JISC

Project Document Cover Sheet

Project Information			
Project Acronym	R4L		
Project Title	The Repository for the Laboratory		
Start Date	May 2005	**End Date**	May 2007
Lead Institution	University of Southampton		
Project Director	Leslie Carr		
Project Manager & contact details	Simon Coles. EPSRC National Crystallography Service, School of Chemistry, University of Southampton, Southampton, SO17 1BJ. t: +44(0)2380 596722, f: +44(0)2380 596723, e: s.j.coles@soton.ac.uk		
Partner Institutions			
Project Web URL	http://r4l.EPrints.org		
Programme Name (and number)	Digital repositories programme 2005-7		
Programme Manager	Neil Jacobs		

Document Name			
Document Title	*Final Report*		
Author(s) & project role	Simon Coles, Leslie Carr and Jeremy Frey; Project Directors		
Date	09/12/2007	**Filename**	R4L final report.doc
URL	http://r4l.EPrints.org/r4lfinalreport.pdf		
Access	☐ Project and JISC internal		X General dissemination

Document History		
Version	**Date**	**Comments**
1.0	20/11/2007	For comment
2.0	09/12/2007	Final version

R4L: The Repository for the Laboratory

JISC Final Report
December 2007

Simon Coles, Jeremy Frey & Leslie Carr

Table of Contents

Acknowledgements

The Repository for the Laboratory is a joint project between the schools of Chemistry and Electronics and Computer Science at the University of Southampton and is funded under the JISC Digital Repositories Programme. We would particularly like to thank the Royal Society of Chemistry (Richard Kidd), Chemistry Central (Bryan Vickery), Bruker AXS, Rigaku and Oxford Diffraction.

Executive Summary

The Repository for the Laboratory (R4L) project aims to leverage current innovations in digital libraries research area of institutional repositories to impact on laboratory based science, however this work is focussed around the operations of a scientific laboratory and is not concerned with]the dissemination aspects normally associated with this field. The project presents a proof of concept of a novel data and information capture, management and discussion framework for experimental, laboratory based science, in particular the area of chemical analysis. The concept is centred around the notion of an repository as the mechanism for the data storage and management and builds tools around this focus for ingest, discussion and report generation.

The laboratory repository is a separate entity from the institutional repository not out of architectural necessity, but in order to emphasise a difference in purpose and to ensure the development of appropriate policies. The policies and processes that need to be considered are data storage, access, backup, archiving of raw [proprietary] data, management, format for publication and timely and appropriate release into the public domain.

The R4L demonstrator repository (http://r4l-dev.EPrints.org/), developed as a deliverable of this project, is capable of ingesting, storing, managing and presenting a cross section of some ten different types of data holding arising from different analytical techniques. The ingest processes have been carefully designed, following detailed analysis of laboratory workflows, in order to ensure complete capture of the raw, derived and descriptive data and thus provide a full provenance trail and support a comprehensive preservation process. A probity service developed as a project deliverable provides a reliable and unique process for unambiguously registering the experimental data in a legally sound fashion. The structure of a record in the R4L repository has been configured so as to be centred around a chemical entity, to which data from various supported analytical can be associated. This novel structure allows management of data and records in the repository based on chemical structure, which is an essential requirement of the chemistry community. The R4L demonstrator is completed by the data discussion/analysis and report generation processes. Blog technology has been employed to facilitate discussion and collaboration with respect to repository data by enabling ‘live copy’ transfer of data from the repository to the blog space. The blog allows collaborating scientists to post data, make it available to particular collaborators and discuss it. Additionally, as is common in the blogging community, the data can be made public at the time of posting if desired thus supporting the emergent concept of ‘open notebook science’. The same live copy approach has been employed to demonstrate the writing of reports by pasting data into publication ‘templates’, which are held in a repository maintained in parallel to the R4L repository and accessible to particular collaborators.

Background

An important aspect of scientific research is concerned with laboratory experimentation, data collection and data sharing for analysis. The Institutional Repository (IR) community has been concerned with the dissemination of experimental descriptions (in the form of articles) and now, more recently the dissemination of finalised experimental results (in Institutional *Data* Repositories, as pioneered by the JISC eBank project) to supplement the IR documents & papers. The R4L project addresses the gap between the actual experiments and the publication of papers. Importantly this includes the infrastructure required to disseminate results while affirming priority (the scientific claim of being the first to achieve, a claim currently supported by publication dates and appropriately counter-signed log books). This follows on from the consideration of the documentation of the experimental procedures, the experimental workflow, the results collected and the analyses performed, which eventually becomes a journal paper. A direct integration can be imagined of the e-Science approach to capturing the laboratory functions and the IR data collections with the document production environment through data description standards together with semantic relationships, i.e using semantic web technologies, such as RDF. This would allow the automated production of tables, figures, statistics and descriptions of process together with links to the archive to provide the necessary scientific (and legal) provenance. These reports would be linked or incorporated as part of the scientific study that is presented as an article in the conventional IR.

Scientific publications, particularly those in the physical science disciplines, invariably report findings that are built upon results gained from experimental data gathering exercises. The processes of gathering the data that underpins a publication can often be very expensive, involved and time consuming, but also information-rich and highly valuable to the wider scientific community. In addition, a number of different experiments may be necessary to acquire all the information required to perform a thorough study for publication. The management of data and results from different analyses is currently performed in isolation from each other and as a result comparison, cross reference and identification of common features is time consuming and unreliable and hence seldom performed.

Current publication protocols and procedures in the data-based scientific disciplines do not suit the dissemination and sharing of data. A journal article describing the results of scientific work is typically a distillation of experimental data aimed at a wider audience than the immediate peers of the authors. Generally inferences are made only from the most pertinent results, which are reported in a summary format, and journal publication is detached from the production of the experimental data. This renders replication or reuse of the data impossible and results in severe information loss. In addition, access to all the underlying data is either hindered or impossible, again prohibiting further reuse of the data in value-added or further studies. A further barrier to unhindered access to scientific data is the ‘licence’ problem, where only researchers in subscribing institutions may access the data held by the publishing body.

In the laboratory environment the researcher will perform multiple analyses, as part of a single study, which must be compared and contrasted in order to make deductions. The traditional approach of treating each analysis separately and recording data on different media and formats is both laborious and unreliable, especially when attempting to draw conclusions from comparison between different analyses. Modern computational and scientific instrument technology now allows rapid analyses to be performed, providing the scientist with vast amounts of experimental data, which is becoming increasingly difficult to manage. The emergent field of e-Science has the potential to address some of these issues, through the development of Grid-based environments for laboratory experimentation. The EPSRC funded e-Science testbed project, CombeChem (http://www.combechem.org) in which a number of project partners were involved, sought to integrate existing structure and property data sources into an information and knowledge environment. To this end analytical instruments and even synthesis labs were ‘put on the Grid’, enabling the digital output from these operations to be efficiently managed, processed, staged and curated using Grid technologies.

The technological advances outlined above have caused an explosion of scientific data over the last few years, allowing results to be derived at an unprecedented rate. However, across the scientific domain only a small proportion of the data generated by experimentation appears in, or is

referenced by, the published literature. The cause of this shortfall is clearly identifiable as the inability of the traditional publication protocols to take the complete dataset through this process, coupled with an increasing burden placed on the peer review system by the inclusion of just the fraction of the dataset that is conventionally required. This problem may be demonstrated by the current situation with the publication of crystal structures arising from chemical crystallography experiments:

A postgraduate student in the 1960's would have typically investigated around three crystal structures, whilst with the modern technologies available today this may be achieved in a single morning. Despite these advances, the publishing protocols for reporting this work are essentially unchanged and in 40 years just 400,000 crystal structures are available in subject specific databases that harvest their content from the published literature. There are around 30 million chemical compounds known today and it is estimated that approximately 2 million crystal structures have been determined in research laboratories worldwide. Hence less than 20% of the data generated in the crystallographic area is reaching the public domain.

