Organic Process Research & Development **2004**, *8,* 1024–1035

Dark Lab or Smart Lab: The Challenges for 21st Century Laboratory Software

Jeremy G. Frey*

School of Chemistry, University of Southampton, Southampton SO17 1BJ, UK

Abstract:

A personal view of some of the challenges for modern laboratory software is described together with some of the techniques that can be used to solve them. New techniques to investigate and to understand the requirements in different laboratory settings are explored, as well as the use of semantic knowledge technologies coupled to pervasive and grid systems to create the necessary support for collaborative chemical laboratory investigations.

Introduction

Chemistry and the Chemical Sciences have always made extensive use of the developing computing and information technologies and have been avid consumers of available computing power. The uses of this technology include activities such as modelling, simulation, and chemical structure interpretation. New procedures in chemical synthesis and characterisation, particularly in the arena of parallel and combinatorial methodologies, have generated ever-increasing demands on both computational chemistry and computer technology. Currently, the way in which networked services are being conceived to assist collaborative research pushes well beyond the traditional computational chemistry programmes, towards the basic issue of handling chemical information and knowledge (see Figure 1). The rate at which new chemical data can now be generated using increased automation such as combinatorial and parallel synthesis, combined with high throughput screening processes, means that the data can only realistically be handled efficiently by increased automation of the data collection and analysis. Nevertheless, automation is not the answer to all of chemistry, and the need to integrate people and equipment is paramount in most laboratory situations.

The Paperless Office or Paperless Laboratory Syndrome

When we consider the application of software that supposedly will assist or replace tasks currently undertaken by hand, then comparisons will always be made with the claim that word processors, databases and file servers, etc. would by now have yielded a paperless office. They may do many of things and are indispensable to a modern office, but there is certainly more paper consumed now than ever before. Nevertheless, the promise is not quite dead as there are organisations that have shunned paper to a large degree, showing that if the right tools are available then this can be done. However, the right tools for the job are not usually available, or the cost of tailoring them to the situation is too high.

* To whom correspondence should be addressed. E-mail: j.g.frey@soton.ac.uk.

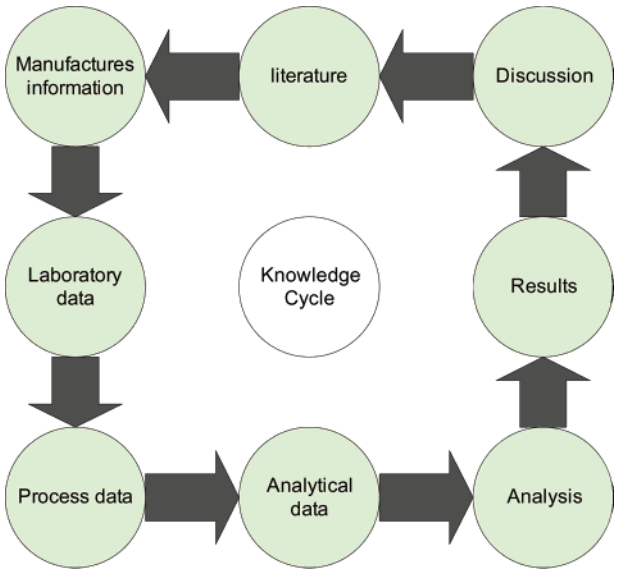

Figure 1. **Flow diagram illustrates some aspects of the chemical knowledge cycle with emphasis on the knowledge discovery or creation processes revolving around the laboratory. To ensure proper provenance of the information the ability to link back through all these steps is essential. This can pose problems, even when all the sources of the information are internal to the organization, but even more difficulties when organizational boundaries, such as to suppliers, are crossed. The analysis phase frequently requires integration from several data sources, including many from the laboratory environment and others that may be external such as databases of physical and structural properties. A more general discussion of the importance the flow of data from sources to the literature and back again and its facilitation by *e*-Science is given in the *e*-Bank project.**

Just over 10 years ago in the *Newsletter of the European Photochemistry Association*, this was discussed by the association chairman, with specific regard to chemistry. (I am grateful for Tony Rest in bringing this article to my attention.) In his thoughts he excluded large-scale computing and considered desktop systems, with the note that of course increase in computer power would mean that large-scale calculations such as molecular modelling, quantum calculations, and processing of computationally intensive multidimensional NMR spectra would, no doubt in due time, move from mainframes to desktop systems. This prediction has certainly come true, and almost all analytical tools require only a desktop or workstation.

For my purposes, the part of his discussion pertinent to working inside a laboratory (as opposed even to working at a desk outside after the actual experiment) was his consideration of the time required to sketch (and print if not actually drawn on paper) a relatively simple organic molecule (6 s

Published on Web 11/05/2004
10.1021/op049895g CCC: $27.50

for freehand, 30 s using a template, and 150 s using standard software) and asking if the increase in time corresponded to an increase in information content, even if the output was clearly much nicer looking.

An overall conclusion was that items such as word processors may be important items in the office-paperwork productivity race but do not often add significantly to an increase in scientific productivity. Partly this is because once shown that we can do the office work, we are required to do the office work. Work originally done by others is transferred to the researcher (a situation all too common in universities and I suspect more generally, if to a lesser extent). A similar situation arose with the production of camera-ready copy which transfers expense from the publishers to the authors.

This discussion highlights a number of significant issues that may not have been clear to software designers. Chemists, especially synthetic organic chemists, love to sketch. It also shows that we were then, and still to a great extent are, stuck between two worlds; the computer was seen as a way to make the drawing on paper rather than being able to use the image on the screen as such. I believe that the former is not something that is going to change, visualisation is so key to chemistry; thus, as with arm-waving, sketching is here to stay. Our interaction with screen and paper and the reliance on the linear form of a paper, book, or report are not so entrenched in chemistry, at least no more than in any other discipline. The hypertext approach possible in a well-designed web environment is taking ground. In the future grid world, with computational services linked into this arena so that finding the programme and resources to run a necessary calculation is automatic, the approach has a very good chance of becoming prevalent. We already know that to many casual users if it comes up on Google it must be right. Books are becoming less important, and journal styles are changing as well in response to the same pressures.

Computers in Chemistry

The use of computers to enable noncomputational aspects of chemistry, for example to assist synthesis, has a surprisingly long history. In the mid 1960s Corey linked his principles of strategic disconnections (1989 book but taught earlier to his students) to software (*Organic Chemical Simulation of Synthetic Analysis*) on computers that had resources significantly less than a modern calculator.[1] A notable feature was the use of a stylus to "draw" a chemical structure (2D) on an electrostatic tablet; the era of sketch input had, in principle, arrived before a mouse had been sighted. This, I believe, shows the importance of the input from practicing chemists to the design of computer interfaces and resulted in something that has only recently resurfaced with the tablet PC.

This computer-based research would perhaps now be considered part of *e*-Science; certainly this is true in Corey's grand vision of a global database of all chemical information to bring to bear on a synthetic problem. An unanswered question was where this information would come from. The huge increase in the rate of production of chemical knowledge in the period since Corey's first publications points to the need for much more automated systems to capture the knowledge. In my view, the data must be converted to a digital form as early as possible, as close as possible to its inception, and capture as much as possible of the physical and chemical circumstances and conditions that led to its creation; only in this way can the provenance of the data be provided for future interrogation. Those with an interest in using chemical information are led right back to the laboratory where data is created. If there is anywhere that can be said to be the origin of chemical data, it must be the laboratory (although the virtual versions in computational chemistry are providing a vast collection of new types of input which in the end need to be tested in a laboratory). The wide range of systems that a researcher can expect to interact with in a laboratory-based experiment is illustrated in Figure 2.

