ELSEVIER

The value of the Semantic Web in the laboratory

Jeremy G. Frey

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

The Semantic Web is beginning to impact on the wider chemical and physical sciences, beyond the earlier adopted bio-informatics. While useful in large-scale data driven science with automated processing, these technologies can also help integrate the work of smaller scale laboratories producing diverse data. The semantics aid the discovery, reliable re-use of data, provide improved provenance and facilitate automated processing by increased resilience to changes in presentation and reduced ambiguity. The Semantic Web, its tools and collections are not yet competitive with well-established solutions to current problems. It is in the reduced cost of instituting solutions to new problems that the versatility of Semantic Web-enabled data and resources will make their mark once the more general-purpose tools are more available.

Jeremy G. Frey
Jeremy G. Frey obtained his Chemistry degree at Balliol College Oxford. He obtained his DPhil in 1982 for work on experimental and theoretical aspects of van der Waals complexes, in Oxford's Physical Chemistry Laboratory, under the supervision of Prof. Brian Howard. A NATO/SERC post-doctoral fellowship (1982–1984) took him to the University of California Berkeley & Lawrence Berkeley Laboratory to work with Prof. Yuan Lee on molecular beam studies of reaction dynamics. In 1984 he took up a lectureship in the School of Chemistry at the University of Southampton, where he is now professor of physical chemistry and Head of the Structure & Materials Section. He is fellow of the Royal Society of Chemistry, fellow of the Royal Statistical Society. He is committed to a collaborative and interdisciplinary approach to chemical research. The interactions with the Schools of Physics, Electronics and Computer Science and the Opto-Electronics Research Centre have been particularly fruitful. He is an associate member of the Southampton Statistical Science Research Centre.

Introduction

In this article I will briefly outline the ideals and concepts of the Semantic Web as they are being (or could be) applied in the modern chemistry research laboratory. For the details of the Semantic Web I refer to papers published in *Drug Discovery Today,* as well as other scientific and general interest journals. The discussion is set in the context of the chemistry research lifecycle. In particular the ways in which the Semantic Web technologies will play an increasing role in dealing with the data, and literature produced by the modern digitally enabled chemical science research laboratory are explored. Some specific projects and software are described and the applications discussed for both the capture of chemical data and the subsequent storage of the data generated by the experiments. Having described the advantages of the Semantic Web and how data can usefully be generated in a way compatible with the Semantic Web the limitations of the data available in the Semantic Web at present are discussed. How these limitations may be overcome in the future forms the concluding sections of the article.

The Semantic Web

Sir Tim Berners-Lee, originator of the World Wide Web, first expressed his vision of the Semantic Web in 1999 [1],

Corresponding author: Frey, J.G. (j.g.frey@soton.ac.uk)
URL: http://www.soton.ac.uk/~jgf, http://www.comebchem.org

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2009.03.007

"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize."

The principles of the Semantic Web (http://www.w3.org/2001/sw/Activity) are a powerful call to make the Web work the way it should [2,3], seamless global integration of data and resources described for use by humans and machines. The vision is one in which information technology (IT) based on the web does work and works for us. The questions this poses are for example: What exactly, however, is needed to ensure that computers can handle all these transactions and operations for us? How different are the demands of computers from those of a new entrant to a field who does not know the jargon, or who has yet to know the reliable members of a community? Very importantly, where does all the data and information come from to allow all this automated processing?

Publications in this journal highlight the role and importance of the Semantic Web and ontologies and how they can serve to facilitate data integration [4], particularly as applied to the biological and bio-informatics computational communities. Williams describes the role and importance of internet-based tools for communication and collaboration in chemistry to communicate, disseminate and discuss their ideas [5]. Tetko has described the differences between the adoption of Web services by the bio-informatics [6] compared with chemical informatics communities, stressing that the issues were the quantity of data involved, and the scale of public funding to the bio-informatics area [7].

Biological classification, with its long history, was in a position to understand and take advantage of this approach. The large-scale funding of bio-informatics, DNA sequencing and related work, has led to very rapid developments that have left the rest of the physical sciences in a shadow. Unfortunately, the business case for the coordinated development of ontologies for chemistry (and other non-biological disciplines) has been more difficult to make. The pace of development is increasing rapidly, however, and it is an area where the 'network effect' applies, which means that the more people use ontologies to aid communication the more people see the advantage of being involved. While the use of semantics increases the ability to share information, it simultaneously runs up against the sociological issues of research ownership and the wish to maximize the direct personal gain from a research discovery (either in terms of publishable output before others or actual direct financial gain).

The Semantic Web for Life Sciences and Health Interest Group's (http://www.w3.org/2001/sw/hcls/) efforts within the W3C are widening the scientific fields in which the Semantic Web is being exploited and the web site http://www.semanticgrid.org is a useful collection of the wider Semantic Web and Grid articles and resources. Focusing on the Chemical sciences, the underpinning of the applications of the Semantic Web in chemical publications has usually been built on the XML-based mark-up descriptions of chemistry using Chemical Mark-up Language (CML) [8]. This led to projects that extract and mark up chemical data from the literature. Examples of this extraction are the SemanticEye [9], OSCAR (http://www.rsc.org/Publishing/ReSourCe/AuthorGuidelines/AuthoringTools/ExperimentalDataChecker/index.asp) from Peter Murray-Rust's group in Cambridge and the ChemXSeer project at Pen State (both discussed later in the article). They have also fed into the Royal Society of Chemistry's (RSC) 'Project Prospect' (http://www.project prospect.org/) to publish literature with much more extensive semantics and is possibly the most advanced system in commercial use.