As high-throughput technologies, automation and e-Science become embedded in scientific working routines the publication bottleneck can only become more severe.

The Open Access movement provides a potential solution to this problem through the use of Institutional Repositories (IR) to manage, curate and disseminate academic research output. The scientific papers contained within an IR are either the authors' final version of a paper submitted to a journal or the reprint provided by the publisher. The information regarding these papers that is disseminated is purely bibliographic metadata and access, albeit unhindered, is provided only to the content of the paper as it would appear in a journal article. So whilst dissemination of, and access to, academic research articles is greatly facilitated, the paper is still detached from its underlying data.

The issue of dissemination and open access to the scientific data underpinning a research publication via an IR has recently been investigated by the JISC funded eBank-UK project (http://www.ukoln.ac.uk/projects/ebank-uk), two partners of which are investigators on this project. This project has developed a prototype Institutional Data Repository, populated with finalised crystallography results according to a specific and individual schema, in order to investigate mechanisms by which scientific data and experimental results may be disseminated. The Institutional Data Repository makes available metadata regarding the author, affiliation, dataset type and a number of chemical identifiers; the eBank UK project is involved with aggregating this metadata with related studies in the public domain, such as journal articles. Thus the dissemination of scientific data is enabled, but only in parallel to the associated journal article and implicit linking and aggregation between the two is difficult to achieve.

The eBank UK project is concerned with disseminating experimental results (via the development of individual schemas), whereas R4L is about collecting and depositing data and subsequently writing up the results. Currently, eBank supports only a single experimental procedure (crystallography) whereas chemists perform many analyses on compounds (e.g. Spectroscopic: Mass Spectrometry, Nuclear Magnetic Resonance, Ultra-Violet, Infra-Red, Raman; Crystallographic: Single Crystal or powder diffraction; Elemental Analysis; Thermal Analysis and ab-initio quantum mechanical calculations). An eventual goal is to link together, in the Institutional Data Repository environment, all the separate analyses on one particular compound in order to increase the scope of the scientific analysis. It is most likely that only a single repository is involved, as a particular institution/investigator is normally responsible for, interested in or co-ordinating the activities of interest on a specific, novel compound.

The laboratory repository is a separate entity from the institutional repository not out of architectural necessity, but in order to emphasise a difference in purpose and to ensure the development of appropriate policies (e.g. data storage, access, backup, archiving of raw [proprietary] data using the ATLAS datastore). The figure below presents a schematic of the Repository for the Laboratory project concept, depicting how the chemical compound characterisation and analysis process is supported by this approach.

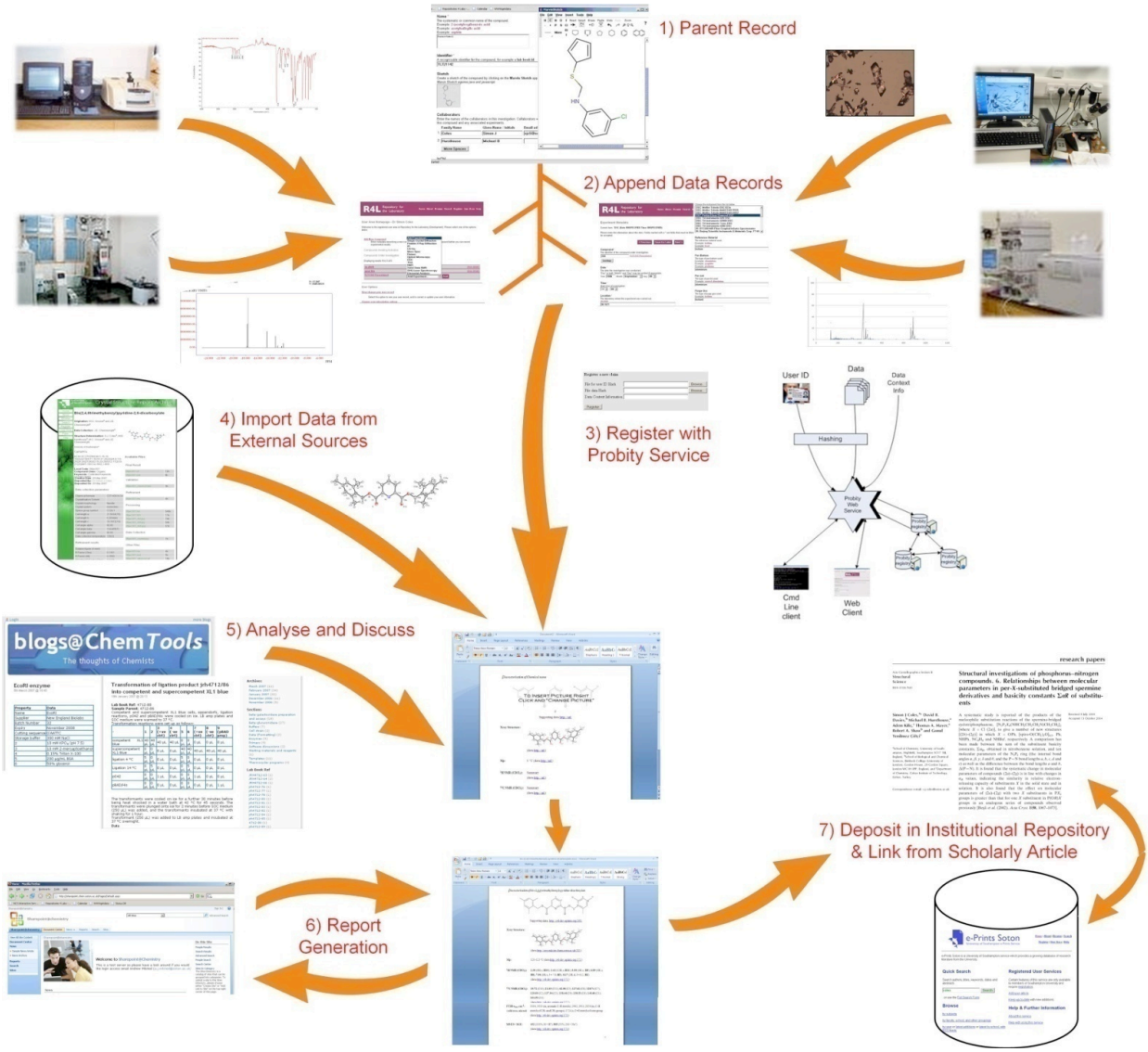


Figure 1

project, is capable of ingesting, storing, managing and presenting a cross section of some ten different types of data holding arising from different analytical techniques. A parent record in the repository (1) consists of high level chemical and identifier metadata for a particular chemical compound. Data records (2) may be appended to this parent record so that a researcher can drill down from the compound level to the underlying analytical data. The ingest processes have been carefully designed, following detailed analysis of laboratory workflows, in order to ensure complete capture of the raw, derived and descriptive data and thus provide a full provenance trail and support a comprehensive preservation process. A probity service (3), developed as a project deliverable provides a reliable and unique process for unambiguously registering the experimental data in a legally sound fashion. The R4L demonstrator is completed by the data discussion/analysis and report generation processes. Blog technology (5) has been employed to facilitate discussion and collaboration with respect to repository data by enabling 'live copy' type of transfer of data from the repository to the Blog space. The same live copy approach has been employed to demonstrate the writing of reports by pasting data into publication 'templates' (6).

