scientists, simulation code developers, computer scientists and grid specialists. In order to develop a grid infrastructure and associated tools that are genuinely useful to the simulation scientists, it has been essential that the *e*Minerals team operates as a VO rather than as a collection of individuals.

Progress to date: the *e*Minerals minigrid and VO support tools

A first stage of the *e*Minerals project has been concerned with setting up the prototype grid infrastructure for the *e*Minerals VO. This infrastructure, called the "*e*Minerals minigrid" [2,4], is an integrated distributed compute and data grid environment, supported by the use of collaborative tools [3,5]. The *e*Minerals minigrid is shown schematically in Figure 2.

The resources are distributed between the Universities of Cambridge, Bath and

Reading, University College London, and CCLRC's Daresbury Laboratory. The minigrid is purposely designed using heterogeneous components, not least because the science applications have heterogeneous compute requirements. The hardware consists of PC's within Condor pools (two sites), Linux clusters (three sites), an Apple Xserve cluster and an IBM parallel computer. The full suite includes Linux/unix, Microsoft Windows and Apple OS X operating systems, and job managers include Condor, PBS, Sun Grid Engine and IBM's Load Leveller.

Access to the *e*Minerals minigrid is through Globus, using standard GSI authentication based on the use of X.509 digital certificates. Where possible, users of the *e*Minerals minigrid will obtain UK eScience certificates, but we have created our own certificate authority to generate certificates for collaborators who are unable to obtain UK escience certificates.

To solve a number of problems with managing data within the minigrid environment, we make use of the Storage Resource Broker (SRB), which was developed by the San Diego Supercomputer Centre [9]. In the eMinerals minigrid there is an SRB data vault at each compute resource, with the central metadata catalogue (MCAT) sever based at Daresbury. The whole SRB gives the project around 4 TB of distributed data storage. We have developed Condor-G scripts, based on Condor's DAGman workflow tools, to enable jobs to pull down data and script files from the SRB, and to place all output files back into the SRB on completion of jobs [2,4]. Users interact with their data directly via the SRB.

The SRB also plays a role in sharing data across the project. Collaborators can give each other permission to access files. This is much more straightforward to facilitate collaborative sharing of data than the traditional methods of placing data on an FTP or web server. The SRB gives access to individual files, which will be organised within directories (called "collections" in the SRB parlance).

To make management of data collections easier, the project uses the *e*Minerals metadata editor and DataPortal [10] to gather data together with appropriate metadata and then share these collections of data across the project. Sets of data files that are arranged in collections within the SRB are associated with specific scientific studies, and metadata is attached both to studies and individual files. The DataPortal then enables colleagues to browse collections based on keyword topic descriptions and metadata in the form of notes, and then download a complete set of files within a single study. In addition to enabling users to add and modify metadata through the metadata editor, we have developed a set of scriptable commands that access the metadata database [11], thus enabling metadata to be generated semi-automatically at the time of running jobs.

To enable data files to be shared in a useful form, we are making use of XML to provide data annotations. Specifically, we use the Chemical Markup Language (CML), for which we have helped develop a number of tools and language features specifically for our project applications [12]. Not only does CML enable data interoperability, it is also possible to perform platform-independent XLST (eXtensible Language StyleSheet Transformation) transformations of the CML-encoded data into other useful forms. For example, it is possible to transform the data to HTML for easy viewing, and also into SVG (Scalable Vector Graphics), an XML application for drawing, which enables us to graphically display data on the same web page as the