The following sections discuss first the nature of the modern research laboratory and the impact of computers and digital systems on the nature of research. The high level view of the modern research life cycle is outlined as a guide to seeing the way use of the Semantic Web could impact on the planning, dissemination and laboratory phases of a research project. The core issues of the way data is, or should be, captured, handled and stored in the laboratory are explored with some examples of Semantic Web laboratory projects highlighted. Finally, some of the limitations of the current Semantic Web are presented with indications of how these are likely to be overcome in order to ensure that more scientific data flows into the Semantic Web.

The Modern Research Laboratory

This review will be mainly concentrated on the chemistry and physics research environment. Even with this limitation, the 'laboratory' takes many different forms in modern research, some dominated by fume cupboards, others by collections of small devices, large-scale equipment and perhaps even a series of automated robotically driven high-throughput lines. With the rise of inter-disciplinary work, many chemical laboratories will contain instrumentation that previously would have figured in biology laboratories in addition to apparatus related to nanotechnology, engineering and physics.

Computing and computers are now all-pervasive in chemical laboratories. Almost all equipment is controlled by at least in-built microprocessors, or by PCs for slightly larger scale, or more complex equipment (e.g. spectrometers). A problem for the development of integrated smart laboratories has been the (quite understandable) initial reluctance of equipment manufacturers to provide adequate links from these processors to external networks, either as a result of the lack of a suitable interface (RS232, GPIB, Ethernet) on the embedded device, or by the use of proprietary formats for the information and files produced and held on the PC systems. Fortunately, both of these problem areas (hardware and software) are being addressed and rectified by manufacturers, who see the advantage of networking their systems or buying into much more open software and information processing systems (Open Source Software or Open Standards). In these cases, the community shares the cost of building and maintaining the software needed, for example, to visualize, analyze or curate data.

The scale of the research footprint may vary from the traditional bench synthesis, through University medium size joint equipment (NMR, mass spectrometers, electron microscopes, clean rooms and X-ray diffraction), up to larger-scale central and international facilities (e.g. synchrotrons, neutron scattering, ultrahigh-power

lasers). Data from all these sources, experiments and methodologies need to be validated, integrated and modelled. The combination of experimental and theoretical modelling and simulation is often the way projects are now conducted. Semantic Web technologies can help to glue such projects together and improve the access to information and bring together information from within a project and from the literature.

Shifting the emphasis for data handling from laboratory equipment to the perspective of the researchers mirrors a similarly significant rise in the role of computers. Researchers record, analyze, communicate and discuss the experimental data and the resulting interpretations using computers. Even traditional voice communication may now be made over a data network using a voice over IP (VoIP) system. The tea room may be about to be replaced (or certainly enhanced) by social networking phenomena as a new generation of researchers join the workforce.

The research lifecycle

In considering the impact of the Semantic Web on the laboratory, the impact on researchers in their interactions with other researchers, with their experiments (i.e. the recording of experiments) and the literature (i.e. with the wider community) should be considered. A typical research project involves:

(a) phase 1, the planning phase, defining the project, involving literature searches, discussions with a range of experts, arranging funding and resources;
(b) phase 2, the execution phase in the laboratory, carrying out the laboratory work and recording exactly what was performed, collecting and describing data and undertaking the necessary analysis;
(c) phase 3, the final dissemination phase, in which a considered story is constructed for dissemination to the company or the world, depending on the environment in which the research was undertaken.

It is useful to consider first the Semantic Web support for the planning phase, then to the dissemination phase and finally, the most difficult problem, the support in the laboratory, bearing in mind that information needs to flow from each stage to the next and that the published information feeds back into the planning an execution of subsequent research.

Planning (Phase 1)

One of the main uses foreseen by the Web community for the use of Semantic Web technology is to speed up the planning phase of a project by allowing researchers to find others interested in similar problems, people who have skills needed in a project, or provide funding. The semantics involved are based on people and topics but may include more specific information on the research area. The semantics improve the selectivity of a search and aim to provide more accurate and relevant information. The same semantics then support the community set up to undertake the support. The e-Science/Cyber-infrastructure community has paid significant attention to the formation and support of virtual organizations formed across traditional administrative boundaries. Addressing the issue of how researchers can access facilities at different sites and exchange information without the need for separate credentials for each site has a direct impact on the research community. There are, however, very few sites where this system has actually been implemented.

Publishing/dissemination (Phase 3)

The influence of the Semantic Web on the dissemination side of research is growing. The key feature of the Semantic Web that takes it above the XML mark-up is the ability to refer to 'objects' by unique identifiers, as highlighted by Williams [5]. The development of a computable URI for organic molecules, the IUPAC InChI (the International Chemical Identifier, http://www.iupac.org/inchi/and for an informal FAQ http://wwmm.ch.cam.ac.uk/inchifaq/) and the InChI key (http://www.iupac.org/publications/ci/2008/3001/iw2_bioit.html), has been a key development in the application of Semantic Web technologies to chemistry [10,11]. It enables linking between entries in the literature, from an electronic laboratory notebook (ELN) to other information known (or calculated) about that molecular structure [12]. The push to ensure that data can be found and re-used (which, for a publisher, translates into greater profile as indeed it does for researchers also) means that ensuring that in addition to traditional publication metadata (authors, title, abstracts, key words and so on), it is highly advantageous to ensure that other material within a paper is made available in a structured manner.

