Curation of Chemistry from Laboratory to Publication

"*The curation of laboratory experimental data as part of the overall data lifecycle*"

Simon Coles, **Jeremy Frey**, Andrew Milsted

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

Abstract

The paper will illustrate the "CombeChem Project" experience of supporting the chemical data lifecycle, from inception in the laboratory to organization of the data from the chemical literature. The paper will follow the different parts of the data lifecycle, beginning with a discussion of how the laboratory data could (or should) be recorded, and enriched with appropriate metadata, so as to ensure that curated data can be understood within its original context when subsequently accessed, as it is generated (the ideal of "Autonomic Annotation@Source"). Intrinsic to our argument is the recording of the context as well as the data, and maintaining access to the data in the most flexible form for potential future re-use for purposes that are not recognised when the data was collected. This is likely to involve many routes to dissemination, with data and ideas being treated by parallel but linked methods, which will influence traditional approaches to publication and dissemination, giving rise to a Grid style access to the information working across several administrative domains summarized by the concept of "Publication@Source".

1. Introduction

e-Science[1] is about global collaboration in key areas of science and the next generation of infrastructure that will enable it. It involves the "end-to-end" linking of data and information in the face of the data deluge created by emerging experimental techniques. CombeChem[2], an EPSRC e-Science pilot project, took this vision as its focus and involved a significant number of collaborators, spread over several disciplines, based in multiple departments at Southampton, together with several other academic and industrial concerns. The project concentrated on Grid-enabled chemistry, involving synthetic, laser and surface chemistry, and crystallography, as examples of the development of an e-Lab, using pervasive computing technology to record information on all aspects of laboratory work and carry this information forward through the whole chain of generation of chemical knowledge.

We aimed to provide the digital support for an end-to-end knowledge sequence in which an experiment produces data, from which results are derived, then searched for patterns, from which conclusions are drawn, leading to further experiments. The progress of science depends on each scientist building on the results produced by others; re-use of data, in both anticipated and unanticipated ways, is vital. E-Science techniques, as demonstrated by CombeChem, enable more data to be more freely available to scientists worldwide from heterogeneous sources, a problem with which industry has wrestled for years. CombeChem has successfully addressed these problems in both practical and theoretical ways. The scientist is crucial and thus the emphasis on usability.

At first glance it might seem that our applications are domain specific but the approach taken has produced generic applications and demonstrated that the software developed for one application can be re-used in another context, i.e., in university, and even in secondary (cf e-Malaria)[3], education. In the interests of widespread applicability and use of our systems, we have made creative use of the Web and developing Grid, and of open source and free software, which has enabled us to deliver data and user-friendly tools, and we have consistently emphasised the importance of standards (e.g., InChI[4]) and have been early adopters of such standards. Over its lifetime, the CombeChem project consistently increased its international visibility and reputation. Building on established expertise in Grid computing at IT Innovation, CombeChem led the world in adopting Web Services as its base platform from the earliest point in the programme, a position subsequently adopted by the wider UK e-Science community.

CombeChem was the basis of the original Semantic Grid report[5] and showed how to achieve full integration of laboratories and experimenters into an e-Science infrastructure based on pervasive and Semantic Grid technology. Many phases of the knowledge cycle were explored in CombeChem, from user interaction with Grid-enabled high-throughput analysis, fed by smart laboratories (notebooks and monitoring), together with modern statistical design and analysis, to utilization of semantic techniques to accumulate and disperse the chemical information efficiently. The way these investigations inform digital curation and dissemination are highlighted in the following sections. We consider the generation of data from instrumentation in section 2 and the Crystallography supplies the major example used here to demonstrate different approach to data dissemination discussed in section 6. The need to capture and link metadata in this and other chemical laboratories is covered in section 3 and examples provided in sections 4 and 5. Wider community interaction is considered in section 7 and conclusions drawn in section 8.

2 Instruments on the Grid and the National Crystallography Service (NCS)

In the Grid-enabling of research in structural chemistry, we focused on chemical crystallography and relating this to the chemical and physico-chemical properties of the target materials, adopting a “high-throughput philosophy”, processing large families of compounds while embedding the protocols for capture of metadata, date-stamping and adoption of agreed standards for archival and dissemination.