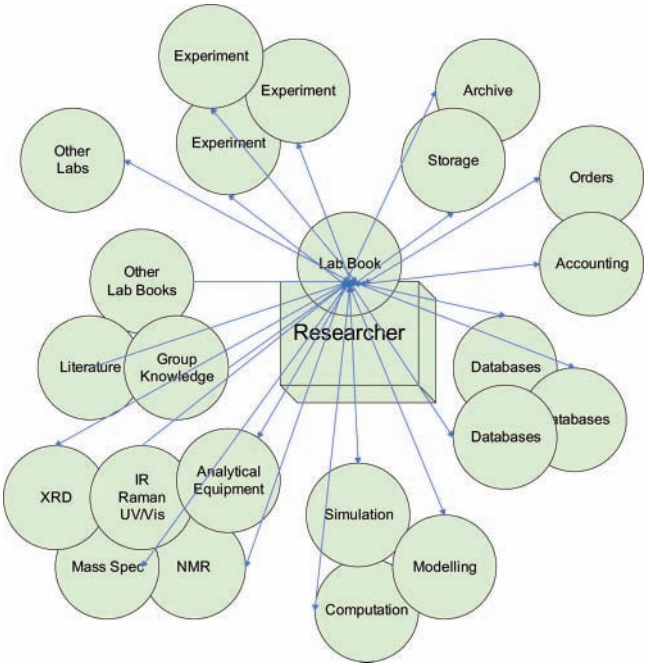

Figure 2. **On optimistic view of the range of experimental, informational, and computational services typically now available to the chemistry researcher. Currently most of these services require a different and unique interface which needs to be learnt by the researcher. Integration of data from these data sources is usually required to answer the chemical problem under investigation. The integration has to take place unaided by software and centered on the researcher lab notebook and computer system.**

The Laboratory Context

Context is a very significant concept for moulding software to users' changing requirements and provides much of the basis for the use of inference from known information to adapt to the users current needs. The context can have a wide degree of granularity, for example planning an experiment vs execution in the synthetic laboratory, or the subsequent analysis or specificity, such as working with a specific piece of equipment to make a particular measurement. I should also explain my context and some of the

(1) Above and Beyond Organic Synthesis. *Chem. Eng. News* **2004**, 37−41.

reasons why we have an interest in software for a 21st century chemical laboratory.

The UK academic and industrial communities have invested significantly over the past few years in an area coined "*e*-Science". As ever, definitions vary, but at the heart of the concept is the facilitation of collaborative research endeavours and the provision of the necessary computing infrastructure, software, and hardware to enable the collaborations. The infrastructure is sometimes referred to, at least in the United Kingdom as "the Grid"). The United Kingdom has taken quite a data-centric view of the Grid. In parallel with the necessary computational facilities and network systems the UK *e*-Science programme has focused on understanding the functions needed by a software middleware layer to provide this collaborative environment in a manner transparent to the users. To understand the requirements, several pilot projects have been set up across the whole range of research activities in the United Kingdom (from science, engineering, social science, medicine, and the arts). Several projects involve chemistry in one form or another, and one, Comb*e*Chem in particular, addresses some of the areas involved in laboratory chemistry.

The Comb*e*Chem Project

The Comb*e*chem project[2,3] is concerned with several aspects of the chemical knowledge life cycle and in particular how to handle the rapidly increasing production of knowledge associated with combinatorial synthetic techniques and high throughput analysis and still be in a position to make use of this knowledge. As part of the project we have explored extending the UK EPSRC (Engineering and Physical Sciences Research Council) National Crystallography Service (NCS) on the grid to enable much greater interaction between the remote users and the NCS staff. We are producing workflow systems to integrate such data sources with structural databases and correlating the data within these databases and with data from and with property measurements and calculations with the aim of improving the chemical and physical models.

In these considerations it became clear that the provenance of the data is of critical importance, especially when the data is being produced at a high rate and the traditional dissemination pathways are creaking and breaking under the strain. The overall effect is that much of the data is either being lost or is tantalizingly available but without sufficient information about its creation so that, when viewed several years later, the only thing to be done is to repeat the experiments. This is not an efficient use of resources. This is not to say experiments should not be repeated! Reproducibility is, of course, crucial to science, but repetition due to inadequate record keeping of material that could have been foreseen is not sensible.

Dark Laboratory Automation

In a fully automated, dark laboratory, the orchestration of all the components is a difficult but relatively clear objective. All the interfaces have to be made to "talk to each other", but it is essentially a machine-to-machine interaction and can thus, with care, be specified and at least, in principle, be recorded to give a clear provenance to the data. The latter is crucial to the subsequent analysis of the information but should not be seen even here as a simple step. Analysts among you will, I am sure, immediately be thinking of the whole chain involved here: It may be easy to record that the instrument was instructed to deliver a particular volume of solution, but how do we know that the solution was the correct one? That knowledge depends on the previous steps in the automated workflow. Furthermore, understanding the measurement accuracy depends on knowledge of the machine's performance and requires access to recent calibration information.

The Web of Information

The rapidly expanding information tree, perhaps better described as a web, goes back to manufacturers, suppliers, etc. Clearly, this is a problem without a solution, as there seems to be no end to this process. A practical solution is to provide as much information as is reasonable, and the concept of linking back to the information sources of the previous steps at least means that organizations are responsible for taking care of data that they understand. Even in this context there are many limitations when an administrative boundary is crossed, such as the boundary between a laboratory and the supplier. For example, it could be very useful to be able to use the bar code batch identifier on materials supplied to my laboratory to link directly to the manufacturer so that any differences in results can be correlated with subtle differences in the starting materials as well as differences in the subsequent processing. Even in a world without viruses and hackers, it is hard to imagine that such linking would automatically be available. We would also require some degree of persistence in this link so that users of the information in the future can fully assess the data if necessary.

Some methods to ensure that this type of interaction can be enabled simply and reliably will be discussed later in the paper. A path, though not a simple one, could be seen to construct an automated flow from the laboratory; creation, storage and curation, analysis and correlation of the data could at least be envisioned, and we have been working on systems to facilitate this. However, it became clear that the true test of this philosophy of ensuring that as much of the metadata as possible is recorded as it is created (which underlies what we refer to as Publication@Source[4]) is in a laboratory where people and equipment interact, rather than in a fully automated laboratory. The mix, of people interacting with changing equipment and processes, which characterizes a chemical laboratory provides a very testing envi-

(2) Frey, J. G.; Bradley, M.; Essex, J. W.; Hursthouse, M. B.; Lewis, S. M.; Luck, M. M.; Moreau, L; De Roure, D. C.; Surridge, M.; Welsh, A. H. Combinatorial Chemistry and the Grid. In *Grid Computing: Making the Global Infrastructure a Reality*; Berman, F., Hey, A. J. G., Fox, G. C., Eds.; John Wiley & Sons, Ltd.: Chichester, UK, 2003; pp 945−962 (Wiley Series in Communications Networking and Distributed Systems).

(3) Frey, J. G. Comb-e-Chem: an e-science research project. In *Proceedings of the 14th EuroQSAR 2002 Symposium*, Designing Drugs and Crop Protectants: Processes, Problems and Solutions, Bournemouth, UK; Ford, M., Livingstone, D., Dearden, J., Van der Waterbeemd, H., Eds.; Blackwell: Oxford, UK, 2003; pp 395−398 (European Symposium on QSAR, 14th).

ronment in which to see if appropriate software constructed with the most modern ideas of knowledge technologies can facilitate the process of chemical discovery.

Multimedia Chemistry: With What Is a Computer Faced When Dealing with Chemistry?