The Royal Society of Chemistry's Project Prospect (http://www.projectprospect.org/) is a major advance in this area, with many aspects of the papers marked up in a rich manner, allowing, for example, searches for chemical structures present in the papers. Project Prospect uses semantic mark-up to highlight the important chemical data, particularly structures in a paper. Currently this information is added after the manuscript has been generated by the authors and submitted to the RSC. The chemical data is located and structures correlated with the relevant text and additional property data. This information could then be exposed to search engines that can, for example, locate papers that discuss a particular structure. The structure is then provided as a CML file than can be viewed for example in Jmol (an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/). The approach taken by Project Prospect demonstrates the power and advantages of having a publication that is compatible with Semantic Web technologies. If the semantics are then exposed to external users, more general services can be built on this semantically rich structure material.

CrystalEye (http://wwmm.ch.cam.ac.uk/crystaleye/) is attempting to capture structures from all available publications and contains over 100 000 entries. These are made generally available and fully marked up in an attempt to provide freely accessible data in structural chemistry to mirror the data available in much of the life science areas. It does not compete for size with the Cambridge Structural Database, ChemSpider (http://www.chemspider.com/), eMolecules (http://www.emolecules.com/), but should provide increased flexibility for future use. A reasonable amount of semantically enhanced data is being collected by the $Chem_X$Seer project (http://chemxseer.ist.psu.edu/) from Penn State (related to the $CiteSeer^X$ project), in which an integrated digital library and database allow for intelligent search of chemistry documents and specifically data obtained from chemical kinetics [13]. $Chem_X$Seer provides a powerful chemical entities search, while TableSeer extracts tables of data into a database that can then be queried.

Together they provide access to content that might otherwise be locked up inside publications.

In this context it is important to highlight that the outputs in the final phase of the research lifecycle are often the inputs to the planning phase of other research, and the way in which the material is prepared should take this into account. Frequently, however, the semantics, the metadata, links to spectra and similar data are lost in preparation for publication and then have to be inferred, reintroduced or reconstructed to produce the marked-up Semantic Web document (see later on: Provenance Explorer). Even when this has been done, the provenance back to the original laboratory notes and raw spectra are generally not available so the provenance chain is lost.

The laboratory and the organization of data (Phase 2)

This discussion suggests that the full power of the Semantic Web approach will only be achieved when tools to capture and maintain the context of data work well in the laboratory. Fortunately, the rise of electronic laboratory notebook and the increased interest in institutional data repositories place the research community in an ideal position to achieve an increase in the ease of capture of high-quality information. This highlights the importance of using systems compatible with the Semantic Web in the laboratory, even for small amounts of data because this information is destined to be integrated with other related data on a global scale. So, while the most obvious uses of the Web have been seen in the first and last phases of the work, the web support for the actual investigation, the key part of the process, lags behind and this is a serious, but understandable, omission that hinders the uptake of web tools in chemical research and one that several groups are now trying to address.

The main objective of a laboratory record is to ensure the data is accurately recorded in context, can be easily found, understood and used by others in the laboratory. The interplay between the researcher's laboratory notebook and the computer file systems that hold the data is crucial to modern laboratory record keeping. In many cases, the data represented by the file can no longer be usefully printed as it needs to be viewed on a computer, and the traditional 'cut and paste' technique where a copy of the print out was stuck in the laboratory notebook fails. Even if the printed (or plotted) version is meaningful, use of a printed copy breaks the link to the raw data in a form that can be manipulated. Better ways to maintain the link between the notebook data and files are needed. This is not an isolated requirement; many areas of commerce and government have similar needs. These needs have been met by using databases, but in the laboratory several areas of resistance are encountered.

In the laboratory environment, scientists often have preferred to store data as a series of flat files. Why is this perceived as a good approach? The superficial simplicity is attractive, but problems in finding, cataloguing, searching and recovering data, especially once the researcher has moved on, are significant. A series of rather random files are not good for data curation and not good for long-term re-use or to ensure that data integrity is maintained. Search systems can go a long way to help if key words exist to search for the data, but data files containing only numeric data with no description, present a major problem. This descriptive data (or metadata) is key to understanding the role of the Semantic Web in the laboratory context.

For long-running experiments, with similar, perhaps slowly evolving methodology, the use of relational database technology, together with the associated database management system (DBMS), looks attractive. They can ensure that data is maintained well, relationships represented and therefore enable easier searching and correlation of data. The DBMS maintains integrity and ensures an audit trail exists. The relational database, however, requires a well-considered schema, which takes time and effort to produce. This overhead in producing and maintaining the schema, the lack of rapid flexibility to changes in the nature of the data recorded, means that it is ideal for well-established experiments and data systems that will not change rapidly with time (e.g. some analytical systems would deploy Laboratory Information Management Systems (LIMS) or large scale international large scale facilities).