The CombeChem philosophy was applied to the operation of the EPSRC Chemistry Programme funded National Crystallography Service (NCS). This facility is a global centre of excellence with a long established service providing experimental structural chemistry resources for the UK chemistry community. The NCS involvement has provided an exemplar of how e-Science methodologies enhance user interaction with an operational service that provides resources to a distributed and varied user base and resulted in a set of recommendations for the construction of such an infrastructure (now being adopted by other EPSRC services). The NCS Grid facility has been deployed as a set of unified services that provide application, submission, and certificated secure, sample status monitoring, data access and data dissemination facilities for authenticated users, and is designed so that components currently under various stages of development from the project may be easily incorporated. Useful lessons have been learned in implementing security features for both the NCS systems, where a balance has been found between cost and a very secure network. In addition, remote and automated data collection procedures, based on a combination of robotics and scripted software routines, and coupled with automated structure solution and refinement, are presented.

2.1 Workflow Analysis

A first step towards designing such a complex system was the identification of the sequence of individual processes taken by users, service operators and samples, from an initial application to use the service to the final dissemination and further use of a crystal structure. All major activities, interactions and dependencies between the parties involved (both human and software components), may then be described as a workflow, from which an architecture that would accommodate all the processes could be designed.

The workflow for a typical service crystallography experiment is quite complex when considered at this level of granularity. For this reason it will not be reproduced here, -a thorough discussion is provided in reference [6]. A typical Grid, or web, service would only involve computing components (e.g. calculations, data retrieval services), hence the workflow involving these services is fairly trivial to derive and can be automated by an appropriate workflow engine. However, the service crystallography workflow also includes many manual operations, e.g. sending a sample to the service or mounting a sample on a diffractometer. From the analysis it was evident that the NCS Grid Service, in common with the few other scientific instruments on the Grid[78910], is server-driven as opposed to purely computational Grid services that are generally orchestrated by the user.

The workflow gives rise to the design of a database which is core to the system and is capable of tracking a sample through the workflow. A sample is automatically given a different status in this database, according to its position in the workflow and each status has different authorisation conditions. The interplay between a generalised form of the workflow and the status of a sample is shown in Figure 1.

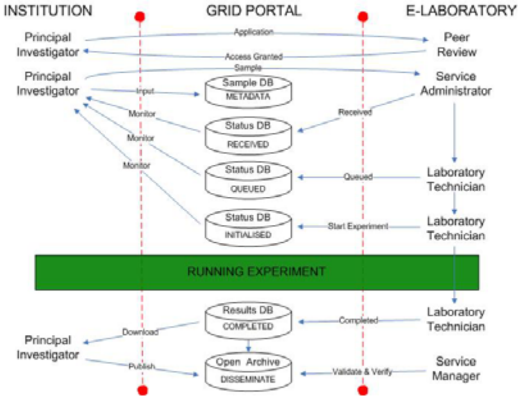


Figure 1

The X-ray diffractometer is normally controlled by bespoke software manually driven by the service operator via a Graphical User Interface (GUI). However, it is also possible to drive the diffractometer using command line calls, via an Advanced Program Interface (API). So, for the NCS Grid Service, scripts have been developed to drive the workflow normally carried out manually, which is essential as the experiment must be run automatically. As the experiment progresses raw data are deposited into a unique working directory on the NCS subnet system, to which the user has no direct access. The necessary experimental data are made available to the user by copying to a secure location on the DMZ server. The control script also makes calls to the sample/status database, at various key points during the experiment, to change the status of the sample being analysed.

2.2 Security and registration procedure

The NCS Grid service security infrastructure is designed in accordance with a Public Key Infrastructure[11] (PKI) policy. This requires the validity of each party involved in a transaction to be verified and authenticated using a system of X.509 digital certificates issued by a Certification Authority (CA) and a Registration Authority (RA). The issuing of certificates conforming to the X.509 specification requires adherence to a strictly-defined procedure[12]. Initially this was adopted, but credibility with the users required a slightly different approach. An alternative approach was devised[13] to avoid the requirement of users to install and use the relatively complex software used for the sign up and key management processes. The modified approach retains the software mechanisms, but handles the key generation centrally at the NCS. This deviates from a strict PKI in that user key-pairs as well as certificates are centrally generated, (i.e. by the NCS CA/RA), signed, and then securely transferred to the user, rather than relying on the user to perform the Certificate Signing Request (CSR) generation. the identity of the requestor and also to transmit the signed certificate and its corresponding passcode. As user generation of private keys becomes more commonplace and the supporting software more user friendly, the NCS intends to adopt standard CA/RA CSR practice. Users are required to re-register annually to obtain an allocation and new certificates are issued accordingly. It is therefore possible to update the security infrastructure at the same time, should it be considered necessary to update or integrate with other schemes.