It is perhaps not common to attach the name multimedia to chemistry. If this conjures up an image of a mad chemist giving an inspired talk on MTV, that is not what I mean. Intrinsically, chemistry and chemical information is highly multimedia; two- and three-dimensional representations, structures, formulae, multidimensional spectra, and diagrams are used extensively to convey ideas. As chemistry is all about change, representation of the temporal aspects are now coming to the fore with video images being used just as fast as suitable software becomes available. In many cases the result of an experiment is an image! Software systems need to handle such diverse material and both present them for human interaction whilst preserving the underlying numerical aspects for subsequent computational processing. Keeping and coordinating the link between the image and the numerical data is far from simple. Without care, the computer equivalent of copying, enlarging, and remeasuring a spectrum can happen. It is even more infuriating to be given a jpeg file containing the image of the spectrum and somehow have to re-interpret it to get back the numbers for further processing; yet again, we see that humans and computers have differing requirements, and the needs of both must be allowed for. All the relevant material to service these differing requirements needs to be kept coordinated and linked together.

The Smart Tea Project

Having established that a chemistry laboratory populated with people and equipment was going to be one of the most challenging environments for the development of collaborative software we needed to explore a way in which we could facilitate the necessary collaboration between the chemists and the computer scientists. We were aware of a number of products in the market that look to provide some aspects of the laboratory integration and of many industries that were looking for similar solutions. At the time we were only partially aware of the significant resources that were being expended by companies to try and assess their needs in this area. This expenditure vastly exceeds the amount we could bring to bear on the subject. However, in the university research environment the presence of a wide range of expertise, from synthetic chemistry through semantic knowledge technologies to expertise in human computer interaction (HCI), brings the hope of providing the pointers to an innovative approach to the problem of providing a suitable laboratory software environment.

It is perhaps fitting that, in the anniversary year of the "Orwell discussion" on how to make a cup of tea, we found ourselves (chemists and computer scientists) not only discussing but also performing, recording, monitoring, analyzing, and modelling the process of making a cup of tea (and of course its derivatives: a cup of coffee and a cocktail). Perhaps only in a UK university could making tea take on such significance!

The Smart Tea model provided a very important system in which both the chemists and computer scientists could exchange ideas and understand the different approaches.[5,6] The understanding of user requirements using such a metaphor turns out to be a highly unusual approach, but we believe very cost-effective. Papers presented at HCI conferences discussing the methodology have excited considerable interest in the community.[7] The resulting interfaces have proved highly successful with graduate students who work on synthetic chemistry in the laboratory and are currently based on a tablet PC, communicating with a whole suite of services to provide the necessary functions. The supporting middleware software is currently being designed with the use of semantic techniques (more of this later) and is, to a considerable extent, a research topic in its own right.

The interaction of people with equipment facilitated by software is one of the central problems to be solved in the next few years. We can integrate software in the office environment to provide people with access to literature and computational systems. We can integrate software to run a fully automated ("dark") laboratory, but the challenge is to integrate software, people, and experiments in the much more flexible and exciting environment of people working in a laboratory, in other words, to provide the support directly to the researcher doing front-line innovative experiments, to ensure that information is collected without huge user overhead, and to provide simple and intuitive interfaces, making the subsequent access to the data (in all the different contexts) secure and simple.

How To Keep the Data?

Once the techniques for acquiring the data along with the absolutely crucial metadata have been established, compatible techniques for storing the information for subsequent retrieval and analysis need to be considered. These techniques must ensure that the data and metadata continue to be closely associated and are both available to subsequent researchers. It is probably not surprising that the issues of efficiency and flexibility tend to drive the software in opposite directions. If large amounts of data of a similar form are being created (such as that expected from CERN), then considerable effort can be expended in producing a tailored system that can cope with the huge data volumes. Such solutions are not typically

(4) Frey, J. G.; De Roure, D.; Carr, L. Position Paper: Publication at Source: Scientific Communication from a Publication Web to a Data Grid; Session: The Web and the GRID: From e-Science to e-Business. In *Proceedings of Euroweb 2002*; Hopgood, F. R. A., Matthews, B., Wilson, M. D., Eds.; EuroWeb 2002 Conference, St Anne's College, Oxford, UK, December 17, 18, 2002; British Computer Society: Swindon, UK (Electronic Workshops in Computing).

(5) Frey, J. G.; De Roure, D.; Schraefel, M. C.; Mills, H.; Fu, H.; Peppe, S.; Hughes, G.; Smith, G.; Payne, T. R. Context Slicing the Chemical Aether. In *Proceedings of First International Workshop on Hypermedia and the Semantic Web*; Nottingham, UK, 2003. Millard, D., Ed.

(6) schraefel, m.c.; Hughes, G.; Mills, H.; Smith, G.; De Roure, D. C. Smart Tea Project. In *eScience Pilot Projects Meeting*, NeSC, Edinburgh, UK, Southampton University, Southampton, UK, March 25−26, 2004; 2 pp.

(7) schraefel, m. c.; Hughes, G.; Mills, H.; Smith, G.; Payne, T.; Frey, J. Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment. In *Proceedings of CHI 2004*, Vienna, Austria.

suitable for a chemistry laboratory, although they may apply to experiments at synchrotrons and similar central facilities if there is sufficient commonality across experimental stations.

Many analytical laboratories (where, again, procedures can often be similar for a large number of routine samples) have implemented laboratory management systems frequently based around the software provided with the analytical equipment. These solutions are ideal for the appropriate laboratory; they are efficient and provide storage and retrieval for the specific data measured in that laboratory. They can, however, be difficult to integrate with other equipment and do not usually have the flexibility to integrate them directly in the synthetic laboratory. Nevetheless, the relational database structure employed by these laboratory systems has much to recommend it.

The rise of the relational database has been dramatic in almost every field except the physical sciences. It seems that many physical scientists like to keep their data in flat files where they can see where they are—of course the usual problem is that after a while they do not know where they are, and even when they are found, they have become detached from necessary metadata. The relational database together with its associated database managements system (DBMS) can take care of much of the association of metadata, update, versions, distributed storage, and curation of the data; this is a very attractive solution. Access to large quantities of data (for example in the form of a set of coordinates from a molecular dynamics simulation) can be inefficient in a traditional relational database, and this has in the past been sited as reason for their restricted uptake. This problem may be overplayed as the access is often not the time-limiting step (unless it needs to be done repeatedly). Solutions to this problem can be found, and indeed one of the *e*-Science projects (BioSymGrid) aims to store TB of data from simulations in this manner. The bioinformatics community have certainly taken this to heart.

We investigated using a relational database to store the data and the analysis from a laser experiment (second-harmonic generation at liquid interfaces). This required a small-scale database as the total quantity of information flowing from the equipment was limited. There was no problem in creating a schema to describe the data and metadata involved, but any attempt to track the analysis met with the problem of ever-changing needs of the analysis procedures. In short, this approach would have required us to change the schema very frequently, leading to all sorts of consistency issues. We concluded that for such variable steps the formal relational database was not the most suitable procedure. The solution we found lay in the use of RDF (resource description framework) to provide the semantic framework sitting upon XML descriptions for the analysis. The correct metadata was then directly associated with the data, and new methods could be added easily and flexibly stored in an RDF "Triple Store". In this case the additional flexibility more than out-weighed the performance and efficiency issues. A similar RDF approach to the semantics is being undertaken to provide the "back-end" for the Smart Tea project.