What is needed is something that covers the middle ground between the uncontrolled flat files and the rigid relational database. Effort will need to be expended on providing some metadata descriptions, but with more flexibility than the relational database. The Semantic Web provides such a technology. It relies on giving objects unique identifiers (these may be samples, molecules and so on) and subsequently making a series of statements (relationships) about these objects. This is a flexible approach that allows new relationships to be added and inferences to be drawn from the collection of relationships held in the Triple Store—the 'database' for the list of statements about these objects. These statements are in the Semantic Web language RDF of triples. A triple has the structure, subject, predicate, object, which might be something like {(molecule), (has systematic name), (ethane)}. Although currently searching large triple stores can be slow, better software is being developed, and large computer memory usually solves most current issues.

The capture of these relationships, essentially the metadata (in a computer readable form), is essential to the whole Semantic Web construct. This has to be achieved in as automated a way as possible, with as much information as possible captured from context. This is where the tension between freedom and control surfaces. In free text annotation a researcher is free to use any description they wish—another user may never be able to find this information as they may well be using different terms to describe the same situation. The use of a controlled vocabulary ensures that everyone uses the same terms, but these terms have to be agreed and workable. The construction of such an ontology is probably more of an effort than the database schema. We have a classic top-down versus bottom-up approach.

In order to exploit the Semantic Web, it is crucial to appreciate that the researcher's view of the content of an information system can be, and usually is, quite different from the 'view' required by a computer system attempting to act for, or with, that human. The key step is to see how to capture from the researcher the information needed for the computer to process the information. The key part is the capture of context. A key reenforcement for high quality metadata capture is the immediate advantage to the researcher of adding even limited additional information in terms of, for example, improved accuracy, easy record keeping and less repetition. The long-term advantage that

may accrue to the group, organization, company or society in general is important but not such a compulsive driver in the short term.

Instruments on the Web are increasingly how data is being collected. Instruments small and large are networked and accessible over the Web, not only to store and move data but also to allow remote interaction with the instrumentation and issues of security inevitably arise. Having suitably rich security models for data, which are not overwhelming, is an important aspect of any wide-ranging digital support for laboratory work. It is important for the protection of intellectual property, for regulatory approval, but also even in the open science context for correct and reliable attribution of work. Ensuring that the data produced by these instruments conforms to international standards, with high quality metadata in a useable form (preferably one which exploits the advantages of Semantic Web) is a battle that has not yet been won.

In summary the need for researchers to be concerned with the correct mark-up of their outputs is encapsulated in the term '*publication@Source*' [14]. Re-use may be either by the originators themselves or subsequently by others. How many of us have spent ages searching for some item of experimental data that we know we recorded but cannot find, or being in the situation where data is incomplete, or we are not sure which records go with which notes? The concern that the experimentalist should have in ensuring that the data record is accurate, complete and usable could be summarized as Curation@Source [15]. The difficulty of maintaining data/metadata coherence in one laboratory is magnified hugely when several laboratories are exchanging samples and information and probably using different information management or LIMS systems. The advantage of using the Semantic Web approach to achieve this in the laboratory is that it then provides data and metadata of high quality that drives subsequent applications in the global chemical information market.

Demonstrator projects for the Semantic Web in the laboratory

The following section outlines how, in order to meet some of the requirements of publication@source, we have implemented both the Semantic ELN and the Laboratory Blog Book (LaBlog) in our own laboratory environment. The Southampton Semantic ELN was built as part of the CombeChem project (http://www.combechem.org) [16] to demonstrate the use of a lightweight semantic model [17] in the planning and execution of a synthetic organic chemistry project [18]. The unique feature of this ELN software (http://smarttea.org) was the use of the language of the Semantic Web to record information on both materials and processes and the links between them. In the project, advantage was taken of the advance planning of experiments, which is necessary in order to comply with the regulations on the control of substances hazardous to health (COSHH). This plan was used to produce, a digital framework for the experiments that acts both as a guide to the sequence of experimental steps in the laboratory and a framework on which to hang the record of the observations. The ELN exists 'in the cloud' and is accessible via a web interface (e.g. for planning and recall) and via a tablet interface for ready access and recording in the laboratory during the experiments (see examples shown in Figure 1). The ELN has been very well received by those using it in the laboratory and has led to a significant increase in the quality of the notebooks, even in the parallel paper notebooks. It has led to much clearer and more thoughtful approach to planning and performing experiments. The initial trials have shown that the underlying RDF model is useful but not yet quite powerful enough and a new version is currently being developed.

By contrast, the Laboratory Blog Book [19] was built by Andrew Milsted, designed initially to be very free with almost no semantics. The Lab Blog Book (http://chemtools.chem.soton.ac.uk/projects/blog/) was built to support chemical biology experiments in Cameron Neylon's group as part of an Open Notebook Science project (Science in the Open at http://blog.openwetware.org/