2.3 User interaction during the experiment

The core of the NCS Grid service is the sample status database, which contains information on the position of the sample in the experimental workflow that may be updated by the system as processes are completed. A Status service written in PHP and running on the server visible to users, determines the DN of a user requesting access from their certificate and uses this to query the sample status database to obtain only the sample data owned by that DN.

At the point when the experiment is ready to start the service operator starts the experiment script, which automatically updates the status to *running* and provides a hyperlink in the status service that enables the user to participate with the experiment through the control service. The service operator may now leave the experiment for the user to monitor and/or guide and download the acquired data. On completion, the service personnel are responsible for the transfer of the complete archived dataset (raw and derived data) to an archival service for long term storage and retrieval if necessary. The NCS grid service uses the Atlas Datastore[14], based at the Rutherford Appleton Laboratory, UK. Approximately 1 Gb of data per day is transferred to the Atlas Datastore, who store the data with an off-site fire safe backup and migrate it on to new media as their service develops. Currently the data transfer is via FTP, but other front-ends to the Atlas Datastore are also provided, e.g. SRB[15].

3 Linking Experimental Data, Statistical Analysis and Publication@Source

We have used this combination of experiment and simulation as an exemplar of the advantages offered by linking publication to the original data. In a traditional journal publication, only the processed results are available, and there is no opportunity for the reader to examine their reliability or provenance. Indeed, there is usually not even the opportunity to see the exact numerical values plotted in a particular graph. A publication illustrating the linking of the original data to the published paper in the context of our Second Harmonic Generation experiments is in press; using the web version of this paper, it is possible to see and repeat all operations performed on the collected data to yield the final reported results.

4 Metadata, RDF, Smart Store and the Semantic Chemical Grid

Pervasive computing devices are used to capture live metadata as they are created in the laboratory, relieving the chemist of the burden of metadata creation. These data then feed into the scientific data processing. All usage of the data through the chain of processing is effectively an annotation upon it. By making sure everything is linked up through shared URIs, or assertion of equivalence and other relationships between URIs, scientists wishing to use these experimental results in the future can link back to the source (i.e., the provenance is explicit). CombeChem inherited traditional relational database technology to store experimental data and discovered the limitations in an environment where the user requirements are ever and rapidly changing.

As a Semantic Grid project, CombeChem has successfully deployed the latest Semantic Web technologies, including the RDF triplestore developed by the Advanced Knowledge Technology Interdisciplinary Research Collaboration (AKT IRC), to address the integration of diverse datastores. This was and continues to be a testing deployment, directly reflecting real world requirements, as different stakeholders own the different stores. There are significant gains to be had by this approach (reflecting Hendler's maxim: "a little semantics goes a long way"). While other groups are adopting XML formats (such as CML) for data interchange, CombeChem has taken advantage of these ideas but has moved to a higher level through widespread adoption of RDF. Knowledge technologies were applied through collaboration between Collaborative Advanced Knowledge Technologies in the Grid (CoAKTinG) and CombeChem, where the tools enhance the collaboration environment for chemists and help provide a complete digital record of the scientific process: the collaborative Semantic Grid.

5 SmartTea, Electronic Laboratory Notebooks and User-Centred Design

The pervasive computing aspects have focused on collecting data "at source", either from the handheld devices or from sensors recording experimental conditions. The former has been tackled through studying chemists at work in a laboratory and then designing a new device to support their work. Throughout the project, effort has been put into usability and representation of results for use by humans (HCI) as well as by machines; the SmartTea system and e-Bank are examples.

ELNs are currently attracting much interest. In a daylong symposium at an ACS National Meeting in 2004, the presentations from Southampton stood out from all the others in terms of scientific content and system design. "Smart tea" has been highly cited in the ELN and user design literature as an example of user-centred design by analogy. The requirements study was carried out in such a way that the designers and the users could communicate effectively and produce a truly user-friendly and effective system. The design approach, also used in the statistics teaching packages, demonstrated the need for context-sensitive systems. The monitoring of the laboratory environment has been made much more flexible and responsive to change by the adoption of the publication/subscribe model facilitated by the use of IBM micro-broker and MQTT middleware, for which we obtained significant publicity (including the BBC) for the use of mobile phones to monitor the laboratory.