Aims and Objectives

The project sought to enable the scientific reporting process to keep up with the speed of modern scientific analysis and to improve the accuracy, quality and reusability of scientific reports. It applied repository technology to experimental data capture, analysis and reporting processes to enable linking between datasets and articles, and also between related datasets. The primary outcome of this research is an exemplar system demonstrating the impact of an Institutional Data Repository on the analysis and dissemination of experimental scientific data in a subject that is crucially reliant on such studies. Requirements studies from commercial stakeholders from opposite ends of the scientific experimental process (i.e. equipment manufacturers and learned society publishers) will lead to the development of these pilot services.

1) Consult with scientific equipment manufacturers (as represented by the project partners), data analysis software developers and 'instruments on the Grid' eScientists to derive methods and protocols to make raw experimental data available and richly annotated with metadata, as it is generated in the scientific laboratory.
2) Develop an automated OAIS INGEST process which deposits the experimental data and metadata directly into the Laboratory Repository.
3) Establish a pilot 'Priority Assertion' service to provide a legally sound guarantee of priority for 'first to invent' protection. This service is a scalable, co-operative service designed for an academic context and solving the problems of trust and openness inherent in the alternative (commercial) solutions. A full description of the service is given below and more information and the service itself are located on the project website.
4) Configure a Laboratory Repository to be capable of managing large numbers of heterogeneous scientific datasets.
5) Consult with of an advisory panel composed of scholarly society publishers (as represented by the project partners) to capture requirements for data oriented publishing, including data reference and citation.
6) Build a report editing tool which can integrate repository data into a journal article.

During the course of the project a number of findings and outcomes dictated some alterations to the emphasis of the aims of engaging instrument manufacturer and publishing communities. This arose from difficulties in demonstrating concepts, benefits and widespread uptake and therefore it was agreed to develop a demonstrator system to show case to these communities and promote adoption by practising researchers.

Methodology

The principal aim of the project was to develop a digital repository to enable seamless and effective capture of experimental chemistry data, as it is produced by analytical instruments in the laboratory. A further aim was to make this experimental data available in a fashion that integrates well with current data acquisition and publishing procedures and the project sought to develop tools, based around the repository to facilitate this.

The overall approach that was adopted is outlined below in a stepwise manner:

1. A hypothetical scenario (Appendix 1) was developed to provide an understanding of the problem to be addressed and inform the design process.
2. A prototype repository was constructed using a standard configuration of the EPrints software. This was done in order to enable laboratory scientists to deposit data, which would allow an analysis of the workflow process and range of data types. Experience gained from this initial implementation and analysis work would then allow design of generic record structures and ingest processes.
3. From the standard configuration repository a detailed analysis of laboratory workflows (Appendix 2) was then performed.
4. Further detailed analysis of file types and formats(Appendix 3) that are possible to capture was then performed. Publishers, experienced practitioners and research workers in the School of Chemistry (University of Southampton) were then consulted on the required file formats for publication and dissemination and a survey (Appendix 4) of open standards available for formats for different techniques performed and the translation to these formats planned.
5. The development of a generic framework for a repository record was then undertaken.
6. A generic set of metadata for a repository record was then derived.

7. The style and presentation of a record was then designed so that it could present the required information to enable the researcher to assess a dataset 'at a glance'.
8. The registration (probity) process was then designed, written and implemented.
9. The repository was then tested by further, more extensive, population.
10. Development of tools associated with the repository for the analysis and publication of data records.

The primary issue that was addressed by the methodology was that of usability, i.e. the design of a record to be compliant with publishers requirements, whilst also providing a useful mechanism for scientific analysis of the data. A deposit process that conformed to OAIS Ingest standards was developed through this methodology and addressed the issue of workflow analysis.

Implementation

A workplan based on the following key areas was developed at the outset of the project:

1) Manufacturer discussions
2) Analyse workflows and design an ingest service
3) Design and build a service to assert the fact an experiment had been conducted and by whom
4) Construction and management of a data repository
5) Requirements capture with the publishing community
6) Design and develop a report generation tool

The initial project plan included gathering views and requirements from both instrument manufacturers (ingest) and publishers and this was a key factor for the design of the repository. Instrument manufacturers are ideally placed to inform the design of the ingest process, whilst the publishing community would provide guidance on the metadata required for dissemination and the general structure and content of records and abstract pages. However, given that no prior work had been done in the field the project was primarily conceptual at the outset and only high level requirements capture from these stakeholders was possible. Additionally, after desk based research and discussions with researchers it became clear that there is a general lack of standards, a large number of file formats (Appendix 3) for a single technique and in some cases a large number of different instruments available for the same experiment resulting in the widespread use of many types of proprietary software and formats. These issues were evidently too large a problem to be overcome by a project of limited scope and accordingly it was decided that the approach to carrying out the project shifted to providing a demonstrator that would deliver a proof of concept to be used to fully engage the publishing and instrument manufacturing communities. Therefore the requirements capture for the proof of concept design was performed through surveying the chemistry community (collaboration with the JISC SPECTRa project) and conducting experiments specifically for the purpose of workflow and ingest process analysis.

The project therefore adopted a 'build one to learn and throw away' philosophy to initially provide a rudimentary repository for analysis by project staff, use and feedback from researchers and invite comment from other interested parties. To achieve this, an 'out of the box' implementation of the EPrints repository software was deployed with a number of configuration modifications. This repository had most of the conventional EPrints functionality removed and permitted depositors to provide some basic metadata (authors, title, chemical identifier) and upload any number of data files of unspecified format. This repository enabled the data deposit process to be trialled and experiences, workflows and requirements to be assessed, which provided experience allowing us to assess and define workflows, assess and define file formats and define the metadata to be captured. From this information gathering exercise it was possible to design an architecture for a generic data record and integrate laboratory workflow into the design of a generic ingest process. The new ingest process and record structure was then implemented into a second configuration of the standard EPrints software. It was the possible to fully populate the repository with test series of datasets, thereby testing the design and better engaging researchers, instrument manufacturers and publishers. The R4L repository was show cased to these communities at a workshop (eBank/R4L/SPECTRa Joint Consultation Workshop, London Hilton Metropole, October 20th 2006 - see http://www.ukoln.ac.uk/events/ebank-r4l-spectra/ for more details), where feedback was provided and requirements and socio-political issues discussed and analysed.

With a functioning and populated repository it was then possible to consider the services and cultures that would depend on it. Firstly a service to guarantee the priority and provenance of data was addressed. An architecture, based on an efficient cross-registration mechanism using different Web Services and a number of distributed probity registries, was designed and assessed with respect to the data contained in the repository. This service was then constructed and implemented as a stand alone web service and an API to enable integration with EPrints code designed.