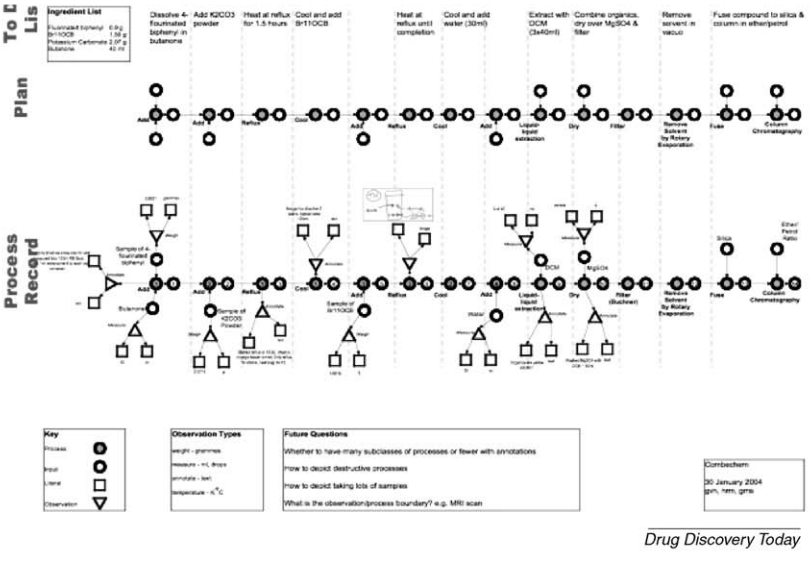

FIGURE 1

A representation of the RDF from the Southampton Semantic ELN showing the planning and process record stages highlighting the links between the steps.

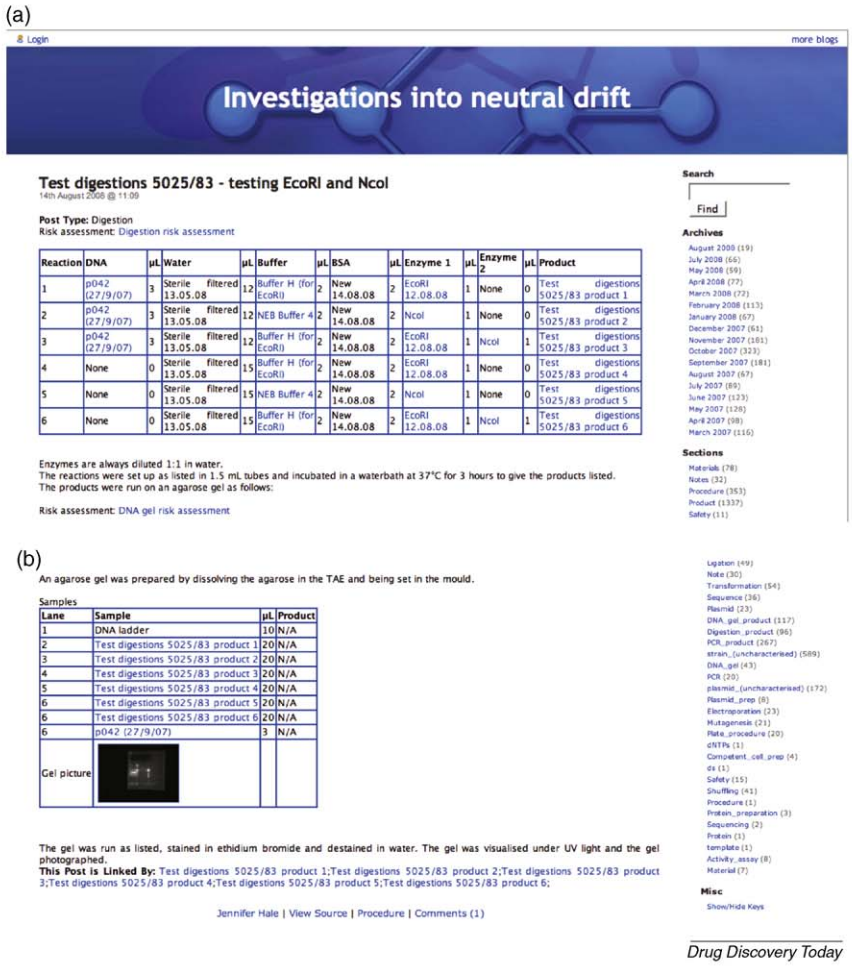

FIGURE 2

Showing examples of **(a)** highly interlinked data from the laboratory blog book **(b)** including links to images, taken from the Open Notebook Science experiments on Investigations into Neutral Drift which can be seen at http://chemtools.chem.soton.ac.uk/projects/blog/blogs.php/blog_id/13.

scienceintheopen/). The Lab Blog Book is now deployed in a number of research groups at Southampton in Chemistry and Physics and also at the Science and Technology Funding Council facilities at the Rutherford Appleton Laboratories (RAL). The purpose-built blog supports all the usual blogging functions and has extended functionality to enable the recording, linking and discussion of experimental data (see the Investigations into neutral drift blog at http://chemtools.chem.soton.ac.uk/projects/blog/blogs.php/blog_id/13), with an example shown in Figure 2. As the project developed, it became clear that the addition of some limited semantics added significant power for the classification of blog posts. The semantics are in the form of user-specified key/value pairs that form an uncontrolled ontology. For example, the key could be material and the values could be DNA primer, solvent, or enzyme. The value of the metadata in simplifying the linking and associating of blog posts is very high. The metadata simplifies the construction of templates for tables of linked experimental data. For example, in a template a table column for DNA primers will produce a drop-down list with only the blog entries marked as DNA primers simplifying the users choices. This encourages the re-use of the same terms by all those using a particular blog-book. The flexible linking structure that results enables navigation through experiments as materials or samples are first the product of some experimental step and then themselves used in subsequent steps.

