UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Chemistry

Facilitating Chemical Discovery: An e-Science Approach

by

Andrew J. Milsted

Thesis for the degree of Doctor of Philosophy

February 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Chemistry

Doctor of Philosophy

FACILITATING CHEMICAL DISCOVERY: AN E-SCIENCE APPROACH

by Andrew J. Milsted

e-Science technologies and tools have been applied to the facilitating of the accumulation, validation, analysis, computation, correlation and dissemination of chemical information and its transformation into accepted chemical knowledge.

In this work a number of approaches have been investigated to address the different issues with recording and preserving the scientific record, mainly the laboratory notebook.

The electronic laboratory notebook (ELN) has the potential to replace the paper notebook with a marked-up digital record that can be searched and shared. However it is a challenge to achieve these benefits without losing the usability and flexibility of traditional paper notebooks. Therefore using a blog-based platform will be investigated to try and address the issues associated with the development of a flexible system for recording scientific research.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Andrew J. Milsted , declare that the thesis entitled *Facilitating Chemical Discovery: An e-Science Approach* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:.....

Date:.....

Acknowledgements

The author would like to thank Professor Jeremy Frey, Doctor Simon Coles and Doctor Colin Bird for their support and advice during the completion of this thesis.

Most importantly a huge thank you to Sarah for the love and endless patience over the years to help him through this work.

Chapter 1

Introduction

Any practicing scientist is taught to pay considerable attention to the planning and recording of all of their observations[2]. A laboratory notebook is the primary record of research and the medium in which these records are kept. Paper books have taken on this role for hundreds of years and have allowed scientists to collate their research into one place, i.e. plans, spectra, sketches of experiments, results and notes on observations.

In addition to observations some, but not all, researchers use a lab notebook to document their hypotheses[3], initial analyses or interpretations of their experiments. The notebook serves as an organisational tool, a memory aid, and can also have a very important role in protecting any Intellectual Property (IP) that arises from the processes and outputs of research. Within this ideal scientific record enough detail should be given to provide another researcher, who is assumed knowledgeable, sufficient information to be able to repeat the work.[4, 5]

To quote from Day:

> Faraday's hand-written notebooks... have long been of interest to historians and philosophers of science because of the extraordinarily direct insight they give into the way his thinking developed... They are also remarkable in the amount of detail that they give about the design and setting up of experiments, interspersed with comments about their outcome and thoughts of a

> more philosophical kind. All are couched in plain language, with many vivid phrases of delightful spontaneity...

As more and more analytical chemistry techniques are becoming computerised (Nuclear magnetic resonance (NMR), Mass spectrometry (Mass Spec), Infrared spectroscopy (IR) etc) it has become obvious that there is a need to store this data in a secure repository. This stored, archived and managed data can then be linked by the individuals 'e' labbook. With all the systems being networked this can then eliminate the risk of data becoming lost through poor understanding or approach to personal data storage and management. In addition to storing this data in a structured and managed form, it can then become part of the formal publication process - either by incorporation into a journal article or by linking to it.

Using paper is an adequate curation method of the 'Permanent Record of Science'[6], but paper is fragile, hard to replicate and requires large amounts of physical space. Can the computer help? Technological advances have led to the development of tools that enable the scientist to carryout their work in an a safe digital environment. There are many additional benefits that come from using the electronic medium such as backups, linking, data management and curation[7].

Computers and technology expand the possibilities for this information to be stored[8] and ultimately shared for future use. Using tools like blogs and wikis, the discussion between researchers can be supported, sharing this information appropriately enables effective collaboration between colleagues overcoming distance/time zone barriers. With all this discussion being recorded it can add value to current practice by making it part of the permanent record of science. Researchers shouldn't have to do extra work to use these tools as this will hinder take up, but by ensuring they compliment the current working practices of the scientist, they are more likely to be adopted.

As an example, as little as 15 years ago analysis instruments (eg NMR, Mass Spec, IR) only outputted the results data on printed paper and many of these are still in active use. The mindset of using this printed copy to analyse the results can definitely still be seen today. Researchers still print out their spectra and take their ruler to measure peaks,

even though computerised solutions exist. In many cases this is probably because the young researchers are being taught and expected to replicate their supervisor's methods. When this is working practices it is easy to forget the preservation of the underling data, which also an important part of the permanent record not just the copy they have in their lab note books.

Recording what 'has been done' and 'with what', is another important aspect of the lab notebook. In chemistry in particular, when working with hazardous materials, noting what safety steps were implemented is important for any possible future health related legal issues, the downside of this is that this information has to be stored in excess of 40 years[9], which can present a physical strain on institutions. Employers have to provide space for this storage, whilst also ensuring these records are searchable and readable long into the future.

1.1 Electronic Lab Notebooks

Electronic Lab Notebooks could help provide a solution to the problems amentioned above, there are three main themes that permutate the general discussion of ELNs:

- whether they represent evolution or revolution
- the replacement of paper notebooks
- and, albeit to a lesser extent, the pros and cons

While important, technology for the implementation of ELNs appears not to be a significant issue. Williams et al. provide comprehensive guidance regarding the expected content, organisation and format of a paper notebook, together with extensive advice about recording experiments, from planning through running to data analysis and conclusions. Their treatise also includes a brief introduction to ELNs and ends with a discussion of intellectual property issues.[10]

Back in 1994 Borman took the view that ELNs could revolutionise how scientists record their research, manage their data, and share their information with others[11], but more

recently, Lass adopted a more cautious view when discussing best practices for implementing ELNs[12].

> If not done correctly, moving from paper to an ELN will be perceived by scientists as a revolutionary activity. When the outlined process is followed, daily routine will be fully mapped to the ELN functionality, enabling scientists to continue documenting their experiments with minimal interruption. The movement to the ELN will be evolutionary and not revolutionary.

Hice asserts there is no single definition of an electronic notebook[13], owing to differing requirements in different areas: Hice gives specific consideration to instrument interfacing. His view is that ELNs will evolve to meet market demands and that the current line of demarcation between the different flavors of electronic laboratory notebooks may be a moot point one day with convergence with other forms of record keeping in a digital space being the highly likely outcome of current research and commercial efforts as an example some organisations use Microsoft SharePoint^{TM}for their record keeping requirements and more open collaborations often use Google Docs^{TM}).

Early opinions on the replacement of paper notebooks were comparatively radical: a 1998 study by the Collaborative Electronic Notebook Systems Association (CENSA) gave a list of reasons why paper notebooks are obsolete.[14] A 2003 editorial in Drug Discovery Today gave reasons for moving from paper to ELN [15]; other articles published at around the same time continued the radical view, one describing paper as *fundamentally flawed in the ability to share and manage data.* [16] Mullin concluded: Once researchers are forced to use ELNs, they will likely never go back to paper even if they are allowed.[17]

Taylor acknowledges that CENSA was ahead of its time, but credits it with providing the first definition of an ELN: it is notable that the CENSA definition would still be acceptable today. He also includes a timeline showing the evolution of ELNs from before 1990 through 2010, going on to outline the benefits of ELNs from the perspective of a scientist's desktop.[18]

In 2007, Du and Kofman published a technology review, using a measured comparison, based on return on investment (ROI) calculations, of paper and electronic notebooks to produce a list of requirements.[19] and Bruce made the interesting recommendation that *adoption of an ELN be voluntary certainly in the pilot phase, but often in roll-out, too. This really forces an ELN to prove value to those that will use it, and demonstrates faith in people to decide what works best for them.*[20] In October 2011, MacNeil blogged a comparison, citing a PLoS One paper published with an entire paper notebook as supplementary data, and making a sound case for the ELN in collaborative research. His reasons included flexible organisation, linking to other aspects of the experiment, and sharing with both the team and the wider scientific community. Unsurprisingly, he also used this opportunity to extol the virtues of the iPad ELN.[21] In a similar vein, Elliott demonstrates how paper notebooks inhibit knowledge sharing[22]. His Webinar sets the scene by outlining the characteristics of paper notebooks, and then proffering a view of knowledge management that involves "a cultural migration to sharing, reusing and creating knowledge". His examination of explicit and tacit knowledge, and the four possible transitions between these types, leads him to conclude that paper notebooks obstruct all four such transitions, whereas ELNs assist them.

In the same year, Butler considered the pros and cons of electronic notebooks.[23] Successful adoption will undoubtedly depend on meeting the personal as well as the technical needs of researchers, as illustrated by an ethnographic study of scientific record keeping.[24] Among its conclusions were the words: *The experiences of the group of scientists studied in this article suggest that standardization of data entry is not the only aim of scientists when creating records.* The empirical evidence supports the stronger view that standards are far from being the top priority of those creating records in the laboratory.

Several writers have reviewed the evolving use of ELNs by pharmaceutical companies.[25, 26, 27] Recently, Kopach and Reiff have considered how ELNs can facilitate the calculation and reporting of green chemistry metrics associated with the development of new drug candidates[28]. This is an example of how new services can be built on top of

digital data in a way that would be far too time consuming to do based on traditional paper records.

1.1.1 Social Media

Over the last decade, digital technology in general and social media in particular, has changed the way people interact and communicate. This has not only affected people socially by using such tools as Facebook, Twitter and blogs to communicate amongst friends, but has allowed researchers to interact using the same tools. The content may be very subject specific but it still allows researchers to openly share ideas and results with their peers. An effect on the publication process that blogs and other self published items have lead to the bluring of the lines about what has been pre-published. Does publishing your lab notebook online through a blog or ELN count as pre publication?

1.1.2 Open Science

Macneil has discussed collaboration from an Open Research perspective and in the context of data publication in a post about the Encyclopaedia of Original Research (EOR).[29] Group openness in pre-competitive research is also relevant in this context. The Pistoia Alliance came into being in 2009, following an earlier meeting in Pistoia, Italy, with a mission to facilitate collaboration and innovation at the precompetitive stage of life sciences research. Data management and sharing issues are of particular interest, leading, amongst other topics of mutual interest, to collaborative consideration of the requirements for ELNs.[30]

For some, the apotheosis of collaboration and sharing in the ELN context is Open Notebook Science. The UsefulChem project, led by Jean-Claude Bradley, illustrates this concept.[31] Bradley is also a co-author of a book chapter entitled “Collaboration Using Open Notebook Science in Academia”, ten pages of which are devoted to the history and evolution of the UsefulChem project.

Dial-a-Molecule is a Grand Challenge Network that aims to generate a transformation in the speed of molecular synthesis: the vision is to make molecules in days rather than

years. Open access to the results of synthesis experiments, successful or otherwise, is critical to realising this vision, as is the ability to process those records automatically.[32, 33] The open publication of scientific work in this manner will not suit all researchers, and can prevent applications for patents based on that work. On the other hand, companies might still find this form of collaboration advantageous if carried out inside a corporate firewall. [18]

Chapter 2

Repositories

2.1 eBank and eCrystals

The eBank project was an initiative set in the context of the JISC Information Environment development fund, supporting end-users to discover, access, use and publish resources as part of their teaching, learning and research activities. eBank UK brought together chemists, digital librarians and computer scientists in an interdisciplinary collaboration which explored the potential for integrating research datasets into digital libraries by using common technologies such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). eCrystals was a repository solution built on ePrints[34] to make available crystallography structures openly.

Technological advances in computing, instrument manufacture, and now e-science over the past three decades have led to an acceleration in the rate at which crystallographic data are generated. In addition, the general route for the publication of a crystal structure report is coupled with, and often governed by, the underlying chemistry and is, therefore, subject to the lengthy peer review process and tied to the timing of the publication as a whole. Mechanisms for the publication of a crystal structure report alone exist through the Acta Crystallographica series of journals [35], but these still remain fairly time-consuming procedures, as a full report must be written and subjected to peer review and editing. These journals are open access and require authors to pay a

submission fee to finance the journal, this is also a hurdle to publishing crystallographic data as it requires authors to 'want' to publish their data given the cost.

A solution to having to author a full crystallographic report and to have to pay for a publication is to adopt the Open Archive Initiative (OAI)[36] approach to the dissemination of information. To improve dissemination of published articles, this method allows researchers to share metadata-describing papers that they make available in institutional or subject-based repositories. [37] Building on the OAI concept, an institutional repository was devised that makes available all the raw, derived, and results data from a crystallographic experiment [38], with little further researcher effort after the creation of a normal completed structure in a laboratory archive. Not only does this approach allow the rapid release of crystal structure data into the public domain but it can also provide mechanisms for value-added services that allow rapid discovery of the data for further studies and reuse, while ownership of the data is retained by the creator. For publication without the peer review process, it is essential that all the necessary provenance information is provided so that users can access all the data generated during the experiment and then use this to self-assess its validity and determine the exact processes used derive the crystal structure report.

2.1.1 An Open Access Crystal Structure Report Archive

The archive is a highly structured database that adheres to a metadata schema[39] which describes the key elements of a crystallographic data set. The schema requires information on bibliographic and chemical aspects of the data set, such as chemical name, authors, affiliation and so forth, which must be associated with the data set for validation and searching procedures. Because 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 data set to the archive. The elements of the schema and a brief description of their purpose are given in Table 2.1.

Table 2.1: Metadata Elements in the Open Archive Schema

metadata element name	content description
EPrintType	type of entry (e.g. crystal structure)
Subject	subject discipline (e.g. crystallography, chemistry)
Title	IUPAC chemical name
Creator	author(s)
Affiliation	Institution(s) of author(s)
Publisher	Publisher of a dataset (usually the institution)
ChemicalFormula	Formula of compound or moieties (according to IUPAC convention)
InChI	International Chemical Identifier (unique text identifier for a molecule)
CompoundClass	Chemical category (e.g. bio organic, inorganic)
Keywords	Selected keywords (provided as a limited ontology)
AvailableData	Stages of the experiment/determination for which datafiles are present (e.g. data collection, refinement, validation)
PublicationDate	Date when entry was made publicly available
Rights	Intellectual Property Rights exercised by the publishing institution

The metadata presented to the OAI interface falls into two categories. Institutional repositories are a mechanism for disseminating articles published in peer reviewed journals, and a protocol, Dublin Core (DC),[40] has been developed for describing the bibliographic metadata that is made publicly available. The metadata that are disseminated by this protocol are EPrintType, Subject, Title, Creator, Affiliation, Keywords, PublicationDate, and Rights. The ePrintType is set to crystal structure, the Subject is chemistry, and the Title is an International Union of Pure and Applied Chemistry (IUPAC) chemical name. It is important to note here that the generation of an IUPAC chemical name is not a trivial matter, and a combination of chemical expertise and software routines are currently required to perform this task. The recommendations in the guidelines and documentation for usage of this archive follow the current IUPAC conventions for generating a chemical name, as given in the Colour Books. [41]

A protocol for describing bibliographic information is insufficient for the dissemination of metadata regarding data sets. Fortunately, the DC protocol contains a route around this problem by provision of Qualified DC, which allows for the description of terms not contained within the kernel DC. The descriptions of terms falling into this category

are made publicly available as an extensible markup language (XML) schema, so that any third party wishing to make use of this metadata may understand its meaning and incorporate it into their schema and processes. The metadata that is described in this manner are ChemicalFormula, InChI, CompoundClass, and AvailableData and are included as identifiers to be utilized for subject-specific services in the areas of discovery, harvesting, aggregation, and linking (See Table 2.1).

The chemical formula is included, with guidelines for its composition, as a specific identifier to enable search and retrieval. The InChI (International Chemical Identifier)[42] is a unique identifier which encodes the molecular structure as a simple text string, with considerably more levels of description than any of its predecessors. In a recent development,[43] the scope of InChI has been extended to include the phase of a compound, and the crystalline phase descriptor may now be included to denote the fact that a particular InChI has been derived from crystal structure data. The InChI is very useful however this is a development project, and there are still problems to overcome (e.g., description of polymers, complex organometallics, and polymorphs) before an InChI can be used to describe all crystal structure data[44]. As a result it is necessary to check the validity of a machine-generated InChI. The archive deposition tool automatically generates an InChI string from a crystallographic data set (via conversion to a MOL file.) As a text string InChI is easily machine-readable and is included in an archive entry for the purposes of highly specific discovery and linking in the broader chemical literature. Initial studies[45] with linking data in different public databases,[46, 47] on the basis of an InChI, have proven that indexing by the Google[48] search engine can give an exact match and may therefore potentially be used as a means of aggregating chemical information. The compound class element is for broad aggregation of data sets within the area of chemistry and is defined as organic, inorganic, bio-organic, or organometallic.

The “available data” declares what categories of the experimental process have files associated with them, and these are defined as stages thusly as follows: processing, solution, refinement, validation, and final result and other files (where any files not recognised by the schema are placed).

On completion of the refinement of a crystal structure, all the files generated during the process are assembled and deposited in the archive, a process that will be automated as part of future developments. The metadata to be associated with this data set 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 (see Figure 2.1) and an interactive rendering of a chemical markup language[49] (CML) file added for visualisation purposes.

For conventional publication purposes, a crystal structure determination would normally terminate at the creation of a crystallographic information file (CIF),[50] and this file would be all that is required for submission to a journal. However this archive enables publication of all the files generated during the experiment and moreover during deposition a number of additional processes are performed which provide added value to the study and enable discovery and reuse of the data. These processes are seamlessly performed by uploading all the files, up to and including the CIF, to a toolbox on the archive server which can perform the necessary additional services required for a full archive entry.

At this point validation of the structure is performed using the web service CHECK CIF.[51] The generation of the InChI and translation of the structure into CML format generates files for the final results stage which are machine readable, and therefore allow automatic processing of an entry by third parties. When deposited the new archive entry is queued to be further checked and signed off by an editor. A trained crystallographer would assume this editorial role and provide further validation of the data prior to making it publicly available.

In order to fulfill obligations of making data public by the Research Councils[52] a policy whereby all crystal structures determined will be made publicly available (unless specific reasons have been provided for withholding the data) on an open archive if the results have not been published within three years of the date of data acquisition. This policy ensures that a researcher has sufficient time to consider the results and prepare a publication (three years is deemed suitable as it is the timescale of a UK PhD

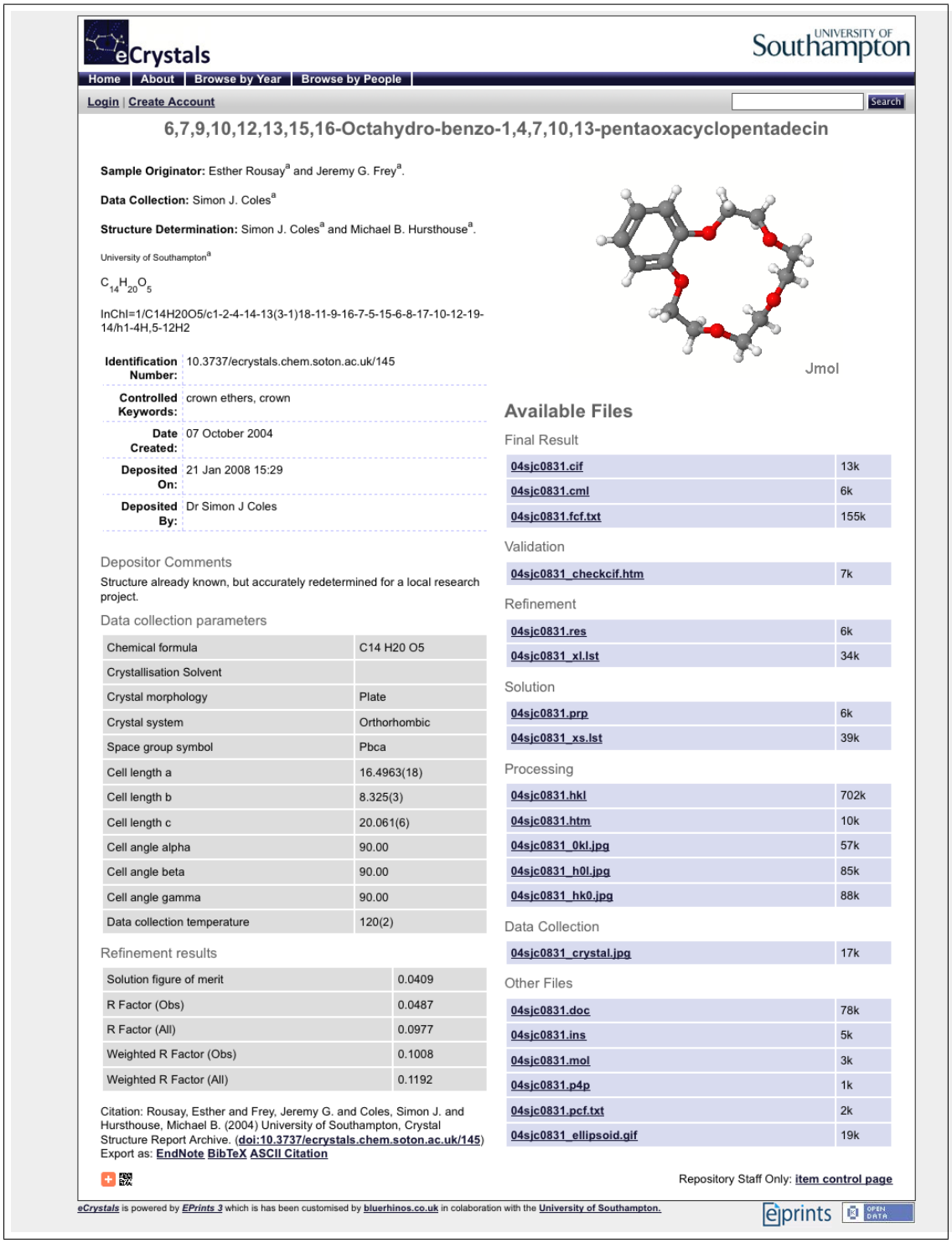


Figure 2.1: An archive entry for one dataset http://ecrystals.chem.soton.ac.uk/145/

studentship or a postdoctoral position), while enabling rapid dissemination if the result is not destined to be included in a traditional publication. The repository automatically releases the data after the embargo period ensuring the data is not forgotten. This policy is clearly stated on the repository website[53].

2.1.2 Upgrading to Eprints 3

eCrystals was developed using an iterative and incremental development methodology. This allowed users to get using the software as soon as possible, enabling feedback to be collected and fed back into the development cycle. A development beta copy of the repository was set up to enable users to freely test the functionality of eCrsytals. The feedback was collected and managed by informal meetings with the developer and the crystallographic team (users), this enabled the development to happen rapidly and efficiently

The eCrystals archive is built on top of EPrints, when EPrints released a new version of their software it became clear that eCrystals needed to be upgraded as well. The new EPrints architecture is built on a framework of plugins, which enables the development of crystallography specific plugins which can then simply be added into the EPrints framework.

There are some other repository software solutions available, DSpace[54] and Fedora[55]. These two are designed to handle the bibliographic information of digital objects, where as the requirement for eCrystals was that it had to maintain the data as well, not only for publication but also for local service management. EPrints was also a front runner as it is being developed by the University of Southampton and the local link helped with the development.

One of the main advantages of upgrading the archive was that the old version had a separate toolbox that was used to prepare all of the files. This meant that the user had to switch between different software tools when depositing the data. The new plugin setup enabled the integration of the toolbox functionality into the main archive software.

The embargo requirement also stretched the old archive, as not only did the access to the data files have to be restricted (which EPrints 2 supported) but all the metadata had to be hidden as well. This is because the metadata alone could also reveal the science being studied. The repository therefor supported two separate archives within it, one being the outward facing archive that only has the open records in it whilst the other is an internal one containing all records and management tools. The internal archive had all the web based security to protect the content and a series of scripts synchronised the two archives.

A major difference between the two versions of EPrints is that version 2 ran only off static pages, this meant that all of the webcontent was rendered once and only updated when metadata changed, but this results in only one copy of the content. EPrints 3 still has the static section of every record, but it runs a small amount of dynamic content for each page request, this means that if you are allowed to view content you can, but it is hidden from public discovery.

For security access to the archive must be considered on three levels. Firstly if you are the archive administrator or one of the editors then you can read the record. Secondly if you are a logged in user and your email address matches that of a creator of that record then you can view it. The last level, is if you have a 'magic url' with a secure key (eg http://ecrystals.chem.soton.ac.uk/500/?key=9be6fab34bc1f97252d801f77dfd8f0c2), this will then unlock the record for anyone not logged in. For example this may be used when a reviewer who needs to look at the data pre-publication. This last security measure is only as secure as the people that use it, but is no different than passing the data in an email and then forwarding it on.

Reuse of the data is a selling point of open repositories and just making the data public is not enough to make it searchable. Web search engines such as Google[48] enable finding records by Title, Author and even the InChI, but they don't offer an in depth domain specific search of the actual data. Therefor the archive has supplemented the OAI[36] interface used by most Institutional Repositories to share bibliographic information with some of the chemical identifiers and links to the data, so that domain specific harvesters can crawl the repository and feed a rich search engine.

The upgrade has resulted in a streamlined deposit process and a more secure archive. eCrystals has been running for a number of years since 2008 and now contains over 1000 structures, this means that it provids a significant contribution of many structures that would of otherwise not be published benefiting the crystallographic community.

2.2 Repository for the Laboratory, R4L

The chemistry subject domain like with many disciplines has been very insular when it comes to using software and technology. There have been many software tools that have worked with analytical instruments to help the researcher with their work, but unless it is part of a large analytical centre or part of a national facility it unlikely that it centres around any data management plan.

These bits of kit are typically described as bench top instruments and if the researcher is still using a paper lab notebook to record their research, it is common practice for them to print the results and attach them to the lab notebook. If the researcher is organised then they will save their data for their own archives, but it is very likely that the data will only be stored on the users personal computer with no back up op data management plan.

Providing a repository for this underling data enables and ensuring it fits in with their research workflow, that data can then be kept safe and preserved.

2.2.1 Loss of Data

Most publications in chemistry journals require the support of experimental and characterisation data, which is often supplied as electronic supplementary information. When data is published in the body of a journal article it is often only summarised in a few words in order to highlight the authors point. Often the rest of the data is ignored, as it may be deemed to have little relevance to the article. However if this data is required for reuse by a third party then it would be unusable unless the original data was made available. It is therefore crucial for the scientific dissemination process to be

able to make explicit links between experimental data and the article built upon that data. This objective is perfectly possible if the process is initiated in the laboratory and carried all the way through to the final article.

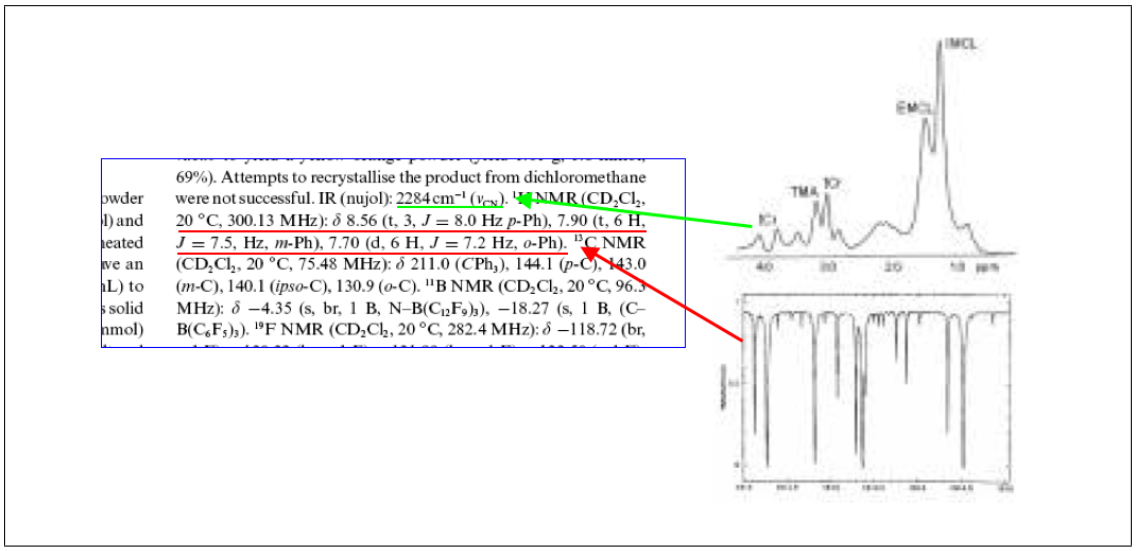
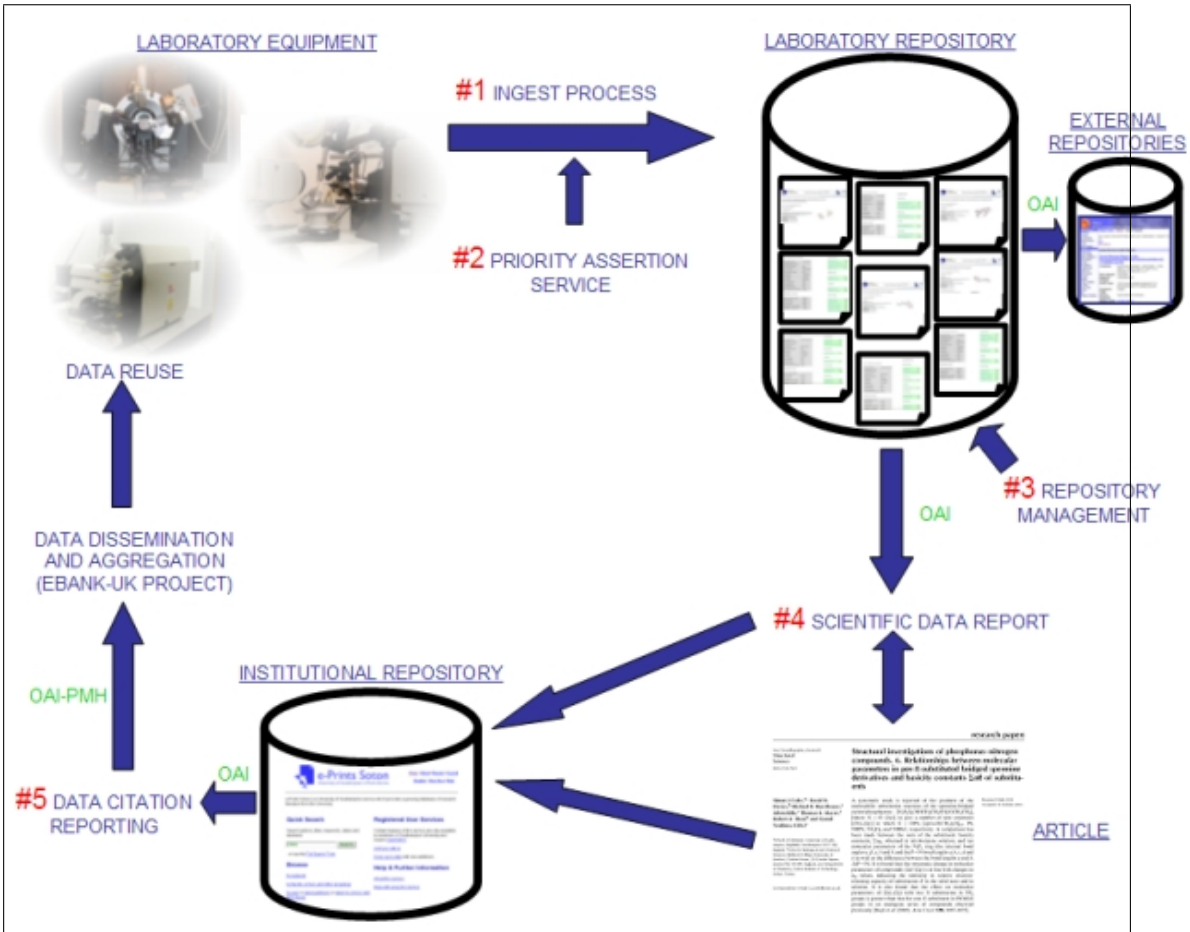


Figure 2.2: An example of data loss in a journal article

The Repository for the Laboratory (R4L) project was concerned with applying repository technology to experimental data capture, analysis and reporting processes in the Chemistry domain to enable linking between datasets and articles, and also between related datasets. The eBank project was extended to cover the whole scope of 'Chemistry Data', the repository would have to store many different data from many sources (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 exemplar system was developed to demonstrate the impact of an Institutional data repository on the capture, preservation, analysis and dissemination of experimental scientific data in a subject that is crucially reliant on such procedures.

