Experiences with Repositories & Blogs in Laboratories

Simon Coles, School of Chemistry, University of Southampton, UK. S.J.Coles@soton.ac.uk
Leslie Carr, School of Electronics and Computer Science, University of Southampton. lac@ecs.soton.ac.uk

Keywords: scientific data repositories, e-science, cyberinfrastructure, preservation.

Introduction

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) to supplement the IR documents & papers. The *Repository for the Laboratory* (R4L) project attempted to address the gap between these two areas: 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 collection with the document production environment through data description standards together with semantic relationships. 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 numerous data gathering exercises. 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 processes of gathering the data that underpins a publication can often be very expensive 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. 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. As part of the UK CombeChem project (http://www.combechem.org) analytical instruments and even synthesis labs were ‘put on the Grid’, enabling the digital output from these operations to be efficiently managed using Grid technologies.

These technological advances 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. 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. Some of these issues have been tackled by the EBank UK project (Warr 2006; http://www.ukoln.ac.uk/projects/ebank-uk/) by establishing a community-agreed standard for crystallographic data transfer built on OAI-PMH, promoting the use of data repositories in the field.

The issue of dissemination and open access to the scientific data underpinning a research publication via an I*D*R allows the dissemination of scientific data in parallel to the associated journal article, however implicit linking and aggregation between the two forms of information is difficult to achieve. 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

The R4L project builds on top of the experience of the Comb-e-chem and EBank projects to provide unaddressed repository functionality. 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 a national service). Figure 1 presents a schematic of the Laboratory Repository project concept, depicting how the chemical compound characterisation and analysis process is supported by this approach.

The laboratory repository should be 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 system 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).

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 were to 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.

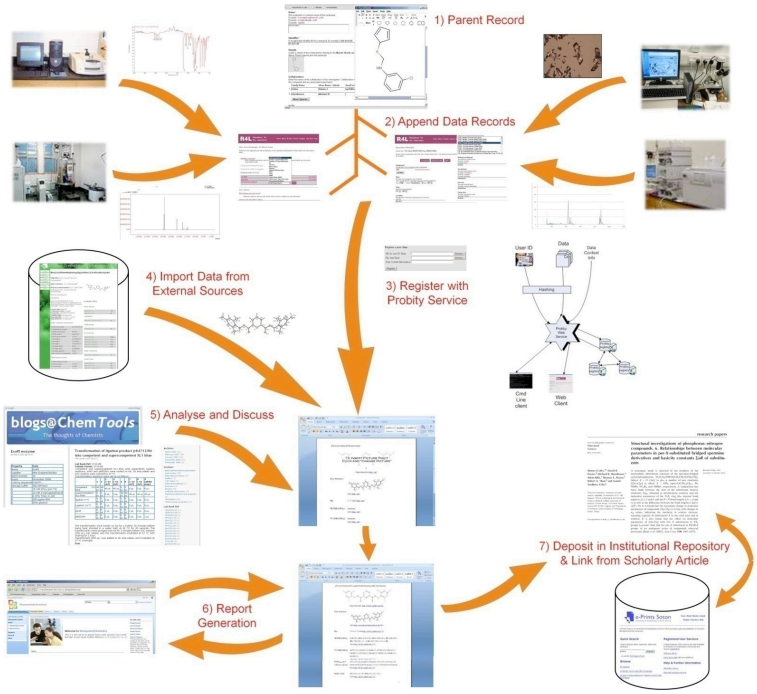

Figure 1: Processes surrounding the Laboratory Repository

The project initially provided a 'rapid prototype' repository for analysis by project staff, use and feedback from researchers and to invite comment from other interested parties. Built on an EPrints platform, 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. Workflow and file format analyses then modelled the ingest process, which in turn enabled the determination of the metadata it would be necessary to capture. 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 (see figure 2).

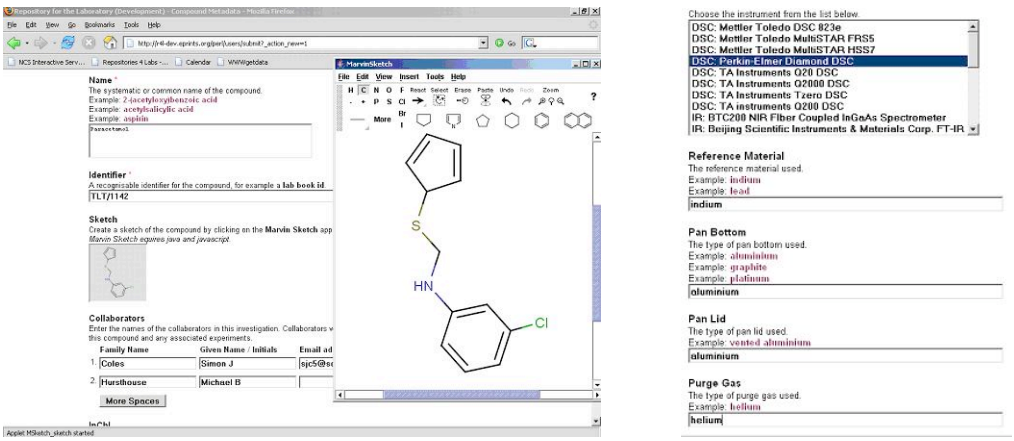

Figure 2. Screenshots of the repository deposit and ingest process

With a repository in place it was then possible to implement the design of the 'Probity' service, 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.

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.

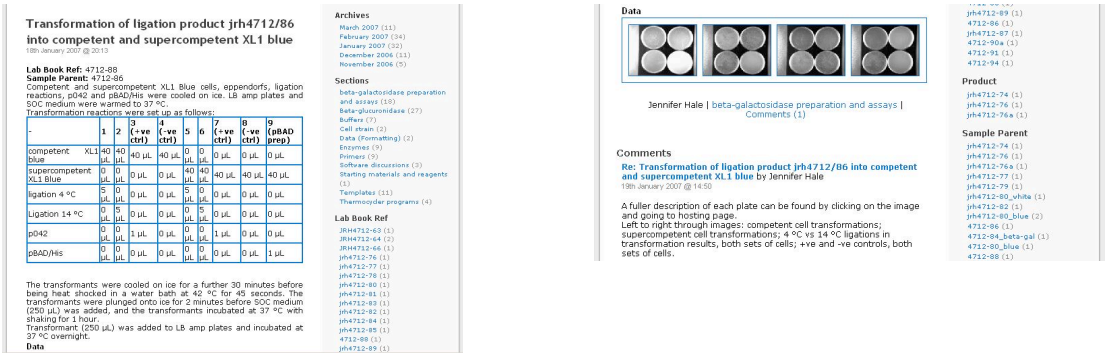

Figure 3. Screenshots of a typical Blog record.

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', but data, threads and discussions may be kept private within a specified group of users. 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.

Conclusions

The R4L project 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.

The work of this project demonstrates a 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.

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.

Bibliography

Coles, S. (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. http://eprints.soton.ac.uk/49395/

Coles, S. (2007) The Repository for the Laboratory (R4L) Project. D-Lib Magazine, Volume 13 Number 3/4, ISSN 1082-9873, March/April 2007.
http://www.dlib.org/dlib/march07/03inbrief.html

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. http://r4l.eprints.org/publications/docs/ebankspectrar4l_workshop.pdf