The Blog system has also been used for a larger scale collaborative project on laser generated ultra fast soft X-rays. The X-Ray blog operates at a higher level than the individual notebooks and has been an excellent system for enhancing communication and collaboration. In both test cases, the quality and accessibility of the information record is very significantly enhanced from the previous paper notebooks. Even project meetings go more smoothly and require much less preparation time as material is already available for discussion. Finding and then modifying figures for presentations is now much easier, as the Blog makes the figures and the underlying data easily available to, for example, the principal investigator, even when the student who made the original is unavailable.

As the Semantic ELN and the Laboratory Blog-Book projects move forward, they are being linked with the description of, and guide for, a synthetic procedure based around the ELN and with the discussion and wider data environment for the experiment being recorded in the blog. The RDF from the enacted experiment

from the ELN is accessible as part of the blog record. The third part of the triangle is to provide a space in which to host the experimental plans (in effect, a type of workflow) and allow them to be exposed and found. For this, a social networking approach is being taken and use is being made of the MyExperiment (http://www.myexperiment.org) Web 2.0 user system, initially developed for bio-informatics workflows [20]

We are certainly not the only group investigating the use of the Semantic Web in chemistry laboratory; there is still relatively little semantically based laboratory software. The 'Collaboratory for Multi-scale Chemical Science' (CMCS, http://cmcs.ca.sandia.gov/) project [21] was one of the first projects to build a significant semantic infrastructure for collaborations and involved a semantic ELN [22] for multidisciplinary chemical sciences applications. The theme of provenance underlies much of the emphasis of the projects described above, and Jane Hunter's group at University of Queensland has looked at several chemistry and materials applications. They have developed several Semantic Web tools to support laboratory science. Provenance Explorer is a secure provenance visualization tool, designed to dynamically generate customized views of scientific data provenance that depend on the viewer's requirements and/or access privileges [23]. Using underlying RDF representations that generate connection graphs, it enables scientists to view the data, states and events associated with a scientific workflow in order to understand the scientific methodology and validate the results. The hypermedia user interface enables drilling down from simple high-level views to detailed views of complex subactivities. The connections between the objects that result from the workflow are inferred and so the system does not rely on this information being captured when the workflow was run. The inference system was implemented using Semantic Web Rule Language (SWRL) rules and the Algernon inference engine (http://algernon-j.sourceforge.net/doc/overview.html) [24].

It is a pity that one of the few examples showing rule-based inference in this area is being used to put back in places links that should never have been lost! If the data management systems used by researchers properly maintained the links between, for example, data, results, graphs and documents, the provenance task would be much easier. This is, of course, something that we all know we should do but rarely accomplish. The Target Information Management Tracking And More project (TIMTAM, http://sourceforge.net/projects/timtam/) does attempt to keep track of research data and the links. It is a system that implements a secure knowledgebase supporting the management, tracking and analysis of experimental data associated with structural biology projects. The project uses Semantic Web technology to support the management of protein crystallography project following all the steps from the choice of protein, through expression, isolation, purification, and crystallization, to the actual X-ray crystallography step.

One commercial example of a semantic ELN is the ELN from Recentris for chemistry and biology (http://www.recentris.com/products.html) and while interest from other commercial software suppliers in the Semantic Web is growing, it is not clear how much software will come from this route. The public domain has been a major contributor to such software and, for example, while not specifically aimed at laboratory work, the semantically aware software produced by the SIMILE project at MIT has proved extremely useful in visualizing data and manipulating data that is, or using their tools can be cast into RDF to join the Semantic Web (http://simile.mit.edu/wiki/).

Functional repositories

Having collected the data in the laboratory, the next problems are to know what to do with it, where to put it and how to keep it. The model builders experience this problem from the other side. A problem with modelling is often a lack of data or information about molecules. In many cases this information may have been measured, but not explicitly published (e.g. in a melting point list), but be present as part of the data provided with a paper on the synthesis of a molecule or the measurement of apparently unrelated properties. In the past when the corpus of chemical literature was more manageable, such information would be extracted by human readers and collated and, most importantly, considered and validated, and combined into a collection, such as the JANEF Thermodynamic Tables.

Some collections of spectra (NMR, IR, MS and so on) exist and are still currently maintained (at considerable cost, which may limit their use). These collections are smaller when compared with the number of spectra that have been recorded. The literature standard requires a limited analysis of the spectra in terms of peak positions but does not usually require the actual spectrum to be deposited. So these spectra are not available to be collated in a database. The more systematic use of repositories by research groups would make it much easier to provide the spectra to the community on publication for example as supplementary data. If this data is properly marked up it would become much more available to search engines.

The key issue here is that the repository needs to be, or should be, aware of the nature of the information held in the repository. This can be done to a varying degree. It may be that the data is held as a file and the file type designated by an extension, which is commonly understood and indicates what type of information is held in the file and what programs should be used to open the file. In its simplest form this could be a .txt file that can be read by a simple word processor, or a .jpg file rendered by image software. It may indicate that the file represents the output of an X-ray diffraction (XRD) structural study (.cif) or that it is another type of molecular structure file (.mol) or that it is an XML structure file in CML. The more information available about the data in the repository, the more use that can be made of the data and the more easily its validity and provenance can be checked. By defining an ontology and having links to definitions of the file type the Semantic Web principle extends the power of these file labels and allows all those accessing the Web to be able to interpret the meaning of the file.

