Grid/Web Enhancements to the National Crystallographic Service: Experiences with an interactive *e*-science demonstrator

M. B. Hursthouse, J. G. Frey, S. J. Coles, M. E. Light
Department of Chemistry, University of Southampton, Southampton, Hampshire, UK

M. Surridge, K. E. Meacham, D. J. Marvin
I.T. Innovation, University of Southampton, Southampton, Hampshire, UK

D. C. De Roure, H. R. Mills
Electronics and Computer Science, University of Southampton, Southampton, Hampshire, UK

Abstract

We report on a Demonstrator Project on the use of Web/Grid Services to enhance the user participation in the National Crystallographic Service (NCS). The coupling of a Web/Grid based set of services to an existing physical service (the EPRSC funded NCS) raised issues connected with the types of multimedia interaction perceived as necessary for the more efficient utilization of both the expensive NCS equipment and personnel in a secure manner. The techniques required to provide a demo service through existing network and security infrastructure are discussed. As well as audio, video and data links between the user, x-ray expert and the equipment, data staging was provided for subsequent added value calculation services. The experience learnt from the demo system led to significant re-design for a real service.

1. INTRODUCTION

Chemical crystallography is the determination of solid state structure by an X-ray diffraction experiment on a single crystal. A crystal oriented in a particular diffracting position, by a 4-axis robotic instrument known as a diffractometer, will reflect X-rays from planes within the crystal. After determination of the position and intensity of the reflected ray, recombination of thousands of 'reflection data' allow one to derive the molecular structure. In addition the manner in which molecules pack and interact with each other may be unequivocally characterised.

The massive increase in computing power in the last decade is one of the primary causes for an exponential increase in the number of entries in the Crystal Structure Database (CSD)[1]. Moreover, this leap in technology has enabled considerably decreased data collection times on increasingly challenging crystals. The Engineering and Physical Sciences Research Council (EPSRC) funds a UK National Crystallography Service (NCS), housed in the Department of Chemistry at the University of Southampton, to which any academic eligible to apply to the Chemistry Panel may subscribe. In this respect the EPSRC funds 'state of the art' instrumentation and expert personnel to run a service for those whose facilities are lacking or inferior.

The users of such a service therefore range from expert crystallographers to those with no experience at all. Hence, in many cases, it is desirable for the service to be able to interact with a user, who might have more knowledge about the sample under investigation than the operator of the instrumentation who provides the expert experience, knowledge and facilities in X-ray diffraction methodology. Moreover, an 'expert' user may control the experiment entirely on his or her own, enabling the local expert to devote time to other activities. To this end the NCS are currently developing robotic sample changing and automated data collection such that human interaction in the laboratory may be minimised if the crystal or user permits.