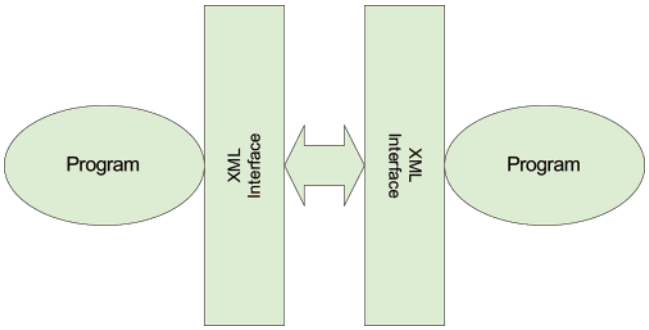

Figure 3. **XML interfaces enable more general machine-driven data exchange between programs and services in general. The XML output is not something in general that is designed to be read by people. The XML exchanges are not always the most efficient way to achieve the communication; faster, specifically written interactions are usually conceivable. It is, however, much more general but still not sufficient to provide the necessary semantic richness to support the data integration we require in a smart laboratory.**

With an XML-based description the standards for data exchange can be made more tractable. The standards do not have to be as prescriptive as they were in the past (line 2 col 3 must be the name). The XML schemas and related concepts mean that the standard makes it possible to describe what is in the document in such a way that a computer can attempt to interpret the content (Figure 3). It is like shipping the explanation of the standard with the standard in a computer-readable form. This makes it much more reasonable to convert one standard to another and places much less emphasis on formatting but more on the actual meaning of the content. XML does not, however, solve everything (although it is a great help), and XML-based systems are now in wide use behind many of the common software tools. In this we have some advantages as chemists: chemical Mark-up Language (CML) was one of the first XML systems.[8−12]

To implement an RDF approach it is useful to consider, as services, all stages involved in processing the data. Ensuring that the services have a common interface means that they can be discovered, linked, and managed as part of a workflow. In a grid environment these services can be located anywhere on the grid and accessed in a uniform manner. The ideas behind exposing a semantically rich interface so that computers (and humans) can more readily reuse the output of one program as the input to another have been gaining ground over the past few years; the rise of XMLs for different aspects and science and mathematics is a testament to the use of these techniques for improving

(8) Murray-Rust, P. Chemical Markup Language. In *World Wide Web J.* **1997**, 135−147
(9) Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1124−1130.
(10) Murray-Rust, P.; Rzepa, H. S.; Wright, M. Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content. *New J. Chem.* **2001**, *25*, 618−634.
(11) Murray-Rust, P.; Rzepa, H. S.; Wright, M.; Zara, S. A Universal Approach to Web-Based Chemistry Using XML and CML. *Chem. Commun.* **2000**, 1471−1472.
(12) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1113−1123.

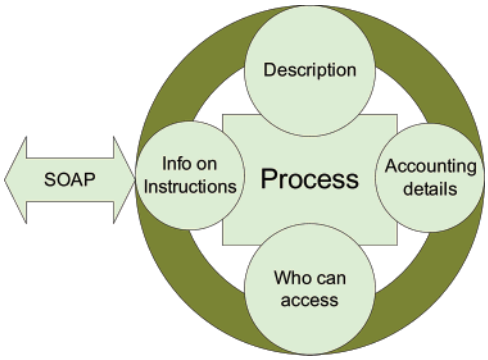

Figure 4. **Web Services. Web services are a key ingredient in the construction of our laboratory grid. They should not be confused with a web site or web pages. A web service is created when the process, which may be a database, a computation, an instrument, or even a human process, is wrapped up to expose a standard interface. The interface provides a description of the process, the functions that can be invoked, and the types of data stored in the process. Access control and accounting complete the details needed to understand how to use the service. These descriptions are provided through the use of a language that other services and workflow systems can interpret so that services can be discovered and linked together in an automated manner.**

dataflow. The result of this thinking led us down the path of web services (Figure 4) as these already had a commercial footing and thus could be expected to survive.

The Smart Tea investigation highlighted the significant gain that could be leveraged from the COSHH (Control of Substances Hazardous to Health) assessment that is carried out by every chemist prior to undertaking a particular procedure. By slightly extending the details in this assessment, the assessment could form the basis for the digital plan for the work. This, together with inference from other input (such as the person and what equipment they are working with), ensured that an active and responsive smart lab book could be created. Paralleling this plan would then be an expanded process plan, together with the process record (Figure 5). Underlying these layers is an even more detailed provenance record, which details all the services invoked. This demonstrates again the power of arranging all functions to be considered as services. The plan is a key phase in organizing the different processes, but different types of plan are relevant to different laboratory activities. As outlined in Figure 6, when only a restricted set of plans is needed, it is relatively easy to construct the necessary digital framework. Even if the plans are highly variable and process driven (for example data-driven analysis of a physical chemistry experiment) and if the variability is occurring "in silico", then a digital record is created almost by default (but the effort is expended in ensuring this can be kept and reused). The interactions in a synthetic laboratory are the most difficult types of plan to capture and enforce.

The Collaborative Culture and the Part To Be Played by the Grid

One way the Grid is sometimes introduced is by analogy with the electricity grid. To use electricity we can simply plug in to the mains and then not worry about where the power is generated on the grid. Somehow it arrives, is measured, and is accounted for. Some restrictions may be placed on its use, and we may sign up for particular supplies, but it is all managed without the active involvement of the user. In some cases we may even be able to supply energy back to the grid. A grid (for example for computing and data services) would have a similar structure with access to systems across different administrative domains being transparent to the user. The complexity of access is pushed into a layer called middleware that sits between the user and the operating systems and the network. From the user perspective the overall system then looks like that shown in Figure 7 rather than the network of Figure 2. As far as software developers are concerned, this still allows for services to talk to each other behind the middleware layer, i.e. not using the middleware layer for communication. However, as services are built up over time by different organizations, it is most likely that the middleware layer will, in fact, extend further to coordinate the interaction between many of these services, as then the workflows and provenance frequently required by the users are easier to enforce. The key here is the way in which the middleware can not only help with access to information but also smooth the collaboration between different users. Such worldwide access does imply that data can be uniquely addressed. This will necessitate extensions to the practice of assigning unique identifiers (see Figure 8) to all the information (and objects) in a laboratory; the pervasive agenda is seen in action again.

Security

There are a great many aspects of collaborative interaction and support that were possible many years ago and indeed very simply. Machines with dial-up connections reporting back to the manufactures and allowing remote diagnostics were first on the market decades ago. Now all this is done over the network which promises a much more scaleable solution for the manufacturer but seems to leave us open to all manner of infiltrations. When accessing these resources we now require a single sign-on (rather than passwords for each service). In Comb*e*Chem we have adopted a certificate-based access control system to achieve this end.[13]

Firewalls are a fact of modern computing life (Figure 9). Information is, in principle, readily available over the network, but enabling access to the information (in either direction) can lead you to be open to many types of attack. The firewalls do not protect against all types of attack, and the attempts to protect the laboratory from outside and inside can lead to a ridiculous cascade of castle building (Figure 10). A software and hardware infrastructure is required to deal with this and needs to be maintained. Risk assessments associated with network and software infrastructure are now as necessary as COSHH assessments for laboratory work. While it is yet another process to be undertaken, it may have benefits in planning similar to those that COSHH and other safety assessments have provided.

(13) Surridge, M. A. Rough Guide to Security. NeSC Technical Report; September, 2002 (http://www.nesc.ac.uk/technical_papers/RoughGuidetoGridSecurityV1_1a.pdf).