Taking a repository-based approach provides numerous benefits to all the researcher roles involved in the research data lifecycle defined by R4L, as shown and outlined in figure 2.3. The demonstrator repository was built on existing repository software EPrints[34]. As with eBank, EPrints provided a very stable and easily configurable base for the R4L repository to be constructed.



#1 Data capture - A simple, if not entirely automatic, OAIS ingest process integrated into the researchers' workflow provides a seamless and comprehensive approach to recording provenance and capturing data that is properly indexed according to a prescribed metadata schema.
#2 Protection of IPR - Once the data collection and repository ingest processes are complete, a registration process provides a legally sound guarantee of priority, or 'first to invent'.
#3 Data management and preservation - A repository offers a simple solution to the issues of managing many different types of digital objects and ensures that they are available for future generations.
#4 Scientific data analysis - the ability for a scientist to aggregate various different analyses in a report form and link through to the data provides an extremely powerful form of correlating and cross referencing.
#5 Added value and availability for reuse - a scientist reading an article is able to access all the underlying and supporting data as well as the experimental processes that generated the data. Thus a scientist would be able to assess and understand the data and reuse it for further value added studies.

Figure 2.3: The R4L data capture model.

For R4L to work within the lab its workflow has to reflect and integrate with the scientists natural working process or it wouldn't be adopted by the scientific community. R4L's interfaces and menus were specifically and iteratively adapted to work with the scientist. For this reason the repository is molecule centric, where the researcher starts an investigation on a particular compound and then adds analyses to its record. The repository can then make all the results searchable by type and by compound, in addition to the conventional bibliographic terms, provided be the underling software.

2.2.2 Integrity

When data is stored digitally it has the age old problem of maintaining its integrity. This is especially true when data is stored within the laboratory, as it can be susceptible to doctoring by the user. The chances of this is greatly reduced if the repository is hosted at the Department or Institutional level but can never be guaranteed. This element of any data repository needed to be outsourced to a third party service that could ensure the validity of the data. Therefore a 'Probity' service was designed, through a cross-registration process provides a mechanism to assert that an experiment has been performed on a particular compound, when and by whom.

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 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

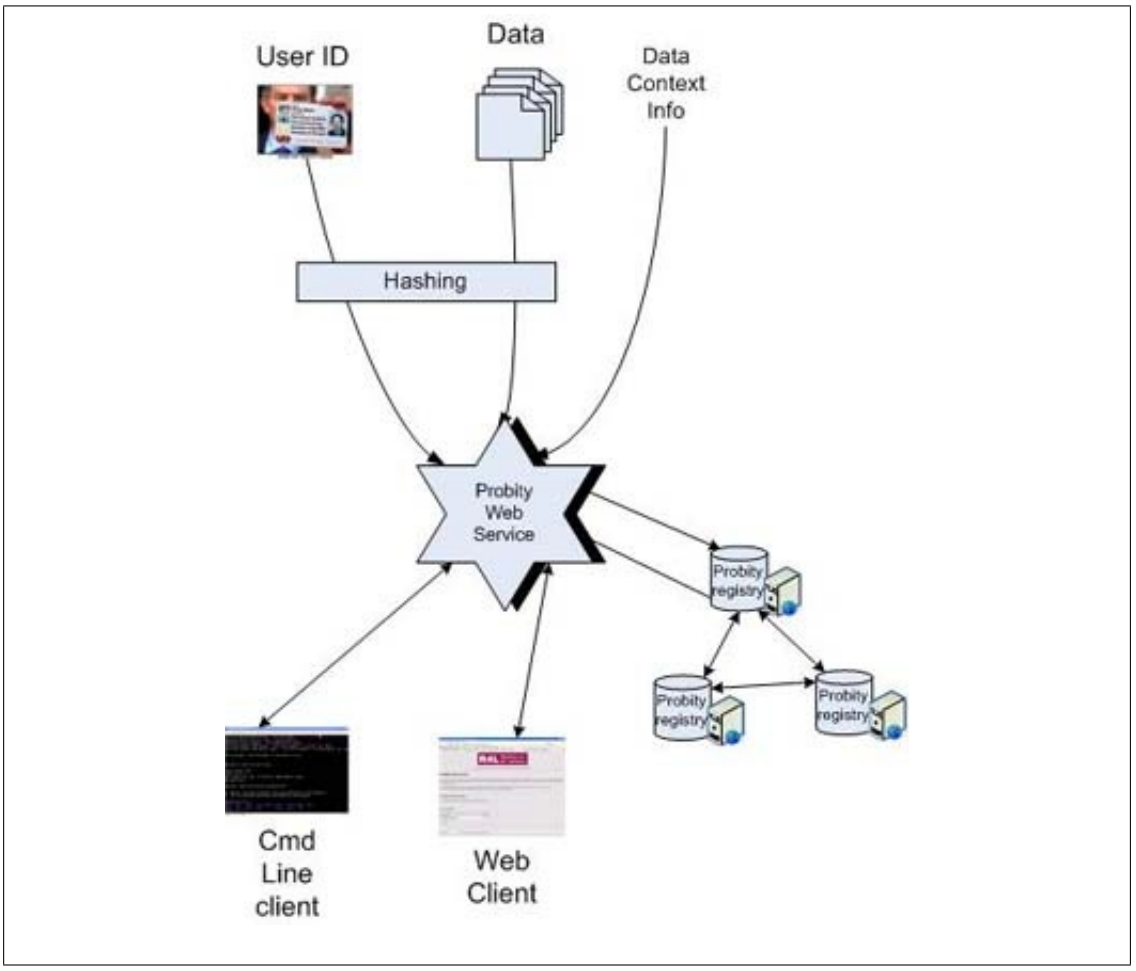


Figure 2.4: A schematic of the Probity service

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.

2.3 Conclusions

eCrystals enabled the release of hundreds of crystal structures that would other wise not of been accessible to researchers, this meant that many new scientific discoveries could be now made. This would not be possible if the records remained unpublished, as much research doesnt have the resources to speculatively create results but could easily perform analysis on the published results, especially if it is looking at many datasets.

R4L was taking this principle further making the repository accept many types of analytical data from different analytical technics. R4L also tried to address the problem of assertions of digital assets, in terms of ensuring records were as they were when the archive says they are with a probity service.

The author was responsible to writing all of the modifications and new code required to redesign eCrystals for the ePrints 3 upgrade. Where as for R4L the author led the design with team from ePrints Services (commercial part of ePrints).

Chapter 3

chemTools

3.1 Overview

As chemistry becomes more reliant on the need for computer-aided tools a challenge exists to make these tools easily available to researchers and students. By delivering these tools over the web it is possible to interact them with the minimum of software installation and configuration as all the researcher needs is a web browser and an Internet connection. It was decided that the chemTools[56] website should be created, in order to be the one stop presence for tools generated by the research groups projects. The site provides a directory of the tools and the software made available to the group, university community and the general public, an example screen shot in Figure 3.1. By hosting this on its own web-server under our control we are able to easily maintain and update the software tools allowing the users to take advantage of the upgrades.

In terms of usage the site gets about 2000 visitors a month with around 40000 hits. (Results From September 2008) Most of the traffic is from the University, but we have had visitors as far as Vietnam and Fiji. A lot of new user traffic comes from search engines, e.g. google, with common search terms such as, chemtools, sortase cloning, southampton, neutral drift, malaria projects, transformation buffer pipes. Most of these search terms refer to blog posts, but these results show that chemtools is ascending the

Figure 3.1: A Screen shot of the chemTools project page.

page rankings[57]. Another source of new visitors is via direct link from other sites, (eg http://blogs.nature.com/thescepticalchymist/ http://usefulchem.blogspot.com).

3.1.1 Single SignOn

To provide a security system to access some of the services on chemtools a sign on service was devised, it was initially intended to use the university log on system, LDAP[58], however this would restrict access to members of the University and the intention was to provide an open set of services for the global chemistry community, the advantage of the LDAP system is that the only personal data we our system is required to store is a username. Passwords, email address and full name are all stored on the University server. Employing this would have solved many data protection issues, and would of meant the end user didn't need to remember one more password. A hybrid solution was developed that allows all University users to use their LDAP login details and would also store the login information for external guests within the chemtools system.

3.1.2 Marvin

The Marvin[59] software suite is a freely available software suite for molecular drawing and manipulation. For example use see the The eMalaria Schools Project (Section 3.4.) The Marvin package comes with a very sophisticated user front-end that entirely runs as a Java applet. An up to date version of the Marvin package is made available on the chemtools site so that it was accessible to students to use as part of an Informatics course.

3.1.3 Web Services

After investigating the usefulness of web services in the eMalaria project, and the subsequent services being wrapped in SOAP, it made sense that all the underling services for manipulating molecules should be made available for other projects. To complete this task, the first services to be made available were the Marvin set, after which the following were added:

- **mol3d:** This performs a 3d optimisation, firstly used to produce well formed 3D structures of molecules, it is used also to produce models in the emalaria project in a form ready for docking.
  wsdl: http://phobos.chem.soton.ac.uk/soap/mol3d/mol3d.php?wsdl

- **mol2d:** This performs a 2d optimisation, which is suited to producing polished 2d diagrams.
  wsdl: http://phobos.chem.soton.ac.uk/soap/mol2d/mol2d.php?wsdl

- **moljpeg:** This creates a jpeg image from a submitted mol file, and when used in conjunction with mol2d can be used to make 2d structure sketches
  wsdl: http://phobos.chem.soton.ac.uk/soap/moljpeg/moljpeg.php?wsdl

- **molsmile:** This simple service returns the daylight smile[60] string from an input mol file.
  wsdl: http://phobos.chem.soton.ac.uk/soap/molsmile/molsmile.php?wsdl

- **smilemol:** This does the exact opposite of the above service by producing a mol file from a smile string.

  wsdl: http://phobos.chem.soton.ac.uk/soap/smilemol/smilemol.php?wsdl

Two non marvin services where also made available:

- **molopt3d:** This uses the original optimisation tool from the emalaria project, It uses more accurate semi empirical methods than Marvin but can only deal with core organic elements.

  wsdl: http://phobos.chem.soton.ac.uk/molopt3d/molopt3d.php?wsdl

- **inchi:** This produces the InChI[42] from an input mol file.

  wsdl: http://vanthoff.combechem.org:8180/inchi/wsdl/inchi.wsdl

```php
<?php
/*
 * mol3d test SOAP webservice
 * by Andrew Milsted (ajm3@soton.ac.uk)
 */

require_once('nusoap-0/lib/nusoap.php'); //get soap library

//input params
$param = array('molfile' => base64_encode(file_get_contents('test.mol')),'fast' => 1);

//initiate soap call
$client = new nu_soap_client('http://phobos.chem.soton.ac.uk/soap/mol3d/mol3d.php?wsdl', true);

//get result by calling the method
$result = $client->call('mol3d', $param);

?>
```

Figure 3.2: An example of how to call a soap wrapped web service.

To call any of these web services one can take advantage of the Web Services Description Language (WSDL), for example in php see figure 3.2 The process only takes 4 lines because the whole configuration is defined by the WSDL.

3.2 eLearning Tutorial on Regression Methods

In order to help students learn about regression and other statistical methods another teaching tool that was developed, the eLearning Tutorial on Regression Method. This

is an interactive site, that accompanies the notes on regression, which allows students to run and adjust many examples of statistical models.

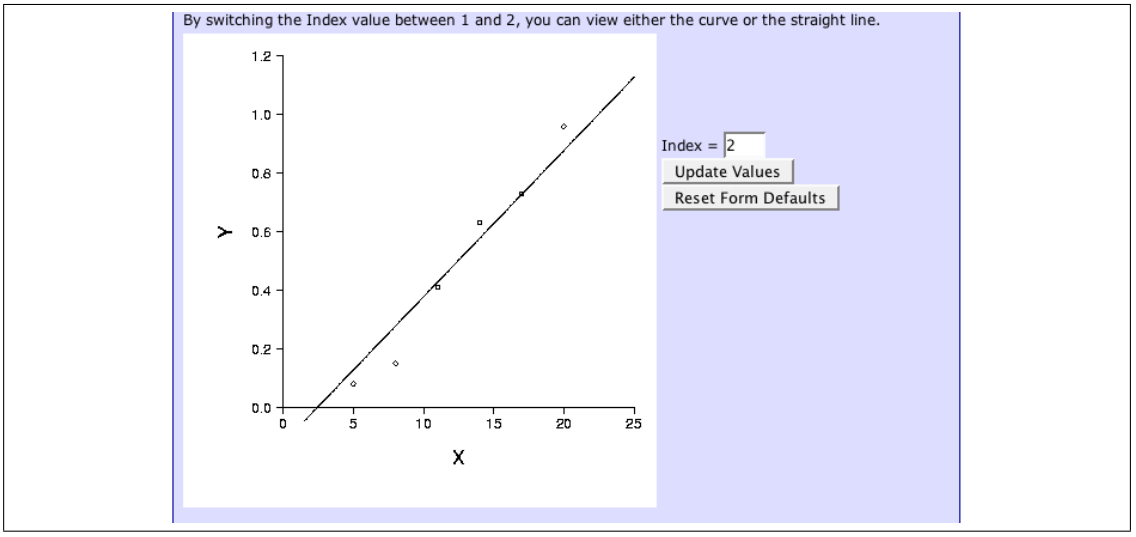


Figure 3.3: This shows the interactive part of the stats site.

This utilises the R[61] statistics modelling package that is hosted on a secure separate server protected by the group firewall. This allows the computational part of the site to be separate from and therefore to affect the main website feature. Under the heavy loading of a full class the web front-end remains responsive even if the computational results are delayed.

3.3 Proflocate

In order to explore the idea of location aware services we developed a tool that can potentially track a users location around the building. The main idea of this project was to collate data already available without any specialist hardware, eg Global Positioning System (GPS) or Radio-Frequency Identification (RFID). Currently there a few uses at the moment, the chemistry hosted LabTrove can select the closest uri printer to you to make printing labels easier (see section 6.2.3). Information relating to an experiment can be displayed on screens only when relevant interested person are in the room and when they are not the screen can be switched off saving energy.

The tool that tracks the location of electronic devices and has two parts. The first is a macro version that tracks personal computers. It does this by making a web request to

chemtools every 6 minutes, then the script on chemtools can then work out where the device is by the ip address. This can resolve locations down to a room in the chemistry department, to buildings within the University and which country when you get further afield.

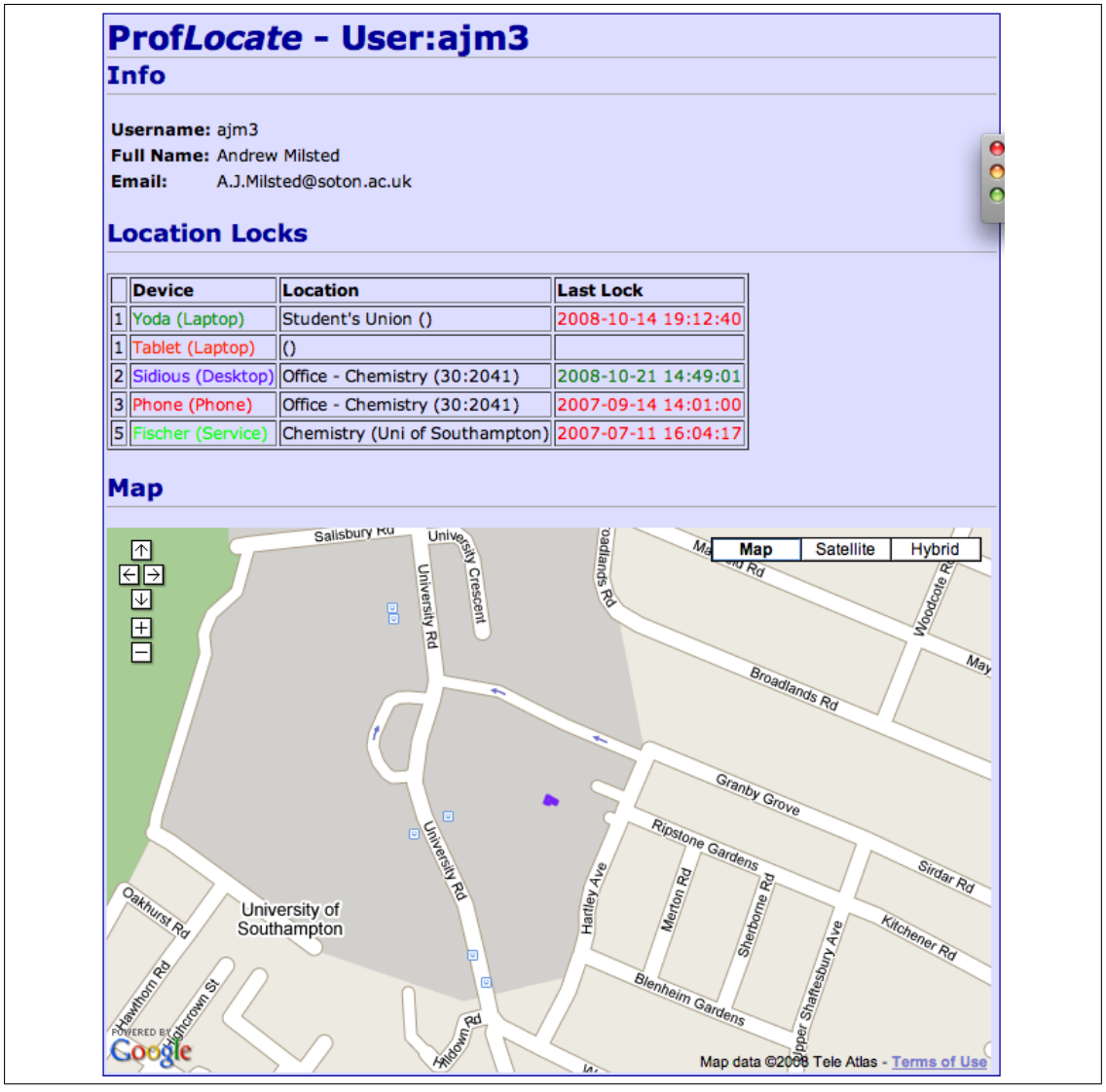


Figure 3.4: An example output page for proflocate

The second part tracks bluetooth devices around the department, there are a series of bluetooth beacons in the office and the laboratory As a device comes into range it is registered with proflocate. Then when the beacons loose the device, it registers the item lost.

To make sure this project operated ethically it only stored information on devices owned by members of the group and all other bluetooth information was discarded.

3.4 eMalaria

When the eMalaria project was first implemented[62] it had a narrow role, to educate school children about the devastating effects of the international malaria problem, its current treatment methods and a little about how drugs could be designed and tested. Technically delivering the accompanying teaching materials to the children was a relatively simple task. They were designed by an education expert and then placed into static content pages. The challenge was to develop an interface that allows the inexperienced user to design a drug molecule, optimise it into a plausible 3D structure and then dock the molecule with the active site in question, in this case the malarial Dihydrofolate reductase protein (DHFR). The ligand docking is achieved using the genetic algorithm docking software GOLD[63], from the Cambridge Crystallographic Data Centre (CCDC). The system comprises a fairly minimal spherical scoop taken from the active site of the DHFR enzyme.

There were few known problems with the system, one major issue was the United Devices (UD) system. The UD system is a 'cycle stealing' distribution server tool. It allows the project to run the docking jobs on desktop machines which are dotted around the department, saving us from purchasing dedicated docking machines. The problem with the UD system lies in the fact that is is designed for high quality screening where individual docking jobs taking hours on the client machines, and not minutes. This led to over polling of the UD server, which caused it to be very resource hungry.

Part of the task of updating the site was to make the components of the system more accessible for other projects. This was achieved by opening up the components with an open standard interface, in this case Simple Object Access Protocol (Simple Object Access Protocol (SOAP)) was used (Section 3.1.3.). An additional aim was to design an interface that would allow more advanced users to use the system as tool rather than a teaching aid. This was shown by a project student investigating the effectiveness of peptides with the malarial DHFR.

3.4.1 Optimization

3.4.1.1 Molopt And Mopac

The original system used two programs to obtain a 3D structure from the user inputed 2d sketch. Molopt is a custom program written specifically for the eMalaria project and Mopac[64] is a standard quantum mechanical geometry optimisation package. This approach is well suited to the original scope of the of the project as it provides quick geometry optimisation for small molecules. However, requirements change and it was observed that the system was being used as a screening tool. To test this functionality a library of natural amino acid bipeptides was built. It was noted that a few of the larger bipeptides were taking too long (5 minutes+) to optimise using the semiempirical methods.

Accordingly it was decided that a rougher geometry optimisation, or "clean", would be acceptable. The Marvin[59] package was investigated as it includes a 3D clean function that can be tuned for our specific requirements. It also includes molconverter, a command line interface to the package, which enables it to be easily wrapped with a SOAP interface. A series of tests revealed that we were getting comparable results with both packages, but there was a noticeable increase in speed with the Marvin software. There is another advantage with Marvin, in that was that it can handle more exotic elements, where Molopt can only cope with core organic ones.

Unfortunalty no quantitive data was collected for this during the development, only the qualitive observations.

3.4.1.2 Marvin and Smiles

Marvin also proved to have another use, as a library of dipeptides (and later the tripeptides) was required. Daylight Smile[60] strings were to be used to represent the monomers. The conventional starting point, as defined by Daylight[65], was ignored and a set of strings were designed so the C-Terminal and the N-Teminal were at the ends of the Smile string. For example, one of the simplest amino acids, Glycine, has the Smile

string N[C@@H]C(=O) and therefore to make the bipeptide Glycine-Glycine it is only necessary to concatenate the two Smile strings thus N[C@@H]C(=O) N[C@@H]C(=O). As amino acids are generally represented in zwitterionic form, as proton must be transfered from the hydroxyl group to form an [NH3+] group on the amine function, therefor an [O-] was placed on the terminus, and the nitrogen protonated. Once we have the final Smile string, [NH3+][C@@H]C(=O)N[C@@H]C(=O) [O-], we can then use marvin to generate the 2D and then 3D structure of the molecule, see figure 3.5

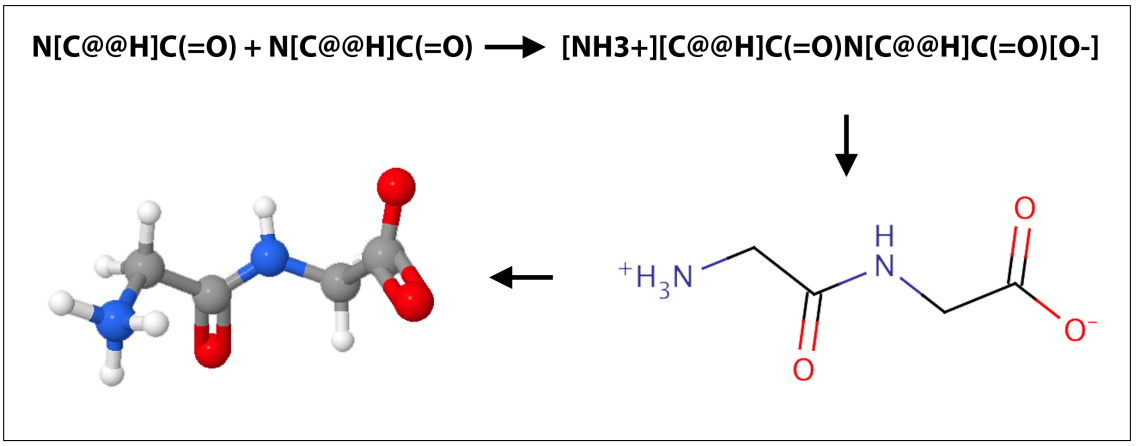


Figure 3.5: The Optimisation process going from smile to 2D then to 3D also showing the systematic construction of peptides

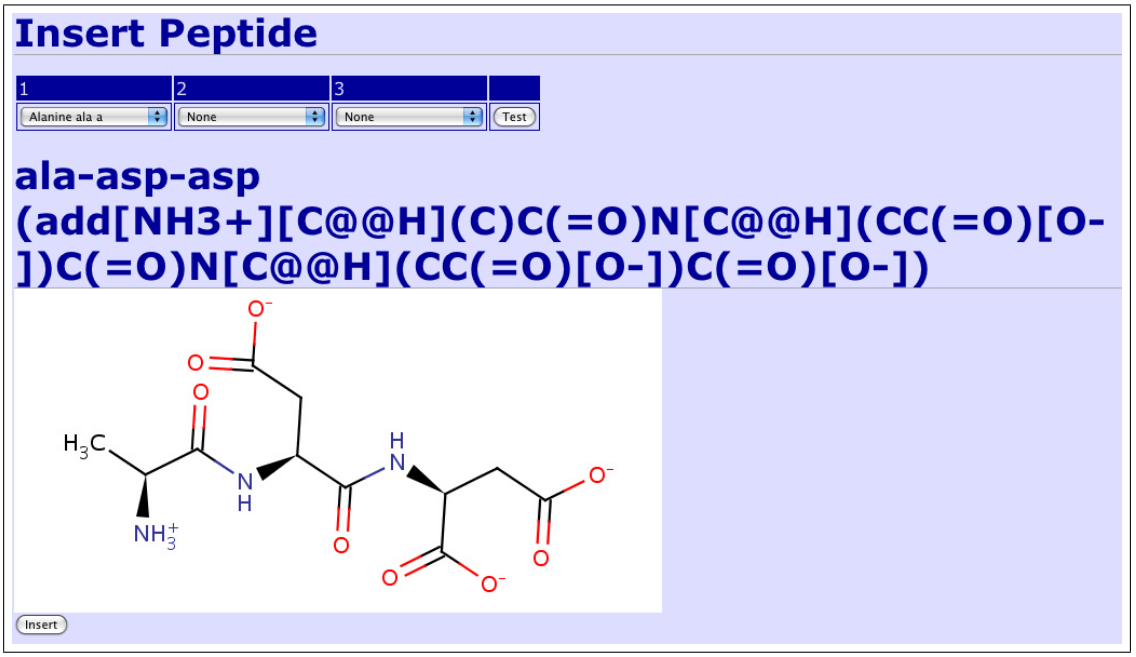


Figure 3.6: Online peptide interface

To aid students in the construction of their peptides a interface was developed to enable them to construct an n length peptide, an example of the result of constructing an

ala-asp-asp tripeptide is shown in figure 3.6.

3.4.2 Testing the Interface

The great thing about young children is that if there is a problem that was over looked then they will find it. To demonstrate and test the system the project took part in National Science and Engineering Week 2006. Participants were invited to design a drug molecule and have it docked, with an associated explanation of the science behind drug design.

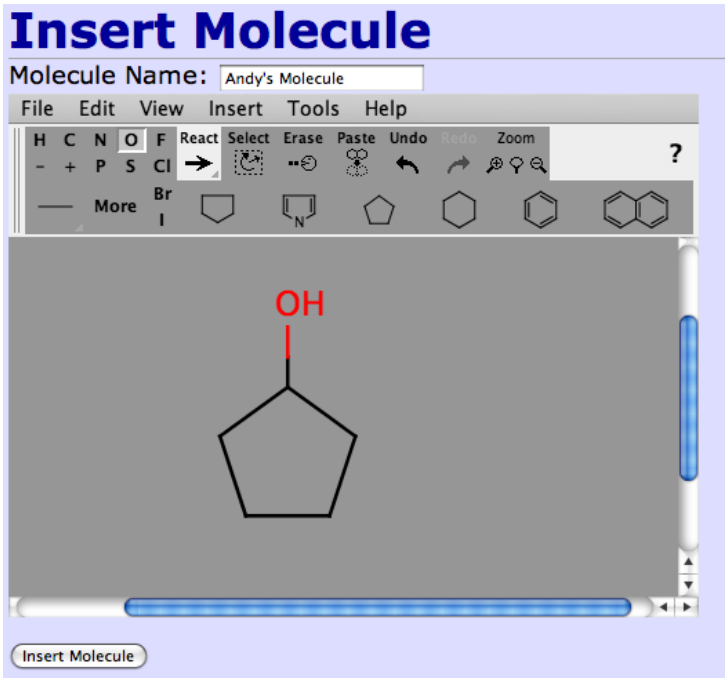


Figure 3.7: Simplified Interface for eMalaria

The simplified Marvin sketch package was used as the molecule drawing tool, however the interface was too advanced for the age group of children (7 - 12). Additionally the molecules that the children where designing were overly large and chemically absurd. This was causing a few problems with molecules taking longer to optimise but the simple geometry optimisation of Marvin appeared to cope, see figure 3.7.

The real problem that was faced was when very large molecules, which had passed the optimisation process, was with the GOLD docking server. With these large molecules the GOLD clients were taking a very long time (10 minutes+) to return a result and

eventually the UD server became unresponsive to the point where sticking jobs couldn't be terminated. This required the server to be restarted, and some human checking of molecules that the children were submitting was then put in place. From this point on the system held up.

3.4.3 Distributed vs Dedicated Computing

Whilst using the distributed computing suites is a very attractive solution for gaining cheap computing cycle by 'stealing' them from other computer downtime. Another option is just to use dedicated local computing power in order strip back the overheads involved with distributing the tasks by computing them locally.

Keeping with the distributed model the system can be developed further in two main areas. The first would be speeding up the UD system, which can be done a number of ways, a) installing on more appropriate hardware, making it 'fit' the server it is intended for, or b) migrating to an alternative software solution. A potential replacement for UD is the BOINC[66] software suite to distribute the docking jobs. Unfortunately the licensing of GOLD makes the use of BOINC unworkable as it doses not offer the same industrial protection of the underlying code that you get from UD.

In testing the interface it also became clear that on a dedicated node, without any of the UD wrappings it could single handedly process more docking jobs than could be done by the current UD setup. This option was only limited to one node as we only had one licence to run in this configuration but it very quickly demonstrated dramatic speed increases. It also meant in relatively quiet times the users were seeing near realtime results.

3.4.4 Undergraduate Studies

A undergraduate course, Chemical Informatics, used the eMalaria system as a course-work component. The students had to investigate a set of possible anti malarial drugs

and their effectives in docking with the haemoglobin degrading aspartic proteases plasmepsin II as the target protein. The drug candidates were provided by the work of Ersmark Karolina et al [67].

By calculating/finding a number of molecular descriptors, including the docking results, the students had to build a model to predict the activity of the above set of molecules. Despite given the opportunity to collaborate the results from eMalaria the students still worked independently clocking up in excess of 3000+ jobs, with a class size of 20 and a molecule set of 25, and that many of the students ran the docking in their own time, it did not matter that only a single node was computing the results.

3.4.5 Computing Libraries

One piece of side work was to calculate large sets of molecules against the malarial Dihydrofolate reductase (DHFR), since discovering the relatively simple way to go from a peptide to a tripeptide using SMILES string, we could then construct a large library of all the tripeptides made up from the 21 naturally occurring peptides. This created a library of 9261 tripeptides in 1D, for which we need to calculate 3D structures and a number of molecular descriptors;

- Atomic mass
- Protein acceptor count
- Protein donor count
- logP
- Wiener polarity
- Van der Waals surface area
- Tpological polar surface area
- Water accessible surface area

We chose to use the Marvin library once again as there no licence restrictions on the number of instances we could use. We could also take advantage of Iridis [68], the university's grid supercomputer. This allowed use to take 2 weeks of compute time for the 3D structure calculation and shrink it down to less than 24 hours. We did this by running 16 instances of the 3D optimisation running in parallel. Like wise with the molecular descriptors 11 days of predicted compute time was also managed in less than a day.

For calculating the docking we also used Iridis, not because we could run this task in parallel but a single Iridis node was 25% faster than our current docking server, this still resulted in 5 weeks of docking to do the calculation.

Unfortunately Iridis can't be used to for running the eMalaria web site as time has to scheduled by a queueing system and average wait times are a couple of hours, but if there ever is a backlog on the eMalaria website, Iridis time can be booked and then it can get to work clearing the back log. An ideal solution would be to obtain more licences for Gold but this is financially prohibitive.