The "back-end" for these laboratory systems, recording the processes undertaken, uses the same RDF technology as in the recording of information about the molecular species highlighted above. This shows the way to integrate the process information captured by the pervasive technology with the knowledge base about the materials, all using the Semantic Grid approach. The further analogy between the publication/subscribe approach and the publication@source provenance model suggested new models for the dissemination of data as well as ideas, building on the e-Print OAI approach pioneered in Southampton, which CombeChem is taking forward (e.g., e-Crystals), together with the JSIC-funded e-Bank I, II and Repository for the Laboratory (R4L) projects.

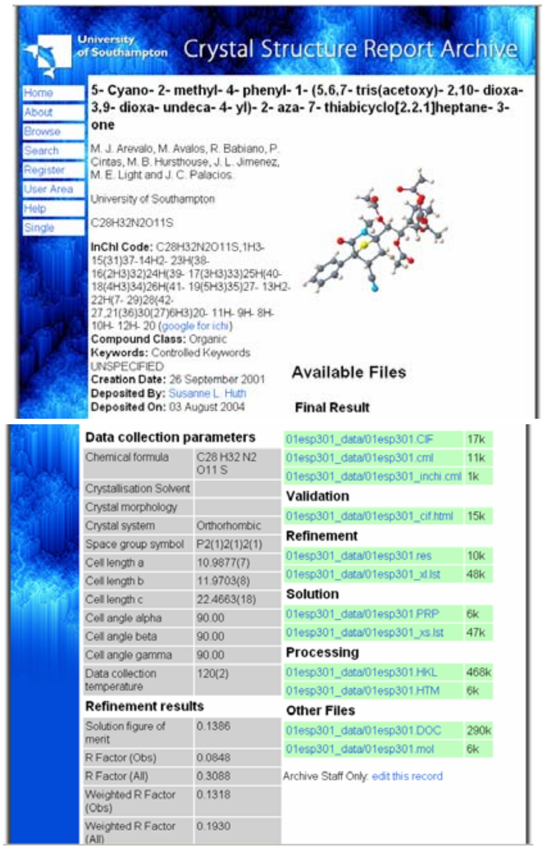
6 OAI Repositories for data capture, management and dissemination

An end-to-end process demands a publication procedure at one end and the CombeChem project has built on this with the e-Bank project (http://www.ukoln.ac.uk/projects/ebank-uk) to produce the e-Crystals archive. This crystallographic e-Prints process obviates the need to pack scientific papers with vast amounts of data. The papers can concentrate on the presentation and discussion of ideas and the data is consulted only when required. An Open Archive has been developed to disseminate all the data accumulated during the course of the crystallographic experiment. This archive not only allows free and unhindered access to the data underpinning a scientific study but also publicises its content (including existence of the data) through established digital library protocols (e.g., OAI). The dissemination of results, held at the NCS service, that would otherwise remain hidden, will benefit structural researchers worldwide. However, not all data is "good data" (from a reputable source), so the provision of provenance tracking, as in e-Crystals, is essential for potentially un-refereed data.

6.1 The OAI crystal structure archive

The archive is a highly structured database that adheres to a metadata schema which describes the key elements of a crystallographic dataset. Current details of this schema can be found at http://www.ukoln.ac.uk/projects/ebank-uk/schemas/. The schema requires information on bibliographic and chemical aspects of the dataset, such as chemical name, authors, affiliation etc, which must be associated with the dataset for validation and searching procedures. As standards must be adopted in order for the metadata in the archive to be compatible with that already accepted and available in the public domain a tool for aiding the deposition process has been built. This tool performs the necessary file format transformations and operations necessary for presentation of the dataset to the archive. The elements of the schema and a brief description of their purpose are given in reference 6.