Initial desk based research into the process of analysing, discussing and reporting the outcomes of analytical chemistry experiments and a survey of possible modern technologies to facilitate this indicated a number of possible approaches:

Experiment discussion tools were evaluated and included looking into chat, forum and blog approaches. Indications were that the blog approach would provide sufficient functionality to enable the researcher to perform a number of these tasks and blog software was built and configured to be suitable for storage, presentation and discussion of the outputs of scientific experimentation. This resulted in the provision of a blog service with a structure for holding experiment results (tabular format), live clipboard, image data support, discussion and annotation tools. The blog (http://chemtools.chem.soton.ac.uk/projects/blog/) was trialled in two ways – for storage, presentation and discussion of chemical biology data (conditions, reagent types and quantities, images, sequences, etc) for experiments where supervisor and student are working at different sites; automatic capture of data from machines and sensors (SHG experiment data and environmental laboratory data).

Requirements capture for visualisation, analysis and report generation tools was achieved by an analysis of the current instructions for authors of a number of chemistry journals, plus discussions and surveys with researchers and an evaluation of possible technologies. Initially it was essential provide a suitable display and layout for a record jump off page so that it could stand alone as a report and provide rudimentary tools for visualisation of deposited chemical structures and associated experiment data files. This resulted in the implementation of open source applets for visualisation of 2D structure drawings and 3D crystallographic structures. Additionally an applet for visualisation of different spectra and interaction with the underlying data was developed – this was due to a lack of suitability of open source software that could support different experiment data types for incorporation into the EPrints software. These additions to the repository abstract pages allow a record to act as a stand alone ‘experimental data report’ to such an extent that they can be linked to from formal documents and be immediately understandable by the research chemist.

As part of the R4L suite of tools to support analytical chemistry and the new project directive to provide an end-to-end demonstrator for the experimental process, it was necessary to investigate the provision of ‘formal report generation tools’. The first attempt in this area was to provide an environment for the analysis and comparison of a number of different analyses on the same compound. The first approaches were based on the IBM ‘Eclipse’ developer software, which allows a number of different feeds to be visualised in the same workspace and an investigation into a browser-based tool to be built as part of the project. Unfortunately, despite investing a considerable amount of time into these approaches, these routes were both found to be unsatisfactory. The Eclipse tool would have involved a considerable amount of work to configure to support the different data types in the repository and would be a heavyweight component requiring installation and configuration on the clients’ machine. The browser approach would involve more developer resources and considerably more time than the project could allow. Therefore, in order to achieve the revised goal of providing an end-to-end demonstrator we used a Microsoft Sharepoint server for the storage of ‘report templates’. This approach provided immediate integration through the use of a proprietary, but ubiquitous, environment and made it possible to produce a demonstrator that enabled the use of familiar Microsoft Office tools and methodologies to share reports and easily live copy data into report templates and thus complete the end-to-end proof of concept within timescale of project.

Outputs and Results

A survey of chemists use of information and communications technology tools and methods in their research was performed. This survey was broadly based on that performed by the SPECTRa project and was conducted as a questionnaire, through both paper-based and online approaches, and posed to the faculty staff and research workers in the School of Chemistry, University of Southampton. The questionnaire and raw results of this survey are available in Appendix 4 and a summary of the most salient outcomes is given below:

a. The respondents principally comprised postgraduate students (55%), postdoctoral workers (18%) and faculty staff (19%) and totalled 110 people.
b. The primary use of computers and the internet is for information researching, writing papers and reports, working up data and instrument control.
c. Computers are used regularly for everyday work, but much less so for social networking and other 'modern' uses.
d. Mainly highly established community standard applications, software and file formats are used, with less use of modern data sources.
e. There is a predominance for self teaching use of software, as opposed to being taught professionally.
f. There is still extensive use of printed paper copies of PDF files. These files are generally stored on personal computers without any structure or use of reference software. A researcher will have 100-1000 PDF files on their computer and prefers to communicate them by sending PDF's to collaborators.
g. About 66% have had to generate electronic supplementary information to supplement their journal articles.
h. Supplementary information is mainly generated and stored in proprietary formats, although there is considerable use of 'popular' formats (eg Microsoft Office).
i. There is a preference, or requirement, to keep a hardcopy of data as well as an electronic one.
j. Experimental and analysed data are generally kept on a group or instrument controlling computer, however there is often a need to keep a hardcopy (eg in a lab notebook).
k. About 66% have not heard of 'InChI', 'metadata' or 'JCAMP' format, whilst around 50% have not heard of 'DOI', 'Open Access', 'Semantic Web' or 'RDF'.
l. There is a considerable lack of awareness surrounding repositories and their function.
m. There is a requirement for search and discovery to be based predominantly on structure, formula, author or keyword.
n. The most attractive purpose of a repository would be for the storage of a 'permanent record'.
o. Most chemists would comply if deposit in a repository was a mandatory requirement of funding or publication, however virtually all are ignorant of what the position of these organisations is with respect to open access and deposit in a repository.

The results of this survey indicate that, although chemists regularly use IT and must prepare and provide supplementary information for their journal articles, there is little knowledge of data capture and management and only moderate interest in novel approaches to data publication and sharing. Standards are generally adopted in a de-facto fashion or when the publishing process demands their usage and few are aware of open or exchange formats for data. It is also apparent that knowledge of open repositories, particularly for data, is limited. Research data is only 'published' when a journal requires it in support of an article or when it is mandatory to deposit with a central database (the latter is relatively uncommon). However, research chemists do find the prospect of a system for data capture and management and the storage of a permanent record appealing. The survey proved to be very useful in understanding attitudes towards the use of IT in the open publication of research data and the primary conclusion from the outcomes is that it will be necessary to demonstrate the ease of use and benefits before adoption of this approach is likely to occur.

The primary output of the project is the populated data repository. The first implementation of the repository was a standard configuration of the EPrints software with most metadata fields removed and the remaining ones renamed and described so as to enable the deposit of experimental data. At this level the repository was capable of storing a data file in any format and associating with it some rudimentary metadata (authors, chemical name/title, experiment type, identifier). This functionality then provided a testbed for analysis of the requirements of a data repository, whereby practitioners in different fields of analytical chemistry were able to experiment with depositing data and provide feedback to inform the design. The prototype repository was also invaluable in the consideration of the design requirements arising from the experiment workflow and file format analyses. The experience gained through enabling chemists to deposit data in a prototype repository and conducting a number of associated studies provided a full set of requirements for the design of the demonstrator. An architecture for a generic record structure was devised which uses a chemical entity as a 'parent' which has a number of 'child' records linked or associated to it, where these subsidiary records are different analytical experiments which have been performed on the compound. The workflow and file format analyses then provided a modelling of the ingest process, which in turn enabled the determination of the metadata it would be necessary to capture. A new (demonstrator) repository was then constructed according to a model derived from the exercises outlined above.