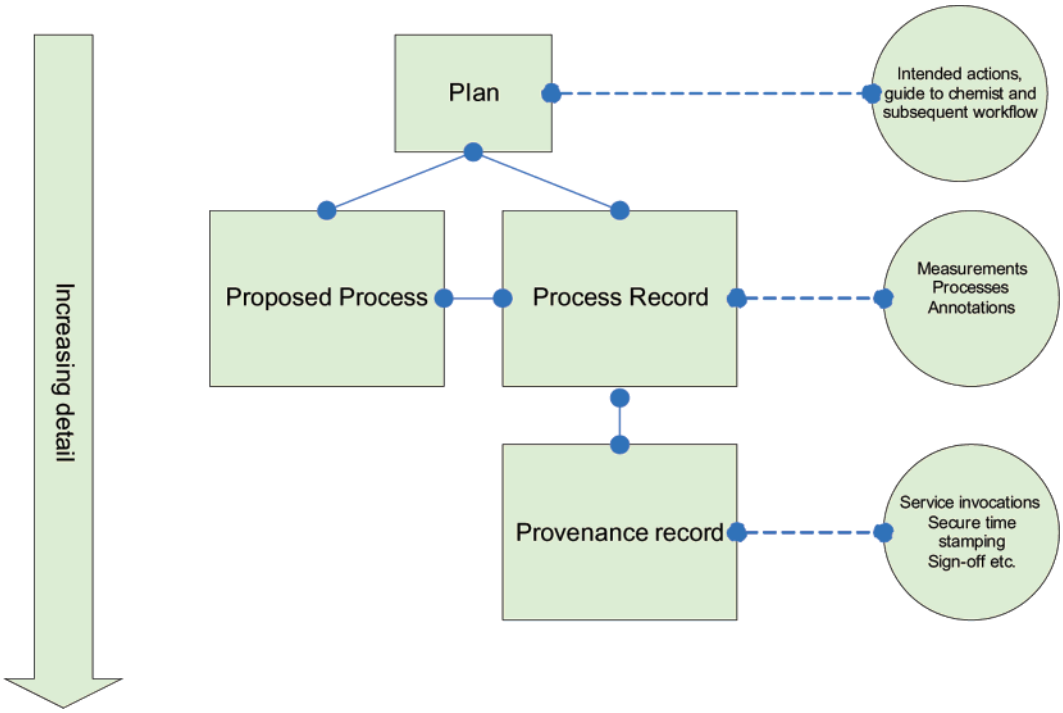

Figure 5. **Data model that is used behind the smart lab project showing the importance of the plan and the different levels of detail that follows from the plan. It is important also to distinguish between the intended actions derived from the high-level plan and the actions and process that actually are performed and the observations made (adapted from the Smart Tea project).**

Pervasive or Portals

The web laboratory, or the more grown-up Grid laboratory (that includes a greater range of resources, e.g., computational and other services), provides a version of a virtual laboratory notebook and could form the heart of the researcher's virtual organization. In principle it provides all the information that is currently available in a paper notebook (and more, see later on) without imposing a geographical constraint. Transporting the paper notebooks to meetings is easy, but only when the most current notebook is considered. Carrying round the total output of even one researcher would be an exercise in weight-lifting, not research.

There remains a question of how best to distribute the computational components. Distributing this power by giving everyone a laptop overcomes the total storage requirements and may make it easier to find a file (but how often have you had to wait while a colleague or student searches for a file, actually, I suspect that students waiting for their supervisor to find a file are the most frustrated!). Even more importantly, the laptop on its own in this manner is simply a replacement for the lab notebook. What we wish for is a much more all-encompassing system that allows access to a wider range of the information and data created in the laboratory (let alone the literature). I suspect this implies that the computing systems are connected to at least an intranet and thus there is probably no need to carry the laptop about at all, at least as far as meetings where a connected system can be provided. A solution is then to server up the information stored on the laboratory notebook systems to connected computers, usually via a web interface as this is one of the most general interfaces that can be expected to work on almost any platform (but not always with quite the expected results!).

These ideas have led to the construction of Portals as a possible solution to distribute and integrating information. The current vogue is for portals to enable relatively uniform access to widely different resources for example the CCLRC Data portal and the Collaboratory for Multi-scale Chemical Science (CMCS) systems[14,15] from the DOI labs in the USA). Portals are great for the analysis of information once collected, location and integration of literature, data. The integration is often achieved by using a series of Plug-ins, but how many will be needed in the future? How will they interact with each other? Can security be maintained? All these issues can be dealt with and as long as a proper semantic underpinning is provided it should be possible to migrate the underlying data to new systems in the future.

Instrument of the Grid – The UK EPSRC National Crystallography Service

As well as investigating the application of modern semantic technologies and pervasive computing to the synthetic laboratory under the Comb*e*Chem project we have looked at one of the next stages of the knowledge cycle, structure determination using crystallography. We have investigated the application of Grid technology to a high throughput instrument.[16,17]

The EPSRC-funded National Crystallography Service (NCS) is a facility available to the entire UK academic Chemistry community. The EPSRC funds a team of experts and "state of the art" instrumentation, based in Southampton

(14) Myers, J. D.; Chappell, A. R.; Elder, M.; Geist, A.; Schwidde, J. Re-integrating the Research Record. *Comput. Sci. Eng.* **2003**, *5*, 44−50.
(15) Hoyt, D. W.; Burton, S. D.; Peterson, M. R.; Myers, J. D.; Chin, G. Expanding Your Laboratory by Accessing Collaboratory Resources. *Anal. Bioanal. Chem.* **2004**, *378*, 1408−1410.

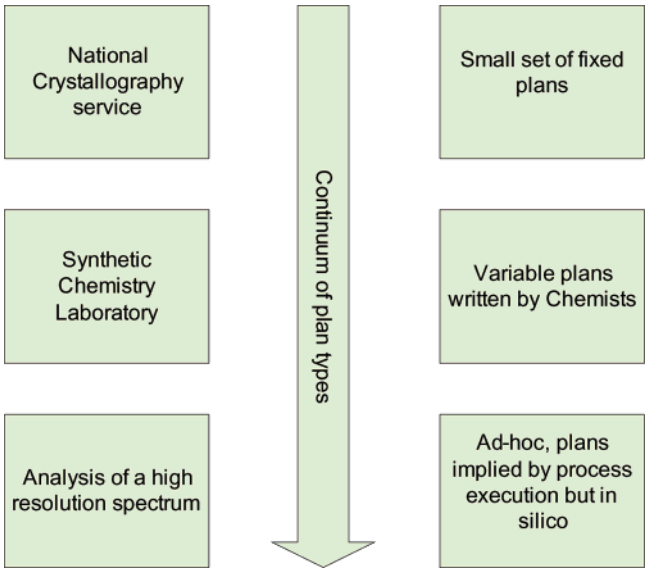

Figure 6. **Illustration of the different types of plans that can be expected for different styles of experiment. The National Crystallography Service deals with difficult crystal samples and has a number of well-defined procedures for this (although each one has many parameters relating to preparation, the experimental settings, and the analysis procedures). In contrast, the frequently somewhat ad-hoc analysis of a complex spectrum or other physical chemistry data has a very high degree of flexibility, but much of this occurs once the data is already in a digital form from which process capture is much easier. The plans here are similar to those seen in a bioinformatics applications (see, for example, the MyGrid Project). Between these limits are the plans written by chemists to guide synthesis procedures. The variability of these coupled with the human and equipment interactions makes this situation the most difficult to provide the necessary semantic support; this is what makes the smart lab such a challenging environment for the computer support.**