3.5 Conclusions

The previous set of tools where used as part of the undergraduate teaching programme allowing them to be extensively tested. This included both heavy load testing when entire classes run processes at once to enabling edge cases that may cause the errors and spot bugs in the tools.

The experiences gained from developing these tools and the identification of problem of recording the scientific record it was decided to investigate the need to develop our own electronic lab notebook.

Chapter 4

The Problem

4.1 Issues

There are many issues, with regards to recording the scientific record, facing the modern researcher, many can be helped by develops in the Electronic Lab Notebook area.

4.1.1 Backup

A drawback of maintaining a paper lab notebook is is not easily backed up, in one extreme buildings do burn down[69] destroying all the research record, but work loss could be as simple as spilling to wrong solvent onto the book in the lab. This leads onto an easy win win for the digital realm, as once information has been digitised it can easily be copied, backed up and kept safe.

This is not to say that much of research data isn't lost through poor management of the data, it becomes easy for the researcher to rely on the laptop/usb stick to store their data, but this can just as easily, if not easier, loose the data through loss, it getting stolen or simply a hardware failure (figure 4.1).

This is where a well designed ELN can excel as it should be able to offer the user the abillity to back its content up onto a remote server, then if setup properly this server would then have the proper backups and redundancy implements.

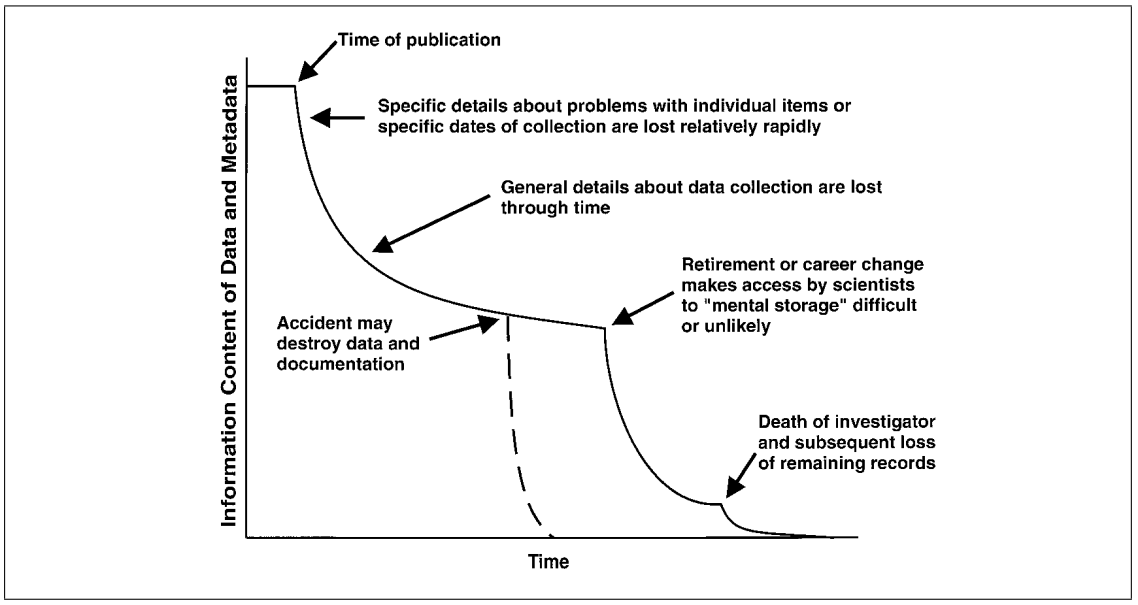


Figure 4.1: An example of the degradation of information content with metadata and original data of time; information entropy. Accidents or changes in storage technology (dashed line) may eliminate access to remaining raw data and metadata at any time[1].

4.1.2 Competitiveness

Science today is competitive, with research spending being reduced due to economic pressures, also the need for big commercial companies to remain profitable. Most researchers have to make sure they protect their own research from misuse by others. This is where other researchers in the field can take ideas from their piers. if they can successfully copy the research and publish first, they would then gain the credit for the work.

This is especially fierce in the world of patents where most countries have a first to file policy, which means that it doesn't matter if the researcher invented an idea first, it is the one that registered the idea with the relevant patent office first that will get the patent. The most secure way of protecting ones research is to continue to use the paper lab notebook, this is very secure, it can't be remotely hacked, if the ELN has any chance of protecting the end user it needs to be secure.

4.1.3 Collaboration

Much of the current research requires collaborations with other researchers, this can be several researches from one subject area but spread across many geographical location, or you could just as easily have research from many subject areas coming together to work on multi-disciplines.

Sharing work and data through conventional means would be possible through fax/email, but this could be cumbersome as the researchers would have to initiate each of the shares, but if they where using an integrated ELN, the system would automatically make available the resources to all members of the research group as soon as it was uploaded.

4.2 Open Research

Not all science domains are as competitive as others, biochemistry and 'off-patent' drug discovery are two such areas. In these areas some researchers are happy to have their work made freely available, an ELN can easily facilitate this be make it available online. The motives for this to be done could one of funding as there is increasing pressure that all output from publicly funded research should be made freely available, and for an ELN to be able to provide this feature then it makes the funding requirements easily for filled. Another reason for working in the open is as an invitation for collaborations, by making the researchers lab note book open, other researchers who could help with the work could come forward, this is particularly advantages in the open drug discovery arena which is discussed in section 7.3.2.

4.2.1 Publication at source

Another driving reason for researchers wanting their work to be made public could be that they which to seek out collaborations with others, or more appropriately have collaborators find them, on such case is with the 'Resolution of Praziquantel' [70] where the researchers made their ELN fully available online. The work was easily discoverable

by the search engines, they then found that they we getting offers of support and help the some of the problems listed in their notebook. This support then led to collaborations which also allowed for the work to continue.[71]

4.3 Provenance

A bound paper lab books has also that the quality of provenance, this is because because if it is well organised and was regually signed off by a colleague, it has the inherent integrity of being tamper proof. It can allow the author to assert that the work that has been stated to be done, was done, and more importantly when it was done. As a well bound book by its very nature has an inherent timeline of order, so if any tampering had been done, then the notebook would clearly show signs of it as pages couldn't simply be added or removed.

There have been recent high media profile exchanges which have highlighted some of the issues when not making raw data and workings available, which quickly unravelled the asserted conclusions. Climate Gate in particular brought this to ahead with the old teaching mantra "show you working" [72] If it had become common practice to make raw data available for derived pieces of work then issues such as Climate Gate just couldnt happen as any results that raised questions could easily be discovered.

Chapter 5

Electronic Lab Notebooks

5.1 What is an ELN?

In its simplest form an ELN is a digital representation of a lab notebook, usually using a computer of some form as a platform. ELNs aim to improve the way experimental practises and results are recorded. A fundamental aim of an ELN system is to provide a standardised, reproducible and reliable replacement to the traditional paper notebook. It has been suggested that in the chemistry domain, laboratory notebooks often do not contain all the required information to repeat an experiment[73].

Normally a notebook will contain an experimental plan, with relevant safety information, written before the experiment is carried out. During the enactment of that plan, key properties such as weights of samples and observed changes are recorded. Once an experiment has been completed, printed copies of analytical instrument output, such as mass spectrometer results, are stuck into the paper book.

Through the use of an ELN, this process can be greatly improved. The ELN software must therefore include the traditional functions of the notebook, storing the plan and thoughts of the researcher, while capturing and linking data from a multitude of sources[74]. By linking this data the provenance can easily be captured. In recording the ownership of the data, the parties IP rights are protected.

A key feature which extends the ELN from a paper notebook is the ability to search records. Should the experimental metadata have been recorded and linked correctly during the experiment, more advanced searches can be carried out and produce more relevant search results.

Another aspect of an ELN system is data archiving, in the pharmaceutical industry a drug can take up to ten years from concept to reaching the market[75] and in health and safety, data records must be made available for up to fifty years[76]. Over this time-scale, standard office software can undergo a number of revisions that may not be backwards compatible; There can be a trade off as data could be converted in to a well accepted preservation format eg the portable document format (PDF). It can overcome this problem as its representation is independent of the hardware and software. The use of Extensible markup language (XML) can also help with future proofing, as well as aiding interoperability[77].

There have been major advances in hardware design that has aided the uptake in ELN software. Devices such as tablet PCs and Smartphones allow users to take touch and pen sensitive devices into the laboratory and record experimental procedures as they are carried out. As many ELN systems use a centralised database to store the data, should the device be damaged no data is lost. In contrast, should a traditional paper notebook be damaged, such as through a solvent spillage, a considerable volume of data may be lost.

5.2 Example ELN's

There are many ELN's available both commercially and opensorce. Some of the commercially available which include;

- Labtronics's Nexxis ELN^{TM}[78]. This system is aimed at highly structured experiments which are repeated regularly. An experiment is presented as a web form, with boxes to fill in the relevant experimental details. The process of completing

this form can be automated, through retrieval from experimental apparatus and the data repository.

- IDBS's E-WorkBook™[79], an ELN developed for any science domain. The software offers a framework for developing specific solutions while promoting data sharing and re-use. E-WorkBook also provides a high level of IP protection. This is achieved through use of SAFE signatures[80], an independent digital authentication association, and capture of document audit trails. IDBS also offer specific domain ELN's for example they provide a chemistry extension; offering functionality specific to the chemistry domain.

- E-Notebook™[81], from CambridgeSoft™, follows a different approach to acquiring the scientific data. Data is collected from existing document types such as Microsoft Office files, PDFs and images. Sharing and collaboration is supported through shared drives and extensive search routines. As E-Notebook is part of a larger suite of software, including ChemDraw, chemical data can be automatically generated and imported into the system with the associated metadata.

- iLabber™[82] form Contour Software offers a cost effective ELN, offing both a client and web based interface giving users the ability to record their work in a non subject specific environment. It has the ability to integrate with Microsoft Office allowing data to be processed in familiar tools, for example excel along with many other tools.

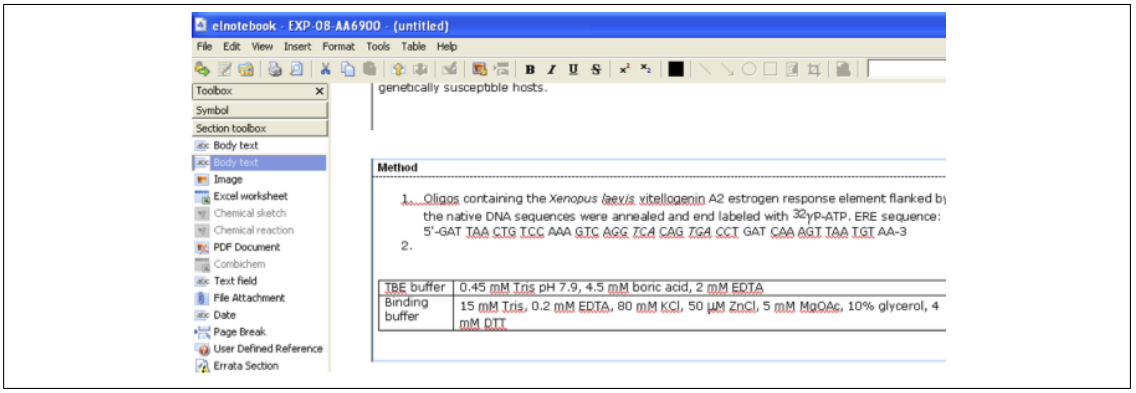


Figure 5.1: An example commercial ELN, iLabber, being used to record the method for an biochemistry procedure. Showing the use of the free text entry and tables

These ELNs share many common features; For any experiment that is repeated many times, it can be beneficial to have the ability to produce structured forms, this promotes consistency in the way the experiments are recorded, making searching, sharing and collaboration easier between users. Audit trails of all the work being recorded allows the review of all edits and any deletions. This makes for a more complete record of the work ensuring the provenience of the work.

There is a split between using a software client on the researchers computer,e.g. E-WorkBook^{TM}and E-Notebook^{TM}, and delivering the application through a web browser, e.g. Nexxis ELN^{TM}and iLabber^{TM}. The advantages of delivering through web browsers is that the reseracher doesn't need to install any software, and could potentially access their research from anywhere.

Using the commercial offerings can inhibit the take up for a researcher as they have to invest significantly in infrastructure and hardware to get many of these products working. One exception in the above list is iLabber^{TM}offers a Software as a Service (SaaS) provision. This allows a researcher to signup for a small fee per researcher per month on their Cloud based service. This gives the researcher the chance to use the ELN and only have to pay for what they use.

5.3 Semantic Importance

ELNs can be simple by providing an electronic analogue of paper, supplemented by data storage facilities, While such basic systems are straightforward and do enable text-based searching they do not exploit the full potential of a computer system to systematise and catalogue data. One way to over come this is to implement a fully semantically aware ELN that would be the extreme opposite and example of this would be the The Smart Tea project [83, 84, 85], This implementation of an ELN aimed both to guide a synthetic organic chemist through a synthesis and to produce a fully semantically annotated record of what had occurred captured directly as an Resource Description Framework (RDF) graph [86] the data description component of the Semantic Web[87, 88, 89, 90].

The metadata framework for the Smart Tea process was based on the assessment form used for the Control of Substances Hazardous to Health (COSHH), a health and safety requirement related to the handling of potentially hazardous materials. The Smart Tea architecture extended the framework to include an RDF representation of the experimental plan, which was interpreted to provide prompts to the chemist with a place provided for adding experimental details and observations (i.e. metadata in advance). The result was a series of RDF statements (triples) that described the procedures undertaken and acted as a provenance chain for the materials produced.

The project also investigated the use of specific hardware by trailing the use of a tablet PC in the laboratory, and the project has been continued via the More Tea[91] work, which provides a richer set of semantics that more accurately reflect the nature of the work of the synthetic chemist.

5.4 The 'Un' Semantic ELN

Experiences with Smart Tea, and other heavily semantic rigorous systems, found that its approach could be too heavyweight and prescriptive and, potentially, significant work would be needed to adapt the approach to other domains of experimental science.

A response to this was to investigate using modern Web 2.0 principals involving the addition of minimal unrestricted semantics to otherwise unstructured data, rather than going for a full semantic web system. There was also a need to develop the notebook from the perspective of the individual researcher, and to avoid being constrained by the requirements of specific subjects, such as chemistry or biology.

At the time, blogs were becoming a component of science communication familiar to, and popular with, a growing number of researchers [92, 93]. Blog systems allow almost complete freedom in which to record, as the empty post is perceived as a blank piece of paper, similar to the traditional paper labbook. Also it would allow no restriction on attaching data, bettering the original cut and paste.

Moreover, the use of blogs would be familiar to incoming future tech savvy generations of scientists, so a laboratory blogging platform with the idea of a laboratory blog as an Electronic Laboratory Notebook was developed. This software would eventually evolve into the LabTrove system.

For the initial development we modeled the activities of a bioscience research lab, mainly the work of Dr Jennifer Hale postgraduate research "Investigations into neutral drift" [94]. This work was very data centric, producing many images of DNA polymerase chain reactions (PCR). The work also needed an ELN as the researchers' supervisor was not based full time at the same institution, they needed an easy way to review and support the research.

Chapter 6

LabTrove

LabTrove is a blog-based system for recording laboratory processes and objects. This chapter describes the architecture of the implementation, the principal components of its design, and the primary aspects of the operation of the system.

6.1 Architecture

The LabTrove system employs a client-server architecture, with a PHP server running under Apache and a MySQL database. For the development system the Apache web server runs under Debian/Linux. This software combination is commonly known as LAMP.

Firstly the operating system (OS) is Linux, Debian[95] was chosen as the development environment as it is fully open source and has a fully package management tool. Another reason for choosing Debian is that it fully adheres to the philosophies of Unix[96], and uses a community of developers, keeping it and all of its packages secure and up to date. LabTrove can run on many operating systems as there have been successful installations on a few different flavours of linux and even Apple Mac OS X^{TM}(MAMP). Unfortunately, whilst most of the flavours of *nix (including OS X^{TM}) all share POSIX[97] as an application programming interface (API), Microsoft Windows^{TM}doesn't not share this feature. For this reason, currently, Windows^{TM}based operating systems can not host

LabTrove. This would only restrict a researcher from running the LabTrove locally on their own machine for which it is not designed to do so, it is more suited to being running on server infrastructure, but if the user wishes so LabTrove can be run on a virtual machince running on a windows machine. This can easily be achievied with software products Windows Virtual PC or VMware Workstation as they allow the running a Linux guest operating system on a windows computer.

Apache[98] is the web sever software this handles all of the client requests and the internet communication. It uses HTTP (Hypertext Transfer Protocol), which is the basis for all web-browsed content. While LabTrove itself doesn't encrypt the data we can use the HTTPS feature of apache to secure all of the traffic to and from the server, securing passwords and content.

All of the data for the blogs is stored in a MySQL[99] database. This means that all of the indexing, searching and backups are handled by the database, large files are handled slightly different and depending on configuration may be saved directly on to the file system.

PHP: hypertext preprocessor (PHP)[100] is the server side processor that actually runs LabTrove. PHP is a scripting language, which means that all the functions of LabTrove are created by a series of PHP scripts that then produce the outputted Hypertext Markup Language (HTML).

The browser sends a request as a input url, for which the webserver (Apache) turns into parameters that then the php scripts interprets these, then fetches appropriate data from the database, before producing the requested output (see figure 6.1).

For example, when the user clicks an example trove, http://blogs.chem.soton.ac.uk/, LabTrove will then generate a list of blogs the user is allowed to see, then present this data in the form of some HTML, which is then returned to the user's web browser and displayed, (see figure 6.2).

6.1.1 Versions

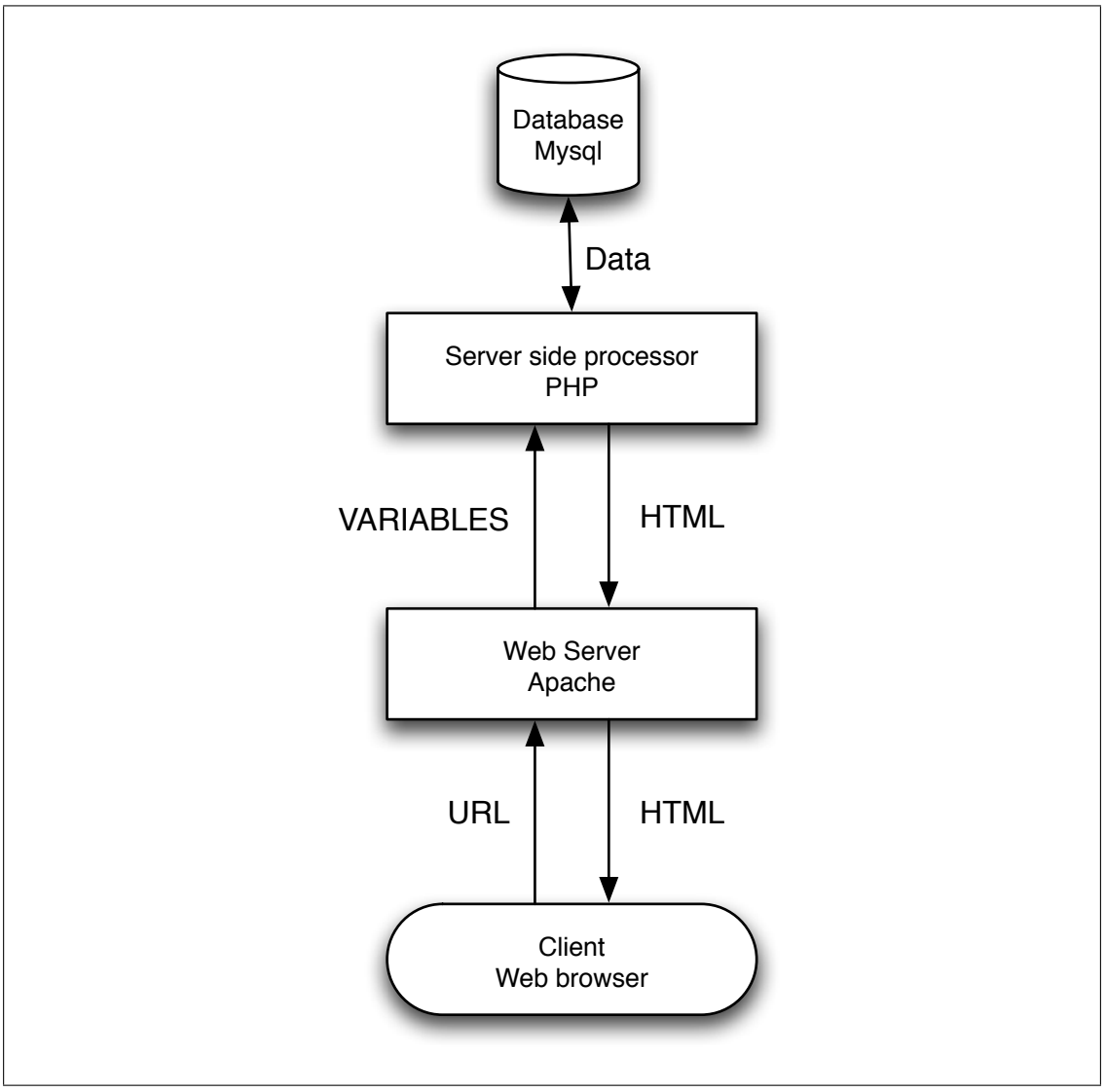


Figure 6.1: Showing the flow of data with in a classic LAMP installation.

LabTrove has gone through a number of versions from prototype to a stable industrial strength product. This has been done through a series of releases that have since focused on producing a stronger product. The author spent 6 months working with the OMII-UK group gaining experience in ways to strengthen the LabTrove product. LabTrove has been since been evaluated by a large industrial company, Unilever, along with several other research groups, this has been discussed later in the thesis in section 7.3.3. The blog was originally based on a GNU Public Licence (GPL) blog called μBlog[101].None of the back-end elements of the original software remain, but the methodology behind style templates and layout still exists. One of the reasons this was dropped was the GPL licence as it is copyleft in nature, this means that any follow on work, ie LabTrove, would have to GPL.

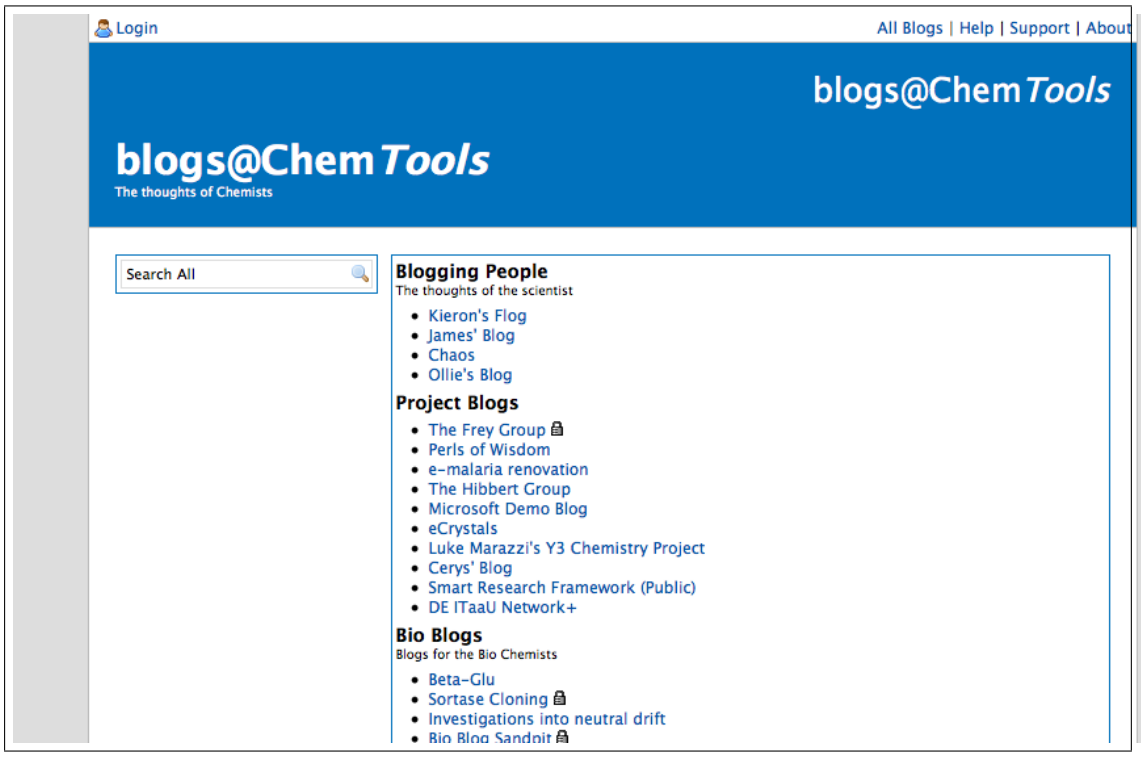


Figure 6.2: Showing the resultant request, http://blogs.chem.soton.ac.uk/, to the user[Accessed: 01/11/12]

The default style, which is called kubrick[102], which will be a familiar design to many users, This is because it is open source and is the default design that is packaged with WordPress[103], a very popular free blogging solution. After the blog had first been used by a set of biochemists a need arose for it to be customisable at the individual blog level. This was because they were drawing tables too wide for the fixed width of the kubrick design. This lead to creating a specific full-width style for the biochemists blog. It called for a configuration file to be made for any single blog. In this case, the config file selected the non standard style template for this blog.

Owing to the evolution of the LabTrove system, PHP script names and database table names do use the label blog, recognising that the technology that underpins a LabTrove e-Notebook is a blog.

6.1.2 LabTrove objects

The data model was designed around a classic relational database design building on the previous blog software. The need to store labtrove objects in a relational database

dictated the need for the objects to be stored in separate tables, they were then referenced and linked together with primary keys. These keys, in LabTroves case, where an incrementing number and in the case of posts, data items and comments would then give the identifier for the URL accessing the object. Troves use a short text field for the url but still had a integer used to associate posts attached to a Trove.

Figure 6.3 illustrates the principal LabTrove objects and their relationships with each other, summarised as follows:

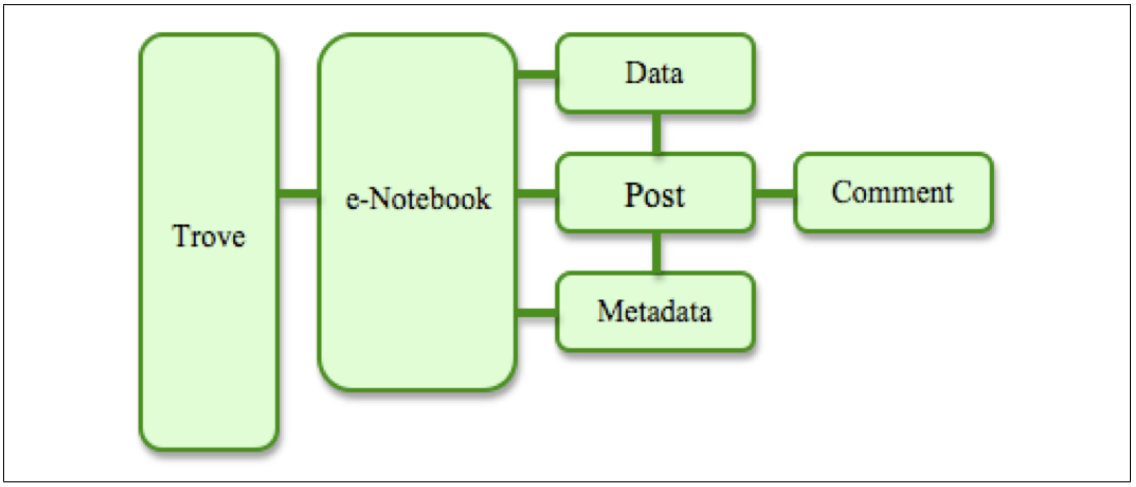


Figure 6.3: The principal LabTrove objects

- A Trove is a single LabTrove installation comprising any number of e-Notebooks grouped under one of three headings: Project Lab Books; Discussions; or Blogs. LabTrove manages access control to the e-Notebooks at the Trove level.
- An e-Notebook comprises any number of posts, any number of data files, and a collection of metadata keys that must include one or more values for the Section metadata key. LabTrove manages write access control to posts at the e-Notebook level.
- A Post can have any number of attached data files and any number of associated comments. Each Post must have a value for the Section metadata key and can be associated with any number of other metadata key-value pairs.
- Data comprises any number of data files, which are managed by the e-Notebook, and inherit the security and visibility settings of the particular e-Notebook. A

data file can be attached to one post only and may be embedded by reference within that post. Optionally, the data file may be embedded within other posts.

- Metadata comprises any number of key-value pairs, of which only the Section key is mandatory for each post.

Each post consists of a title, an author, a date-and-time stamp, content, and metadata. Each post revision also has an 'Edit Reason', which is set to 'First Post' when the author creates the post. The author then supplies the 'Edit Reason' for any subsequent revision of the post. Each post has a unique numeric identifier that LabTrove supplies when the author creates the post. Each revision has a discrete identifier, which for the 'First Post' is the same as the post identifier. Each revision also has a discrete date-and-time stamp, but to ensure the correct sequencing of posts displays the date-and-time stamp of the 'First Post'. All references to the post use the post identifier.

Each post can have any number of associated comments, which also have an author, a date-and-time stamp, a body, and an 'Edit Reason', but have no metadata.

The post body consists of free-form text that might include BBCode tags , for example to mark up lists, text highlighting, or images. The post body can also include links to other posts, using their post identifiers, and to external resources, using their URLs.

Links to LabTrove posts are not strictly not objects themselves, but are nevertheless fundamental to LabTrove as an ELN. Each link is has URI, which provides a unique form of identification for every element of the research process. Links between posts are aggregated and then presented with in the content where they are link, or as backward link by providing a 'Linked By' component to the post.

LabTrove provides post information to the client in formats other than HTML, those currently supported being XML, PNG, ATOM, and JSON Users wanting to view posts, comments, or revision lists in XML format, for example, modify the request URL in their client browser, replacing the '.html' extension with '.xml'.

Viewing LabTrove in XML format is the method for 'read' access to the API, the write/edit part of the API will be described later in section 6.5.

Format	File Extension	Use
HTML	.html	This is the default output, and the format used by web browsers
XML	.xml	A xml representation of the content, used as the read part of the api.
PNG	.png	Creates a screen shot of the content, good for thumbnails when the content is embed in other tools
ATOM	.atom	Used by screen syndication feeds as it is much more capable than RSS (eg supports pagination)
JSON	.ejs	A common web script format, is used internally to provide timelines feature.

Table 6.1: Formats for view LabTrove content

6.1.3 LabTrove Components

The figure 6.4, illustrates the main control flows between the components of the PHP server process. The LabTrove download establishes a directory structure in which the '/docs/' directory contains all the scripts that contribute to serving pages to the client. The other directories hold the scripts for all internal operations. The folder '/docs/' becomes the root of LabTrove for which the webserver will serve LabTrove.

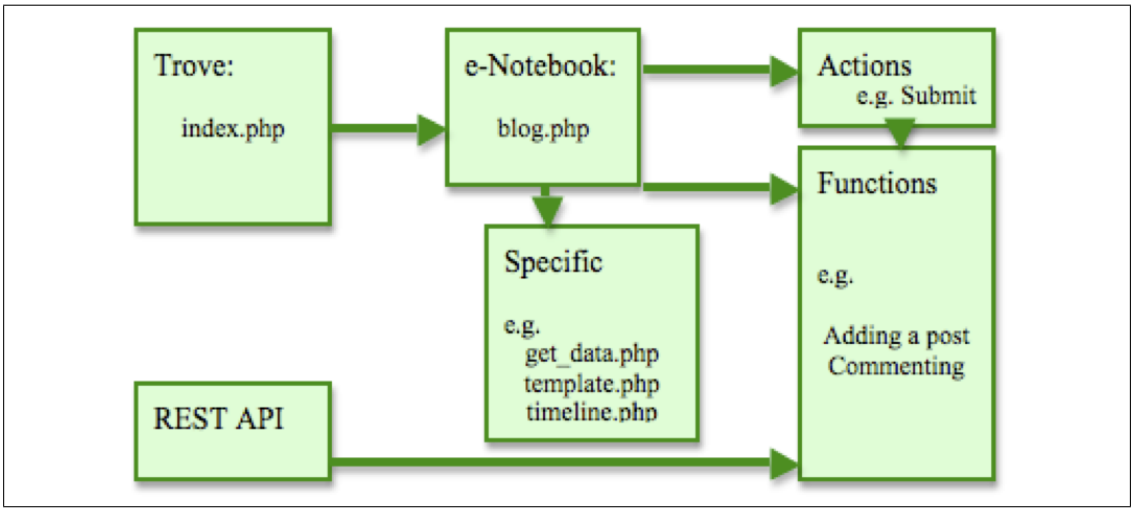


Figure 6.4: Schematic diagram illustrating the principal components of the PHP server process and the main flows of control between components

The index.php script serves the main page for a Trove and handles initial requests, including the index of all the posts and when the user logs in their dashboard.