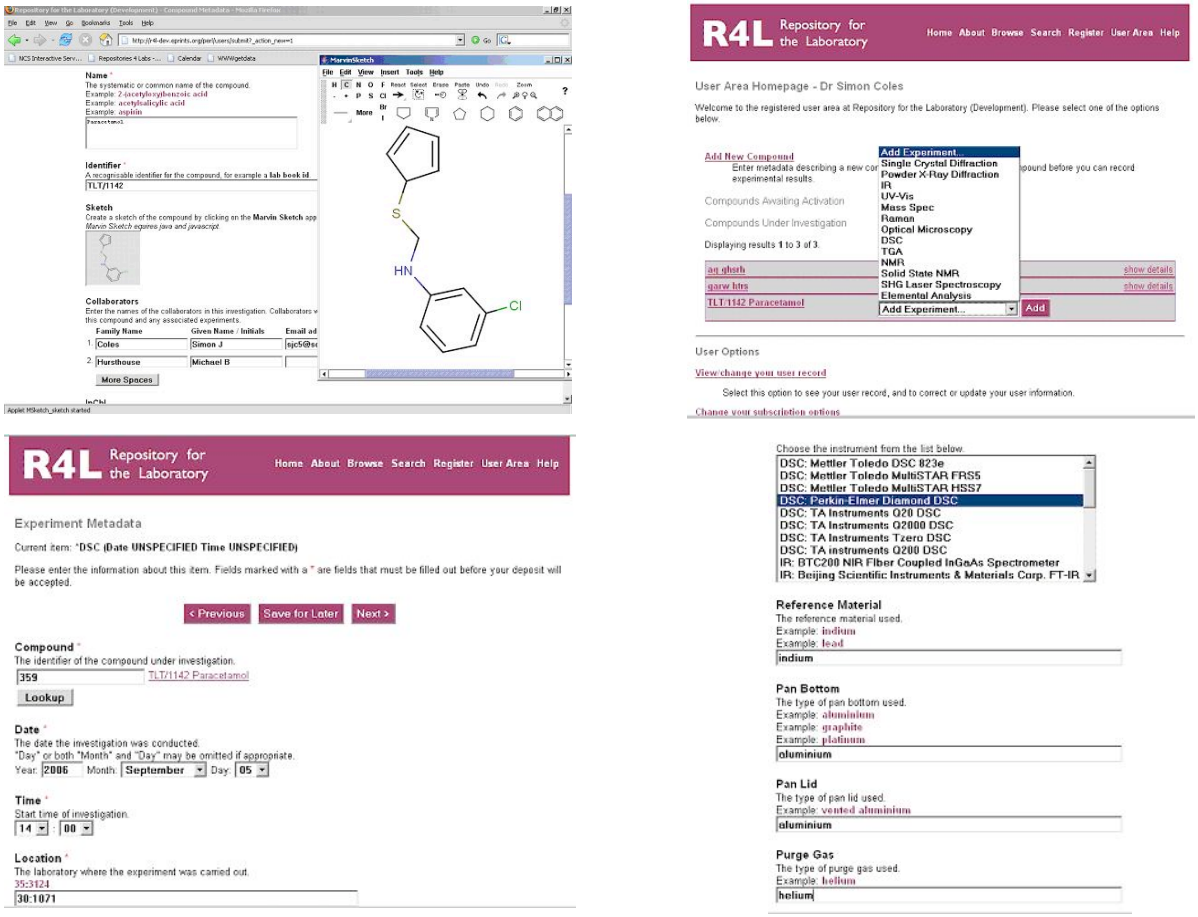


Figure 2. Screenshots of the repository deposit and ingest process

Refinement of the repository design was done through feedback from a selected group of research students asked to deposit their data as they were generating it. An original aim of the project was to incorporate the deposit, ingest and metadata capture processes into the workflow of the experiment performed on a particular instrument. To achieve this it would be necessary to fully engage individual instrument manufacturers so that 'hooks' could be put into their proprietary software to provide both data and metadata at the appropriate points. The project worked hard to engage this community, but it did not prove possible to achieve this aim, due to the significant amount of work that would be required from the manufacturers to provide these hooks and the difficulty in demonstrating the financial benefits of investing this time. It is clear that further progress needs to be made on the open

access agenda before scientific instrument manufacturers will buy into the concept, however the demonstrator repository produced during the course of this project is now proving to be a useful tool in engaging chemists, who would in turn provide the demand to influence the instrument manufacturers.

With a repository in place it was then possible to implement the design of the 'Probity' service, which, through a cross-registration process provides a mechanism to assert that an experiment has been performed on a particular compound, when and by whom.

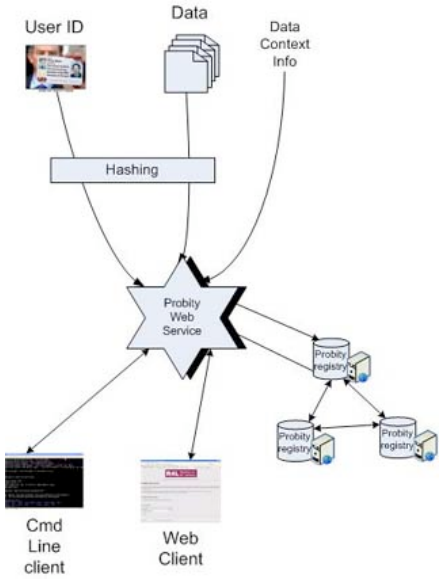


Figure 3. A schematic of the Probity service

The R4L Probity Service is a secure provenance service for laboratory-based experimental data and results. It enables researchers to register their findings and can guarantee the priority and provenance of registered data through an efficient cross-registration mechanism which uses a number of distributed probity registries. The service is implemented using a service-oriented architecture with different Web services interacting with probity registries in the registration and cross-registration processes. The main functionalities provided by the probity service include standard registration, browsing and querying of the registries, and the background cross-registration. A standard registration service takes a unique registrant (user) id (e.g. a scan of passport photo page) and the data which it encrypts and stores in a probity registry along with any relevant context information. The cross-registration process then occurs in the background between different probity registries and supports potential cross verification of data/results ownership claims by users. The priority of registrations is guaranteed not by the time but by the cross-registration mechanism. At its current stage the R4L probity service has several registries set up in the School of Chemistry and the School of Electronics and Computer Science, which it manages via command line and Web interface clients to store, query and browse claims in probity registries. It is also being integrated with the GNU EPrints software as a provenance service on R4L eprint archives.

The demonstrator repository was presented to select members of the publishing community, The Royal Society of Chemistry and Chemistry Central (as learned society and open access publishers respectively) for feedback and comment. The concept received a warm response and there was clear interest in the possibility that data could be 'self published' by an author, thus allowing all supporting information to be linked to from a journal article rather than it be missing or unnecessarily reproduced (with considerable information loss) in printed form in the manuscript. Concern was voiced over the longevity of such repositories, but there was considerable interest in the possibilities that could be opened up in the information and knowledge engineering fields by making data available in this manner. It was also noted that the abstract page of a repository record should be self explanatory and present the data in the form of a concise 'report', i.e. pertinent metadata displayed, graphical rendering of data and 2/3D visualisation of the chemical compound to which the data relates. These recommendations gave rise to some redesign of the abstract page and an investigation into technologies available for the rendering of spectroscopic data. After some considerable investigation and experimentation with various candidates it became clear that there was no suitable generic

software that could perform this for all data types in the repository. The project therefore designed and developed a browser applet that can display spectroscopic (and other) data in JCAMP-DX format and provide some querying of the underlying data.