Extending this idea to descriptions of the data contained within the file, a schema can be used to supply the information needed to ensure that the content of the file can be used. The Southampton e-Crystals site (http://ecrystals.chem.soton.ac.uk, see Figure 3) demonstrates the power of such a schema. The schema, used in the eCrystals system to indicate the different input and output files used in the crystallography process can be found at http://www.ukoln.ac.uk/projects/ebank-uk/schemas/xsd/ebankterms.xsd. The advantages of making the stages of a

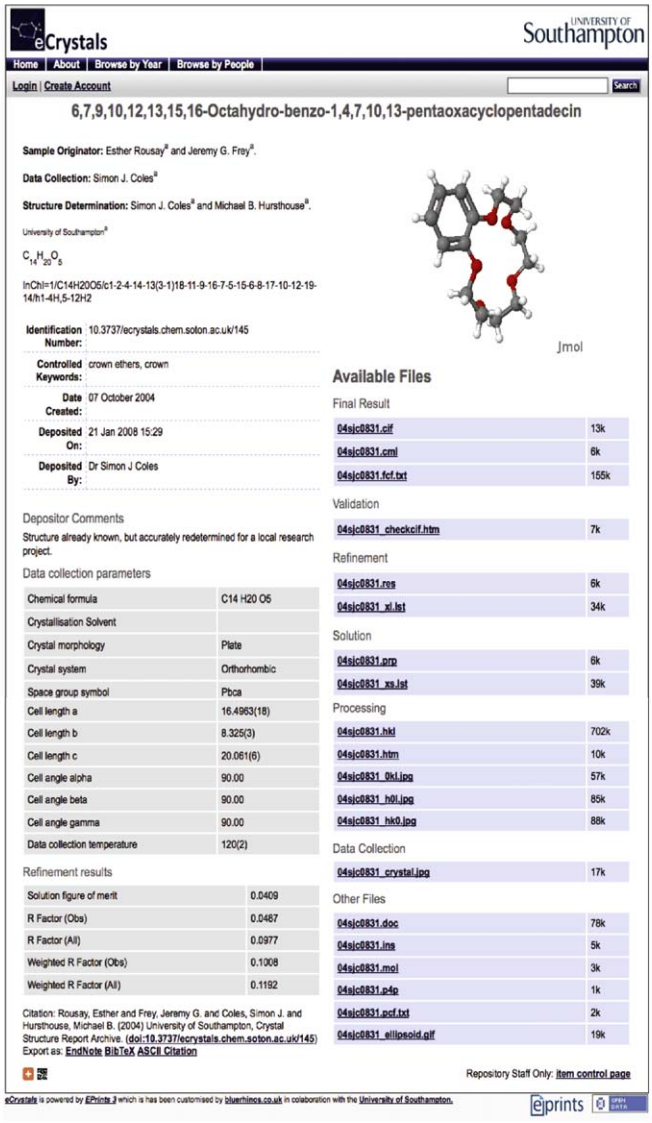

FIGURE 3

The figure shows a screenshot of the e-Crystals entry with the URI http://ecrystals.chem.soton.ac.uk/145/. For this set of results the files inheriting this base URI so that they can also be addressed individually and the URI resolves to deliver the file, for example, http://ecrystals.chem.soton.ac.uk/145/1//04sjc0831.cif and http://ecrystals.chem.soton.ac.uk/145/1//04sjc0831.hkl. The information on the 'splash page' is populated automatically from the various files that are produced in the XRD experiment and analysis when the files are uploaded. Using the XML schema (an augmented Dublin Core description) together with the URIs (http://www.ukoln.ac.uk/projects/ebank-uk/schemas/xsd/ebankterms.xsd) provided by e-Crystals is the contribution to the Semantic Web, with the information being made available via the OAI-MPH (Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, http://www.openarchives.org/) interface.

single crystal X-ray structure determination available, as well as the final structure file are clear when making unpublished data available and to maintain a usable provenance chain [25]. The final structures can then be found, using the Open Archive interface and extracted automatically for deposit in the international archive CSD (http://www.ccdc.cam.ac.uk/products/csd/) or any of the other related databases for example, the protein databank (PDB, http://www.rcsb.org/pdb/home/home.do), or the inorganic Crystal Structure Database (ICSD, http://www.fiz-karlsruhe.de/icsd_content.html *inter alia*). The novel aspect is that a link to the raw and processed data can be maintained and the details of the structure refinement are available for scrutiny. While the original intention for the eCrystals site was purely outward facing, in an attempt to deal with the dissemination of the large number of unpublished structures, together with the associated analysis data, it has now become an internal tool as well, used by the UK EPSRC National Crystallography Service (NCS, http://www.ncs.chem.soton.ac.uk/). The NCS uses the eCrystals systems along with other CombeChem tools, to manage part of the laboratory dataflow for the structures it records. It is ideal for keeping track of all the processed files generated in a crystallography refinement. It has proved difficult to encourage other crystallography groups to take up the eCrystals ideas.

The R4L (Repositories for the Laboratory, http://r4l.eprints.org/) and Spectra Projects (http://www.lib.cam.ac.uk/spectra/) were projects looking how to extend these ideas to other laboratory data. While a thread running through these projects is to facilitate the availability of data via Open Access routes, the fundamentals being addressed are important for ensuring data is properly curated and potentially available within any organization or available more widely though 'Open' or more commercial routes. In a similar vein of creating an international repository for combustion kinetics data with significant quantities of supporting laboratory data is the PriMe database system (http://primekinetics.org/) [26], which is built on an ontology, but uses more conventional database technology. It nevertheless has increased interest in the corresponding support for the experimental data within the laboratories that are planning to contribute data to the PriMe database; it will, in principle, parallel the eCrystals/CSD developments.