When a user selects an e-Notebook, control passes to the blog.php script, which serves all pages relating to rendering posts and other elements of the e-Notebook. A lot of actions are performed by these frontline scripts are actually implemented as functions which are in the '/lib/' folder. This is done because the function being performed may be required by other areas of the blog, examples of this is the function 'ender_blog_link'. It takes in the id of a post and will return the fully formed url for the post, this action is required in many places throughout LabTrove.

The error codes that the server process returns are as defined by the HTTP protocol. For example, a generic processing error returns code 500, supplemented by a reason string, or a code 404 for a resource not found. LabTrove does not maintain its own logs, relying instead on the Apache web server to track accesses and PHP errors.

The LabTrove architecture relies on plugins to provide features that are standard but system administrators might choose to customise. Access control is described in section 6.1.3 including how the authentication plugin works.

Currently LabTrove is distributed with 3 authentication plugins:

- **login_local** Provides local database user access.
- **login_openid** Provides user access via openid.
- **login_ldap** Provides LDAP driven user access.

Also LabTrove also has an Uniform Resource Identifier (URI) plugin, by default LabTrove uses a local database table manage URIs, but has the capability to use a service to produce identifiers, eg Digital Object Identifier (DOI):

- **uri_samedb** Provides URI service through the same database.

LabTrove currently has a number of plugins in development:

- **claddier** Enables inter webservice linking using the claddier service[104], i.e. the ability to link between troves even if they are hosted on differnt servers.

- **elndd** Allows the export of a ELN Data Description of individual notebooks from LabTrove.

- **chemspider** This will parse posts for chemical terms using OSCAR[105], chemical language processing, to link found substances to Chem Spider [106].

To help development of plugins an example plugin is provided (see figure 6.5)

```
<?php

$ct_config['hooks']['on_post_render'][] =
    array("function"=>"hook_example_post",
    "params"=>array("bit_id","bit_cache"));

function hook_example_post($bit_id,$bit_cache)
$return = $bit_cache;
$return .= "<br/> Bit ID: $bit_id";

return $return;


?>
```

Figure 6.5: An example plugin hat demonstrats the user of the plugin system, in this example it will alter the HTML content of the post by adding the post idenitfier to the bottom

6.1.4 Database

The MySQL database consists of twelve interconnected tables, some of which are for operational convenience. Figure 6.6 illustrates schematically how the main tables are connected.

In addition to administrative and descriptive information, the blogs table maintains a cache comprising the monthly archives for the e-Notebook, post references categorised by user, and a compilation of all the metadata associated with the posts in the e-Notebook. By using this cache LabTrove improves system performance.

LabTrove maintains metadata in the posts table, as a single field holding an XML string containing the metadata in the form of key-value pairs. The metadata field always contains at least one such pair, owing to the Section key being mandatory.

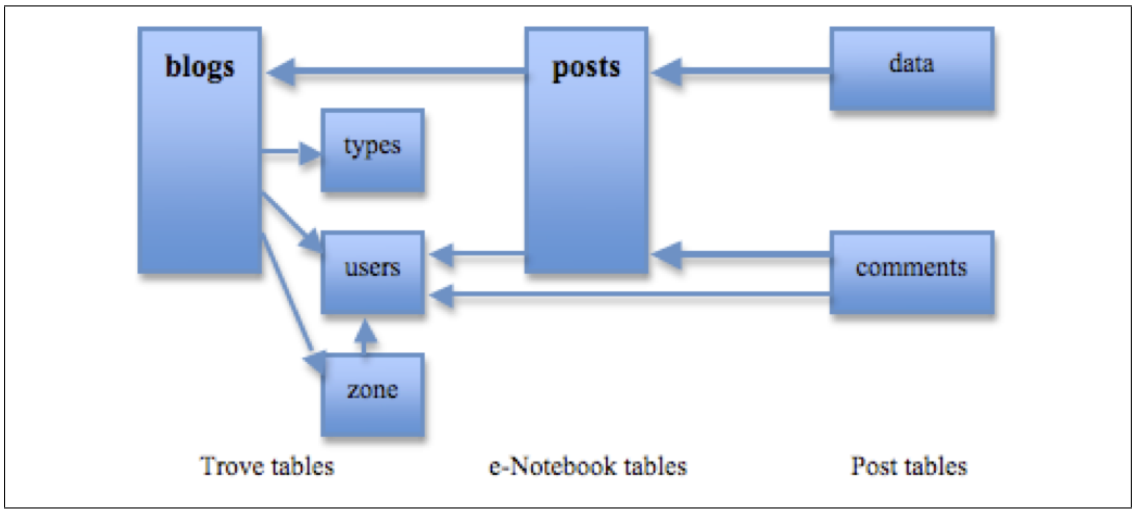


Figure 6.6: Schematic diagram illustrating the main database tables and their interconnections

The option of a discrete table holding the metadata in separate key and value fields was considered, but decided in favour of the single field because it provides the flexibility to add new keys and values arbitrarily, without modifying the underlying database schema. Moreover, keys and values can both be changed at any time; keys can take any text value. However, this implementation does require the text string to be read with an XML parser to extract the key-value pairs. LabTrove optimises metadata filtering by updating the cache in the blogs table each time a user creates or modifies a post.

The posts table has a field that holds a list of any other posts that link to the post identified in that row of the table. When LabTrove saves a post, it checks the body of that post for references to other posts, using the link to add the source post to the Linked to list for the referenced post.

Once a record has been entered into the database, no mechanism is provided to alter or delete it, thus ensuring that the provenance trail is complete, reliable, and reproducible. However ELNs do require a revision mechanism. When a user edits a post or a comment LabTrove updates the relevant table, storing a new record, which includes fields holding the reason for the edit and the name of the user. It modifyies the previous record to reference the new revision. The post and comment tables therefore contain a complete history and version control for each post or comment.

The current version of LabTrove stores data items as binary large objects in the MySQL

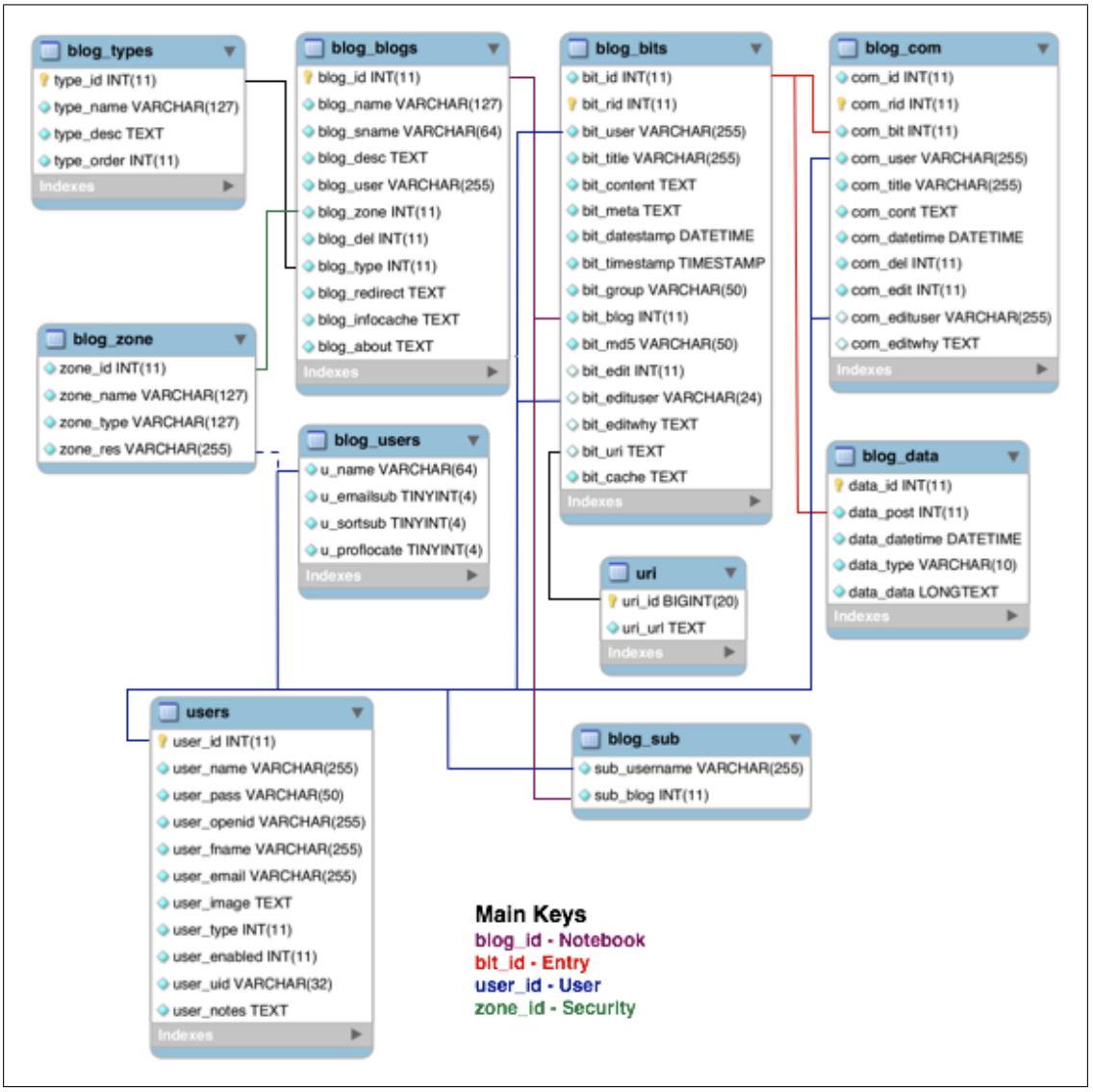


Figure 6.7: Detailed schematic diagram illustrating the main all the tables and indexed keys.

```
<METADATA>
  <META>
    <LAB_BOOK-EXPT._CODE>1111-001</LAB_BOOK-EXPT._CODE>
    <DATE_OF_EXPERIMENT>29.06.09</DATE_OF_EXPERIMENT>
    <PROCEDURAL_STEP>Safety</PROCEDURAL_STEP>
    <POST_TYPE>Safety</POST_TYPE>
  </META>
</METADATA>
```

Figure 6.8: XML used to store metadata for a post in LabTrove

database, subject to a threshold that is configurable by the system administrator. Above the threshold, LabTrove stores large data items in the file system. Storing all but very large files in the database has the advantage that a backup of the database contains

all the Trove data; restoring that backup with a fresh copy of the code rebuilds the entire Trove. Future versions will include a user-configurable option to store data items separately.

Even though the LabTrove was developed for MySQL databases some work has been completed separating out the SQL functions from the core LabTrove code and moving it to a plugin, this is so that other database engines could be used. As a demonstrator, a plugin was created for PostgreSQL[107], using this initial work to separate the database from the code as model there is no reason for other databases not to be used including Oracle™or Microsoft SQL Server™.

Once the MySQL database has been setup, for which the LabTrove installer does, there is very little administrative load, only majors version changes would require any adjustment to the database. All administration that is required by the LabTrove can be done from within the software so there is no need for a systems administrator to get involved in the day to day running of a labtrove service. This will also protect the integrity of the data contained with in the database as all changes made from with in LabTrove are tracked and checked by LabTrove, for example Post Edits will make sure revisions are kept.

All administration related to maintaining the database (security patches, upgrades etc) should be handled by the operating systems package manager which should also not involve very much system admins time.

6.1.5 Security

The realm of security and IP (intellectual property) protection is a very important aspect to any research environment. In general, access to a Trove for any purpose other than browsing requires users to have a system account, for which each user must have an identity. The exception to this would be for closed troves, where obviously access control is applied to all of the trove content. LabTrove holds user identities in a database table: the fields include user name, e-mail address, and the authorisation level (described in detail later). For each e-Notebook LabTrove also maintains a zone table, which holds a group of lists of users who have specific authorisation level. In operation, the level

of authorisation that a user has is the greater of the zone level and that users personal level. LabTrove always has one zone, in practice labeled 1, this zone allows any one who is logged on to view the trove. For other zones the trove owners can set list of users that can have access to that Zone.

The system administrator establishes the authentication method when setting up the LabTrove instance, so each Trove has a single method, determined by the plugin installed. For example, for users at the University of Southampton, the authentication plugin uses the University Lightweight Directory Access Protocol (LDAP) service. Access and authentication via OpenID[108] has also been implemented, which allows users to maintain a consistent identity over multiple systems. OpenID not only enables integration with third party services but also a federated model of LabTrove servers, in which the same user can have a single identity across multiple deployments.

LabTrove authorisation levels provide for a range of security and editing rights. The general policy is to make posts visible, but to require a system account for commenting for spam prevention and proper attribution of those comments. LabTrove defines authorisation levels according to the following layered model:

View This is the default level, at which users can view posts but require authentication before adding a comment.

User At this level, users can create posts and also their own e-Notebook. Note that LabTrove will check the identity of a user attempting to change an e-Notebook setting to ensure that the user is the owner.

Editor At this level, users can read everything in the Trove, even if they are not specifically entitled to, but can modify only their own posts or e-Notebooks.

Admin At this level, users can edit anything, although every change is attributed by user name. Only a system administrator can create a new LabTrove instance and set access privileges for that Trove.

Any user can create their own trove with its own permissions. A normal model for setting up a publicly accessible LabTrove instance would be to allow general web signed

up users to have the 'View' level by default. Then the system administrator has the ability to promote that user to 'User' allowing them to create their own troves.

Where identity/authentication is coming from a authoritative source, eg a Institutional LDAP, then LabTrove can be set to create new users directly at the 'User' level, this is safe to do so if the instance policy allows everyone from the identity provider permission to use the LabTrove instance.

When an institution runs a number of LabTrove instances it might can become necessary to operate a Single Sign On (SSO) system to make it easier for users to move between instances with out the need to login to each. At Southampton a system was integrated into LabTrove with ChemTools, Section 3.1.1

6.2 Features

LabTroves initial feature requirements capture was centred around the single use case of a biochemist needing to share their research with their supervisor (discussed in section 7.2). As development started and other users started to use the tool, their requirements where incorporated into the feature list. When version 2.1 was being developed all the required features of LabTrove where laid out in figure 6.9 as part of the requirements capture process with the interested parties, which also included users from the xray group (discussed in section 7.3)

6.2.1 The User interface

Because the technology that underpins a LabTrove e-Notebook is a blog the user interface format is standard for a blog. The content pane displays either the content in list form or a text box for editing an individual item, commonly a post. The content displayed can be at Trove, e-Notebook, or post level.

When a user creates a post, the editing view comprises a Title field, a Text field with a toolbar containing icons for introducing specific markup, and data entry field for the Section and other metadata key fields. When a user edits a post, the view includes additional fields for entering the reason for the edit and, if required, for attaching data files. When a user adds a comment to a post, the editing view includes only the Title and Text fields, with a toolbar for markup. When editing a comment, the view includes an additional field for entering the reason for the edit. The toolbar icon for linking to another post displays a popup window containing a list of posts to which the user is authorised to link.

Initially, for reasons of data integrity, the additional field for attaching data is not present in the new post view when a user creates a post, The editing interface did not create the record in the posts table until the user clicks the submit button, whereas it does create a record in the data table at the time the user attaches a data file. If post creation was not completed, for example because the user discontinues preparing the post, the data

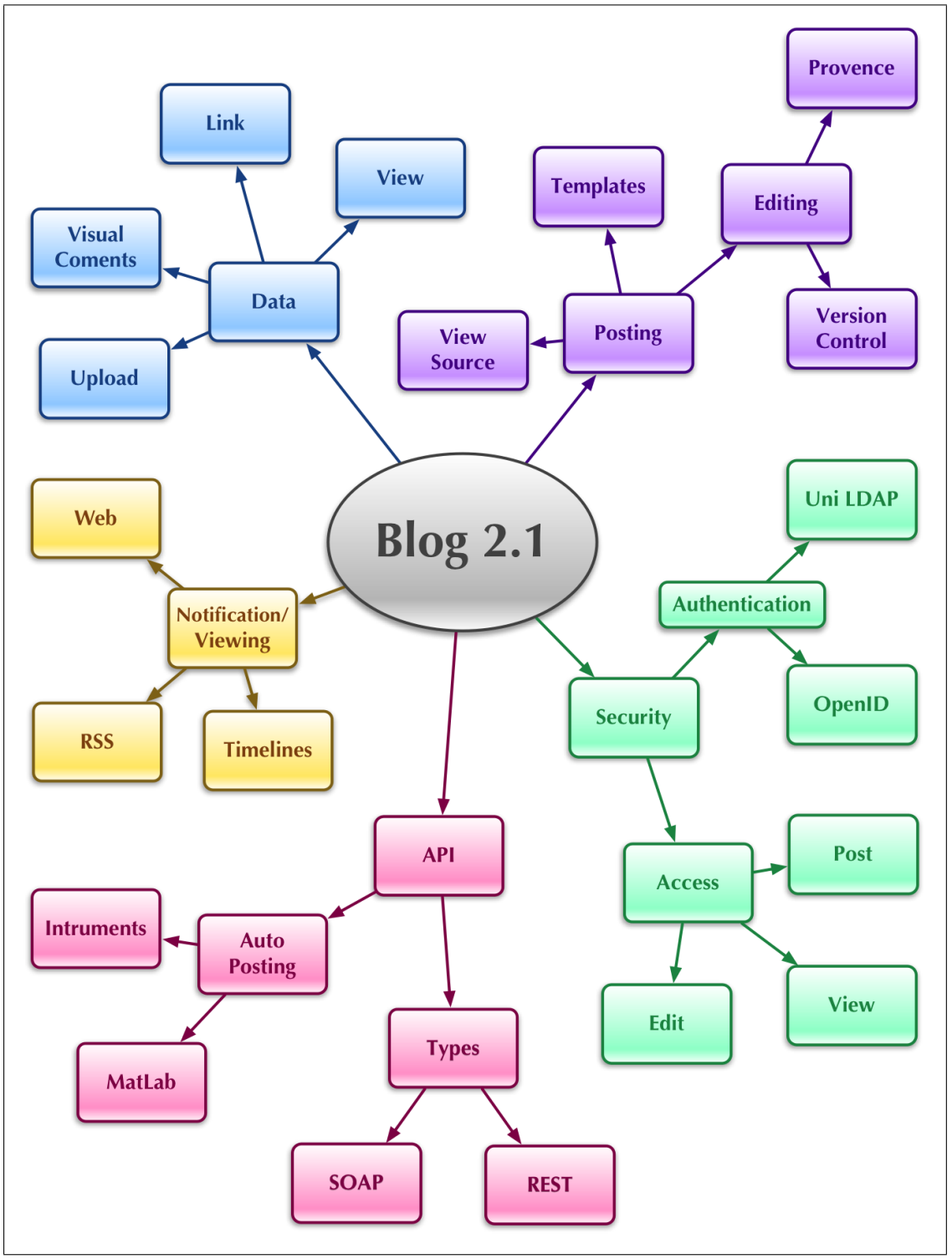


Figure 6.9: A feature overview of Blog2.1

file would be orphaned, in the sense that it would not be attached to a post. Attaching a data field during a subsequent revision ensures that the post does already exist.

A revision of this was to allow the user to attach data to the post as they are creating it, therefore the model of a draft was developed, this allowed the user to have all the

Current user: Andrew Milsted | Log Out

Admin | Dashboard | All Blogs | Help | Support | About

blogs@ChemTools

Investigations into neutral drift

<< Next Post

Previous Post >>

Search

Digestion 5025/29 (experiment 2, round 2)

6th November 2007 @ 11:06

Post Type: Digestion
Risk assessment: Digestion risk assessment

Reaction	DNA	µL	Water	µL	Buffer	µL	BSA	µL	Enzyme 1	µL	Enzyme 2	µL	Product
1	Purified mutagenesis product 5025/28 (experiment 2, round 2)	45	None	0	NEB Buffer 4	6		6	EcoRI	3	NcoI	3	Digestion 5025/29 product 1 (experiment 2, round 2)
2	p042 (27/9/07)	10	Molecular biology grade	5	NEB Buffer 4	2		2	EcoRI	1	NcoI	1	Digestion 5025/29 control product (experiment 2, round 2)

Enzymes are always diluted 1:1 in water.
The reactions were set up as listed in 1.5 mL tubes and incubated in a waterbath at 37°C for 3 hours to give the products listed.
The products were run on an agarose gel as follows:

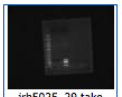
Risk assessment: DNA gel risk assessment

Gel properties

Reagent	Property
Agarose	0.6 g (low melting point)
TAE buffer	40 mL
Gel %	1.5
Voltage	80 V
Time	55 minutes

An agarose gel was prepared by dissolving the agarose in the TAE and being set in the mould.

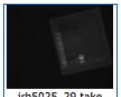
Samples

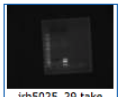
Lane	Sample	µL	Product
1	DNA ladder	10	N/A
2	Digestion 5025/29 product 1 (experiment 2, round 2)	63	Purified digestion 5025/29 gene product (experiment 2, round 2)
3	Digestion 5025/29 control product (experiment 2, round 2)	20	N/A
4			
5			
6			
Gel picture	 jrh5025_29 take 2.jpg	Gel shows left to right DNA ladder, cut out bands of Digestion 5025/29 product 1 (experiment 2, round 2) across 2 lanes, and positive control bands showing a successful reaction	

The gel was run as listed, stained in ethidium bromide and destained in water. The gel was visualised under UV light, the product bands removed and the gel photographed.
The product bands were purified using the Wizard kit.

Linked Posts

Attached Files

 jrh5025_29 take 1.jpg

 jrh5025_29 take 2.jpg

Jennifer Hale | Edit Post | Procedure | Comments (1)

This Post

Permalink
URI
URI Label
Revisions
Export:
XML

This Blog

New Post
Blog Settings
Timeline View
Export Blog

Archives

July 2012 (3)
June 2012 (1)
May 2012 (3)
May 2010 (1)
November 2009 (1)
September 2009 (1)
August 2009 (1)
April 2009 (6)
(more)

Authors

Cameron Neylon (3)
Jennifer Hale (1924)
Wendy Suzanne Smith (237)

Sections

Materials (90)
Notes (47)
Procedure (405)
Product (1578)
Safety (11)
Summary (1)
Templates (32)

Post Type

DNA Gel Product (130)
Digestion Product (122)
Ligation (58)
Note (46)
PCR Product (296)
Transformation (63)
Plasmid (uncharacterised) (210)
Strain (uncharacterised) (687)
(more)

Tools

Show/Hide QR Code
Show/Hide Keys

Figure 6.10: A screen shot showing an example single post. Public URL

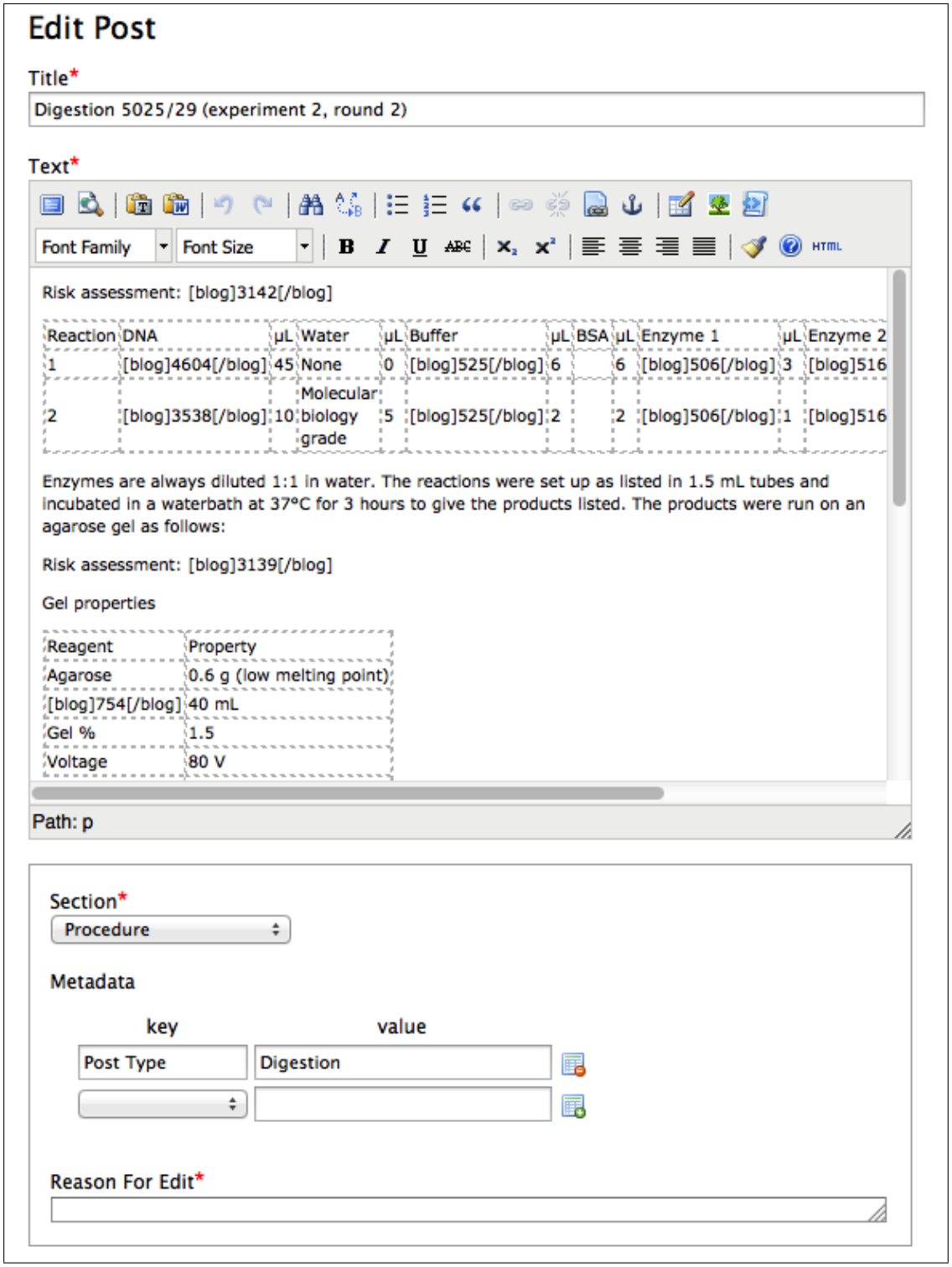


Figure 6.11: A screen shot showing the edit interface of the above post with the new TinyMCE editor

editing features, even when they have just clicked 'Add Post'. As LabTrove now tags new posts as a draft, it also allowed users to keep their post in draft mode until they wanted to publish their work, this allowed them to save for later instead of publishing incomplete posts.

In earlier versions of LabTrove users can format the contents of the Text field with the BBCode markup language, either by marking up the text or by using the toolbar. However, the interface provides neither "what you see is what you get" (WYSIWYG) editing nor HTML markup. We chose BBCode in preference to HTML markup because the latter is more complex. In practice users can readily describe how they want to format a post and not having a WYSIWYG editor offers very little hindrance to users.

With advances in third party WYSIWYG editors it was decided to add such a feature to LabTrove. TinyMCE[109] was chosen because of its mainstream uptake and because it has a very flexible plugin framework to allow integration into some of the other LabTrove Features.

When a user views a post, with or without comments, LabTrove has rendered the BB-Code as HTML, using the style sheets provided, for presentation in the client browser. Additional fields for adding metadata are present when a user creates or edits a post. Clicking the buttons to the right of the Section field and all other key fields provides a drop-down list of existing values and the option to add a new key. A Section value of Templates marks the post as a template, causing the server process to interpret any placeholder markup and render the post for template-style input.

As with most blog interfaces, the navigation pane is to the side of the content pane. At the top, under the heading 'This Blog' are options pertaining to the e-Notebook and, when viewing or editing a post, the options pertaining to This Post. Below these options are the Archives for the e-Notebook and then the Sections, with each of the values assigned to that key appended with the number of instances. Similarly, below the Section list each of the other metadata keys is presented with the assigned values and number of instances. By selecting a Section or other metadata key, users can filter the posts listed in the content pane.

Under 'This Blog', there is also an Export Blog option; under 'This Post' the Export options enable users to serialise a post as XML and to capture an image of the post in PNG format. All views include a search field at the top of the navigation pane, which enables users to search the e-Notebook for occurrences of specific text strings. LabTrove displays the search results in the content as a list of posts, annotated with the author, date, and time of each post. LabTrove uses style sheets in CSS format to control the presentation in the client browser. Administrator users can modify these style sheets to customise the presentation format, for example to increase the otherwise fixed width of a template to accommodate wide tables.

6.2.1.1 Commenting

Any post in the blog can be commented on by any of the the poster's peers. This is a good medium for supervisors to leave feedback on the post. Another use, is the science of a scientist collecting some data from a instrument and that instrument auto blogs:- the scientist could then instantly comment on the quality or usefulness of the data produced.

If the Trove is a closed instance comments would be used by peers, people in the same research group, but if the Trove was being used more publicly, then members of the same research community might be invited to comment with suggestions for further exploration or noticing a point of interest in their line of investigation.

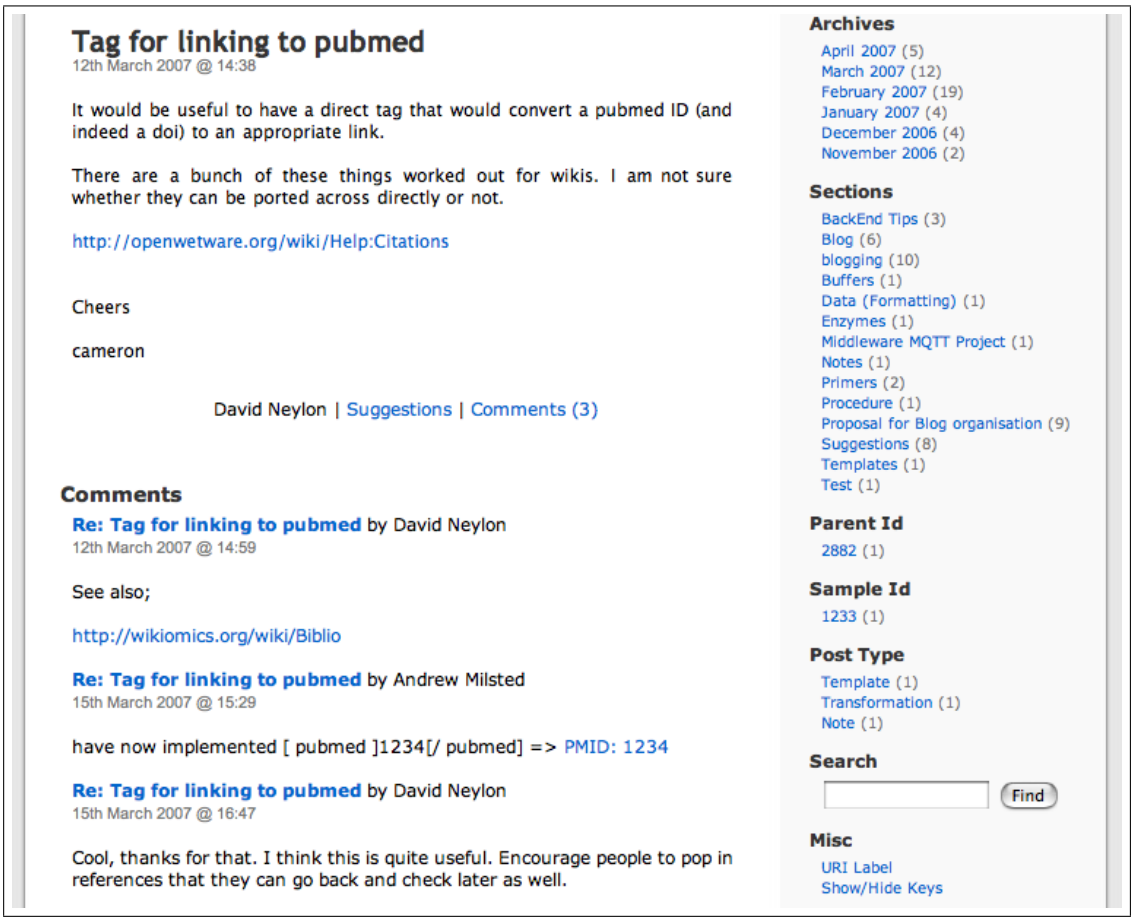


Figure 6.12: Showing a post with 3 comments below it.