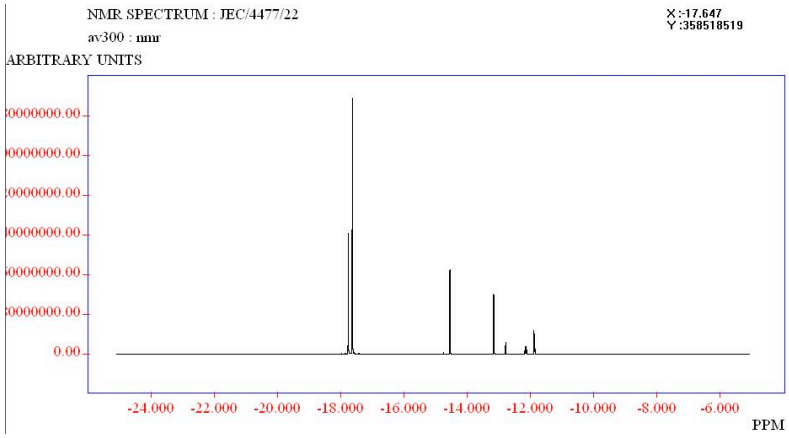
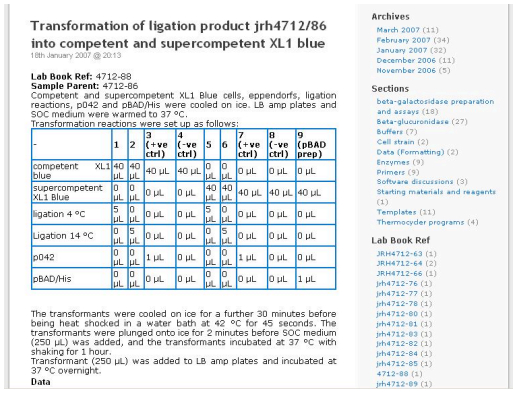
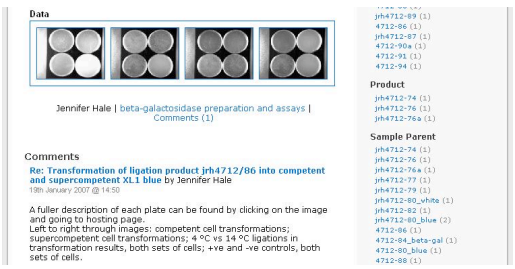


Figure 4. The data visualisation applet incorporated into the R4L repository.

With the repository constructed and capable of displaying and visualising data records in a fashion that allows interaction and interrogation of the data it was then possible to consider the discussion and analysis tools. Considerable problems were encountered when attempting to adopt current technologies or software and alter them to suit the requirements of the project – for example considerable effort was invested in the Bioclipse software, a scientific variant of IBM's Eclipse software, that enables numerous 'feeds' to be concurrently displayed and interactively interrogated. The project also considered development of a bespoke browser-based tool to fill this role, however eventually the blog approach was adopted, based on its suitability for collaborative work and the ability to easily 'upload' data. The project has developed a dedicated blog as part of the ChemTools suite of resources (http://chemtools.chem.soton.ac.uk/projects/blog/) – a School of Chemistry, University of Southampton initiative.





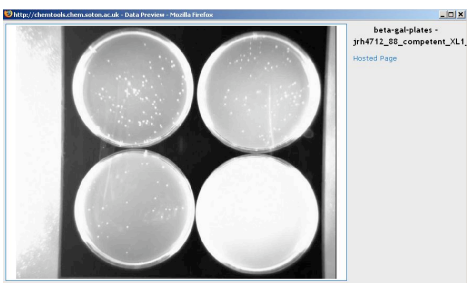


Figure 5. Screenshots of a typical blog record.

The figure above depicts a typical record in the blog for an experiment that closely mimics the common usage of a blog platform, i.e. the data for this experiment is a set of images, which can be uploaded to a 'photo gallery', from where a selection can be 'live copied' into the blog, where tables of 'derived' data can be constructed and the results of the entire experiment discussed with members of the 'group'. This demonstrator blog is now in every day use in this research project, where the supervisor spends 80% of their time at a different location from the group and this platform has proved crucial to maintaining the discourse over experiments conducted by geographically separated parties. The blog therefore in principal provides a platform where different data sets/streams can be pooled together in one record and provide an environment where they may be informally discussed. It is important to note here that posts to this blog may be 'open', as in the conventional blog sense, but data, threads and discussions may be kept private within a specified group of users. Establishing and proving the viability of the demonstrator blog to pooling data and discussion of the scientific experimentation process then enabled a deeper investigation into the flexibility of this approach. Further collaboration tools were then considered, where data posted to the blog were downloaded, annotated and commented on and then uploaded back to the blog for discussion.

This approach has proved very powerful, to the point that further development of these tools will be undertaken so as to provide an online annotation service to enhance scientific discussion through this medium. Further data sources were considered and the automated blogging of data by instruments and sensors targeted as an important approach to investigate. Further demonstrations of the power of this approach were developed, using a second harmonic generation experiment as the exemplar.

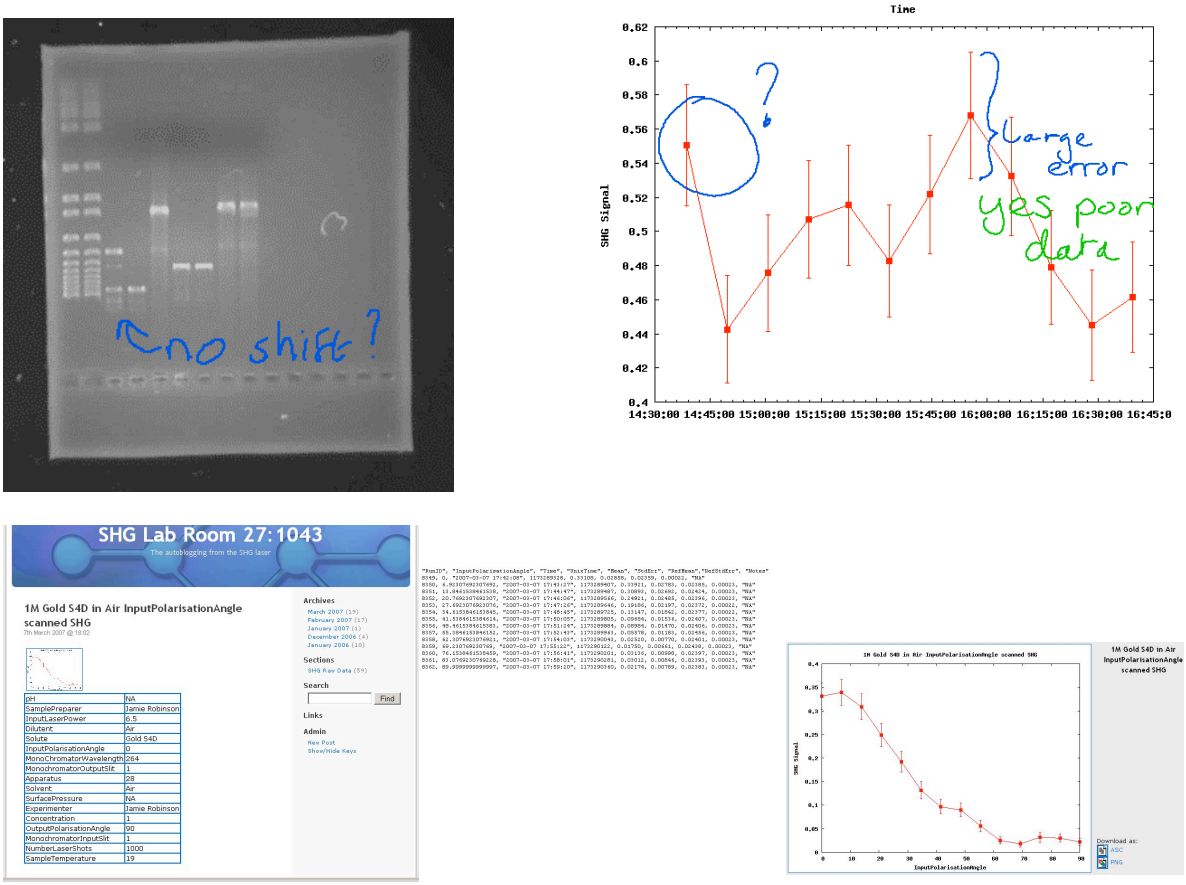


Figure 6 Examples of annotation and automatic blogging by instruments.