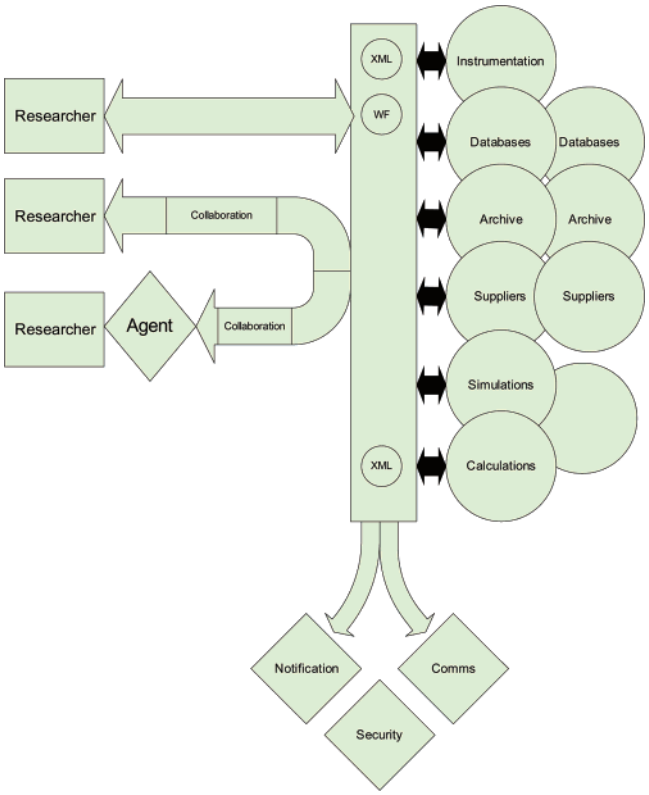

Figure 7. **More coherent view of the services required by a chemistry researcher obtained by the use of Grid middleware. The insertion of a middleware layer between the user and the underlying operation systems allows for a much more uniform style of access to resources that may have quite different underlying systems and are supported on different geographically diverse platforms; when working well, transparent access for the user to these resources is possible. The middleware can enforce workflows and ensure that provenance information is recorded, and the middleware layer can facilitate the interaction between researchers working on the same problem to provide a virtual co-laboratory, as well as facilitating interactions between the user and the services. Geography now only impinges on the collaboration via the issue of time zones.**

University School of Chemistry, to provide this service. The NCS has a broad base of users covering a whole range of chemical applications and whose structural expertise can range from none/novice to expert crystallographers. This service therefore provides support to those who have no facilities, whilst also complementing the less powerful instrumentation available to most experts. The NCS has approximately 150 subscribers who submit >1000 samples per annum. Coupled with providing a departmental service and maintaining the crystallography research group interests, NCS has approximately an order of magnitude higher throughput than the average crystallography laboratory. The NCS is currently at the forefront of pioneering the fully automated crystal structure determination experiment. This involves both hardware and software developments such as robotic sample changing, intelligent computation of a data collection strategy, and automated data processing and workup.

The NCS is perfectly suited to hosting an interactive Grid service for a number of reasons: The submitter of a sample will have a more detailed knowledge and understanding of it and hence be able to purposefully contribute to the experiment. The users of the service are distributed across the United Kingdom, yet need to collaborate closely with NCS staff. The high throughput of the service already demands effective sample management and tracking, especially when a user has multiple samples in the system. A user can monitor "out of hours" experiments to ensure they are completed successfully and that the most effective use of "instrument time" is made. The NCS staff are sometime almost overrun with "demanding" samples, and the ability of a user to manage more routine samples would relieve some of this pressure. The NCS Grid service infrastructure is outlined in Figure 11.

(16) Coles, S. J.; Frey, J. G.; Hursthouse, M. B.; Light, M. E.; Surridge, M.; Meacham, K. E.; Marvin, D. J.; De Roure, D. C.; Mills, H. R. Position Paper: Grid/Web Enhancements to the National Crystallographic Service: Experiences with an Interactive *e*-Science Demonstrator. Session: The Web and the GRID: From e-Science to e-Business. In *Proceedings of Euroweb 2002*; Hopgood, F. R. A., Matthews, B., Wilson, M. D., Eds.; EuroWeb 2002 Conference, St Anne's College, Oxford, UK, December 17, 18, 2002; British Computer Society: Swindon, UK (Electronic Workshops in Computing).

(17) Meacham, K.; Surridge, M.; Taylor, S.; Coles, S. J.; Light, M.; Bingham, A. L.; Hursthouse, M. B.; Peppe, S.; Frey, J. G.; Smith, G.; Mills, H. 2004 National Crystallography Service (NCS) Grid Service. EPSRC eScience Testbed Projects Meeting; NeSC, Edinburgh, UK, Southampton University, Southampton, UK, March 25−26, 2004; 4 pp.

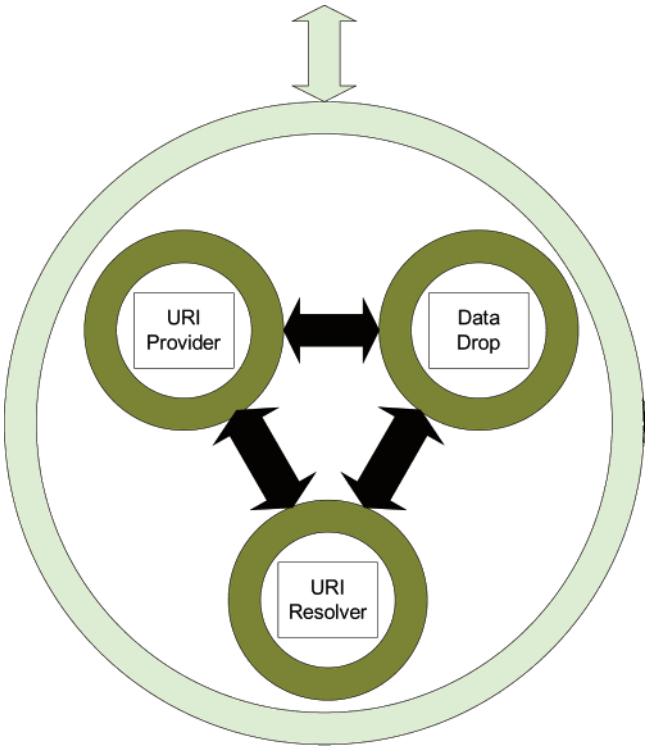

Figure 8. **Assignment of unique identifiers (URI: uniform resource identifiers) to all objects considered in the laboratory is crucial to the success of automated tracking, provenance, correlation, and inference. This applies not only to all records made but also to the physical samples as well. Within limited domains, such as an analytical laboratory, the assignment of such identifiers is now routine, and techniques for scaling up to deal with combinatorial chemistry and high throughput analysis have been implemented. For digital records a service can be constructed that provides a unique identifier and arranges for storage of the record (along with digital signatures and time stamps as needed). The counterpart to this service is the URI resolver that resolves the URI usually to a URL so that a browser or web service can access the information. While any URI should only resolve to one object, it is impractical to ensure that an object always has the same URI, partly as the issue of what is the same object may depend on circumstances (i.e. the criteria used). We propose another set of records that assert relationships (identity, similarity, etc.) between URIs as a scaleable way forward in correlating information. Both the data and the metatdata are distributed.**

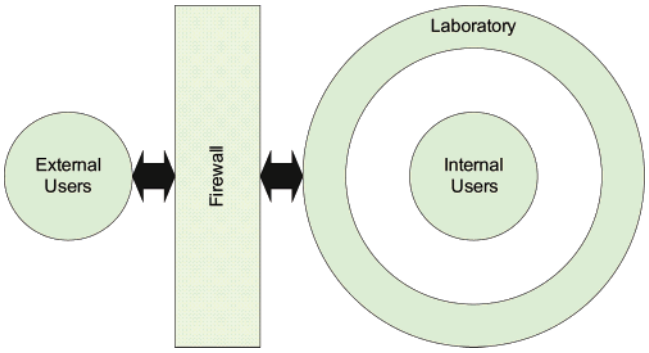

Figure 9. **In the collaborative world it is no longer feasible to isolate the laboratory systems from the outside world. However, exposing the network entails the risk of unwanted interaction. Initial protection from such attacks is provided by firewalls, which restrict the traffic to known types or to and from known sources. This greatly reduces the risks but also restricts the interactions possible across the network, and the firewalls are often the cause of software failing to work across a network.**