6.2.2 Pictorial Comments

One feature of a paper labbook is the facility to draw over diagrams which have been stuck into the book [110]. One technology that simulates this is the tablet PC, as it allows the user to draw on the screen with a stylus. A way of drawing overlays onto an image was implemented.

As with text comments, the system can identify the user that made the comment and allow anyone who is enabled to view the content to select which layers to see. Whether they are a colleague suggesting a redesign of an experiment or a biochemist commenting on activity upon a gel plate, they can simply draw on the image. As the proverb goes 'picture paints a thousand words a pictorial comment could potentionally be worth the same as it can make it easier for the commenter describe their meaning.

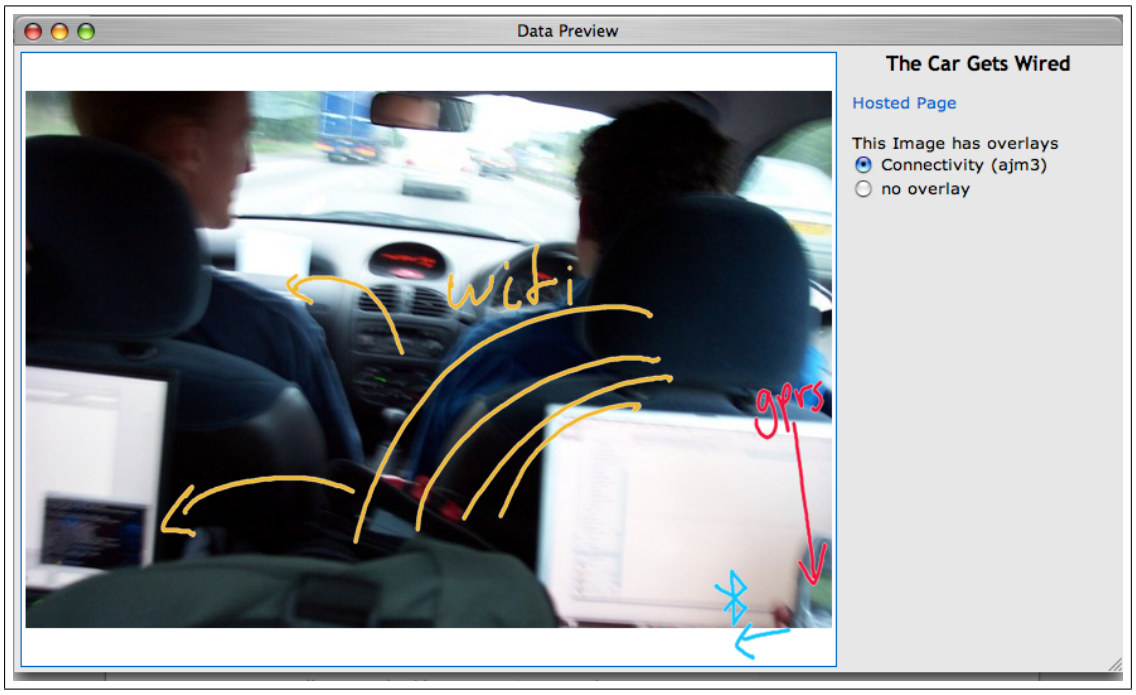


Figure 6.13: An example of Pictorial Comments

6.2.2.1 Revisions

The blog records a new copy of each post when a user commits a save. This is to allow a record of exactly how the blog looked at a given time. This could be solved by not letting users edit the posts once posted, but in practice users always want to correct small mistakes, or save the post half way through the post to ensure their work is safe.

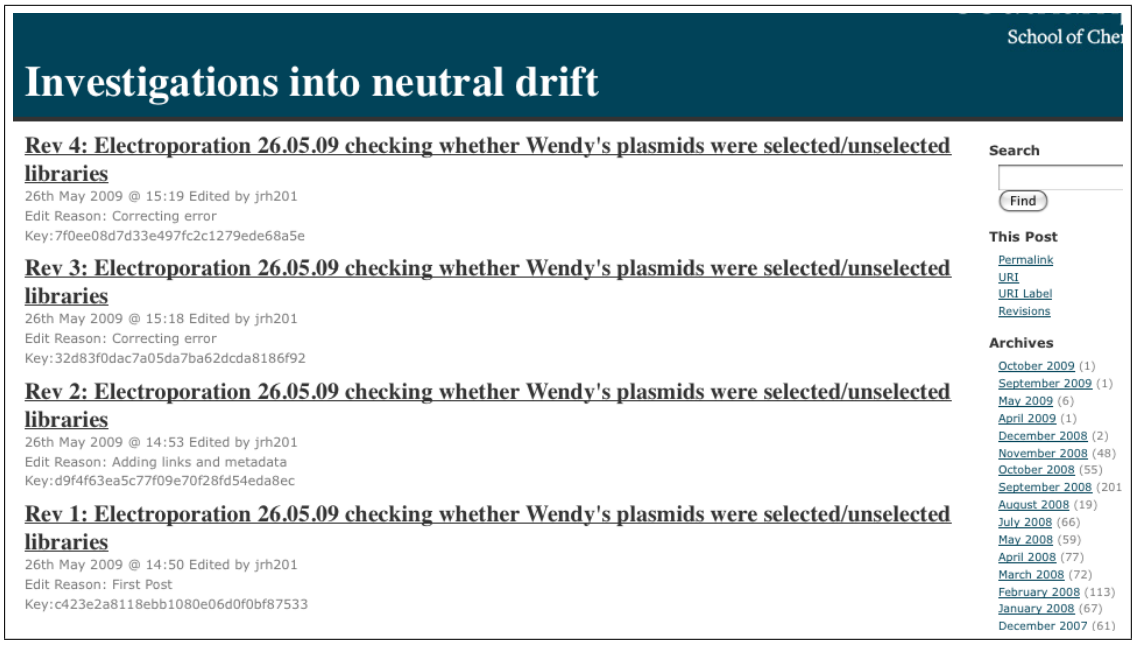


Figure 6.14: A screen shot of the revions page.

The interface allows the user have a look at all the version of the post along with the required reason for the edit.

6.2.3 Provenance

Real labbooks are a very trustworthy system as it is very hard for anyone to change the order or dates of entries within it. This is because if everything is written in chronological order, then any sign of tampering with the book will be very evident. The problem with electronic storage of entries, is that it is possible for someone to change the date or the content of any entry. This could be done to claim that work was carried out before it was or to satisfy drug discovery requirements.

In the case of our blog, even following all security precautions for the server, it would still be possible to change data within it if physical security was breached. Also in a worse case scenario, it would be very easy for a system administrator to make any changes.

A few steps have been taken to help prevent this from happening:

Every time a new post or comment is made a MD5[111] is made of all the data that relates to the post. This is then added to the entry, and contains a date stamp of when the post was entered. If a post or comment is edited, it is saved as a new entry, no matter how trivial the edit is. A new MD5 is then saved with a new time stamp. This in itself doesn't solve the provenance security issue as any attempt at altering the data could simply adjust the MD5 hash to the new content.

An alternative would be for the MD5 to be recorded with a 3rd party service that could be trusted to report that the MD5 hash was submitted at a particular point in time. At the moment there is no such service, because the trust of this service could also be compromised. For a system to work there would need to be a set of services that are self-trusting, which would be achieved by them all communicating with each other.

Another way of recording these MD5 hashes is to have a dot matrix printer with a continuos tractor paper feed that prints an MD5 hash and date stamp every time one

is submitted, It could then to be verified by a person every so often by a signature to check the paper hasn't been broken.

We have decided to implement a slightly simpler method, which is based on a paper labbook. With the use of a label printer the scientist could then print out the MD5 hash as a sticker and place it into a labbook. This would have the same provenance integrity as the original paper labbook, so nothing is lost in the technology shift.

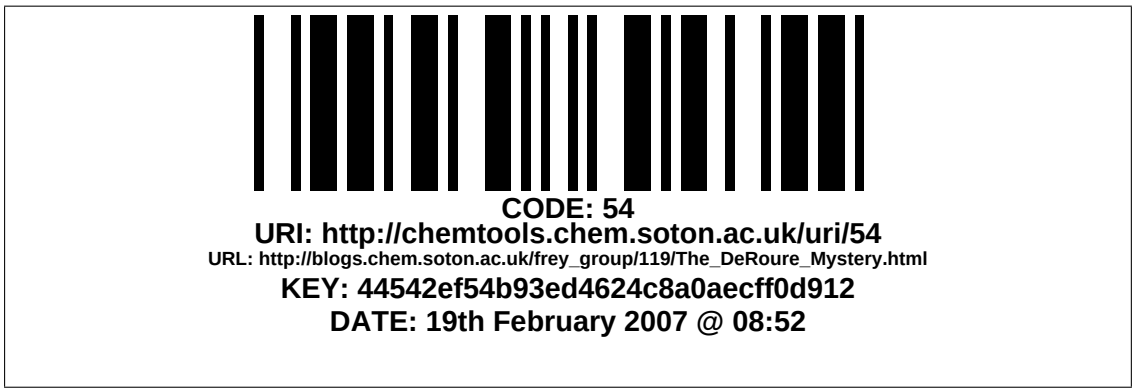


Figure 6.15: An example of a URI barcode label

The labels are printed on a thermal label printer and they also have a URI which is either internet resolvable or by using a barcode reader and chemtools they will link back to the original post.

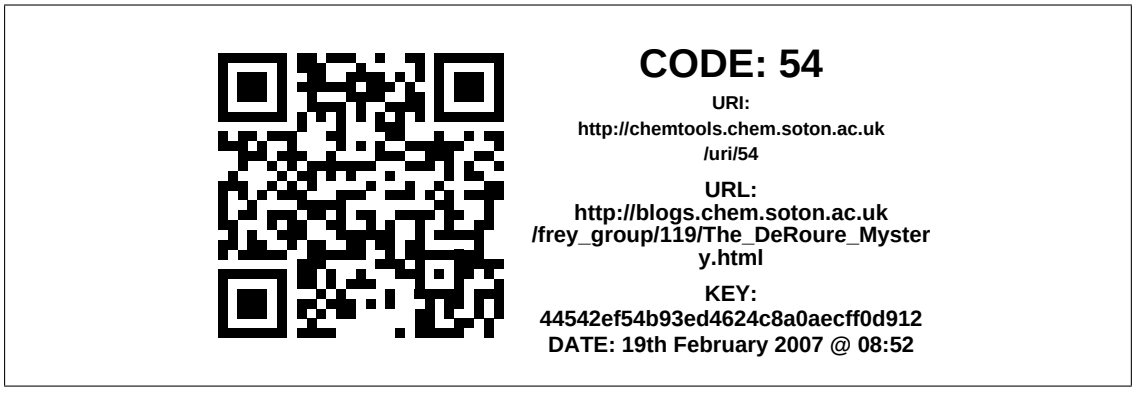


Figure 6.16: An example of a URI QR code label

A QR(Quick Response) code[112] version of the label was produced, The QR Code can be read by any device with a camera, ie a mobile or webcam, which forgoes the need for investment in barcode reader. The QR code is a 2-dimentional barcode which means it can include a small amount of text data, the full resolvable URI is encoded into the QR

code so when read the device will know it is an object that has a web page so will guide the user to the end blog page.

6.2.4 RSS Feeds

Notification of new posts and comments became a requirement for the blog, as it was being used by supervisors to oversee the research of their students and the students to see their supervisors comments. This was identified when users started using LabTrove during the initial development process, users where struggling to notice new posts especially if they where following and engaging with a number of Troves. RSS (Really Simple Syndication) provides a practical solution that is very accessible and there are many RSS feed viewers, including the main stream browsers. When the RSS feed was first implemented, it only included posts. It became very clear that comments needed to be added to the feed as the it is important for the user to be notified when new comments had been made.

The RSS feed is simply an XML page view of the last 30 (configurable) posts. It contains several pieces of information including title, date, content, author, URI and Uniform Resource Locator (URL). It shows the number of recent posts, instead of showing only the new ones (since last RSS request) because this removes the need to track which clients have seen which posts as new and it allows a new client to "Catch Up" on the posts.

6.2.5 Linking

Cross linking between posts was implemented in order to make the navigation throughout out the blog easier. In the example of the biochemists' blogs, this feature is being used to track parallel experiments which have multiple tracks. It is also being used to track starting materials and products, and being able to track lineage of products which are a result of a series of experiments.

When the links have been established between posts, the blog is dynamic enough to fetch the title of the linked post and then create an HTML link. When this has been

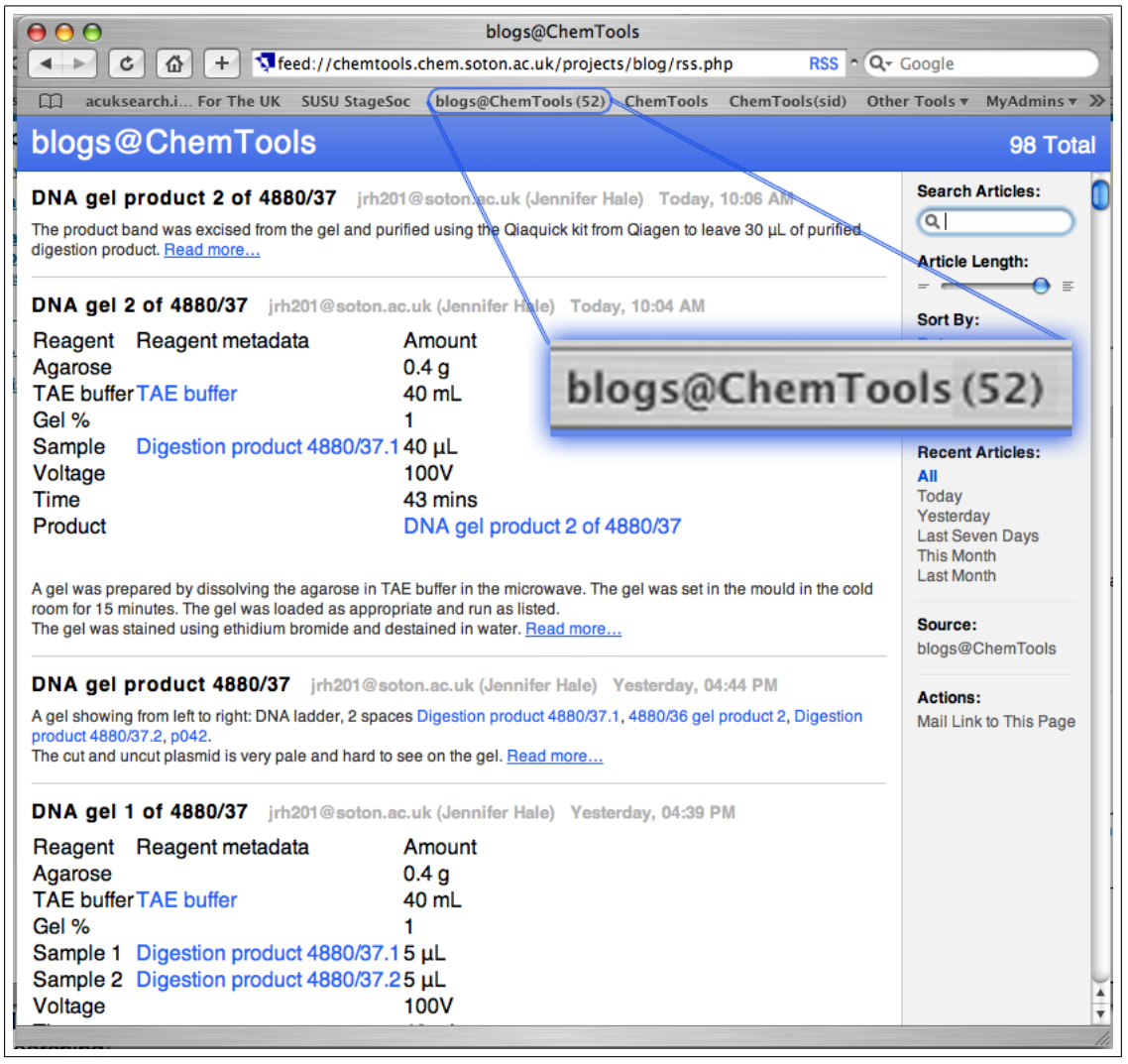


Figure 6.17: Safari's in browser RSS feed viewer

done, the linked post will also get a linked post to make this join bidirectional. This is then displayed by creating a "Linked By" list that is placed at the bottom of the post, this can be seen in figure 6.18.

6.3 Handling Data

Any scientific online resource must have the ability to store data and attach it to posts. It can be done simply by uploading data directly into the blog. This type of submission is designed for one off data, for example a photo of your equipment setup, or a model of a predicted result or raw data.

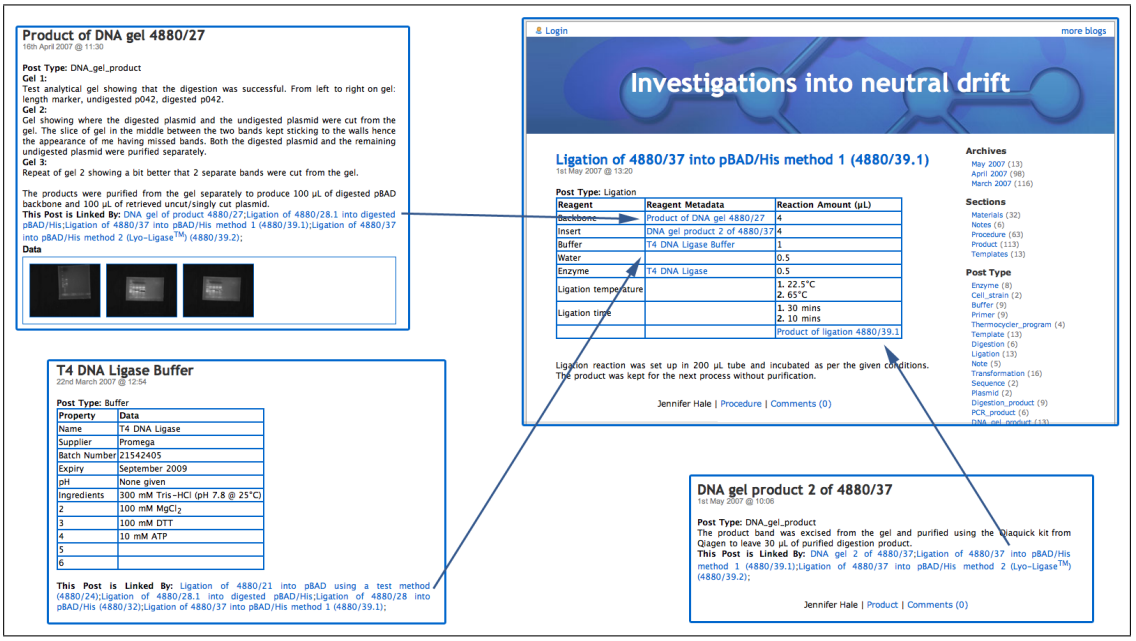


Figure 6.18: Crosslinking within the blog

With advances in browsers, the there is now no realistic limit to the size of the uploaded file so now LabTrove has default limits of 500MBs, but when data is being produced by automated processes the data files can be added via the API (Section 6.5). This can be done in two ways the data can be done in two ways; first the data can be copied in to the blog, this is useful because the data is stored alongside the research inside LabTrove, preventing a possible disconnect with the data. Secondly, if the data is being stored in a managed form, for instance a national laboratory service archive (eg Diamond ICAT [113]) then a URL can be used to link the data to LabTrove. This is especially useful when the data is very large (many GBs) not having to handle the large files when they are being managed somewhere else. But this does rely on a trust in the data provider, ultimately it depends on the situation.

All data items have an XML representation (See 6.19), this offers flexibility in storing the metadata about the data. One feature is the ability to store multiple versions of the data, eg if the original data item is a .BMP then LabTrove can store a JPEG version of the same image. This is done in this case because a .BMP is not supported by many browsers to display as an image, so LabTrove will create the .JPEG version and present that to the user on viewing, but the .BMP is still available for the user to download the original. This is done by default on a number of non browser friendly image formats

(.BMP,.TIFF,.EPS)

The following skeleton xml is used to store the key information about the data in order to help the blog display the data appropriately:

```
<metadata>
    # Some text that describes set. (Required, 1 only)
    <title>1M Test Compound</title>
    # A version of the data (Required, 1 or more)
    <data_png> # the _asc defines file type
       # <type> where the data is stored"(Required,1 of following)
       # "local" stored within the blog
       # "url" stored within somewhere else
       # "inline" stored inside this metadata
          <type>local</type>
          <id>468</id> # if type = local
          <url>http://example.com/data.png</url> # if type = url
          # if type = inline (data base64)
          <data>VZCT1J3MEtHZ29BQUF...</data>
          # the blog should try and render this (Optional,max 1)
          <main>1</main>
    </data_png>
    # optional version of the data(Optional)
    <data_zip>
        <type>local</type>
        <id>469</id>
    </data_zip>
</metadata>
```

Figure 6.19: XML used to store data in LabTrove

6.4 Templates

In the case of biochemists blogging their synthesis procedures, it became clear that this is time consuming process. A feature to assist this is templates, that reduces this effort. These are blank blog posts that are rendered with empty text fields, into which the user fills in the relevant data, enabling the user not to worry about code behind the blog post.

The template system works by creating a post that is given the section “Templates”. They can use special tags that inform the blog as to where to place the text boxes. The user then clicks a “Use Template” link and a new post is created in a “What You See is What You Get” (WYSIWYG) environment.

As the blog has a metadata store about posts the template designer can use this information to make a drop down list of all previous posts in a section that have a particular metadata tag. For example if a user wishes to collate source materials and these have been blogged in the section "Materials" then they can make a drop down box with the code *[[Section>Materials]]* . To extend the feature the user can use wilcards, '%' to allow the selection of multiple groups.

One feature that the template tool doesn't do at the moment is auto create "children" posts with appropriate linking. The problem comes with the order of posting. To make sense, procedure posts should be created before the product posts because that's the order in which it happened. But, then once created, the user needs to go back and enter the links to the children on the parent page. An ideal solution for this would be to use a predefined workflow that can automatically create the products and procedure pages with all the links. For example a Taverna[114] workflow might be able to use the, later mentioned, API to make these posts (Section 6.5).

An in depth description on how the templates can be used in practice is discuss later in Section 7.2.3.

6.5 API

To integrate the blog with other web services, an API is essential. If a user has an automated process and they wanted to record the research process they could use the API to facilitate this.

6.5.1 The REST API

A representational state transfer (REST) API was chosen as it is a style of software architecture designed ideally suited to applications that are web based as it uses Hypertext Transfer Protocol (HTTP) as a protocol.

LabTrove provides two different parts to its REST API, the read and write parts.

Firstly the read part is driven by the main LabTrove system, it uses extention driven URLs, for example to get the api access to the single post

`http://blogs.chem.soton.ac.uk/neutral_drift/41420/Betagalactosidase_work_from_early_2006__part_6.html`

it is accessed by changing the html extension for .xml

`http://blogs.chem.soton.ac.uk/neutral_drift/41420/Betagalactosidase_work_from_early_2006__part_6.xml`

The user will then get an XML representation of the post, which then can be read by their application.

```
<post>
      <id>41420</id>
      <rid>41420</rid>
      <title>Beta-galactosidase work from early 2006 - part 6</title>
      <section>Notes</section>
      <author>
        <username>jrh201</username>
        <name>Jennifer Hale</name>
      </author>
      <content><![CDATA[
  <p>A numerical error in the Beer-Lambert calculation used to convert the
  raw absorbances into concentrations has led to a lot of the data in assay
  4712/2 being wrong. I have now corrected the data and updated the graphs.
   This is now the corrected data file.</p><!--HTML-->
  ]]></content>
      <html><![CDATA[
  <b>Post Type:</b> Note<br /> <p>A numerical error in the Beer-Lambert
  calculation used to convert the raw absorbances into concentrations
  has led to a lot of the data in assay 4712/2 being wrong. I have now corrected
  the data and updated the graphs. This is now the corrected data file.</p>
   <!--HTML--><div class="postTools"></div>
  ]]></html>
      <datestamp>2012-08-31T12:52:21+00:00</datestamp>
      <timestamp>2012-08-31T12:58:54+00:00</timestamp>
      <blog>13</blog>
      <key>317b379a1712008b4ae9868136a846c8</key>
      <metadata>
        <post_type>Note</post_type>
      </metadata>
      <attached_data>
        <data>http://blogs.chem.soton.ac.uk/data/28267.xml</data>
      </attached_data>
      <links>
        <uri>http://chemtools.chem.soton.ac.uk/uri/601d</uri>
```

```
      <permalink>
http://blogs.chem.soton.ac.uk/neutral_drift/41420/Betagalactosidase_work_from_early_2006__pa
</permalink>
    </links>
    <formats>
      <format type="text/html">
http://blogs.chem.soton.ac.uk/neutral_drift/41420/Betagalactosidase_work_from_early_2006__pa
</format>
      <format type="text/xml">
http://blogs.chem.soton.ac.uk/neutral_drift/41420/Betagalactosidase_work_from_early_2006__pa
</format>
      <format type="image/png">
http://blogs.chem.soton.ac.uk/neutral_drift/41420/Betagalactosidase_work_from_early_2006__pa
</format>
    </formats>
    <revisions>
      <revision current="true">
http://blogs.chem.soton.ac.uk/neutral_drift/41420/Betagalactosidase_work_from_early_2006__pa
</revision>
    </revisions>
    <comments/>
  </post>
</posts>
```

To get a full list of posts in a Trove the user can place a index.html onto the end of an the Trove's URL

`http://blogs.chem.soton.ac.uk/neutral_drift/index.xml`

The second part of the API enables the user to add a post, edit an existing post, or add data to a post. This is described in some detail on the appendix B3 under the LabTrove Manual.

6.5.2 Auto Posting

Uses for the API include an auto posting client, where the client watches a folder and as files are added to the folder by some experiment, the client then uploads the data into LabTrove, along with a post that describes the data added that it extracted from the data files. By giving the experiment its own Trove, a full log of the experiment is created, then the user can refer to experiment entries in their own Trove creating a link between them.

Chapter 7

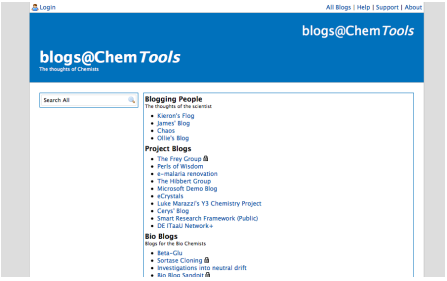
LabTrove Evaluation

7.1 Implementations of LabTrove

There are now several real world examples of LabTrove, this has been predominantly down to the fact that the LabTrove software is open source therefore freely available. Free hosting has been provided to research groups around the world in order to trial LabTrove. This has been done through the ourExperiment initiative provided by the University of Southampton.

Many of these instances are open and their content is openly viewable. Examples include:

1) blogs@ChemTools



A LabTrove instance provided for members of the School of Chemistry at University of Southampton.

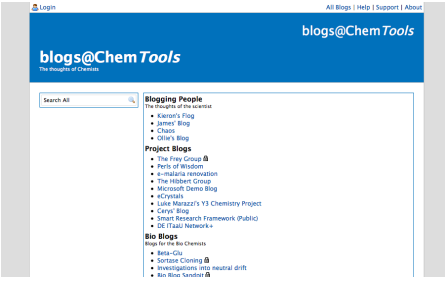
URL: `http://blogs.chem.soton.ac.uk`

Numbers: 84 Troves, 65 Users

Contact: Professor Jeremy Frey, University of Southampton, `j.g.frey@soton.ac.uk`

Visibility: mixed (exemplars are public)

2) UltraFast Xray Group



A LabTrove instance provided for members of the UltraFast Xray Group, Optoelectronics Research Centre (ORC).

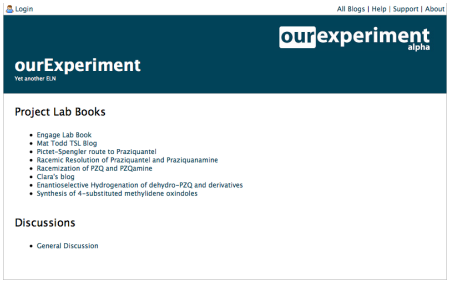
URL: `http://xray.orc.soton.ac.uk`

Numbers: 18 Troves, 25 Users

Contact: Professor Jeremy Frey, University of Southampton, `j.g.frey@soton.ac.uk`

Visibility: private

3) ourExperiment



An open LabTrove instance where anyone can signup and use the software.

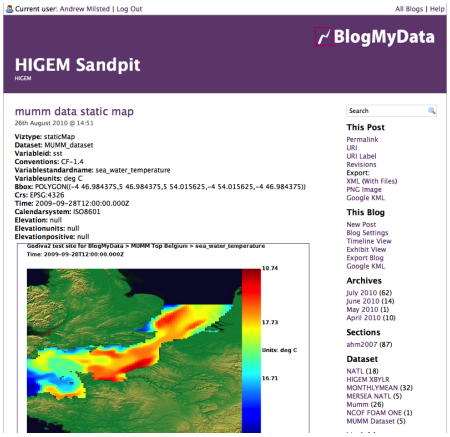
URL: `http://www.ourexperiment.org`

Numbers: 38 Troves, 26 Users

Contact: LabTrove, University of Southampton, `support@labtrove.org`

Visibility: public

4) Blog My Data



A LabTrove instance where users of the Godiva software can post points of interest into a LabTrove notebook for discussion.

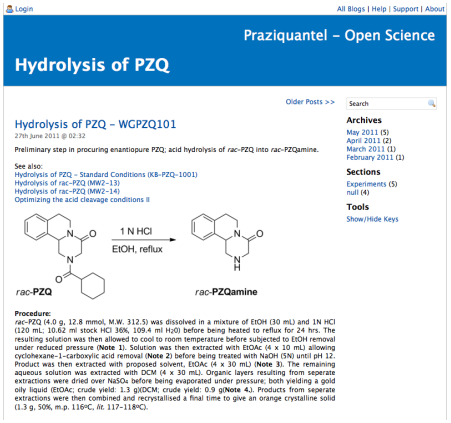
URL: `http://blogs.blogmydata.org`

Numbers: 6 Troves, 4 Users

Contact: Dr Jon Blower, University of Reading, `j.d.blower@reading.ac.uk`

Visibility: private

5) Praziquantel



An Open Science initiative to enable cheap drug discovery. 'Improved Synthesis of an Important Drug via Undergraduate Collaboration'

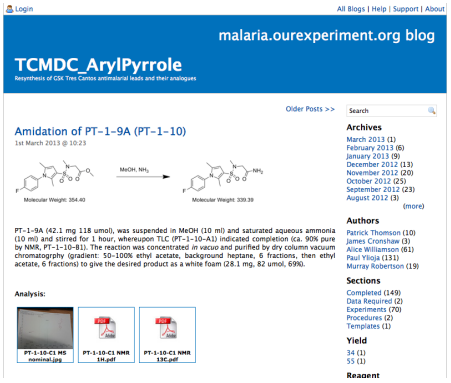
URL: `http://pzq.ourexperiment.org`

Numbers: 12 Troves, 8 Users

Contact: Dr Matthew Todd, University of Sydney, `matthew.todd@sydney.edu.au`

Visibility: public

6) Open Malaria Research



The open lab notebook for a hub for global efforts in open source drug discovery for malaria.