The figure above shows a post that has been automatically uploaded to the blog by the SHG experiment and made available to authorised users for discussion. This automatic posting approach is now being investigated for the recording of sensor data related to the environmental conditions in a

laboratory (temperature, humidity, human presence, etc) so that it may be associated with a particular experiment conducted at a particular time.

The R4L project therefore presents an end to end proof of concept demonstrator that serves as an introduction to the use of a digital repository as an approach to the effective capture, deposit, management, analysis and subsequent dissemination of all the data generated by a chemistry study, laboratory or instrument. See Figure 7 for an overall architecture showing the relationship between the repository, blog and sharepoint server, and how they feed into the ultimate production of the article. A final outcome of the project was the presentation of this approach to the chemical information, instrument manufacturer, publishing and data centre communities at a workshop, the transcript of which is provided in the report given as Appendix 5.

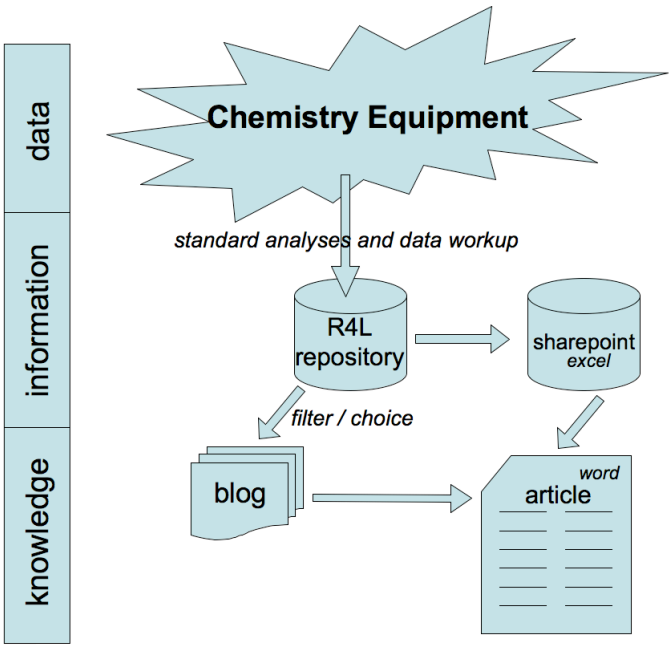


Figure 7. High level architecture and relationships in the R4L system

Dissemination

A list of conference papers and presentations, articles and workshop proceedings arising from the project work is given below.

- Coles, Simon J. (2007) A repository based framework for capture, management, curation and dissemination of research data. In, The 2007 Microsoft eScience Workshop *at RENCI,* Chapel Hill, USA, 21-23 Oct 2007.
- Coles, S. (2007) The Repository for the Laboratory (R4L) Project. D-Lib Magazine, Volume 13 Number 3/4, ISSN 1082-9873, March/April 2007.
- Coles, S. J., Frey, J. G., Hursthouse, M. B., Carr, L. A. (2006) R4L: The Repository for the Laboratory. Main R4L project outline poster.
- Coles, S. J. (2006) Institutional Data Repositories for Chemistry. Presented at the EBank/R4L/SPECTRa Joint Consultation Workshop on "Digital repositories supporting eResearch: exploring the eCrystals Federation Model", held at London Metropole Hotel, 20th October 2006.
- Warr, A. W. (2006) Digital repositories supporting eResearch: exploring the eCrystals Federation Model. Report of the EBank/R4L/SPECTRa Joint Consultation Workshop, London, 20th October 2006.
- Coles, Simon J., Frey, Jeremy G., Hursthouse, Michael B., Milsted, Andrew J., Carr, Leslie A., Gutteridge, Christopher J., Lyon, Liz, Heery, Rachel, Duke, Monica, Koch, Traugott and Day, Michael (2006) Enabling the reusability of scientific data: Experiences with designing an open

access infrastructure for sharing datasets. In, Designing for Usability in e-Science, Edinburgh, UK, 26-27 Jan 2006. Southampton, UK.

- Coles, Simon, Frey, Jeremy and Milsted, Andrew (0018) Curation of Chemistry from Laboratory to Publication "The curation of laboratory experimental data as part of the overall data lifecycle". In, UK e-Science All Hands Meeting 2006, Nottingham, East Midlands Conference Centre, UK, 18 - 21 Sept, 2006. Edinburgh, UK, National e-Science Centre, 8pp, 185-192.
- Coles, Simon J. (2006) Data management in the chemistry domain. In, Getting the Most Out of Data, Making the Most Out of Research, London, UK, 5 Dec, 2006. Southampton, UK, University of Southampton.
- Coles, Simon J. (2006) Developing institutional data repositories. In, JISC Data Cluster Consultation Workshop, Rutherford Appleton Laboratory, UK, 10 Oct 2006. Southampton, UK, University of Southampton.
- Coles, Simon J. (2005) Open Archives as a Route for Capture, Dissemination and Access to Chemical Data and Information. At, eChemInfo Interaction Meeting, Basel, Switzerland, 9-10 Nov 2005. Southampton, UK, , 20pp.
- R4L Introduction poster, at the JISC Program Conference, Cambridge, July 7th - 8th 2005.
- Carr, Leslie, Coles, Simon and Lyon, Liz (2004) Archiving research data and research publications. At, Research Councils UK Workshop on Publication of Research Results, London, UK, 18 Oct 2004.

Outcomes

The primary outcomes of the project are as follows.

- The project has acted as a demonstration of the central role and importance of data repositories in the research data lifecycle to key stakeholder communities

- This work has shown the real complexity of scientific (chemistry) data and moreover the depth of the processes and protocols involved in its generation, exploitation and dissemination, providing an in-depth case study for further work in the area.

- This demonstration work has enabled chemists to understand the benefits of using a data repository and gain insight into new ways of working in the laboratory.

- This work has already informed the metadata categories of the standard EPrints 3.0 distribution, and will be taken up in greater detail by EPrints 3.1

- This work has contributed to the establishment of the Southampton University Data Preservation working group to evaluate the extent of the problem of research data curation and preservation of an entire institutions output and how it is integrated with the processes and policies of an established repository of research outputs.

- A direct consequence of this work is that publishers, such as the RSC (Learned Society) and Chemistry Central (Open Access) are now considering new routes for the publication of supplementary data in their journal articles.

- This proof of concept demonstrator has been used to write proposals for follow on funding. The exit strategy of the project is to seek further funding to embed the concept of a Repository for the Laboratory not only into the working lives of chemists, but also into the research data lifecycle. A number of routes are being followed to ensure that the findings and outputs from the project are taken forward:

    - The project intends to work with the OAI-ORE project and leading researchers in the chemical information field (Cambridge University, PubChem, Indiana University, Penn State University, Cornell University and Los Alamos National Laboratory) to adapt and apply their protocols to enable the description of complex compound chemical objects. This will make the exchange and amalgamation of chemistry data and information a seamless process.

- The outcomes and findings of this project are being fed into the myExperiment Virtual Research Environment project and will form a testbed for the sharing of chemistry data.