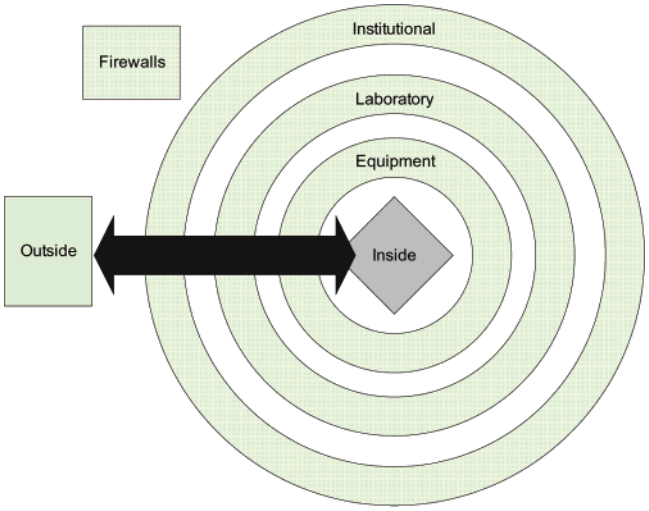

Figure 10. **Potential cascade tyranny of firewalls. Firewalls inside firewalls to protect the laboratory from users both outside and inside the institution. These systems have to be maintained and managed if they are to be effective and add yet more load to the administration of a laboratory. Interactions between the different firewalls adds yet more to the complexity and the likelihood that systems will have communication problems. This is not a stable approach in the long term.**

Service Security

Security is crucial to the successful operation of the NCS Grid Service, both in terms of the authentication of users and in maintaining the integrity of their data. For the NCS grid service, we have set up a well-defined trust network, with its own Certificate Authority (CA) and Registration Authority (RA). Only users with a valid NCS certificate may access the Grid Service. All data transfer is encrypted, and each user is authorised to access only their own data or to monitor their own experiments. This is achieved by mapping the user's credentials to the appropriate authorisation or datasets. The Status Service determines the client's Distinguished Name (DN) from their NCS certificate and then queries the Sample Database for a list of all samples submitted by the client (they may only see their own samples). Samples are presented in the client's browser, showing the status of each sample (or collection). In this way, the clients may regularly track the progress of their respective samples within the NCS system. Once a sample enters the running state, a link is made available to the Control Service, whereby the client may monitor the running experiment. This security infrastructure (Figure 12) provides control not only over data but also over access to equipment; it depends on not only the role of the user but also the state of the process (i.e. whose sample is on the machine at the time).

The Control Service provides the client with a portal to run the X-ray diffraction experiment, providing the opportunity to observe the experiment in progress and to steer it if so needed. The display is continuously updated to reflect the current state of the system as the experiment goes through its various stages (pre-scans, unit cell determination, full data

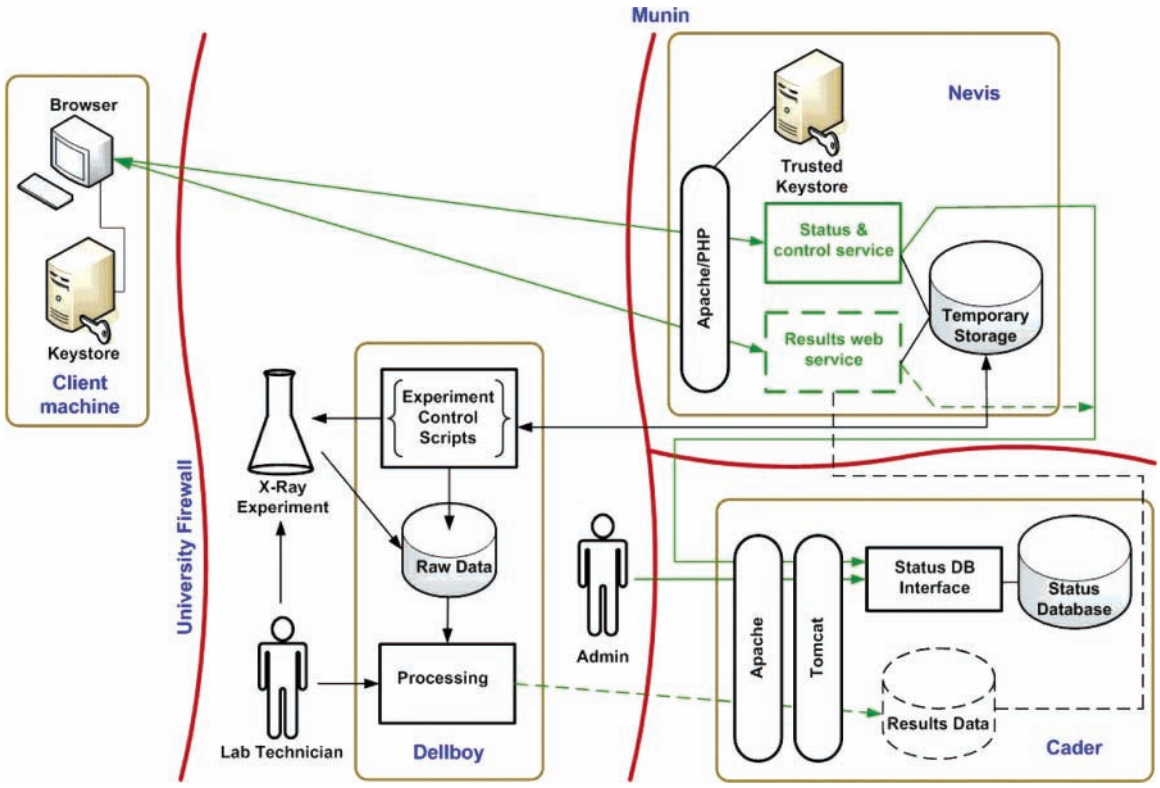

Figure 11. **Outline of the "software & hardware" infrastructure used in the NCS grid surface to support the secure user interaction, the sample tracking and control. The different processes are separated into several regions that are insulated from each other by firewalls. Figure used with permission from K. Meacham and the NCS Grid team.**

collection, and data processing). Scanned images and other raw data are collected by the diffractometer and published via the portal, enabling the client to make informed decisions at each stage whether to continue the experiment, etc.

e-Dissemination: Publication@Source

One of the most frustrating things when reading a paper is to find that the data you would like to use in your own analysis is in the form of a figure, meaning that you have to resort to scanning the image in to obtain the numbers. Even if the paper is available as a PDF, your problems are not necessarily reduced. In some cases the numeric data is already provided separately by a link to a database or other similar service (i.e. the crystallographic information provided by the CIF data file). In many cases if the information required is not of the standard type anticipated by the author, then the only way to obtain the information is to contact the author and hope he or she can still provide this in a computer readable form (assuming it was ever in this form). We seek to formalise this process by extending the nature of publication to include these links back to information held in the laboratories from which it originated. In principle this should lead right back to the original records (spectra, laboratory notebooks). A recent inquiry undertaken by the UK House of Commons Select Committee on Scientific Publication which look at open-access journals actually raised the question of how to ensure that chemists do report (or make available) the original spectra, and would some of the proposed publication methods and software actually make this possible or routine. Trials are underway to ensure this can happen with the crystallographic data from the NCS (the *e*-Bank project[18]).