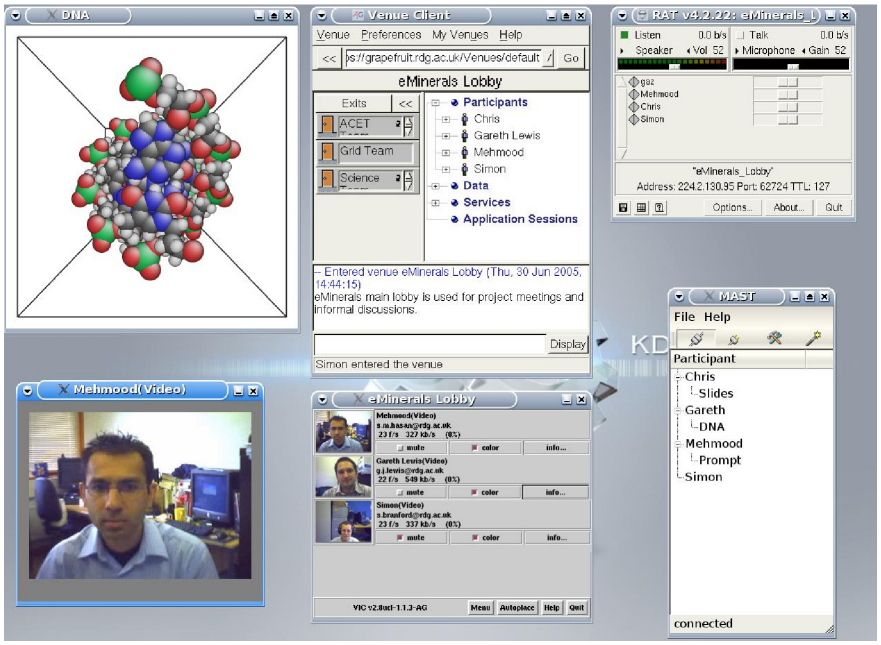

Figure 3: Screen shot of the MAST application sharing tool running within the Access Grid environment, showing a molecular visualisation program being shared by two users (one of whom can be seen in the Access Grid window).

main program output. Molecular-scale visualisation can also be implemented using CML, and embedded in the same web page using java applets.

We have implemented a range of VO support tools across the *e*Minerals project [3,5], including the use of wikis, instant messaging, the personal desktop version of the Access Grid, and the use of a new multicast application sharing tool (MAST) developed within the *e*Minerals project and shown in Figure 3 [13]. Part of the value of these tools has been in the training and support role. As new technologies have been implemented, it has been essential that the scientists have had regular access to the grid specialists. The VO support tools have been critical in enabling the grid team to provide a high level of support to the science team.

These developments have been used by the simulation scientists to carry out a wide range of scientific studies reported elsewhere [6,7,14–20]. The collaborative tools have been used to support many tasks of the *e*Minerals VO, including training, problem-solving, information updates, sharing of results, team working, testing and evaluation, document preparation, and for informal meetings.

Ongoing and future work: The *e*Minerals VO in action

The next stage in the *e*Minerals project is to move the VO concept a qualitative step forward, to enable the team to operate as a task-oriented VO that can pull together its personnel resources and shared infrastructure to work as a collaborating team on large-scale studies of environmental processes, rather than have the scientists work on smaller-scale individual projects within the VO.

To demonstrate our approach, we give four examples of collaboration work across the *e*Minerals project. First, one of our core applications is to use large-scale simulations to study the behaviour of materials that might be used to encapsulate high-level radioactive waste, such as plutonium [16,21]. The challenge is to develop materials that are durable when subject to the effects of alpha decay of the radioactive species. The major effect is the high-energy recoils of the radioactive atoms. These require extremely large simulation samples, and long running times. Part of the work of

the *e*Minerals code development team has been the development of a new version of the DL_POLY molecular dynamics simulation code for this problem. The scientists involved in this work have worked very closely with the DL_POLY code development team, with a closer and more responsive relationship than is common in this field.

Our second example is a collaboration between another group of scientists and the grid team in order to carry out simulations of pollutant organic molecules adsorbed on mineral surfaces [22]. Some important pollutant molecules are members of families that only differ in the number of chlorine molecules; examples are dioxins and PCB's. There are several calculations that are required for each molecule, and each calculation itself will take several hours on a modern commodity PC. Although these calculations can be packaged for high-performance computers, using grid computing methods makes this a much more manageable problem in terms of workflow and data management, as well as in high-turnaround of calculations. This work has been enabled through close collaboration between the scientists and grid team.