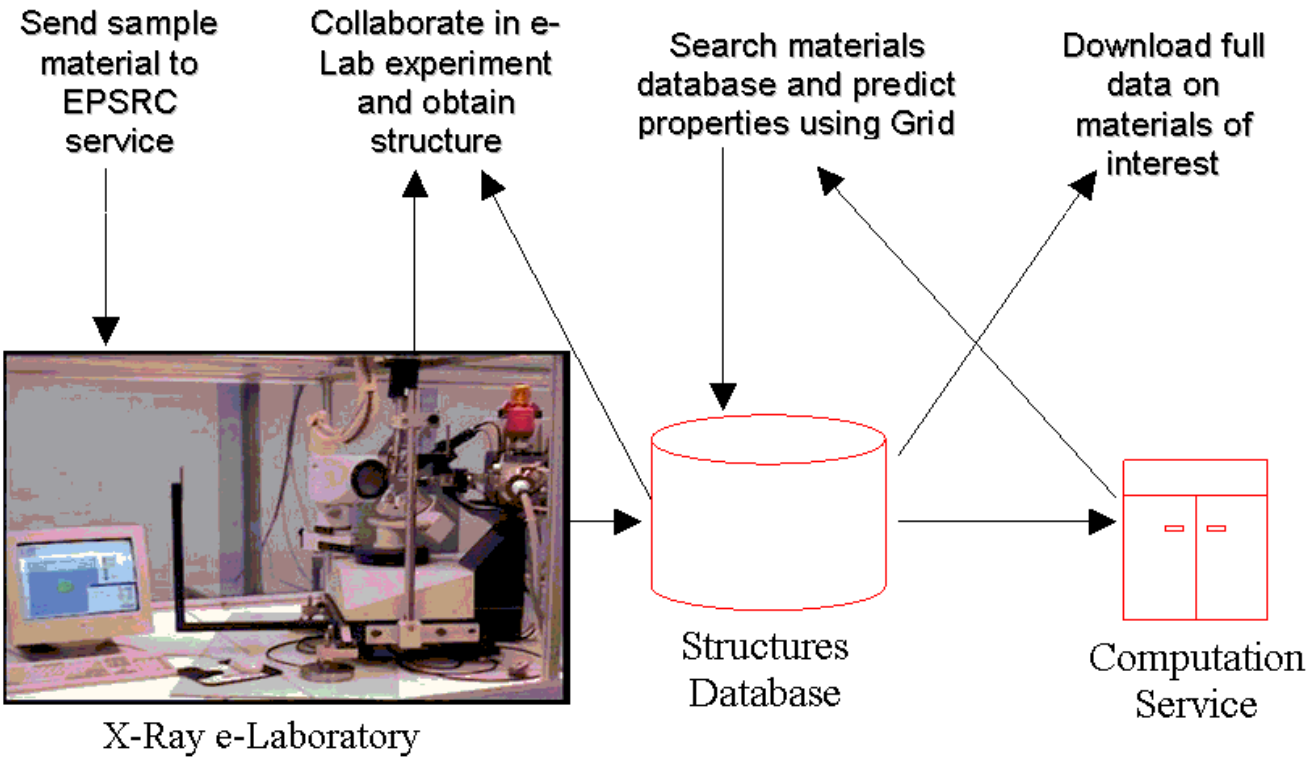
Current practice dictates that the structural result is stored on a local database until such time that the

information may be disseminated to the academic community by means of publishing in a peer reviewed journal at which point it is made available to the public through the CSD. As a direct result of the lengthy and involved publication process it is widely considered that only ca 10% of all crystal structures determined are actually deposited in the CSD. A crystal structure is deposited in a database in the form of a Crystallographic Information File (CIF)[2], which is an international standard means of information interchange in crystallography. A CIF is based on the Self-Defining Text Archive and Retrieval (STAR) procedure which is a general, flexible and easily extensible free format file that may be edited by a simple text editor. The file consists of data names constructed hierarchically so as to form data categories, data items and a loop facility for repeated items. The CIF contains pertinent information about data collection, structure solution, model refinement and accuracy and details of the structure, -such as atomic coordinates, thermal motion tensors, bond lengths and bond angles.

There is a growing area of research that employs database mining to study families of structures, searching for similarities and trends that address questions regarding formation and properties of these materials. In addition there is the opportunity for the computation of properties related to structure by quantum mechanics (QM), molecular mechanics (MM) and geometric relationship calculations. Thus there is a serious shortfall in this work in that the CSD is greatly deficient in the structural data that is potentially available. An explosion of information is likely to arise such that any one database cannot contain all the amassed information and it is therefore essential to explore the use of distributed databasing.

It is intended that GRID infrastructure is exploited in this demonstrator to:

1. increase interaction of local experts and users
2. assist in ‘dark’ laboratory instrument automation
3. increase the output of crystal structures through automated software routines
4. allow rapid dissemination of results to the community and vastly increased numbers of crystal structures to be deposited in databases
5. enable interactive ‘search, calculate and rank’ studies to be performed on subsets of databases
6. facilitate the growth of information in databases by ‘feedback’ from the calculations performed



2. RELATIONSHIP WITH THE WEB

The Web enables documents published in one location to be made available to the wider community regardless of geographical boundary. This is also what we require of the service: a central service which is made available without boundary. This is significant in bringing the appropriate expertise to bear in an effective manner: the "user" knows most about the sample and the technicians at the NCS often know much more about the x-ray techniques and analysis. For expert users the NCS staff could devote time to other activities and let the users get on with the experiment.

The Web increasingly supports *automation*, i.e. machine processing of information rather than delivery of documents to humans. This is key to the e-Science vision; for example, it enables the scale to increase without the bottleneck of human intervention. The automation theme is also strong in our particular application context, from the robot arm through to the delivery of results.