On completion of the crystal structure determination all the files generated during the process are assembled and deposited in the archive. For conventional publication purposes a crystal structure determination would normally terminate at the creation of a Crystallographic Information File (CIF)[16] and this file would be all that is required for submission to a journal, however this archive makes available ALL the underlying data. The metadata to be associated with the dataset is generated at this point, either by manual entry through a deposition interface or by internal scripting routines in the archive software which extract information from the data files themselves. All the metadata are then automatically assembled into a structured report and an interactive rendering of a Chemical Markup Language[17] (CML) file added for visualisation purposes (Figure 2).



5- Cyano- 2- methyl- 4- phenyl- 1- (5,6,7- tris(acetoxy)- 2,10- dioxa- 3,9- dioxa- undeca- 4- yl)- 2- aza- 7- thiabicyclo[2.2.1]heptane- 3- one

M. J. Arevalo, M. Avalos, R. Babiano, P. Cintas, M. B. Hursthouse, J. L. Jimenez, M. E. Light and J. C. Palacios

University of Southampton

C28H32N2O11S

InChI Code: C28H32N2O11S,1H3-15(31)37-14H2- 23H(38-16(2H3)32)24H(39- 17(3H3)33)25H(40-18(4H3)34)26H(41- 19(5H3)35)27- 13H2-22H(7- 29)28(42-27,21(36)30(27)6H3)20- 11H- 9H- 8H-10H- 12H- 20 (google for ichi)
Compound Class: Organic
Keywords: Controlled Keywords UNSPECIFIED
Creation Date: 26 September 2001
Deposited By: Susanne L. Huth
Deposited On: 03 August 2004

Data collection parameters

Chemical formula	C28 H32 N2 O11 S
Crystallisation Solvent	
Crystal morphology	
Crystal system	Orthorhombic
Space group symbol	P2(1)2(1)2(1)
Cell length a	10.9877(7)
Cell length b	11.9703(8)
Cell length c	22.4663(18)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	120(2)

Refinement results

Solution figure of merit	0.1386
R Factor (Obs)	0.0848
R Factor (All)	0.3088
Weighted R Factor (Obs)	0.1318
Weighted R Factor (All)	0.1930

Available Files

Final Result

01esp301_data/01esp301.CIF	17k
01esp301_data/01esp301.cml	11k
01esp301_data/01esp301_inchi.cml	1k

Validation

01esp301_data/01esp301_cif.html	15k

Refinement

01esp301_data/01esp301.res	10k
01esp301_data/01esp301_xl.lst	48k

Solution

01esp301_data/01esp301.PRP	6k
01esp301_data/01esp301_xs.lst	47k

Processing

01esp301_data/01esp301.HKL	468k
01esp301_data/01esp301.HTM	6k

Other Files

01esp301_data/01esp301.DOC	290k
01esp301_data/01esp301.mol	6k

Archive Staff Only: edit this record

Figure 2

6.2 Metadata harvesting and value added services

When an archive entry is made public the metadata are presented to an interface with the internet in accordance with the Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH)[18]. OAI-PMH is an accepted standard in the digital libraries community for the publication of metadata by institutional repositories which enables the harvesting of this metadata. Institutional repositories and archives that expose their metadata for harvesting using the OAI-PMH provide baseline interoperability for metadata exchange and access to data, thus supporting the development of service providers that can add value. Although the provision of added value by service providers is not currently well developed a number of experimental services are being explored. The eBank UK project has developed a pilot aggregator service[19] that harvests metadata from the archive and from the literature and makes links between the two. The service is built on DC protocols and is therefore immediately capable of linking at the bibliographic level, but for linking on the chemical dataset level a different approach is required. The Dublin Core Metadata Initiative (DCMI) provides recommendations for including terms from vocabularies in the encoded XML, suggesting: *Encoding schemes should be implemented using the xsi:type attribute of the XML element for property*. As there are not as yet any designated names for chemistry vocabularies, for the purposes of building a working example the project defined some eBank terms as designators of the type of vocabulary being used to describe the molecule. Thus a chemical formula would be expressed in the metadata record as:

```
<dc:subject
xsi:type="ebankterms:ChemicalFormula">C27
H48</dc:subject>
```

There are currently no journals publishing crystal structure reports that disseminate their content through OAI protocols. In order to provide proof of concept for the linking process the RSS feed for the International Union of Crystallography's publications website was used to provide metadata and crystal structure reports published in these journals were then deposited in the archive, thus providing the aggregator service with two sources of information. Aggregation is performed on the following metadata; author, chemical name, chemical formula, compound class, keywords and publication date, thus providing a search and retrieval capability at a number of different chemical and bibliographic levels. The demonstrator system, along with searching guidelines may be viewed and used at http://eprints-uk.rdn.ac.uk/ebank-demo/.