Some of the escience developments that are being driven in this new stage of the *e*Minerals project to support the *e*Minerals VO have involved the grid team working together closely using the same tools. Our third and fourth examples concern the development of the *e*Minerals compute portal [23], which is designed to enable all users have access to the *e*Minerals minigrid on their desktop, and the development of our multicast application sharing tool [13]. The work on the portal was conducted using regular desktop access grid sessions, following a conventional project-management approach. The work on the application sharing tool actually involved using the tool coupled with the desktop access grid to test and refine it.

These small scale examples, which will have parallels in many other escience projects, illustrate the value of scientists and/or grid specialists working together as part of a VO. Our objective in the second stage of our work is to enlarge this approach across all the teams on the *e*Minerals project. We have chosen as the main demonstration a scientific focus on transport of contaminants in the environment.

Dealing with contaminated land and water resources poses a major challenge to both the developed and the developing worlds. Strategies to remediate contaminated sites or to protect aquifers require robust models of water catchment, sediment transport, and of ground water flow. Such models need an underpinning understanding of adsorption, desorption and transport processes at the molecular level; a significant part of our science programme is now concerned with developing this understanding together with the creation of a well-constrained thermodynamic database for mesoscale modelling in order to rectify the major shortcomings of current data. We are using our integrated grid environment to simultaneously explore a wide range of contaminants and minerals in the environment.

eScience offers two significant advantages for this study. First, the grid infrastructure is required to support the large number of complex calculations that are involved, including support for workflow, handling high-throughput computations, and matching different types of calculations against most appropriate resources. The data management tools of the grid infrastructure are needed to support the vast range of data being produced in this study. Second, such a

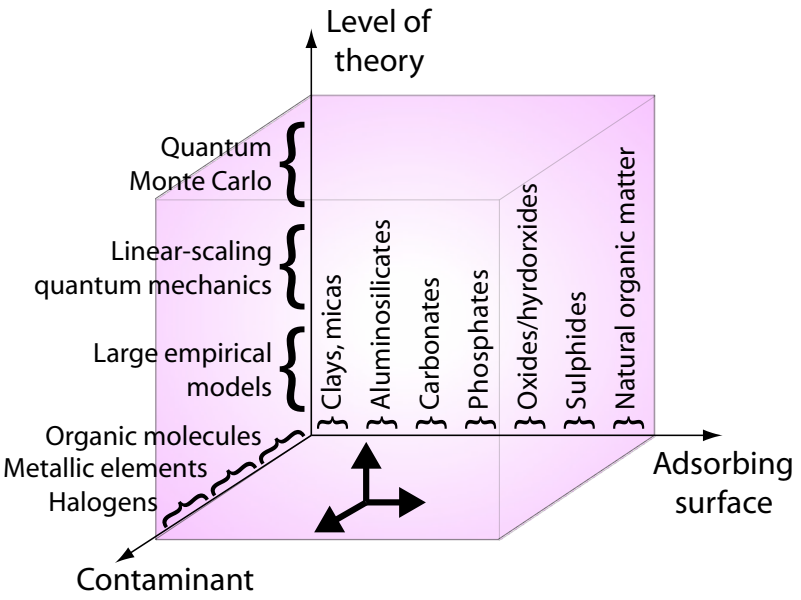

Figure 4: Representation of the range of areas of expertise within the eMinerals virtual organisation. This three-dimensional workspace spans a range of contaminants, mineral surfaces, and simulation methodologies. The range of experience/expertise represented in this workspace far exceeds what will be available within a single research group.

wide study requires expertise drawn from several personnel with complementary skills and experience. This is the role of the VO, and this study is making use of the collaborative tools and compute/data infrastructure developed in the first stage of our work to support the essential close collaboration between scientists working at remote sites.