Hence on the surface the marriage of Grid and Web is a good one. It was tested through the construction of the demonstrator and conducting a series of demonstrations at a number of sites.

3. EXPERIENCES

Security Issues: Controlled access to data and equipment

1. Need to smoothly interact with the current systems employed by the NCS
2. Only the originator of the sample (and the NCS staff) should have access to the data & results, thus a highly dynamic authentication and authorisation system needed down to the file level
3. Commercially sensitive samples (and thus data) are part of the life of the service
4. As an EPSRC service the operation must be fully accountable.
5. Data storage - can old data be retrieved?
6. What data should be kept available in the future

Implementation Issues: GLOBUS vs. Web services approach. The development of the NCS collaborative system has been in two phases. The first phase (Examining Crystal Structures using E-Science - ECSES) was a demonstrator project funded by the DTI. The second phase is part of an EPRSC funded e-Science project called CombeChem.

One of the primary lessons learned in developing the ECSES demonstrator is that firewalls are awkward and unpleasant things to deal with. The experiment collaboration part of the demonstrator consists of a number of services sending data back and forth across several firewalls. It rapidly becomes unfeasible to reconfigure all of the necessary firewalls to allow the ECSES applications to work with arbitrary clients. The use of a DMZ area outside the main host firewall, containing a proxy service machine simplifies one side of the problem -- ECSES uses a Globus interface to communicate between the remote user and the host institution, and Web Services to connect the Globus server in the DMZ to the internal services. This allows access to the vulnerable internal services to be much more tightly controlled. The Comb-*e*-chem project will use a combination of a dynamic web portal and Web Services on the proxy machine, again communicating with the internal services through a tightly-controlled local firewall. The use of Web Services for the external connections provides a much more easily securable first point of access to the service than does the original Globus system of the ECSES demo.

A major driver for this dynamic authorisation is the nature of the interaction required between the crystallographic user community on the web and the operators and experts within the NCS. Access to parts of the system, both computational resources and physical equipment, depends on the ownership of the sample currently being investigated. The turnover of experiment and analysis can be as fast as a 15 min cycle. For future development of the Grid service this fast turnover has significant implications for scheduling and notification services.

A second lesson learned in ECSES was that although the video-conferencing was an interesting plaything, there was actually little advantage in being able to see the experimenter at the other end of the link. At most, only an audio link was necessary. Of much greater interest to the users of the service was the ability to see the results of the experiment as they became available. This allows the user to collaborate more fluently with the laboratory operator, and potentially to change the flow of the experiment in a timely manner should something unexpected happen. Some of the useful interim results can be easily extracted from output files and sent to the client, but

some important results can effectively only be viewed on-screen in the laboratory. For the pilot NCS service, we have found that use of the VNC (Virtual Network Computing)[3] package is highly beneficial, in that it allows the status of the experiment to be monitored across the network. The NCS web service will use a "gridded" version of VNC which tunnels the VNC protocols through Web Services, allowing the remote user to see selected areas of the experimenter's desktop, and hence to observe otherwise opaque interim results.

The crystallographic community has pre-existing protocols and databases with which the demo needed to be compatible. These well developed protocols mean that the transition to XML (from CIF) will be relatively straight forward and thus extend the applicability of the Web Service approach. However the corresponding transition to a distributed relational database for the structural information, required to make the best use of the data staging and computational services, is a more complex problem; political as well as technical issues abound.

A clear lesson both from the implementation of the demo and discussions, particularly with potential industrial and commercially aware academic users where security is paramount, is that the Web Services approach is essential. This places our development pathway as converging with OGSA while maintaining our ability to "sell" the NCS service in the near future.

REFERENCES

[1] F.H. Allen, O. Kennard & R. Taylor, 1983, Acc. Chem. Res., 16, 146-153.
[2] S.R. Hall, F.H. Allen & I.D. Brown, 1991, Acta Cryst., A47, 655-685.
[3] T. Richardson, Q. Stafford-Fraser, K.R. Wood, & A. Hopper, Virtual Network Computing, 1998, IEEE Internet Computing, 2, 33-38.