There are a number of communities that will benefit from the outcomes of this project:

- Chemists will be able to store their research data in a structured way so that it may be accessed in the future. The structured nature of these records will enable rapid and effective dissemination of data alongside conventional journal articles that consist purely of intellectual interpretations and hypotheses and are not cluttered with raw or underlying data.
- Librarians will benefit greatly through gaining an increased understanding of the nature of the problem of capturing and preserving (chemistry) research data.
- Publishers have a new model for publication of supplementary research data.

Conclusions

There are a number of conclusions that can be drawn from this work, which are principally that:

Repositories are well positioned architecturally and politically to serve as the foundation on which the research data lifecycle can be supported in the future.

It is very difficult to alter the opinions and approaches of researchers to storing and publishing research data. There are few drivers and benefits for these processes, which lead to an attitude of reluctance to change. Common responses include 'whats in it for me?' and 'what I do right now is sufficient to get published, so why should I change?' and therefore it is highly important to demonstrate benefits and introduce drivers of this new approach to addressing the problems of supporting the research data lifecycle.

It is imperative to establish a buy-in from the publishing community in order to provide the drivers necessary to engage researchers. A proof of concept demonstrator is a useful tool for engaging this community and provide the necessary evidence to generate the confidence required before they are prepared to contribute to development and adopt such technologies. More importantly they require a guarantee of the longevity of the data if it is to be linked to publications.

The long term preservation of research data is a fundamental question to raise in order to provide the longevity that the research and publishing communities require. This will require a radical new approach by research institutions to the preservation of the data they generate, which has to be underpinned by well thought out policies.

Instrument manufacturers also need to see the benefits that the community will soon be demanding before they will invest in this approach.

Implications

This project addressed a number of different areas and communities involved in the research data lifecycle and as such the future implications of this work are broad and far reaching. The implications for a number of communities are outlined below.

The work of this project demonstrates an entirely new infrastructure for supporting laboratory based science. Working within this infrastructure will provide chemists with a peace of mind and ability to recall all the data that they require to write up their experiments and subsequently make available for verification and reuse. This will change the way scientists work in the laboratory.

With a demand generated by the scientists desire to work in an environment where it is imperative to capture raw data in a structured way, instrument manufacturers will be under pressure to be compliant with this way of working. The R4L project demonstrates the both the need and the benefit of adopting this approach and provides an incentive to alter software to become compliant with open standards and these new approaches to the capture of electronic data in the laboratory.

The R4L project demonstrates a new approach to the storing and dissemination of scientific data supporting the scholarly publication process. The implications and benefits to the publishing community are enormous. Publishers are now showing an interest in this approach as it offers a route whereby all the data supporting a scholarly work may be made openly available alongside the article. This not only has the benefit that the article will be uncluttered with experimental data, but also enables the supporting information to be curated by the community, who have the necessary domain knowledge to know what to keep and how long to keep it for.

A highly significant implication of this work is that institutions must now consider more carefully how to handle the preservation of the digital data they produce. This requires a knowledge of what data is being produced and what must be stored and it is imperative that the research community is engaged and enthused to do this. A further implication is that institutions must consider the financial backing that this process requires and develop policies to support it.

Developing a highly structured architecture to enable the capture, storage and dissemination will have the effect of building a very solid foundation on which third party data services may be constructed. Therefore, one might envisage new types of informatics services based on open scientific data, such as data linking, mining, cheminformatics and follow-on calculations or simulations.

It is clear that for new science based on open data to be performed, the current open access protocols for describing repository content are insufficient. It will be necessary for new protocols to be developed to describe the content of complex scientific data objects in such a fashion that chemical data discovery and mining engines can fully interpret them and automatically assess whether that content is of use for their specific purpose.

Recommendations

There are numerous technical considerations that may be taken into account here, but there are high level primary recommendations arising from this work that are essential to highlight to any project working in this area:

1. It is imperative to engage the scientific community from the very outset when considering the concept of data repositories. Scientists are users of these systems and will not adopt them if they are unable to see the benefits, if interfaces and systems are difficult to use and if the data is difficult to discover or reuse.
2. A consideration of preservation and long term availability issues is imperative for peace of mind in the research and publishing communities and the digital libraries community should be engaged to develop new models for curation of **ALL** the digital research output from an institution.
3. Development of a standard core set of metadata for the high level description of a scientific dataset to enable preservation, dissemination, discovery and reuse is necessary. This can only be achieved by engaging all stakeholders and concerned communities.
4. Within the constraints of the project, SharePoint was a convenient (but temporary) proprietary solution for repository/desktop integration. This needs to be addressed natively in the repository platform, and is a development goal for EPrints 3.1
5. An increased level of integration is also needed between the repository and the blogging software platform.

References

- CIF standard http://www.iucr.org/iucr-top/cif/standard/cifstd1.html
- eBank-UK http://www.ukoln.ac.uk/projects/ebank-uk/
- EPrints http://www.eprints.org/
- IUCr publishing http://www.iucr.org/iucr-top/publ/index.html
- JCAMP-DX http://www.jcamp-dx.org/
- myExperiment http://www.myexperiment.org/
- Project Prospect http://www.rsc.org/Publishing/Journals/ProjectProspect/
- RIN http://www.rin.ac.uk/

- SPECTR-a http://www.lib.cam.ac.uk/spectra/

Appendixes

Appendix 1: The R4L Scenario

Appendix 2: Instrument, experiment and workflow analysis

Appendix 3: File Formats

Appendix 4: Survey questionnaire and response data

Appendix 5: eBank/R4L/SPECTRa Workshop report

Glossary of Terms, Abbreviations & Acronyms

ATLAS	A central facilities magnetic tape store
API	Advanced Program Interface
BLOG	An internet social networking forum for open discussion
CIF	Crystallographic Information File / Framework
CML	Chemical Markup Language
DSpace	An Open Source OAI-compliant institutional repository software platform developed jointly by MIT and Hewlett-Packard
Dublin Core	A metadata standard
eBank UK	A set of JISC-funded projects focussing on the management of crystallographic data
eCrystals	An institutional repository for crystal structures
EPrints	An Open Source OAI-compliant institutional repository software platform developed at the University of Southampton
EPSRC	Engineering & Physical Sciences Research Council
Fedora	An Open Source OAI-compliant digital repository software platform developed jointly by Cornell University and the University of Virginia
GRID	A series of nodes for high end computing on demand
InChI	International Chemical Identifier - a string of characters capable of uniquely representing a chemical substance
IR	Institutional Repository
IUPAC	International Union of Pure & Applied Chemistry
JCAMP-DX	Joint Committee on Atomic and Molecular Physical Data Exchange
JISC	Joint Information Systems Committee
JSpecView	A viewer for spectral data in the JCAMP-DX format
LIMS	Laboratory Information Management System
MDL Molfile	A proprietary file format, created and owned by MDL, for molecular information
METS	Metadata Encoding and Transmission Standard
MPEG21/DIDL	Moving Picture Experts Group standard 21/Digital Item Declaration Language - a metadata format for all digital objects
NMR	Nuclear Magnetic Resonance
OAI	Open Archives Initiative
OAI-ORE	Open Archives Initiative – Object Reuse and Exchange
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
OAIS	Open Archival Information System
Open Access	Free and unrestricted online access to digital scholarly research, including but not restricted to peer-reviewed research papers
Open Data	Digital research data that are accessible, free and without restriction
PDF	Portable Document Format
RDF	Resource Description Framework

RSC	The Royal Society of Chemistry
SPECTRa	A JISC-funded project focussing on the management of chemical data
SHG	Second Harmonic Generation
UKOLN	A centre of expertise in digital information management (originally "UK Office for Library Networking")