The key questions we are addressing are: a) what are the mechanisms involved in adsorption and desorption of contaminants onto mineral surfaces and natural organic matter in soils and rocks, and what is the relative importance of different mechanisms; b) how do contaminants partition between water, mineral surfaces and natural organic molecules; c) how will these processes affect the transport of the contaminants in the environment? Molecular simulations of contaminants interacting with water, mineral surfaces and organic molecules (representative examples with carboxylic acid and other functional groups) are being exploited to calculate the energetics of adsorption processes, leading to the calculations of energy barriers for transport, diffusion rates, desorption rates, and partition coefficients. The end results will be extensive data sets that can be used for mesoscale modelling of the transport of contaminants in environments such as sediments and porous rocks.

Up to now, scientific investigations of these issues, both experimental and theoretical, have focused on individual contaminants and individual potential atomic environments. eScience has now created the opportunity to compare different contaminants and systems in a single integrated study. The representation in Figure 4, which is affectionately called "the cube", shows the range of pollutants being studied, including metal cations, anions, and chlorinated organic molecules such as PCBs, DDTs and dioxins, and the range of mineral surfaces, including hydrated and anhydrous surfaces of clays, aluminosilicates, carbonates, phosphates, oxides/hydroxides and sulphides. The scope of this study encompasses applications to the effects of mine tippings, toxic industrial waste, agricultural pollution from pesticide run-offs, long-term storage of radioactive waste, and large-scale natural occurrences of As and F in regional groundwater systems.

This study requires a range of molecular simulation methods, which give the third

(vertical) axis of the cube in Figure 4. There are two important considerations, namely the level of accuracy and sample size, and there is always a trade-off between these since higher accuracy and larger samples both require increased computer power. Highest accuracy can be achieved using quantum mechanical methods, but the largest samples can only be achieved using empirical models for the forces between atoms

The key point is that collectively members of the *e*Minerals VO have experience and expertise in a wide range of simulation methods, types of pollutants, and mineral surfaces, as represented in the three-dimensional workspace shown in Figure 4. Pulling all this experience together means that collectively the *e*Minerals VO has detailed knowledge and expertise matching that which would be found in a conventional organisation working in a single location. The challenge is to exploit this potential, using the armoury of the escience tools that comprise the *e*Minerals minigrid and VO; this is the territory of the next stage of the *e*Minerals project.

Acknowledgement
We are grateful to NERC for financial support

References

1. Dove MT & de Leeuw NH, *Mol Sim* **31**, 297, 2005
2. Calleja M et al, *Mol Sim* **31**, 303, 2005
3. Dove MT et al, *Mol Sim* **31**, 329, 2005
4. Calleja M et al, *All Hands 2004*, p 812
5. Dove MT et al, *All Hands 2004*, p 127
6. Wells SA et al, *All Hands 2004*, p 240
7. Alfredsson M et al, *All Hands 2005*
8. http://www.eminerals.org
9. Moore RW & Baru C, in *Grid Computing: making the global infrastructure a reality*, ch 11, 2003
10. Blanshard LJ et al, *All Hands 2004*, p637
11. Dove MT et al, *All Hands 2005*
12. Wakelin J et al, *Mol Sim* **31**, 315, 2005
13. Hasan SM et al, *All Hands 2005*
14. Du Z et al, *Mol Sim* **31**, 339, 2005
15. Puneda JM et al, *Mol Sim* **31**, 349, 2005
16. Trachenko K et al, *Mol Sim* **31**, 355, 2005
17. Fernández-Serra et al, *Mol Sim* **31**, 361, 2005
18. Alfredsson et al, *Mol Sim* **31**, 367, 2005
19. Wells S et al, *Mol Sim* **31**, 379, 2005
20. Marmier A et al., *Mol Sim* **31**, 385, 2005
21. Trachenko K et al, *Phys Rev B* **70**, art no 134112, 2004
22. White TO et al, *All Hands 2005*
23. Tyer RP et al, *All Hands 2004*, p 660