URL: `http://malaria.ourexperiment.org`

Numbers: 22 Troves, 33 Users

Contact: Dr Matthew Todd, University of Sydney, `matthew.todd@sydney.edu.au`

Visibility: public

Table 7.1: Timeframes for various case studies of LabTrove

Use Case	No of Notebooks	Users	Length of Use
Jennifer Hale	2	1	4 Years
The ORC Physics Group	18	25	7 Years
Open Drug Discovery	12	7	2 Years
Unilever	*	10	6 Months
Blog My Data	6	4	1 Year
UNSW	*	74	1 Year

*Data not avaiable.

7.2 User Experience:The Biochemist: Jennifer Hale

The following follows the experiences of a PhD bio chemist student, Jennifer Hale, who was willing to trial the LabTrove during its early development. Her experience provided valuable insight into how LabTrove could be used, also it is very fortunate that her work was being released into the public domain and is freely available. She created two notebooks 'Beta-Glu' (`http://blogs.chem.soton.ac.uk/beta_glu`) and 'Investigations into Neutral Drift' (`http://blogs.chem.soton.ac.uk/neutral_drift`), Her supervisor, Dr Cameron Neylon, also created the blog 'Sortase Cloning' (`http://blogs.chem.soton.ac.uk/sortase_cloning`) Using these notebooks as examples we will make references to individual posts in order to show observations and conclusions made.

7.2.1 Using LabTrove as a simple journal notebook

The natural starting point in applying a Blog as a laboratory notebook system is to treat it as a journal, effectively a web based analogue of the paper notebook. As noted previously the digital nature of the Blog provides a number of advantages over paper including automated backup of data and protocols and the ability to do simple text searches. The comment facility, common to most blogs, allows notes to be taken during the process of an experiment or for other researchers to ask questions, comment or offer advice.

One of the problems identified by both these users and other groups [115] is the need for editing of posts, both to correct typographical and other errors but also for adding observations over the course of extended experiments. These can be added as comments but this removes this information from the core of the post. Making changes to the notebook raises a number of significant issues, the most important being that of the reliability of a modified record. It is generally regarded as axiomatic that laboratory notebooks should not ever have material deleted or modified.

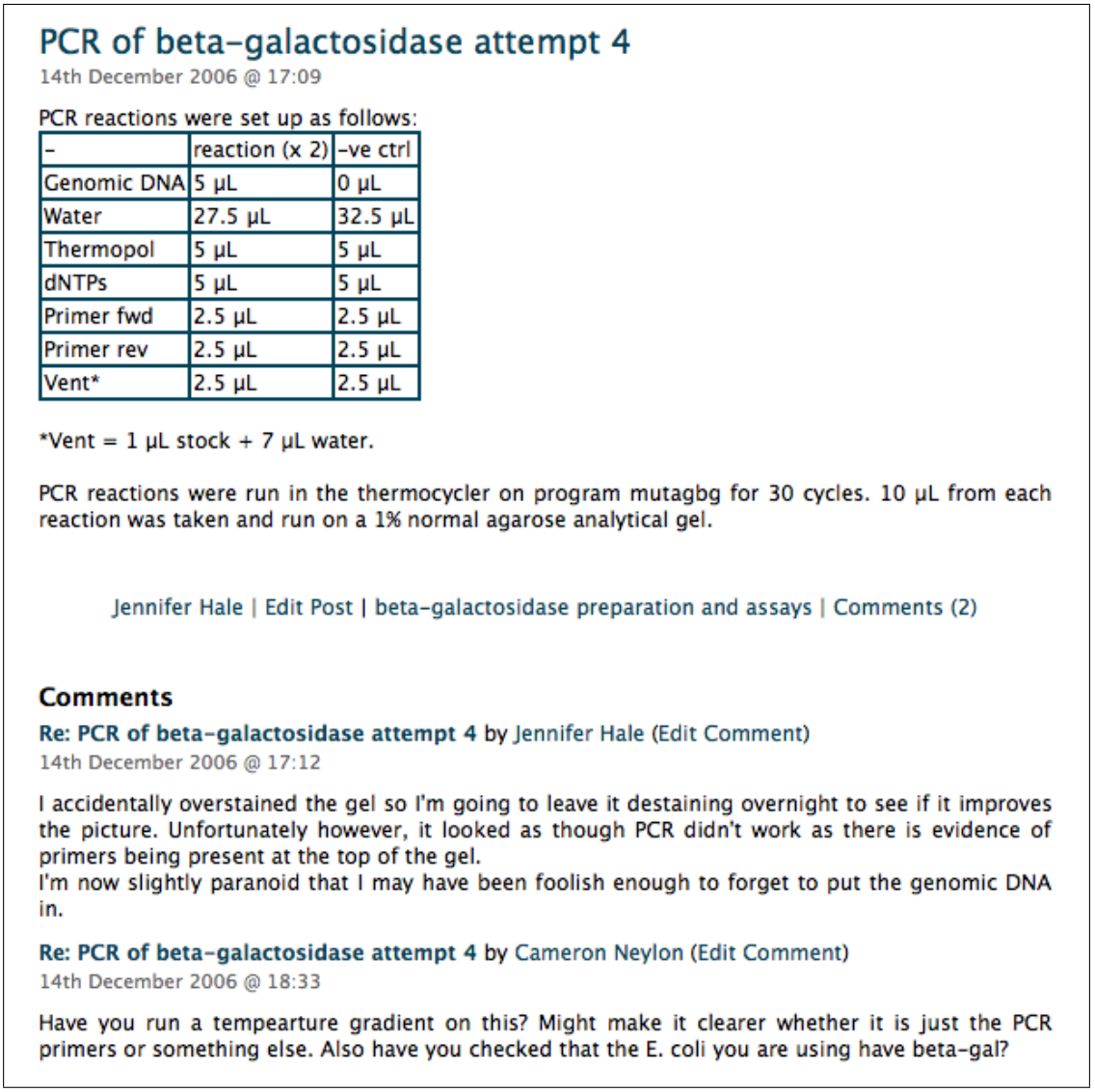
Allowing changes to be made to this primary record is therefore a significant departure. However there is not necessarily a conflict here. The complete and original record is still maintained, even if it is not what is presented by default. In most Wiki implementations a complete record of all changes to a page is recorded making it possible to track back through individual versions, this is traditionally not the case with Blogs. To enable modifications to be made and tracked the LabTrove system required the development of a versioning system.

7.2.1.1 What makes a post

Having made the choice of a use of a blog we implicitly have taken the view that the laboratory notebook will be made up of a series of (interlinked) posts. This raises the question of what is, or should be, the appropriate content, or indeed size, of a single post? The simple answer is 'one experiment' however it is not necessarily clear what 'one experiment' consists of.

The first use of LabTrove as a simple journal can be seen in the early entries (Nov/Dec 06) on the 'Beta-Glu' notebook. Experimental procedures were recorded either in free text or in tables and data, generally images, were uploaded into the procedure post. A post could cover both the preparation of samples, their processing, and their analysis. In some cases it is can be unclear which sample relates to which analysis. There is no informational link between samples, their input materials, preparation, and analysis. While capturing of the laboratory process is enabled the subsequent processing of this information is not facilitated. There is little scope for the utilisation of metadata as

the content and context of a post is difficult to define as it consists of so many different things.



PCR of beta-galactosidase attempt 4

14th December 2006 @ 17:09

PCR reactions were set up as follows:

-	reaction (x 2)	-ve ctrl
Genomic DNA	5 µL	0 µL
Water	27.5 µL	32.5 µL
Thermopol	5 µL	5 µL
dNTPs	5 µL	5 µL
Primer fwd	2.5 µL	2.5 µL
Primer rev	2.5 µL	2.5 µL
Vent*	2.5 µL	2.5 µL

*Vent = 1 µL stock + 7 µL water.

PCR reactions were run in the thermocycler on program mutagbg for 30 cycles. 10 µL from each reaction was taken and run on a 1% normal agarose analytical gel.

Jennifer Hale | Edit Post | beta-galactosidase preparation and assays | Comments (2)

Comments

Re: PCR of beta-galactosidase attempt 4 by Jennifer Hale (Edit Comment)
14th December 2006 @ 17:12

I accidentally overstained the gel so I'm going to leave it destaining overnight to see if it improves the picture. Unfortunately however, it looked as though PCR didn't work as there is evidence of primers being present at the top of the gel.
I'm now slightly paranoid that I may have been foolish enough to forget to put the genomic DNA in.

Re: PCR of beta-galactosidase attempt 4 by Cameron Neylon (Edit Comment)
14th December 2006 @ 18:33

Have you run a tempearture gradient on this? Might make it clearer whether it is just the PCR primers or something else. Also have you checked that the E. coli you are using have beta-gal?

Figure 7.1: An early post showing very little structure, `http://blogs.chem.soton.ac.uk/beta_glu/52/PCR_of_betagalactosidase_attempt_4.html`

The question being asked is what is the 'atomic' portion of a laboratory record. Each post provides a URL, an identifier for a specific entity. The best way to phrase this question is therefore what is the smallest unit of research which we wish to be able to uniquely identify? If our aim is to enable the identification of specific objects via the semantic web then which objects do we need to point at? Clearly individual samples, and individual datasets should have identifiers an to make sense of the relationships we need identifiers also for the process that links these together. This lead to the adoption of the one-item one-post system described in section.

7.2.1.2 Developing a metadata framework

Clearly the choice of how to divide up the research process into posts has a significant impact on the way in which those posts are categorised. The development of using metadata over the initial investigation of blog usage was significant. Most Blogs utilise tags to categorise posts, this is a simple and reasonably effective means of filing. However the key-value pair system implemented in the LabTrove makes more sophisticated processing possible.

A simple example of this is the division of the Blog into sections where each section contains posts with a different function. Common sections are Materials, Procedures, Products, and Notes. Some Blogs also contain Safety and Data sections. This is then kept separate from descriptions of the type of material, the particular investigation a post relates to, or the location where the experiment was carried out. However, while this allows categorisation it does not provide any view of the workflow of a specific experiment; what input materials have contributed to the generation of a specific sample. None of this provides a link from one procedure to the next.

The first attempt to provide this link was to introduce a metadata key 'Sample Parent'. It quickly became evident that this would not work as the system only allows one entry per post for each piece of metadata (i.e. for a given post 'Sample Parent' can only have one value). While this is a limitation of the system it raises a more general issue for procedures carried out in parallel. If a procedure involves several parallel actions on different, but equivalent materials (e.g. five digestions on different PCR products) then associating multiple 'sample parents' with the procedure will not make it (directly) possible to tell which reaction is being done to which sample. As the object was to make these kind of links machine readable this was not the best approach to use. It was clear that there was a need to be able to provide a pointer to each sample, each protocol, and each output to enable the connections between these to be mapped out.

The adoption of the one-item one-post system solved the problem of the connection between samples and procedures via the use of hyperlinks. This enables the use of the metadata system to focus on characteristics of each post rather than the relationships

between them. This drove the adoption of broad categories of post 'material', 'procedure', 'data', as well as more specific characterisation of each. Ultimately the system of metadata organisation was driven by the need for templates to function efficiently (section 7.2.3) and this has lead through a number of stages of development to organisational systems that are appropriate for each user.

7.2.1.3 Conclusions from the initial investigation: Specific requirements

Two important sets of conclusions came out of our initial investigation. Firstly a specific set of requirements were developed. The system must enable and support the publishing and linking of posts that refer to individual research objects, samples, procedures, and data sets. The addition of templates would support this approach to post production and to the linking of posts together to describe relationships between them. At the same time the system must enable a user to enter free text without structure if that is appropriate. The use of the template system requires that posts be characterised in a way that the templating system can use effectively to present the correct set of possible inputs or outputs to the user. The metadata organisation scheme and the design of templates are therefore highly dependent on each other.

The other important result that arose from our initial investigations was that our organisational approach arose naturally from our need to efficiently record and present the data. The use of a blog based system, and the ability to link posts together, naturally led us to use the links that describe relationships as other approaches using metadata did not work. The need for templates lead to a need to maintain metadata consistency, which lead us to use templates to encourage consistency. By applying critical analysis to how we were recording our work the system itself was encouraging us to develop and re-develop our approach. The web based nature of the system actually drove us to an organisational approach which, in retrospect, is similar to that found on the wider web.

7.2.2 Using the Blog as an ELN

7.2.2.1 The one-item one-post system

The key realisation of the first phase of development was the need to uniquely identify each sample and protocol. The use of a journal type approach for recording experiments cannot provide a readily machine readable system because specific samples and input materials cannot be uniquely identified. However a Blog (or indeed any web content management system that provides pages with discrete, stable, URLs) provides a straightforward means of providing a unique identifier through the URL, name, or number of each post/page. For many input materials, i.e. those that had been bought in, this identifier had already been created through a specific post (e.g. XhoI enzyme second batch). What was missing was equivalent posts for each product generated as part of the experiment. Generating these posts lead to the 'one item-one post' system that was then adopted. The relationships between posts are then indicated through the placing of explicit hyperlinks between them.

This approach creates an implicit, if very simple, data model of the form 'Object has relationship to Object'. There is no explicit description of what that relationship is, nor of what the item represents. Metadata can be used to distinguish between different general classes of item (material, procedure, data file) and from the classes of items it is generally possible to extract an understanding of the relationship between two items. However there is no explicit semantic content in this data model.

An important consequence of identifying a post as a representation of a specific item, especially for materials, is the need to be explicit about whether the item is a specific instance or a class of instances. For example an item representing 'NaCl' may refer generically to the material NaCl, to a specific supplier provided bottle of NaCl, or to a specific, weighed out, sample of NaCl. We have adopted the general convention that a post will refer to a specific instance of an item. For materials this means a specific container containing that material that can then be conveniently labelled. This is applied in a pragmatic way to those materials for which it is important to distinguish between

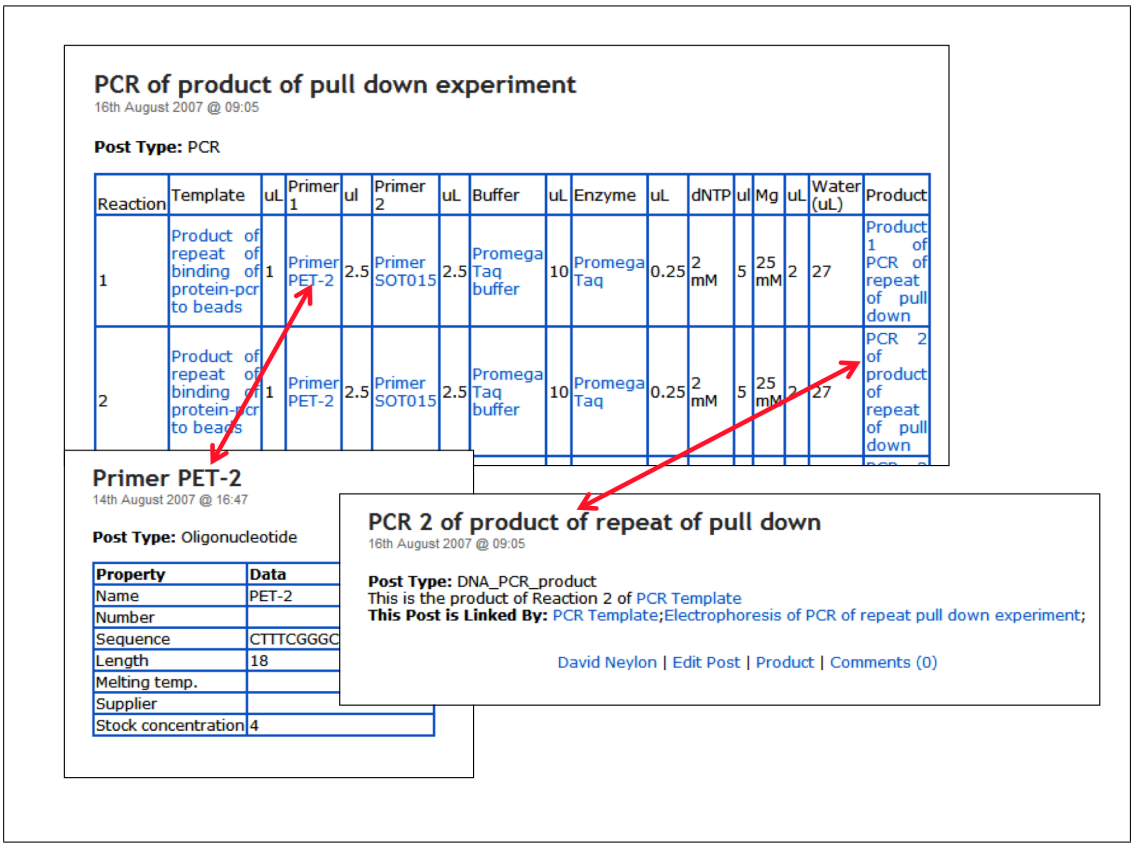


Figure 7.2: Showing a PCR reaction that has inputs, Primmers, and outputs, Products, linked. `http://blogs.chem.soton.ac.uk/sortase_cloning/2689/PCR_of_product_of_pull_down_experiment.html`

containers e.g. each tube of a restriction enzyme will get a distinct post whereas each bottle of NaCl will generally not.

7.2.2.2 What merits its own post?

It is clear that there is a potential here to generate vast number of posts if, for example, every step in a common procedure is separately captured. Broadly speaking the philosophy adopted was that if there was a tube (bottle, container etc.) of material that would be stored or that might be used for a different purpose then it should have its own post. In practice this means that common molecular biology procedures (running a gel, purifying plasmid DNA, PCR) each have their own post and their own set of outputs.

There is a logical argument that if every sample has its own post then every individual procedure (e.g. each of ten parallel PCR reactions) ought to have its own post. This would mean that all connections between samples, procedures, and their products would

be uniquely defined. This is logical from the informational perspective but creates a system, which is essentially not human readable. The decision was therefore taken to retain sets of procedures together and to infer the relationship between inputs and outputs for a specific procedure from the organisation of the procedure post.

Overall this generates a blog in which a human user can read the procedure posts, in essentially the same way as a laboratory notebook. The intermediate posts (input materials and products) are used essentially as place holders and in most circumstances do not need to be directly viewed. They can, however hold information that may be use to other systems, or to the user including chemical identifies, suppliers, physical properties, or safety or chemical incompatibility data.

7.2.2.3 Consequences and applications of the one-item one-post system

Most of the immediate applications of the system arise directly because every item now has a URI. Thus there is a sample management and identification system built in. It is straightforward to identify any specific sample or datafile and its associated and linked posts from a single ID number. A URI based on both the post id and the address of the server is also generated by a local service and a URI resolver can therefore resolve any specific item in any LabTrove system worldwide and redirect the user to the appropriate post.

Because each sample, or data file, now has a URI it is possible to describe a multistep process through a series of procedures and linked products. The user can manually page through a process simply by clicking on the links. An entire LabTrove can also be dumped as an rdf or xml file describing the posts and the links between them. This can provide an entirely new way of visualising a lab book. Such a 'network view' immediately provides a visual representation of the flow of materials and data through procedures and analysis (See Figure7.3). It can also identify work that is not completed or areas where the notebook has not been completed (isolated posts without connections). The useful application of this type of viewpoint remains to be explored in detail but one immediate possibility is the linking of the lab notebook into the wider information on the World Wide Web. Links that point out to published papers, or datasets, can be

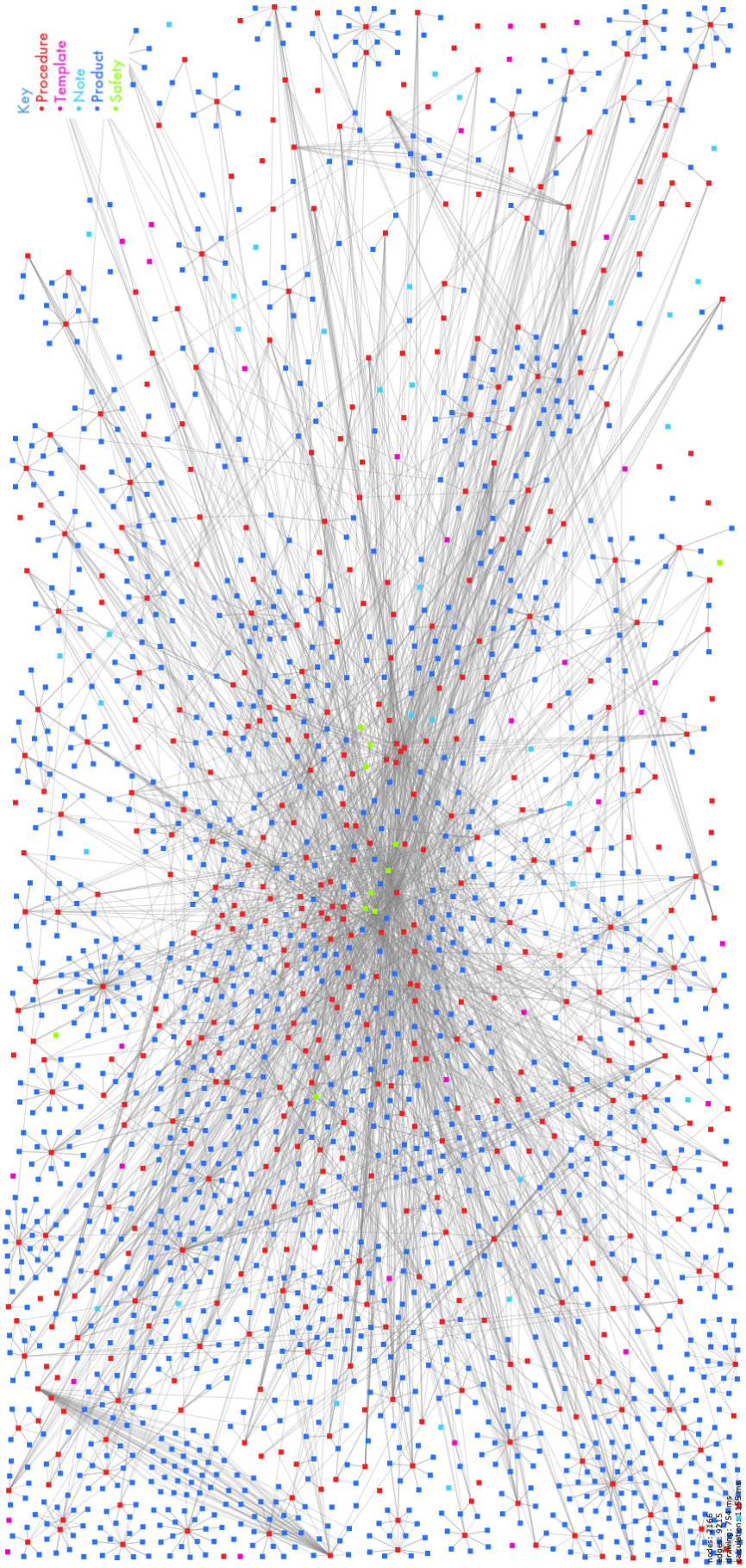


Figure 7.3: A network visualisation of the 'Sortase Cloning' blog from a rdf export using Welkin[116] tool from Simile project

incorporated into the graph, providing the first inkling of how the whole 'data web' could be wired together.

If the metadata is well organised it becomes possible to use the system to provide multiple databases of lab materials and stocks. If all posts referring to oligonucleotides are labelled with appropriate metadata and the formats of the posts are consistent then the set of all oligonucleotides can be extracted from the system along with their properties. These could be provided either as metadata themselves, or in tabulated form, then be processed into a fully featured relational database system and if required and this could additionally be automated in principle.

7.2.2.4 Metadata organisation in the one item-one post system

The successful realisation of the features of the LabTrove system requires a consistent and organised approach to the structuring of metadata. Many different approaches can be envisaged and approaches differ depending on the background of users. In specific scientific domains where extensive ontologies are available it may also be appropriate to structure the metadata to mirror such ontologies. It would be possible to configure an ontology browser to suggest appropriate metadata tags if integration were required.

7.2.3 Using the Templates: Approaches to metadata frameworks

The design of templates depends on the organisation of metadata and the organisation of metadata determines the design of templates. The flexibility of the metadata approach enables the user to adapt both templates and metadata to develop a coherent system that becomes self reinforcing when it works well. While this necessarily involves a 'working out period' this process of allowing a coherent system to evolve is actually one of the main strengths of the system as it means that a local vocabulary is developed that is appropriate and relevant to the work being carried out. Conversely this also allows the user to build templates around an existing controlled vocabulary if preferred. We have explored a number of different approaches and found some more effective and flexible than others.

The simplest approach is to simply use a single metadata key to describe the type of the post. This approach is implemented in the 'Neutral Drift' blog and in a slightly different form in the 'Sortase Cloning' notebook. This is simple but limits the form of templates that can be used. In the 'Neutral Drift' notebook, which was regularly used before the template functionality was developed, the description of different materials and types of procedure is not ordered. For instance both pcr_product and restriction_fragment refer to types of DNA that might be run on an agarose gel. This in turn means that it is not possible to use a template which generates a drop down menu that will provide both types of post. Thus the templates in 'Neutral Drift' notebook mostly require the entering of the post ID number rather than providing a drop down menu. In the 'Sortase Cloning' notebook this problem is avoided by simply adding a generic handle to the front of the value to give DNA_pcr_product and DNA_restriction_fragment which can both be addressed by using the wildcard Post_type:DNA% to provide the desired drop down menu in a template.

This approach which provides additional granularity at the level of values fails, however, when materials cross over between categories due to a lack of granularity at the key value. For instance in the 'Sortase Cloning' notebook the preparation of a protein-DNA conjugates is described. The products of these reactions are both protein and DNA, something which cannot be represented in the metadata approach used in the 'Sortase Cloning' blog. This is remedied in the 'Bio Sandpit' notebook (`http://blogs.chem.soton.ac.uk/bio_sandpit`) by increasing the granularity of the metadata keys. This in turn creates the problem of a multiplicity of keys. Taking this approach to its extreme creates a metadata system that can be represented as a multi-level ontology. In turn this runs the risk of a strict and complex vocabulary not mapping well onto a specific item. In practice a practical balance needs to be struck between keeping a simple and flat representation of objects within the LabTrove and providing a sufficiently detailed structure that maps well onto the work being carried out.

While the process of developing and adopting a system that works well in a specific context can be difficult, experience has shown that it is preferable to the difficulties of imposing an inappropriate vocabulary from the outside. There will be cases where the

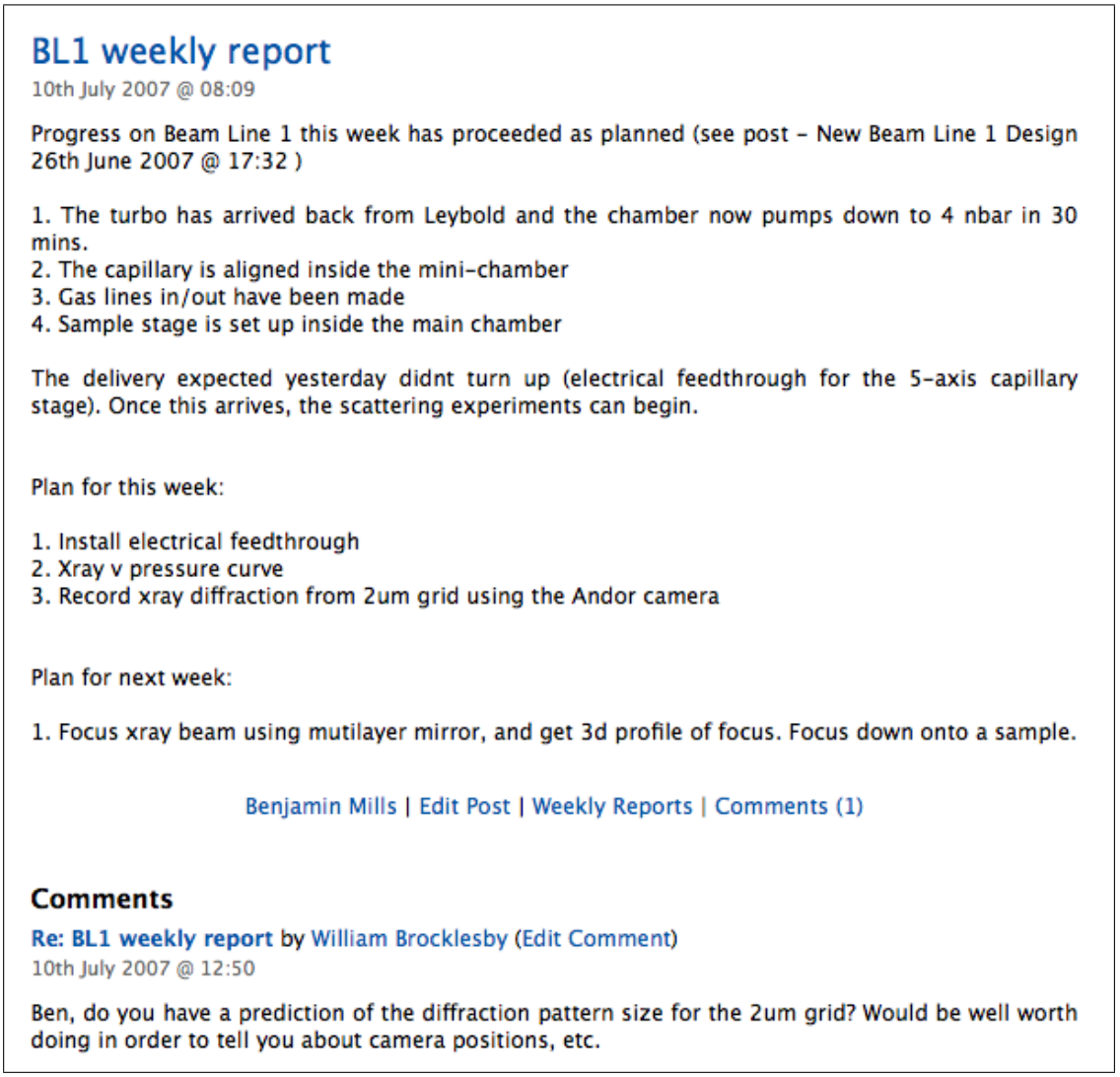
building of templates and design of metadata around an existing controlled vocabulary will be appropriate but to be sufficiently general a great deal of flexibility is required. The key point is that it is the combination of URIs, flexible metadata, and templates that encourages a workable structuring of the data without requiring it.

7.3 User Experience: the Research Group, the ORC Xray Group

As LabTrove offers web based delivery of its content, why should its use be restricted to just a simple laboratory lab notebook. A research group with in the ORC has been using the LabTrove software to manage their research and sharing it within the group. Even though the group members don't necessarily use LabTrove as their primary notebook, working practices now require them to post milestones and discussion points to the group notebook, this allows superiors and colleagues to keep abreast of the work being carried out by the group.

Content can range from papers that have been discovered that would be of interest for the group, conferences that could be attended and latest data from the experiments for review. The group has a weekly meeting where the previous groups posts are reviews and used for the discussion with comments attached to posted items. This saves the researches from preparing content, eg presentation slides, for the group meeting as the content is already available to them.

The group has been using LabTrove now for over 4 years, this now gives a large history of over 1,700 posts (as of February 2013) which can easily be searched using text searched or using the metadata facilities.



BL1 weekly report

10th July 2007 @ 08:09

Progress on Beam Line 1 this week has proceeded as planned (see post - New Beam Line 1 Design 26th June 2007 @ 17:32)

1. The turbo has arrived back from Leybold and the chamber now pumps down to 4 nbar in 30 mins.
2. The capillary is aligned inside the mini-chamber
3. Gas lines in/out have been made
4. Sample stage is set up inside the main chamber

The delivery expected yesterday didnt turn up (electrical feedthrough for the 5-axis capillary stage). Once this arrives, the scattering experiments can begin.