6.3 Data capture and management

Building on the advances made by the eBank project we are now developing repositories to drive efficient and complete capture of data as it is generated in the laboratory. Embedding the archive deposition process into the workflow in the laboratory in a seamless and often automated manner allows the acquisition of all the necessary files and associated metadata. This procedure is determined and driven by the experimental workflow and publication process and the 'Repository for the Laboratory' project (R4L: http://r4l.eprints.org) is in the process of workflow and publisher requirements capture. The repository driven experiment has the advantage that very accurate metadata can be recorded and a registration process has been designed whereby an experimenter can unambiguously assert that he/she performed a particular action at a particular time. A schematic of this project is shown in figure 3.

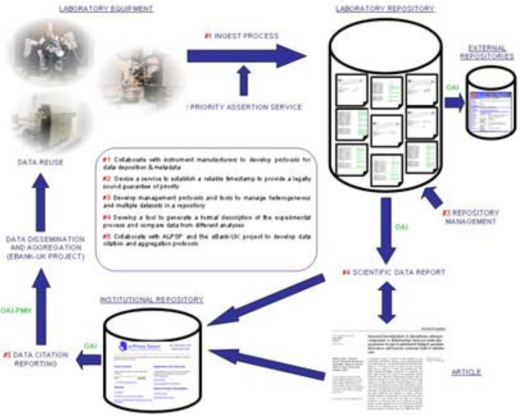

LABORATORY EQUIPMENT
LABORATORY REPOSITORY
#1 INGEST PROCESS
! PRIORITY ASSERTION SERVICE
EXTERNAL REPOSITORIES
OAI
DATA REUSE
#1 Collaborate with instrument manufacturers to develop protocols for data deposition & metadata
#2 Devise a service to establish a reliable timestamp to provide a legally sound guarantee of priority
#3 Develop management protocols and tools to manage heterogeneous and multiple datasets in a repository
#4 Develop a tool to generate a formal description of the experimental process and compare data from different analyses
#5 Collaborate with ALPSP and the eBank-UK project to develop data citation and aggregation protocols
#3 REPOSITORY MANAGEMENT
OAI
DATA DISSEMINATION AND AGGREGATION (EBANK-UK PROJECT)
#4 SCIENTIFIC DATA REPORT
INSTITUTIONAL REPOSITORY
OAI-PMH
OAI
#5 DATA CITATION REPORTING
ARTICLE

Figure 3

7 Outreach: CombeChem and e-Malaria

It is generally acknowledged that the public reputation of science is poor. In schools, science teaching can be uninspiring, science is perceived as boring, hard and irrelevant to people's lives, and the decline in the number of pupils choosing science courses is worrying for the science community and society at large. To address this difficulty, we have, as part of a project jointly supported by CombeChem and

JISC, developed an educational tool targeted at drug design for malaria. An integrated software environment combining web design, database development, and distributed computing has been developed. The software is aimed at A-level students of chemistry; the students are asked to design, using a sketch pad, chemical compounds that they can then submit for docking against a known malaria target. The score from the docked structure may then be used, together with molecular graphics, to refine further a potential drug. This software teaches the elements of molecular structure and intermolecular forces, with the added driver of targeting a serious illness.

Further development of this project through the South Eastern Science Learning Centre based at Southampton is planned. Diseases such as malaria, while being unprofitable to most pharmaceutical companies, make good choices for academic outreach projects. At a system level, software to be used by school students has to be designed differently from that used by university researchers and it must also be robust. The project has taken these factors on board and presents valuable lessons in how to achieve the secure integration of industrial strength programs into a "free" outreach environment.

8 The Pervasive Semantic Grid

Our deployment and use of pervasive computing technologies was informed by working alongside the infrastructure team of the Equator IRC, and this research is directly in line with the Next Generation Grids strategic report from the European Commission, with its attention to ambient intelligence. We have developed a particularly powerful combination of Grid and pervasive computing in that we do not just have the Grid meeting the physical world through the pervasive devices, but rather we have the physical world intersecting with the Semantic Grid: the experimental processes are themselves described as RDF graphs; the devices capturing experimental conditions do so in the form of semantic annotation; and interaction with the information is seen as annotation, that is, as enrichment through use. Hence the CombeChem vision provides a case study in the "Pervasive Semantic Grid". Much has been said about the Semantic Web but we will not see the benefits of Semantic Web technology unless programmes such as CombeChem actually use it.