It may be argued that for publicly funded research we have a responsibility to make all this information available (and not just the highlights pertinent to the points made in the papers). The immediate impact that many people may imagine from this is that it will make the detection of fraud much easier, but this is in fact a relatively minor issue. The main advantage will be the much greater use and reuse of the original data and the consequent checking of the data and different approaches to the analysis.

e-Science and *e*-Business

I hope that this discussion has highlighted some of the specific aspects of chemistry research that can interact with computer science to facilitate the needs of collaborative research. The way in which chemistry can use and catalyse many of the generic aspects of the knowledge grid is not that different from those needed for *e*-Business. In particular there is a common need for security (especially where patent-sensitive information is involved), authentication, and provenance to ensure that the information can be trusted or at least investigated. The research community can hope to use (and maybe informed about) techniques developed in this larger forum. The choice to go with a web services approach was for the Comb*e*Chem project not only motivated by

(18) Lyon, L. eBank UK: Building the links between research data, scholarly communication and learning, Ariadne, 36, (www.ariadne.ac.uk/issue36/lyon/).

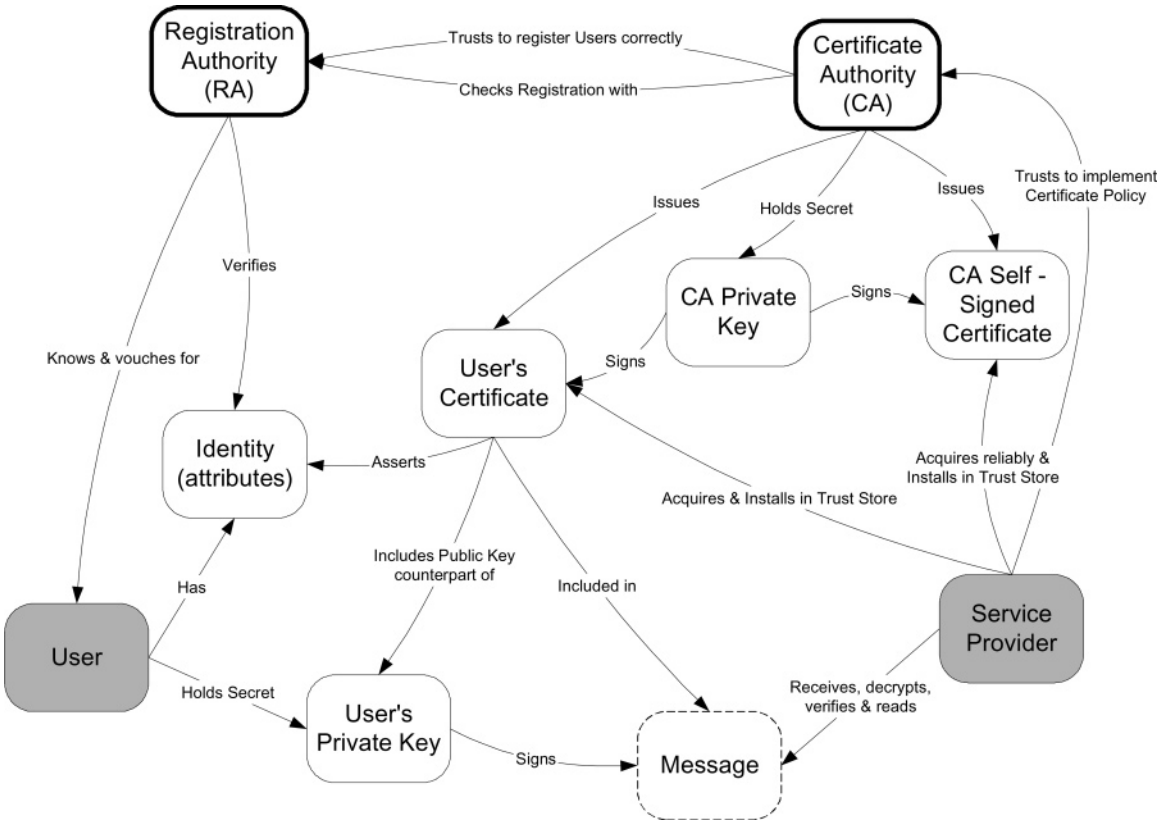

Figure 12. **Some of the interactions required to set up a reasonably secure security infrastructure for an instrument on the grid. Considerable human interaction is required to be assured of people's identity when setting up the certificates. Once this has been done the certificate controls the access to the users' samples and data. Figure used with permission from K. Meacham and the NCS Grid team.**

theoretical views but also by the realization that in this manner we could use and reuse much of the infrastructure deployed (or soon to be deployed?) by e-commerce. Reuse of software is as important as reuse of data.

Conclusions

The Grid infrastructure when fully developed will enable the chemist to sit at the centre of a virtual world with simple, rapid access to a wide range of physical, computational, and informatics resources. The implementation of automatic knowledge handling right from the inception of an experiment, through all stages of analysis and use of the information generated, will enable the single human mind to work collaboratively with others to keep pace with the exponentially growing quantities of chemical information generated by combinatorial techniques taking place in both "smart" and "dark" laboratories. Careful considerations of human computer interactions can ensure that the laboratory notebook is not just replaced by an electronic version which benefits only the subsequent users down the information chain but becomes a fully integrated part of an overall smart laboratory with direct and immediate benefit to all the researchers. Without such techniques to create and integrate the data we run the very real risk of generating even more information that is effectively hidden from the very people who should be using it.

Bibliography. Information on the Comb*e*Chem project can be found at the web site http://www.combechem.org, more details of the Smart Tea investigations at http://www.smarttea.org, for a full list of papers on the human computer interface aspects of the work. The NCS service is described at http://www.soton.ac.uk/xservice, and the *e*-Bank project can be found at http://www.ukoln.ac.uk/projects/ebank-uk which is a joint project with UKOLN based at the University of Bath. Information on the academic departments and research groups contributing to the Comb*e*Chem project can be found at the following web sites: (1) chemistry, http://www.chemistry.soton.ac.uk, (2) electronics and computer science, http://www.ecs.soton.ac.uk, (3) mathematics (statistics) http://www.maths.soton.ac.uk, and (4) IT-innovation (http://www.it-innovation.soton.ac.uk). An account of the UK *e*-Science research programme can be found at http://www.rcuk.ac.uk/escience. For more information and publication of aspects of the application of semantic support to multiscale chemistry the CMCS project web site at http://www.cmcs.org provides many links. The theme of steering computations (and to some extent experiments) underlies much of the work of the Reality Grid *e*-Science project; again, more information is available at http://www.realitygrid.org.

Acknowledgment

The Smart Tea group is G. Hughes, H. Mills, G. Smith, Dave de Roure, and m. c. schraefel; The NCS Grid Team is K. Meacham, S. J. Coles, M. Light, S. Taylor, G. Smith, S. Peppe, A. L. Bingham, M. Surridge, and M. B. Hursthouse.

Other laboratory software investigations in a physical chemistry context involve H. Fu, L. Danos, and J. Robinson. The principle and co-investigators of the Comb*e*Chem Project are J. G. Frey, M. B. Hursthouse, J. W. Essex (Southampton Chemistry), D. C. de Roure, L. Moreau, M. Luck (Electronics and Computer Science), A. H. Welsh, S. M. Lewis (Southampton Mathematics), G. Orpen (Bristol Chemistry). The Comb*e*Chem team also acknowledges the support we have had from the staff at IBM Hursley. Funding for the some of the work described here comes from the EPSRC *e*-Science and Chemistry programmes together with the JISC of HEFCE.

Received for review May 25, 2004.

OP049895G