Plan for this week:

1. Install electrical feedthrough
2. Xray v pressure curve
3. Record xray diffraction from 2um grid using the Andor camera

Plan for next week:

1. Focus xray beam using mutilayer mirror, and get 3d profile of focus. Focus down onto a sample.

Benjamin Mills | Edit Post | Weekly Reports | Comments (1)

Comments

Re: BL1 weekly report by William Brocklesby (Edit Comment)

10th July 2007 @ 12:50

Ben, do you have a prediction of the diffraction pattern size for the 2um grid? Would be well worth doing in order to tell you about camera positions, etc.

Figure 7.4: An example report submitted for the groups weekly meeting, `http://xray.orc.soton.ac.uk/xray_group/285/BL1_weekly_report.html` [Accessed: 10/02/1013]

7.3.1 Integration of the Laboratory

7.3.1.1 Auto Posting

In order to capture all the data from the lab, the researchers where saving all of their files on to a institutional managed file server, but this only stored the files in their raw data formats and not in any searchable format. The current practice is to save the files in a date organised folder.

Using the api and the auto posting scripts as mentioned in section 6.5.2, a set of scripts were created that watched the file server and when a file was uploaded the scripts would

post the file into LabTrove. Once the files were in LabTrove they then could be linked to from the Groups discussion notebook. The data would also now be searchable by using metadata that was attached to the data.

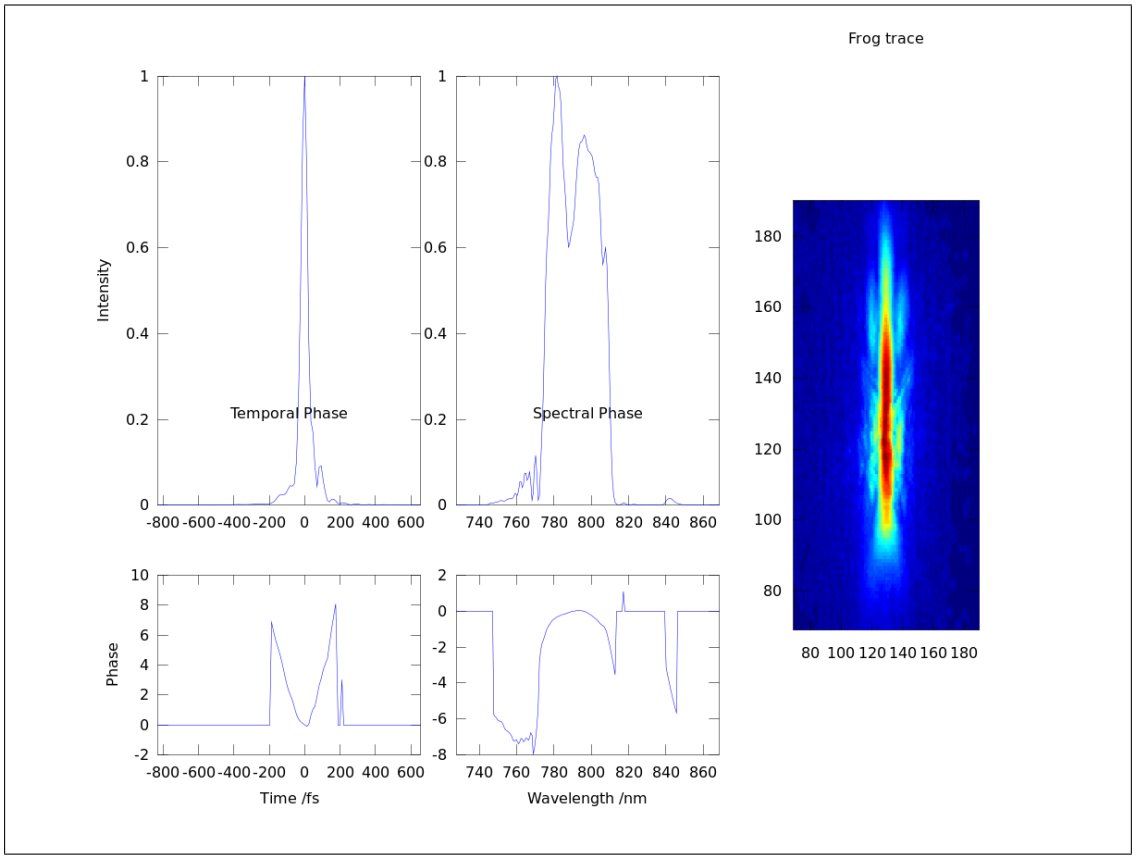


Figure 7.5: An example data set, processed using the auto poster and the uploaded as a image.`http://xray.orc.soton.ac.uk/data/6547.html` [Accessed: 10/02/1013]

Uploading the data was not the only benefit to the users, as the auto posting scripts could be customised, additional processing could be performed on the data before being uploaded allowing human readable forms (e.g. a picture) can be uploaded along with the raw data, allowing researchers to quickly sift through the data finding the results they want.

7.3.1.2 Laboratory Environment

The laboratory that the group use to do their experiments is a semi clean room, and therefore has a controlled environment. The lab uses brokering software to recored and store observations of temperature and humidity from many points in the lab,

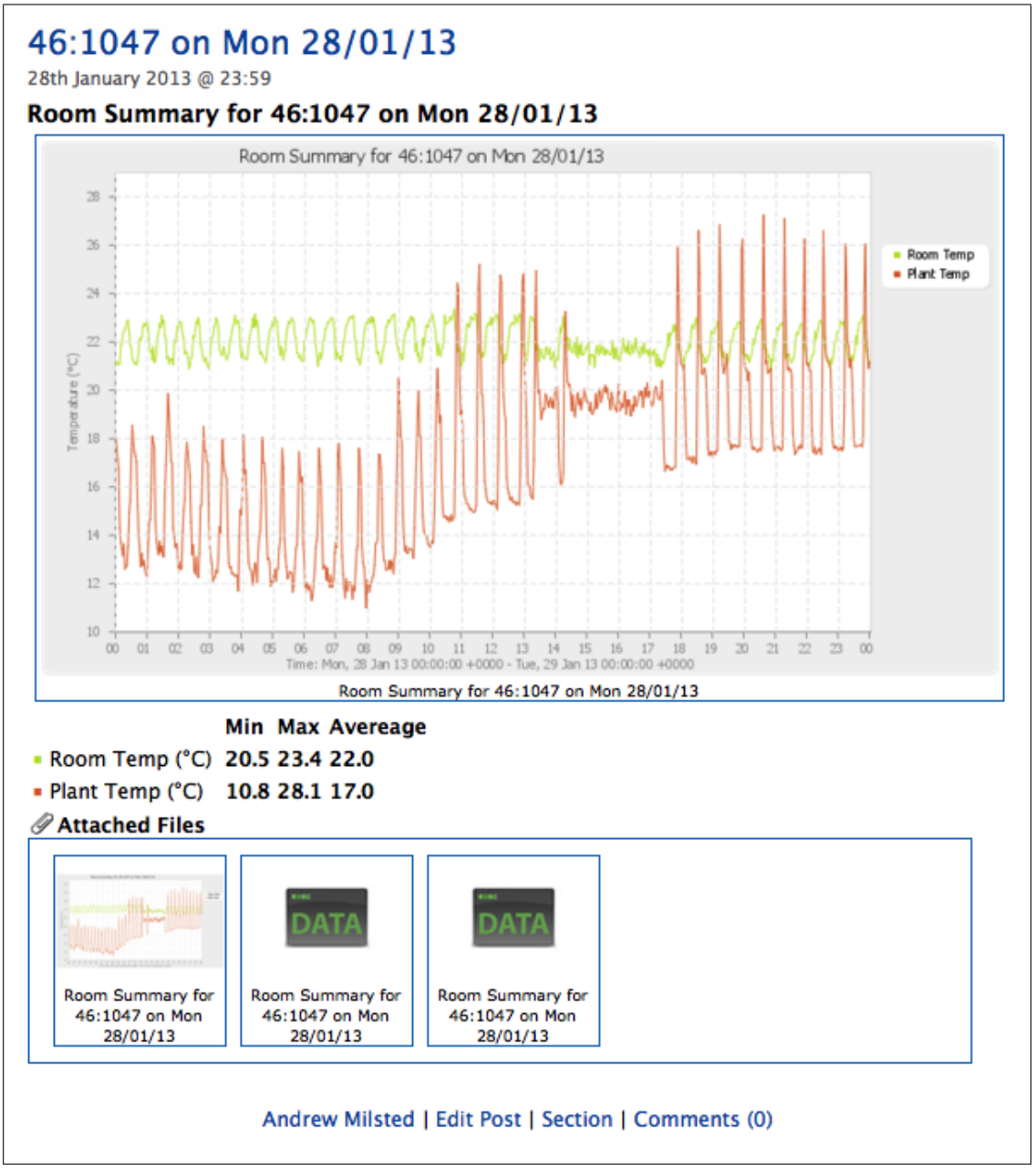


Figure 7.6: An example of a environment summary of the laser lab.

LabBroker[117]. To integrate the broker into the LabTrove notebooks, the broker uses the api to post a summary of the day's conditions into a notebook. Once in labtrove this record is now accessible in the same way as any other post, and can be searched/retrieved back. (See figure: 7.6)

7.3.1.3 MatLab

The research group uses MatLab to perform most of the analysis of their data, as this is a scripted tool set it was possible to enable it to post Matlab[118] code (with all the run graphs) and then store them in a blog. The idea was to be able to record processing runs of the data and any derived data.

MatLab is heavily module based and there is a function "publish()" which already produces an HTML version of the code and saves a PNG version of the graphs. A function was created that packages up the files and auto inserts them into LabTrove. The function was written in MatLab and initiated as the function "publish_blog()", this would then trigger the auto post.

This enabled a record of each run of the code and more importantly a copy of the code, as it was, when it was run, this is important as many of the scripts are shared amongst researches and often tweaked. It is important for any publication of results that that version of the code is available in order to reproduce any of the derived work, as a copy is uploaded in LabTrove, this version of the code is available.

7.3.2 Using LabTrove For Open Drug Discovery

> "The drug praziquantel (PZQ) is used very widely in both animal and human medicine, where it is the mainstay of the treatment of the neglected tropical disease schistosomiasis. The drug is currently manufactured and administered as a racemate (1:1 mixture of enantiomers) but for various reasons the large-scale production of PZQ as the single active enantiomer is very desirable. We describe here the preparation of praziquantel as a single enantiomer using classical resolution. The protocols are experimentally simple and inexpensive. One method was found and validated by an unusual research mechanismopen sciencewhere the details of the collaboration (involving academic and industrial partners) and all research data were available on the web as they were acquired, and anyone could participate. The other route

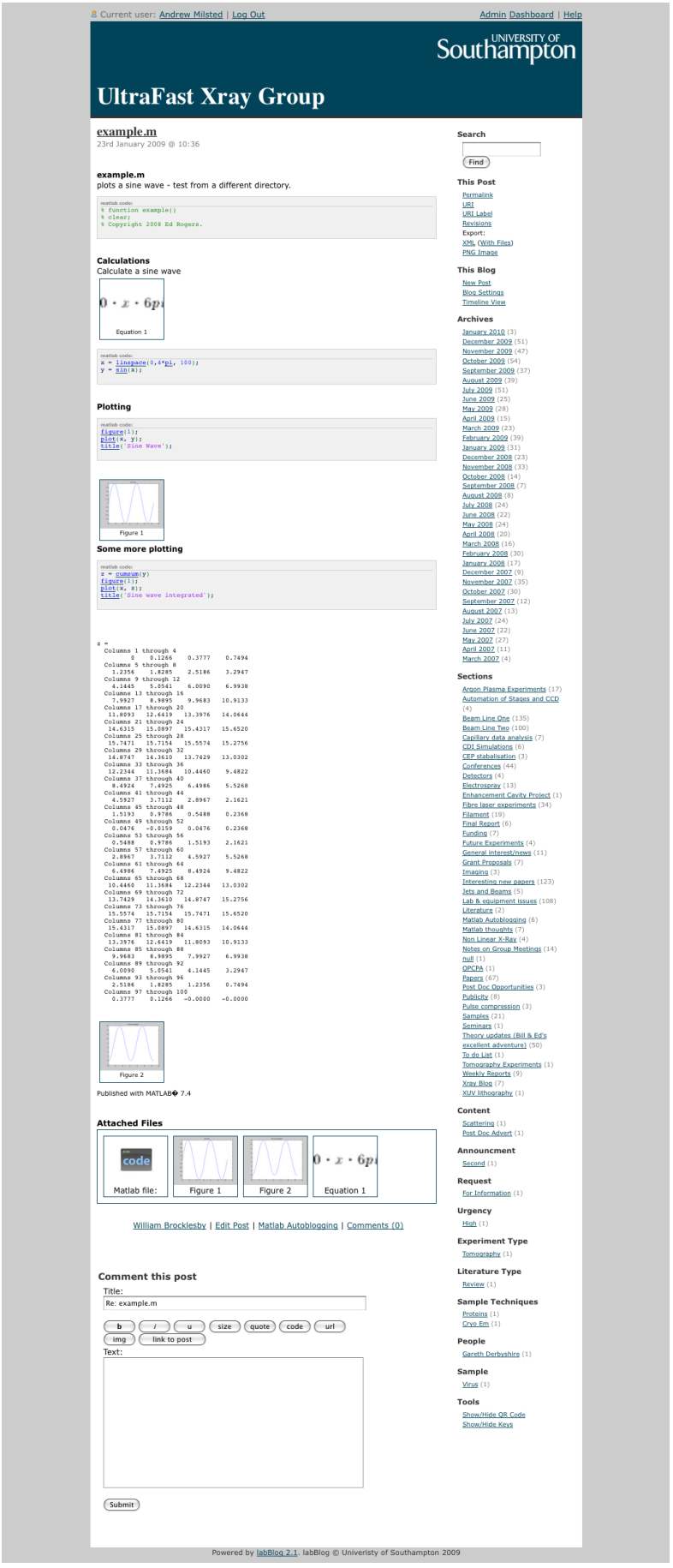


Figure 7.7: An example of a MatLab generated post

> was found in parallel by a contract research organisation. Besides being possible routes by which praziquantel may be produced in large quantities for the affected communities, it is also hoped that these methods can be used for the production of smaller quantities of enantiopure PZQ for pharmacological studies."[119]

An open science initiative based at the University of Sydney saw the potential of using LabTrove to help publish their raw lab notebook entries to facilitate the discovery of the required enantiomers of PZQ.

The research scientists in the lab would use the LabTrove notebook to publish their procedures for a particular experiment, as their notebooks where publicly available their observations/results could be read and discovered by interested researches. Suggestions and even offers of help from community, which resulted in finding a number of methods.

7.3.3 A Commercial Viewpoint

A group at the Unilever^{TM}, R&D Port Sunlight Labs ran a pilot project to evaluate the LabTrove software within a commercial environment. Because of the sensitive nature and commercial interest that relates to the work being entered for the test, the service was set up on a Unilever owned server within their private network. This meant that none of the content would be available to outside of Unilever but should be accessible to all staff.

The test group of users where from a few different groups from within Unilever and also from different sites, this showed the benefits of this ELN as it was provided through the web browser as was accessible to all users with out any set up issues.

The openness by default of LabTrove was seen as an initial undesirable, as the ethos of the working environment was all about protection of IP and even though there was no policy restricting colleges from seeing each others data it still was the users wanted to close of their data, therefore the instance became a set of closed blogs developed with users then users having to manage their own access to their blogs.

Positives from the trial where users liked the fact that its is a open page ELN allowing free text to entered;

> "Extremely easy to use and flexible."
>
> "I do like the fact that I can type in - which make things more legible,"

But users did find that without any preplanning of the structure of posts their entries became very disorganised.

> "Labtrove needs some thought in advance in terms of how to organise entries."

But allowing users to have the free text environment they did find that they could start using the product very quickly, and reaping the rewards of using an ELN.

> "Extremely easy to use and flexible."
>
> "An ELN is a much better system than the current paper version - Labtrove was a good easy to use system."
>
> "Speed, ease of use and intuitiveness."

The more computer literate members of the trial from the computational group did take advantage of the API allowing them to log the progress and results from simulations and the linking them to their own discussions.

This trial did make it clear that the requirements of a commercial environment are very polar to that of academia, certainly open 'academia', as the willingness to share work is different. One aspect that could be deemed similar is the fact that with a very large organisation like Unilever there is a need to openly share research with in the organisation to protect against duplication of work and therefore wasting money.

Unilever or any large commercially sensitive organisation will allways have many factors influencing the decisions governing the use of software products, whilst LabTrove would be a great solution if Unilever wanted to share openly internally then LabTrove would

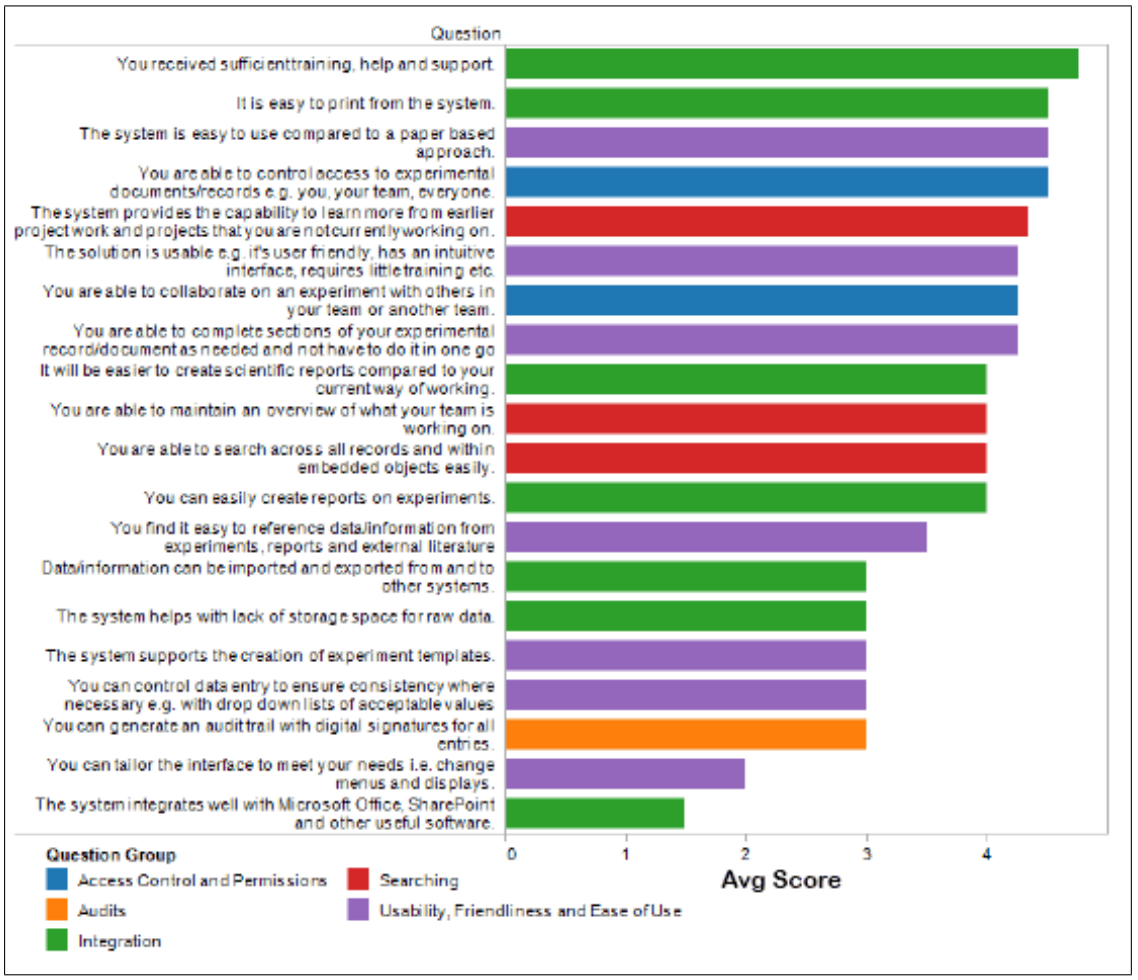


Figure 7.8: LabTrove Feedback from user trial survey. Statements scored on how well the respondent agreed/disagreed. Scale 1-5, where 1 is disagree and 5 agree. (4 participants).

have no problem providing a suitable solution as the content could easily be protected from the outside world by using already in place private corporate networks, there in no requirement for it be on the web. But in the partnership with Unilever this social change was not a likely going to change, there were though very positive intentions to collaborated with particular research groups where access would be granted on a case by case bases. Again LabTrove does have this facility to protect the content by creating the user groups. Unilever could also use their private networks as another layer on top of LabTroves groups to ensure the commercially sensitive data was hidden from outside the organisation.

7.4 Blog My Data

The National Centre for Earth Observation (NCEO) and the National Centre for Atmospheric Science Climate Group (NCAS-Climate) are both high-profile interdisciplinary research centres involving numerous universities and institutes around the UK and many international collaborators. Working with large-scale earth simulations requires the collaborative effort of scientists from many different disciplines and institutions.

Both groups make use of the latest numerical models of the climate and earth system, validated by observations, to simulate the environment and its response to forces such as an increase in greenhouse gas emissions. Their scientists must work together closely to understand the various aspects of these models and assess their strengths and weaknesses.

At the present time, collaborations take place chiefly through face-to-face meetings, the scholarly literature and informal electronic exchanges of emails and documents. All of these methods suffer from serious deficiencies that hamper effective collaboration. For practical reasons, face-to-face meetings can be held only infrequently. The scholarly literature does not yet adequately link scientific results to the source data and thought processes that yielded them, and additionally suffers from a very slow turnaround time. Informal exchanges of electronic information commonly lose vital context; for example, scientists typically exchange static visualisations of data (as GIFs or PostScript plots for example), but the recipient cannot easily access the data behind the visualisation, or customise the visualisation in any way. Emails are rarely published or preserved adequately for future use. The recent adoption of off the shelf Wikis and basic blogs has addressed some of these issues, but does not usually address specific scientific needs or enable the interactive visualisation of data.

7.4.1 Using LabTrove as a possible solution

A Virtual Research Environment is an attractive solution to the above problems. In the JISC-sponsored BlogMyData project a Virtual Research Environment (VRE) was

created by combining the capabilities of two existing technologies that have already seen wide adoption among scientists:

1. The Godiva2 data visualization system (http://www.reading.ac.uk/godiva2) provides a means for scientists to browse interactively in a Google Maps-like fashion through large environmental datasets, including numerical model outputs and high-resolution satellite imagery, using only a web browser. Scientists can produce maps, timeseries and other plot types. This system completely removes the need for the scientist to understand the technical details of how and where the data are stored.

2. LabTrove: will be used as a a collaboration tool that allows discussion between colleagues. For open science work the LabTrove Instance would be opened up to publish its content to the public domain but can also use the necessary access control to keep any private work secure.

Having logged in to the BlogMyData VRE using OpenID, scientists examine output from the latest cutting-edge climate and ocean models using the Godiva2 interface. Upon finding a feature of interest (perhaps an extreme event, or a suspected problem with the model) the user creates a new blog entry that is linked to the current visualisation. The blog entry is automatically tagged with metadata about the feature of interest (e.g. its location in time and space, and the dataset from which it is derived). Colleagues provide input through comments and by linking blog entries together. Through semantic and geospatial tagging, scientists can discover colleagues working on similar scientific problems. The system is augmented by the addition of a geospatially-enabled database, based on the widely-used open-source PostgreSQL database with the PostGIS extensions. This database will associate blog entries with geographical areas and time periods and allow users to discover discussions that relate to particular areas of interest very efficiently (See Figure 7.9).

The system was developed as an iterative process, with regular feedback from users in NCAS-Climate and NCEO. An end-to-end prototype of the system, was created in which users can create notebook entries based upon map-based visualisations (i.e. horizontal

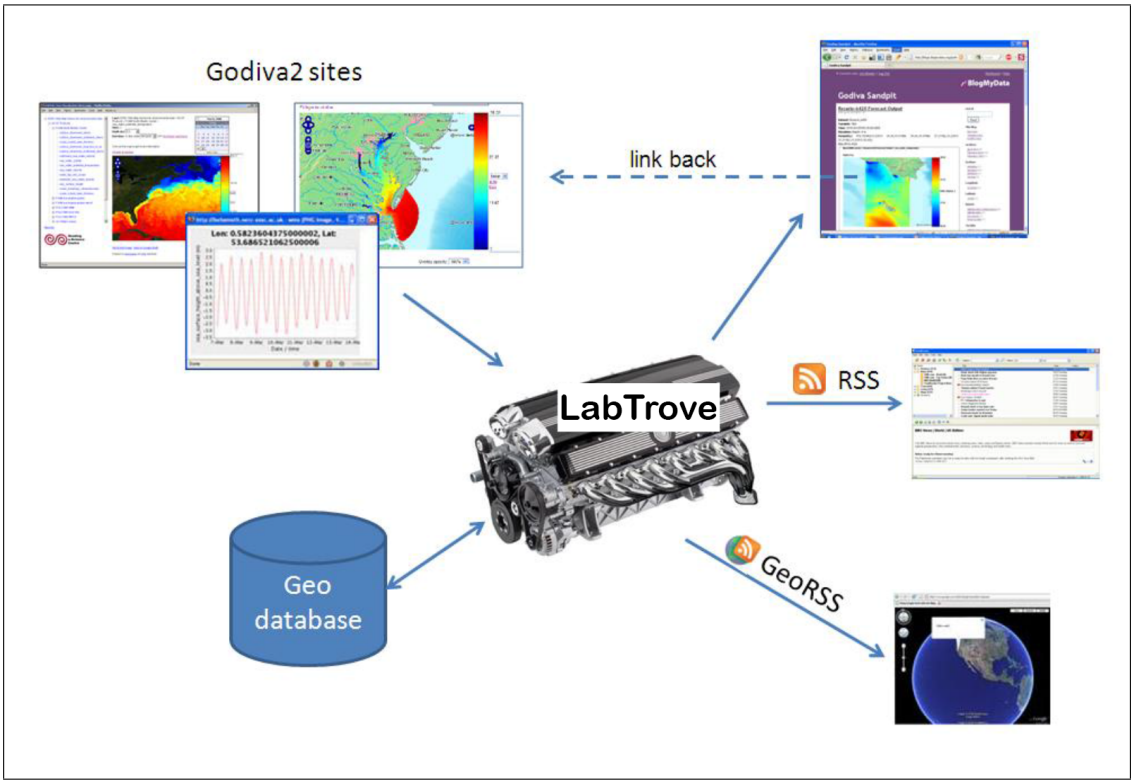


Figure 7.9: Sketch architecture of the BlogMyData system. Users explore environmental data using Godiva2 sites, which project information onto draggable, zoomable maps. Users create blog entries that are linked to particular visualizations, which are stored in the blog engine, which uses a geospatial database to store geospatial and temporal information. The blog entries are displayed on the project website, on which other users can leave comments. Each blog entry links back to the Godiva2 site that created it, preserving the state of Godiva2 at the time of creation, allowing easy further exploration. Content is syndicated via RSS (for standard feed readers) and GeoRSS (for geo-enabled feed readers).

x-y views of the data). Entries are captured in a private notebook, which is only visible to a controlled set of users, thereby maintaining the privacy of the research.

Entries are then syndicated as Geographic Really Simple Syndication (GeoRSS) feeds (GeoRSS is an enhancement to Really Simple Syndication, in which each entry is tagged with geographic information). These feeds can be consumed in standard Really Simple Syndication (RSS) viewers (such as Microsoft Outlook, Google Reader and Firefox Live Bookmarks), or in geo-enabled viewers such as Google Maps (Figure 7.10). These feeds provide a simple means for scientists to discover research activity in related areas.

The prototype was tested on some members of the NCAS-Climate group, who are working on the development of the latest high-resolution climate models, including HiGEM

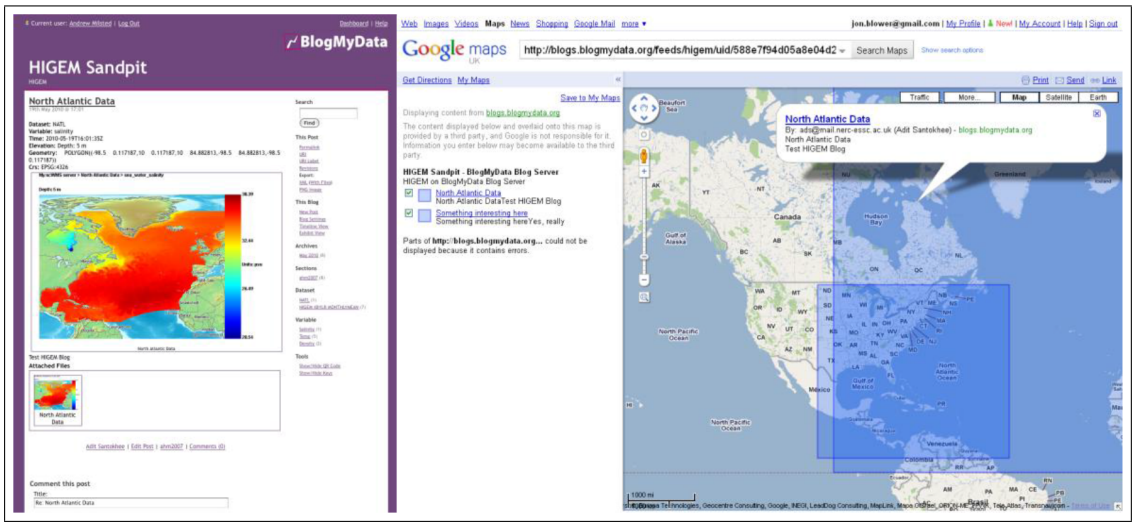


Figure 7.10: Detail figure showing the display of blog entries on the project website (left) and a GeoRSS-enabled feed reader (Google Maps, right).

[120], this has generated some initial feedback. Among the most important items of feedback are:

- The privacy controls are regarded as essential: without these, the users would hesitate before posting their most interesting thoughts.
- Content is king: the VRE must display exactly those data that the users are interested in at the current time (sample test data from other domains is much less engaging). We have therefore gone to considerable effort to ensure that the data are relevant and presented correctly1.
- The generation of animations of data gives scientists a great deal of insight into the dynamics of the Earth system. The Godiva2 system is popular for its ability to generate animations of complex numerical model data quickly and easily. As this was a high priority, the VRE as whole was amended to allow recording of animations of data, not just static images.

7.5 School of Chemistry, University of New South Wales

LabTrove was selected to take part in a user trial of using an ELN as a teaching tool for postgraduate students with the School of Chemistry, University of New South Wales[121]

A trial of academic staff (n = 18) and postgraduate students (n = 56) in chemistry were invited to complete an online survey to determine their willingness to use ELN and their perceptions of the usefulness and limitations of the ELN. In this survey they were asked in which branch of chemistry they were researching.

Using the first survey, some of students and staff where invited to participate in a 12-month trial of the ELN. At the end of the trial, those who participated were invited to complete another survey aligned to the initial survey.

This enables the ability to a) compare the perceptions of those who did and did not want to participate in the trial and may be able to determine the most important sticking point for not adopting new technologies in a science research environment and b) determine whether the perceptions of participants in the trial have changed, and if so, how. During the trial we have asked participants to blog (via the ELN) how they find interacting with the ELN blog.

During the trial 2 pertinent observations we noted:[122]

1. Students are using the ELN as both a place to store and to annotate their dataset. Most of the participants are linking their records to the relevant published literature. It is expected that the practice of linking experiments to the published literature will, in theory, make the process of writing up their work as a thesis and papers much easier for students.

2. Using an ELN as a data portal will allow undergraduate students, wherever they are located, access to a much more diverse range of experimental equipment. The potential of the LabTrove other ELNs to address access issues in the science higher education sector is enormous.