The CombeChem team has achieved a real Semantic Web that enhances the ability to disseminate and curate data with appropriate context. A danger in the Grid community is that it is building within its own clique but the Grid is no good unless other people can join in. CombeChem is the real Grid because it has real users. We are changing the way in which science is done and ensuring that in the future our data and information will be used and useable!!

9 Acknowledgements

We acknowledge the tremendous support from the whole of the CombeChem and e-Bank teams, from the NCS, and colleagues in the Schools of Chemistry and Electronics & Computer Science. CombeChem was funded under the UK e-Science programme EPSRC grants GR/R67729/01 and EP/C008863/1. eBank is funded by JISC under the Semantic Grid and Autonomic Computing Programme.

10 References

[1] T. Hey, A. Trefethen, Cyberinfrastructure for e-Science, Science 308 (2005) 817–821.

[2] http://www.combechem.org

[3] R. Gledhill, S. Kent, B. Hudson, W.G. Richards, J.W. Essex, J.G. Frey, A Computer-Aided Drug Discovery System for Chemistry Teaching, J. Chem. Inf. Model, 2006, ASAP Article, DOI: 10.1021/ci050383q

[4] www.iupac.org/inchi

[5] D. De Roure, N. Jennings, N.R. Shadbolt, Research Agenda for the Semantic Grid: A Future e-Science Infrastructure; Technical Report UKeS-2002-02; National e-Science Centre: Edinburgh, UK, 2001; De Roure, N.R. Jennings, N.R. Shadbolt, The Semantic Grid: Past, Present, and Future, Proc. IEEE 93 (2005) 669–681.

[6] S.J. Coles, J.G. Frey, M.B. Hursthouse, M.E. Light, A.J. Milsted, L.A. Carr, D. De Roure, C.J. Gutteridge, H.R. Mills, K.E. Meacham, M. Surridge, E. Lyon, R. Heery, M. Duke and M. Day, M. (2006). An E-Science Environment for Service Crystallography from Submission to

Dissemination. *Journal of Chemical Information and Modeling* (doi:10.1021/ci050362w)

[7] R. Bramley, K. Chiu, J.C. Huffman, K. Huffman and D.F. McMullen, Instruments and Sensors as Network Services: Making Instruments First Class Members of the Grid, *Indiana University CS Department Technical Report 588*, 2003.

[8] http://nmr-rmn.nrc-cnrc.gc.ca/spectrogrid_e.html

[9] http://www.itg.uiuc.edu/technology/remote_microscopy/

[10] http://www.terena.nl/library/tnc2004-proceedings/papers/meyer.pdf

[11] A. Nash, W. Duane, C. Joseph, O. Brink, B. Duane, PKI: implementing and managing E-security, 2001, New York: Osborne/McGraw-Hill.

[12] R. Guida, R. Stahl, T. Blunt, G. Secrest, J. Moorcones, Deploying and using public key technology, *IEEE Security and Privacy*, 2004**,** *4*, 67-71.

[13] A. Bingham, S. Coles, M. Light, M. Hursthouse, S. Peppe, J. Frey, M. Surridge, K. Meacham, S. Taylor, H. Mills, E. Zaluska, Security experiences in Grid-enabling an existing national service, Conference submission to: eScience 2005, Melbourne, Australia.

([14]) http://www.e-science.clrc.ac.uk/web/services/datastore

([15]) http://www.sdsc.edu/srb/

([16]) I.D. Brown, B. McMahon, CIF: the computer language of crystallography.**,** *Acta Cryst.*, 2002, *B58*, 317-324.

([17]) P. Murray-Rust, H.S. Rzepa, M. Wright, Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content, *New J. Chem.*, 2001, 618-634.

([18]) C. Lagoze, H. Van de Sompel, M. Nelson, S. Warner, The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. 2002, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

([19]) M. Duke, M. Day, R. Heery, L.A. Carr, S.J. Coles, Enhancing access to research data: the challenge of crystallography. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005, 46 – 55, ISBN:1-58113-876-8