7.6 Conclusions

As the evaluation of LabTrove demonstrates, it has been used by many users, this has mainly come about by direct interaction of the project collaborating with others

which result in instances of LabTrove being used by at least hundreds of users. As LabTrove is an open source bit of software, it can be downloaded from many sources, the svn respository, a source forge code package, or as a debian package. Whilst anyone can download the software and install it themselves it can be quite difficult to gauge the impact of the software. Source forge does supply download statistics for the code package, but the svn repository and debian install package does not. Up to April 2014 (from July 2010) LabTrove has been downloaded 1,402 times. Because the end users identity is not provided and whether or not they have installed it, it is unclear how many instances there are. The support email account does sometimes get queries, a couple a year, on instances we have no idea exist. This does demonstrate that the software is being used by users that have discovered LabTrove for themselves.

Chapter 8

After the ELN

The success of any electronic laboratory recording system depends on providing records that are sufficiently rich to allow the detailed reproduction or checking of any part of a reported process. Laboratory systems can and should also enable the reuse of data in new and unexpected ways, the efficient repurposing of materials, and the redeployment of experimental and analysis procedures for modified experiments.

8.1 Where should the research reside

The problem of preserving research long beyond when the work has been completed has traditionally relied on the publish paper in journals. Many journals have now started accepting attached data for their publications, this usually falls short in showing the full research record only including relevant structural analytical files to confirm the author's research.

A cheap approach to perform this longterm archive this preservation is to use one of the many cloud storage providers. These are providers that can offer large amount of storage relatively cheaply as they are just providing the storage. They all enable users to share their data via publicly accessible links for which users can then link to in their research.

A list of curently popular storage providers:

- **Dropbox** provided by Dropbox, Inc. `http://www.dropbox.com`
- **Amazon S3 (Simple Storage Service)** provided by Amazon.com `http://aws.amazon.com/s3/`
- **Google Drive** provided by Google, Inc. `http://drive.google.com`
- **Box** provided by Box, Inc. `http://www.box.com`

Using such providers involves a lot of risk because none of them offer any guarantees to the permanence of the data or in some cases a warrantee if they lose the data. This could be down to the providers going bankrupt, changing their business model/direction or not being able to recover from a data loss disaster. Other issues can be caused by the end user may not have any control on where geographically their data is being stored and therefore could fall foul of legal issues in exporting their data to other countries.

Because of this lack of permanence publishers are unlikely to accept links to such providers as they can't show any persistency, so there is a need to use a provider that can offer some of the required guarantees and assurances. A way of achieving this is to use a provider that can provide a DOI for the research data. A DOI issuing provider has a contractual obligation to make sure that the data is always accessible. Once the research data has been given a DOI the publishers are likely to accept it as a link from the publication.

8.1.1 Dryad

Dryad[123] is an international disciplinary repository for data that underlies scientific, its mission is to promote the availability of data underlying findings in the scientific literature for research and educational reuse. It provides a platform for users who want to upload research data and will then publish it with appropriate metadata and importantly a DOI.

The downside to using Dryad is the cost to the user, or the user's institution, the user will have to pay for each data package they upload, but this is at least a one off cost

which would mean that the data package is maintained for as long as the Dryad project is running.

8.1.2 Figshare

Figshare[124]:

> "figshare is a repository where users can make all of their research outputs available in a citable, sharable and discoverable manner."

FigShare is another repository service, but unlike Dryad, is free for users to upload their research output, it also gives each item a DOI. All data is persistently stored online under the most liberal Creative Commons licence, waiving copyright where possible. This allows scientists to access and share the information without any hindrances with licences.

8.1.3 DataCite

DataCite is an international consortium which aims to improve data citation in order to:

- establish easier access to scientific research data on the Internet, to
- increase acceptance of research data as legitimate, citable contributions to the scientific record, and to
- support data archiving that will permit results to be verified and re-purposed for future study.

It achieves this by providing organisations a method to register DOIs to their own output and therefore not relying on third party services, enabling them to have greater control of their own data. It allows institutions who run their own repositories to then mint a citable DOI. As data is commonly accepted to be owned by the institution, they are then taking on the responsibility to preserve the data.

8.2 Credit and Impact

A method of placing a value on a researchers output and contribution to the scientific knowledge can be managed in a number of ways. One method is to calculate the number of citations that researchers work as accumulated. By preserving and publishing more research data in a citable form, the data its self can be included in these calculations.

Many online tools aggregate the citations from published works and can then provide this a metric of the researchers impact, some include:

- **Impact Story** aggregates altmetrics: diverse impacts researchers your articles, datasets, blog and posts. `http://impactstory.org`
- **Research Gate** similarly aggregates research impact, but adds a layer of social web, allowing research to communicate and share with others. `http://http://www.researchgate.net`

Whilst many of these tools allow the process of sharing and preserving the research data, they are still not perfect and it remains unclear if they are sustainable. For the foreseeable future the traditional publication will remain as the important part of scientific output, but hopefully also publishing supporting data will become as important.

Chapter 9

Conclusions

The research of this project has identified three distinct areas and has the following conclusions: The scientific record, Collaboration and Open Science.

9.1 The scientific record

The success of any electronic laboratory recording system depends on providing records that are sufficiently rich to allow the detailed reproduction or checking of any part of a reported process. Laboratory systems should also enable the reuse of data in new and unexpected ways, the efficient repurposing of materials, and the redeployment of experimental and analysis procedures for modified experiments. The scientific literature as it stands rarely, if ever, provides sufficient detail to enable other researchers to replicate the detail of a published study. Achieving the desired standard will require sophisticated recording systems that integrate human-generated journals with a wide range of instrumental and observational data, and are capable of presenting contents that are useful to, and readable by, both humans and machines. Regrettably, the shift away from paper notebooks has brought about a diminution in the careful journaling of the thoughts and ideas leading up to scientific innovation. We are unlikely ever to see the electronic equivalent of a Faraday notebook.

9.2 Collaboration

In broad terms, scientists and science in general have gained much from the developments in electronic recording, not least for interdisciplinary and international collaboration and for the reuse and repurposing of vital data. Collaboration is vital for progress in all branches of science and technology, but in the digital era we do still need to increase trust in sharing. For collaboration to be effective, record keeping must become more comprehensive and provide good quality, verifiable, data and information. Now with model network connections and supporting tools geography now should not be a barrier to effective colaboration.

9.3 Open Science

Even though open science began in the 1600s with the advent of the academic journal when the societal demand for access to scientific knowledge reached a point where it became necessary for groups of scientists to share resources with each other so that they could collectively do their work. In modern times there is debate about the extent to which scientific information should be shared, the conflict is between the desire of scientists to have access to shared resources versus the desire of individual entities to profit when other entities partake of their resources.

For those researchers wishing to participate in open science having tools that can publish their raw work easily is essential and having ELNs that enable this should be received as a positive thing.

9.4 LabTrove and the future

LabTrove as the main output of the work presented here has shown that it can provide a route towards reconciling the tensions and challenges that lie ahead in working towards these goals. The future of labtrove would be to provide a stable support structure around

it, allowing users to rely on it, LabTrove now has a user base of hundreds, demonstarting this need.

It is inevitable this will mean that LabTrove will have look at its self commercially and may not lie well with the academic routes that it has. This being said the model that is currently being looked at is that LabTrove will remain Open Source, providing a free tool for researchers to use, but the funding to continue the support would come from hosting services and support contracts. Hopefully this will ensure a future for LabTrove.

Appendix A

Abbreviations/Definitions

A.1 Acronyms/Abreaviations

API application programming interface

COSHH Control of Substances Hazardous to Health

DC Dublin Core

DHFR Dihydrofolate reductase

DOI Digital Object Identifier

ELN Electronic laboratory notebook

GeoRSS Geographic Really Simple Syndication

GPL GNU Public Licence

GPS Global Positioning System

HTTP Hypertext Transfer Protocol

IP Intellectual Property

IR Infrared spectroscopy

LDAP Lightweight Directory Access Protocol

Mass Spec Mass spectrometry

NMR Nuclear magnetic resonance

ORC Optoelectronics Research Centre

PZQ praziquantel

PCR polymerase chain reactions

PHP PHP: hypertext preprocessor

RDF Resource Description Framework

REST representational state transfer

RFID Radio-Frequency Identification

RSS Really Simple Syndication

SaaS Software as a Service

SOAP Simple Object Access Protocol

SSO Single Sign On

TCP/IP transmission control protocol/internet protocol

WSDL Web Services Description Language

XML Extensible markup language

WYSIWYG "what you see is what you get"

HTML Hypertext Markup Language

URI Uniform Resource Identifier

URL Uniform Resource Locator

VRE Virtual Research Environment

A.2 Data Sizes

B A unit of digital information

KB Kilobyte - 1024 (2^{10}) bytes

MB Megabyte - 10^6 bytes

GB Gigabyte - 10^9 bytes

TB Terrabyte - 10^{12} bytes

Taken from `http://en.wikipedia.org/wiki/Binary_prefix`

A.3 File Formats

BMP bitmap image file

CIF crystallographic information file

CSV comma separated variables

EPS Encapsulated PostScript image file

JPEG Joint Photographic Experts Group image file

PDF portable document format

TIFF Tagged Image File Format

Appendix B

Supporting Data

The supporting data CD, attached, contains electronic versions of this thesis and some high quality images for some of the figures, as well as copies of the LabTrove at some of its major miles stone releases.

If the attached CD is damaged, corrupted, lost or unavailable, an electronic version can be obtained from the University of Southampton ePrints system at the following address: http://eprints.soton.ac.uk/[TBC]

The CD is laid out with the following folders:

B.1 Thesis

Contains the pdf and images used in to create the thesis.

B.2 LabTrove Software

Contains the archived copies of LabTrove which have been downloaded from the svn on sourceforge:

Version 2.1:- labtrove_2.1.zip

http://sourceforge.net/p/labtrove/code/678/tarball?path=/branches/labtrove_2.1

Version 2.2:- labtrove_2.2.zip

http://sourceforge.net/p/labtrove/code/678/tarball?path=/branches/labtrove_2.2

Version 2.3:- labtrove_2.3.zip

http://sourceforge.net/p/labtrove/code/678/tarball?path=/branches/labtrove_2.3

B.3 LabTrove Manual

Can also be browsed via the web: http://docs.labtrove.org/2.3/lt/Main_Page

References

[1] William K. Michener, James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1):330–342, 2013/07/29 1997.

[2] Dartmouth College ChemLab. How to Keep a Notebook. [web page] http://www.dartmouth.edu/ chemlab/info/notebooks/how_to.html. [Accessed 12 Feb 2013].

[3] Michael Faraday. *The philosopher's tree: a selection of Michael Faraday's writings.* CRC Press, 1999.

[4] Phillip A Griffiths. *On being a scientist: responsible conduct in research.* National Academies Press, 1995.

[5] Howard M Kanare. *Writing the Laboratory Notebook.* ERIC, 1985.

[6] JULIUS H COMROE. Some functions of a scientific journal. *Circulation Research*, 19(1):3–214, 1966.

[7] Jim Gray, David T. Liu, Maria Nieto-Santisteban, Alex Szalay, David DeWitt, and David DeWitt. Scientific data management in the coming decade. *CTWatch Quarterly*, 1(1), Febuary 2005. [web page] http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/ [Accessed 3 Apr 2009].

[8] Brian Hayes. Computing science: Terabyte territory. *American Scientist*, pages 212–216, 2002.

[9] Heath and Safety Executive. Control of substances hazardous to health (Fifth edition). [web page] http://www.hse.gov.uk/pubns/priced/l5.pdf. [Accessed 12 Feb 2013].

[10] B. Dorsey S. Larsen M. Williams, D. Bozyczko-Coyne. *Current protocols essential laboratory techniques. ed. S. Gallagher and E. Wiley.* Wiley Hoboken, 2008.

[11] Stu Borman. Electronic laboratory notebooks may revolutionize research record keeping. *Chemical & Engineering News Archive*, 72(21):10–20, 1994. doi:10.1021/cen-v072n021.p010.

[12] B. Lass. Implementing electronic lab notebooks. [web page] http://www.scientificcomputing.com/articles-IN-Implementing-Electronic-Lab-Notebooks-Part-6-102411.aspx. [Accessed 21 May 2014].

[13] Randy C. Hice. Roadmap to a clear definition of eln. [web page] http://www.scientificcomputing.com/blogs/2009/05/roadmap-clear-definition-eln, 2009. [Accessed 21 May 2014].

[14] Richard Lysakowski and Leslie Doyle. Electronic lab notebooks: paving the way of the future in r&d. *Records management quarterly*, 32:23–30, 1998.

[15] Mats Kihln and Martin Waligorski. Electronic lab notebooks - a crossroads is passed. *Drug Discovery Today*, 8(22):1007 – 1009, 2003.

[16] M. H. Elliott. Are elns really notebooks?, sci. comput. instrum. [PDF] http://www.atriumresearch.com/library/July[Accessed 22 May 2014].

[17] Rick Mullin. Learning to share in the lab. *Chemical & Engineering News Archive*, 81(43):19–24, 2003. doi:10.1021/cen-v081n043.p019.

[18] K. T. Taylor. *Collaborative computational technologies for biomedical research, ed Bingham, Alpheus and Ekins, Sean and Hupcey, Maggie AZ and Williams, Antony J.* John Wiley & Sons, 2011.

[19] Ping Du and Joseph A. Kofman. Electronic laboratory notebooks in pharmaceutical r&d: On the road to maturity. *Journal of the Association for Laboratory Automation*, 12(3):157–165, 2007.

[20] Stephen Bruce. A look at the state of electronic lab notebook technology. [web page] http://www.scientificcomputing.com/articles/2002/12/look-state-electronic-lab-notebook-technology, 2002. [Accessed 21 May 2014].

[21] R. MacNeil. The electronic lab notebook blog. [web page] http://elnblog.axiope.com/?p=956. [Accessed 21 May 2014].

[22] Michael H Elliott. Electronic laboratory notebooks: A foundation for scientific knowledge management edition v, 2011.

[23] Declan Butler. Electronic notebooks: a new leaf. *Nature*, 436(7047):20–21, 2005.

[24] Kalpana Shankar. Order from chaos: The poetics and pragmatics of scientific recordkeeping. *Journal of the American Society for Information Science and Technology*, 58(10):1457–1466, 2007.

[25] Paul van Eikeren. Intelligent electronic laboratory notebooks for accelerated organic process r&d. *Organic Process Research & Development*, 8(6):1015–1023, 2004.

[26] A. Nehme and R. A. Scoffin. *Computer applications in pharmaceutical research and development, ed Wang, Binghe and Ekins, Sean*, volume 2. John Wiley & Sons, 2006.

[27] Gale Dutton. Lab notebooks offer efficiency gains. [web page] http://www.genengnews.com/gen-articles/lab-notebooks-offer-efficiency-gains/1951/, 2006. [Accessed 21 May 2014].

[28] Michael Rubacha, Anil K Rattan, and Stephen C Hosselet. A review of electronic laboratory notebooks available in the market today. *Journal of the Association for Laboratory Automation*, 16(1):90–98, 2011.

[29] R. MacNeil. The electronic lab notebook blog. [web page] http://elnblog.axiope.com/?p=1037. [Accessed 21 May 2014].

[30] . Pistoia alliance. [web page] http://www.pistoiaalliance.org/, 2009. [Accessed 21 May 2014].

[31] Jean-Claude Bradley. Usefulchem wiki. [web page] http://usefulchem.blogspot.fr/. [Accessed 21 May 2014].

[32] Dr Simon Coles, Professor Richard Whitby, Professor Jeremy Frey, Dr Colin Bell, and Dr Aileen Day. Towards publishing semantic descriptions of electronic laboratory notebook records. page 90. Abstracts of Papers, 244th ACS National Meeting & Exposition, 2012.

[33] Dr Simon J Coles, Dr Graham Tizzard, Professor Jeremy Frey, Mr Andrew Milsted, Dr Mark Edwards, Mr Romanus Onyeabo, Dr John Spencer, and Dr Jan Kuras. Opening up elns and repositories to support formal publication. page 71. Abstracts of Papers, 244th ACS National Meeting & Exposition, 2012.

[34] Eprints 3, Digital Respository. [web page] http://www.eprints.org/software/. [Accessed 3 Apr 2009].

[35] IUCR. Acta Crystallographica. [Accessed 20 Jan 2009].

[36] Open Archives Initiative (OAI). [Accessed 20 Jan 2009].

[37] C. A. Lynch. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *Association of Research Libraries*, (226):1–7, January 2006.

[38] eCrystals - Crystal Structure Report Archive. [web page] http://ecrystals.chem.soton.ac.uk/. [Accessed 3 Apr 2009].

[39] eBank UK - Schema for exchange of crystallography metadata. [web page] http://www.ukoln.ac.uk/projects/ebank-uk/schemas/. [Accessed 3 Apr 2009].

[40] Using Dublin Core. [web page] http://dublincore.org/documents/usageguide/. [Accessed 3 Apr 2009].

[41] blue Book Guide: A Guide to IUPAC Nomenclature of Organic Compounds; Blackwell Scientific Publications: Oxford, U. K., 1993. Purple Book: IUPAC Compendium of Macromolecular Nomenclature, 2nd edition; Blackwell Scientific Publications: Oxford, U. K., 1991. Red Book: IUPAC Nomenclature of Inorganic Chemistry, 3rd edition; Blackwell Scientific Publications: Oxford, U. K., 1990. White Book: IUBMB

Biochemical Nomenclature and Related Documents, 2nd edition; Portland Press: London, U. K., 1992.

[42] Antony N. Davies. XML in Chemistry and Chemical Identifiers. *Chemistry International*, 26(4):25, July-August 2004.

[43] I. David Brown, Sidney C. Abrahams, Michael Berndt, John Faber, Vicky L. Karen, W. D. Sam Motherwell, Pierre Villars, John D. Westbrook, and Brian McMahon. Report of the Working Group on Crystal Phase Identifiers. *Foundations of Crystallography*, 61:575–580, November 2005.

[44] Peter Murry-Rust. Unofficial InChI FAQ. [web page] http://wwmm.ch.cam.ac.uk/inchifaq/. [Accessed 3 Apr 2009].

[45] Simon J. Coles, Nick E. Day, Peter Murray-Rust, Henry S. Rzepa, and Yong Zhang. Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic Biomolecular Chemistry*, 3(10):1832–1834, 2005.

[46] Molecules from KEGG. [web page] http://wwmm.ch.cam.ac.uk/data/kegg/. [Accessed 3 Apr 2009].

[47] Distributed Structure-Searchable Toxicity (DSSTox) Database Network. [web page] http://www.epa.gov/NCCT/dsstox/. [Accessed 3 Apr 2009].

[48] Google Internet Search Engine. [web page] http://www.google.co.uk. [Accessed 3 Apr 2009].

[49] Peter Murray-Rust, Henry S. Rzepa, and Michael Wright. Development of chemical markup language (cml) as a system for handling complex chemical content. *New Journal of Chemistry*, 25(4):618–634, 2001.

[50] I. David Brown and Brian McMahon. CIF: the computer language of crystallography. *Acta Crystallographica Section B*, 58(3 Part 1):317–324, Jun 2002.

[51] A. L. Spek. Single-crystal structure validation with the program *PLATON*. *Journal of Applied Crystallography*, 36(1):7–13, Feb 2003.

[52] Research Councils UK. [web page] http://www.rcuk.ac.uk/research/outputs/access/default.htm. [Accessed 3 Apr 2009].

[53] Publication Policy - Full Structure Determination. [web page] http://www.ncs.chem.soton.ac.uk/pub_pol.htm. [Accessed 3 Apr 2009].

[54] DSpace, Digital Respository. [web page] http://www.dspace.org/. [Accessed 3 Apr 2009].

[55] Fedora Commons, Digital Respository. [web page] http://www.fedora-commons.org/developers/fedora.php. [Accessed 3 Apr 2009].

[56] Andrew J Milsted, Jeremy G. Frey, and Jamie Robinson. chemTools. [web page] http://chemtools.chem.soton.ac.uk/.

[57] Amit Singhal. Technologies behind Google ranking. [web page] http://googleblog.blogspot.com/2008/07/technologies-behind-google-ranking.html [Accessed 24 Apr 2006], July 2008.

[58] M. Wahl, T. Howes, and Kille S. Lightweight Directory Access Protocol (v3) RFC 2251. Technical report, The Internet Engineering Task Force, December 1997.

[59] ChemAxon. Marvin. [web page] http://www.chemaxon.com/marvin/. [Accessed 5 Apr 2006].

[60] Daylight Chemical Information Systems, Inc. SMILES. [web page] http://www.daylight.com/smiles/. [Accessed 5 Apr 2006].

[61] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. ISBN 3-900051-00-3.

[62] Jeremy G. Frey, Robert Gledhill, Sarah Kent, Brian Hudson, and Jon Essex. Schools Malaria Project. 2006.

[63] W. Graham Richards. Virtual screening using grid computing: the screensaver project. *Nature Reviews Drug Discovery*, 1:551–555, July 2002.

[64] CAChe Group. MOPAC. [web page] http://www.cachesoftware.com/mopac/. [Accessed 5 Apr 2006].

[65] Daylight Chemical Information Systems, Inc. SMILES - A simplified chemical language. [web page] http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html. [Accessed 5 Apr 2007].

[66] Berkeley Open Infrastructure for Network Computing (BOINC). [web page] http://boinc.berkeley.edu/. [Accessed 24 Apr 2006].

[67] Karolina Ersmark, Martin Nervall, Elizabeth Hamelink, Linda K. Janka, Jose C. Clemente, Ben M. Dunn, Michael J. Blackman, Bertil Samuelsson, Johan Åqvist, and Anders Hallberg. Synthesis of malarial plasmepsin inhibitors and prediction of binding modes by molecular dynamics simulations. *Journal of Medicinal Chemistry*, 48(19):6090–6106, 2005.

[68] University of Southampton: Research support services: Iridis (2012). [web page] http://www.southampton.ac.uk/isolutions/computing/hpc/iridis/. [Accessed 13 Jan 2013].

[69] BBC News. Fire destroys top research centre. [web page] http://news.bbc.co.uk/1/hi/england/hampshire/4390048.stm, 10 2005. [Accessed 19 June 2013].

[70] Michael Woelfle, Jean-Paul Seerden, Jesse de Gooijer, Kees Pouwer, Piero Olliaro, and Matthew H. Todd. Resolution of praziquantel. *PLoS Negl Trop Dis*, 5(9):e1260, 09 2011.

[71] Michael Woelfle, Piero Olliaro, and Matthew H. Todd. Open science is a research accelerator. *Nat Chem*, 3(10):745–748, October 2011.

[72] Mike Hulme and Jerome Ravetz. 'show your working': What 'climategate' means. [web page] http://news.bbc.co.uk/1/hi/8388485.stm, 2009.

[73] G Hughes, H Mills, D De Roure, J Frey, and L Moreau. schraefel, mc, smith, g. and zaluska, e.(2004) the semantic smart laboratory: a system for supporting the chemical escientist. *Organic and Biomolecular Chemistry*, 2:1–10.

[74] Alfred Nehme and Robert A Scoffin. *Electronic laboratory notebooks*. Wiley: Hoboken, NJ, 2006.

[75] Mark C Fishman and Jeffery A Porter. Pharmaceuticals: a new grammar for drug discovery. *Nature*, 437(7058):491–493, 2005.

[76] Public Record Office. Retention scheduling - health and safety records, 1998.

[77] David J Drake. Eln implementation challenges. *Drug discovery today*, 12(15):647–649, 2007.

[78] Nexxis. Labtronics. [web page] http://www.labtronics.com/electronic_laboratory_notebook.htm. [Accessed 20th Jul 2011].

[79] E-WorkBook Suite. IDBS. [web page] http://www.idbs.com/ELN/. [Accessed 20th Jul 2011].

[80] SAFE. SAFE-BioPharma Association. [web page] http://www.safe-biopharma.org/. [Accessed 20th Jul 2012].

[81] CambridgeSoft Life Science Enterprise Solutions. CambridgeSoft. [web page] http://www.cambridgesoft.com/. [Accessed 28th September 2012].

[82] Contur Software. iLabber, the flexible electronic laboratory notebook software. [web page] http://www.contur.com/home/. [Accessed 19th Jul 2013].

[83] JG Frey, GV Hughes, HR Mill, GM Smith, David De Roure, et al. Less is more: lightweight ontologies and user interfaces for smart labs. 2004.

[84] Gareth Hughes, Hugo Mills, David De Roure, Jeremy G Frey, Luc Moreau, Graham Smith, Ed Zaluska, et al. The semantic smart laboratory: a system for supporting the chemical escientist. *Organic & Biomolecular Chemistry*, 2(22):3284–3293, 2004.

[85] Gareth V Hughes, Hugo R Mills, Graham Smith, Terry R Payne, Jeremy Frey, et al. Breaking the book: translating the chemistry lab book into a pervasive computing lab environment. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 25–32. ACM, 2004.

[86] Resource Description Framework (RDF). MySQL. [web page] http://www.w3.org/RDF/. [Accessed 12rd Apr 2013].

[87] Jeremy G Frey. The value of the semantic web in the laboratory. *Drug discovery today*, 14(11):552–561, 2009.

[88] Jeremy Frey. Curation of laboratory experimental data as part of the overall data lifecycle. *International Journal of Digital Curation*, 3(1):44–62, 2008.

[89] Kieron R Taylor, Jonathan W Essex, Jeremy G Frey, Hugo R Mills, G Hughes, and EJ Zaluska. The semantic grid and chemistry: Experiences with combe chem. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(2):84–101, 2006.

[90] Kieron R Taylor, Robert J Gledhill, Jonathan W Essex, Jeremy G Frey, Stephen W Harris, and Dave C De Roure. Bringing chemical data onto the semantic web. *Journal of chemical information and modeling*, 46(3):939–952, 2006.

[91] M. J. Addis K. Meacham J. G. Frey, m. c. schraefel and S. Middleton. More Tea. *In preparation.*

[92] ed: ScienceBlogs LLC. The Blogs in ScienceBlogs. [web page] http://scienceblogs.com. [Accessed 20th Jun 2013].

[93] Zuiker A, Zivkovic B, Munger D (n.d.). Science Blogging Aggregated. Science Blogging Aggregated. [web page] http://sciblogs.co.nz/code-for-life/2010/08/22/science-blogging-aggregated-and-streamed/. [Accessed 21th Jun 2013].

[94] Cameron Neylon Jennifer R. Hale. Investigations into neutral drift. [web page] http://blogs.chem.soton.ac.uk/neutral_drift. [Accessed 28th September 2012].

[95] SPI. Debian: The Universal Operating System. [web page] http://www.debian.org/. [Accessed 20 Oct 2012].

[96] SPI. Debian social contract on debian homepage. [web page] http://www.debian.org/social_contract. [Accessed 20 Oct 2012].

[97] IEEE. POSIX Standards. [web page] http://standards.ieee.org/regauth/posix/. [Accessed 20 Oct 2012].

[98] The Apache Software Foundation. Apache v2.2.3. [web page] http://httpd.apache.org/. [Accessed 24 Apr 2006].

[99] MySQL :: The world's most popular open source database. MySQL. [web page] http://www.mysql.com/. [Accessed 23rd Jul 2013].

[100] PHP: Hypertext Processor. PHP v 5.2.0. [web page] http://www.php.net/. [Accessed 24 Apr 2006].

[101] Marko Štamcar. μblog. [web page] http://www.stamcar.com/projekti/microblog/. [Accessed 1 Apr 2006].

[102] Michael Heilemann. Kubrick Woodpress Style Template. [web page] http://binarybonsai.com/wordpress/kubrick/. [Accessed 3 Apr 2009].

[103] Wordpress Publishing Platform. [web page] http://wordpress.org/. [Accessed 3 Apr 2009].

[104] CM Jones, KA Bouton, JMN Hey, SE Latham, BN Lawrence, BM Matthews, AJ Miles, S Pepler, and K Portwin. Data publication: outputs of the claddier project. *Proc. Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data (PV 2007), Munich, Germany*, pages 09–11, 2007.

[105] Ann Copestake, Peter Corbett, Peter Murray-Rust, CJ Rupp, Advaith Siddharthan, Simone Teufel, and Ben Waldron. An architecture for language processing for scientific texts. In *Proceedings of the UK e-Science All Hands Meeting 2006*, 2006.

[106] Harry E Pence and Antony Williams. Chemspider: an online chemical information resource. *J. Chem. Educ*, 87(11):1123–1124, 2010.

[107] PostgreSQL: The world's most advanced open source database. PostgreSQL. [web page] http://www.postgresql.org/. [Accessed 16th Jun 2012].

[108] David Recordon and Drummond Reed. Openid 2.0: a platform for user-centric identity management. In *Proceedings of the second ACM workshop on Digital identity management*, pages 11–16. ACM, 2006.

[109] Moxiecode Systems AB. TinyMCE. [web page] http://www.tinymce.com. [Accessed 24 Nov 2012].

[110] Stephen Wilson, Andrew J Milsted, and Jeremy G Frey. *Comment by sketch: a picture says a million words.* PhD thesis, 2009.

[111] R. Rivest. The MD5 Message-Digest Algorithm RFC 1321. April 1992.

[112] BS ISO/IEC 18004:2006, Information technology. Automatic identification and data capture techniques. Bar code symbology, QR code.

[113] Damian Flannery, Brian Matthews, Tom Griffin, Juan Bicarregui, Michael Gleaves, Laurent Lerusse, Roger Downing, Alun Ashton, Shoaib Sufi, Glen Drinkwater, et al. Icat: Integrating data infrastructure for facilities based science. In *e-Science, 2009. e-Science'09. Fifth IEEE International Conference on*, pages 201–207. IEEE, 2009.

[114] Taverna Workbench. [web page] http://taverna.sourceforge.net/. [Accessed 3 Apr 2009].

[115] Jean-Claude Bradley. Open notebook science using blogs and wikis. Nature Precedings, doi: 10.1038/npre.2007.39.1.

[116] SIMILE, Massachusetts Institute of Technology. Welkin. [web page] http://simile.mit.edu/welkin/. [Accessed 16th Jun 2013].

[117] The Smart Research Framework (SRF), University of Southampton. LabBroker. [web page] http://www.mylabnotebook.ac.uk/software/labbroker.html. [Accessed 16th Jun 2013].

[118] MATLAB. [web page] http://www.mathworks.co.uk/products/matlab/. [Accessed 3 Apr 2009].

[119] Michael Woelfle, Jean-Paul Seerden, Jesse de Gooijer, Kees Pouwer, Piero Olliaro, and Matthew H. Todd. Resolution of praziquantel. *PLoS Negl Trop Dis*, 5(9):e1260, 09 2011.

[120] L. C. Shaffrey, I. Stevens, W. A. Norton, M. J. Roberts, P. L. Vidale, J. D. Harle, A. Jrrar, D. P. Stevens, M. J. Woodage, M. E. Demory, J. Donners, D. B. Clark, A. Clayton, J. W. Cole, S. S. Wilson, W. M. Connolley, T. M. Davies, A. M. Iwi, T. C. Johns, J. C. King, A. L. New, J. M. Slingo, A. Slingo, L. Steenman-Clark, and G. M. Martin. U.k. higem: The new u.k. high-resolution global environment model—model description and basic evaluation. *Journal of Climate*, 22(8):1861–1896, 2013/03/06 2009.

[121] Rosanne Quinnell, Brynn Hibbert, and Andrew Milsted. escience: Evaluating electronic laboratory notebooks in chemistry research. *Proceedings ascilite Auckland: Concise paper*, pages 299–802, 2009.

[122] Brynn Hibbert, Jeremy G Frey, Rosanne Quinnell, Mauro Mocerino, Matthew Todd, Piyapong Niamsup, Adrian Plummer, and Andrew Milsted. Teaching instrumental science globally using a collaborative electronic laboratory notebook. In *Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference)*, volume 16, 2010.

[123] Bram Luyten, Mark Diggory, and Peggy Schaeffer. Subject repositories for research data-the dryad approach. European Library Automation Group (ELAG) Annual Conference, 2012.

[124] Figshare. [web page] http://www.figshare.com/. [Accessed 14 May 2013].