The Semantic Web does not yet, in general, supply the data needed by Chemical Science researchers. The majority of data sources are both non-semantic and commercial. The Semantic Web projects that have been most useful are being used in side laboratories, and the information is not yet exposed to the outside world. Demonstrator public versions are visible but, for example, the eCrystals sites contain only a relatively few structures (though these are harvested by the CSD). All the structures generated by the NCS are, however, held on a private version of eCrystals because the links to the processed and raw data, as well as the final structures, is so useful for the service. It is the provenance trail to the processing and raw data that makes the eCrystals site interesting and useful. For wide uptake, which could result making the estimated 75% of structures that have been determined, but not published available, it does, however, require a shift in viewpoint of the community and this is proving to be more difficult to achieve than the technical solutions.

Linked data and RDFa: how to get more data on the Semantic Web

It is clear that the aim of the Semantic Web is to enable information to be linked together and eventually to automate the associa-

tion of related data. It is the links that add value; but getting people to add them, or add sufficient information that they can be created automatically, is proving to be hard. In many areas of scientific data relevant to laboratory work, however, it will be sufficient to ensure that data when created are linked to obviously related information. This will ensure that the data is not isolated and the connection graphs will steadily grow. In a further exploration of Semantic Web, Berners-Lee outlined some principles of good practice that would ensure that data structures could be published and used as easily as documents (http://www.w3.org/DesignIssues/LinkedData.html). The 'Linked Data' guidelines (http://linkeddata.org/) are simply about ensuring the links between data from different sources are made in a robust useful way. The available 'Linked Data' (see http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets) in scientific areas is extremely limited and must grow if the Semantic Web is to prosper. Of course, proprietary data held in a company could form 'a Semantic Web' and not necessarily be visible to 'The Semantic Web'.

When the Web was invented, the experts knew that it deviated from best practice—against the principles of the SGML (Standard Generalized Markup Language, http://www.w3.org/MarkUp/SGML/) community: it mixed content and presentation. Against the principles of the hypertext community, it mixed content and links; but it worked and promoted the rich interlined set of documents we now have. We now have to undo the problems of the content/presentation/links mixture, but we have the World Wide Web. The ideal of human and computer-processible information that the XHTML (Extensible HyperText Markup Language, http://www.w3.org/TR/xhtml1/) and RDF gives us, has not yet seen the explosion in content that characterizes the Web, take up has been poor. The RDFa standards (http://www.w3.org/TR/xhtml-rdfa-primer/and http://wiki.creativecommons.org/RDFa) mean that RDF information, to be part of an XHTML document, which while probably violating the 'separation ideals' may well just make the Semantic Web sufficiently appealing to a wider range of content providers. Using a few simple XHTML attributes, authors can mark up human-readable data with machine-readable indicators for browsers and other programs to interpret. A web page can include mark-up for items as simple as the title of an article, or as complex as a user's complete social network. The browser receives information on the meaning of a web page's visual elements. It is easer to understand what meanings need to be added, facilitating the production of Semantic Web materials using current tools, and kick-start the material on which the Semantic Web tools will operate.

SWAN (Semantic Web Applications in Neuromedicine)

One of the most complete scientific semantic systems is SWAN (Semantic Web Applications in Neuromedicine, http://swan.mindinformatics.org/). SWAN is a Web-based collaborative program that aims to organize and annotate scientific knowledge about Alzheimer disease and represents one of the most complete sets of ontologies for a subject area [27,28]. The ultimate goal of this project is to create tools and resources to manage the evolving data and information about Alzheimer disease. The SWAN project is designed to allow the community to author, curate and connect a diversity of data and ideas about Alzheimer disease using the emerging Semantic Web paradigm for deep interconnection of data, information and knowledge. The 'SWAN ontology' http://swan.mindinformatics.org/ontology.html [29] that was originally organized in a single block has been modularized to foster reusability and integration with other existing ontologies in a software ecosystem. The eco-system contains the basic ontology for the neuroscience terms and ontologies for agents to support scientific discourse. Importantly the discourse ontology captures a middle ground between the natural language in which scientific discussions are conducted and the far more controlled, rigorous, and fixed nature of formal ontologies 'about' the science. The SWAN system, extensive as it is, does not reach down into the laboratory, but it does provide an indication of what we can hope to achieve in the future by deploying more Semantic Web technology in the laboratory.

Is the Semantic Web sufficient?

The semantic concepts that underlie the Semantic Web are clearly necessary for the future of research. The examples given above demonstrate that the Semantic Web is powerful enough to ensure that information is recorded in context and can be handled by humans and computers and even be presented to a multi-disciplinary audience with some hope that the data will be sufficiently well-described to be of value. It has to be admitted even by the enthusiasts, however, that running Semantic Web applications can be a challenge. In many cases there is a significant overhead in loading the right versions of JAVA and other base software. A great deal of the software comes from demonstrator projects and the code is not robust and unlikely to work for long, as operating systems are upgraded. Improved software is becoming available as either public or commercial interests see the demand for the Semantic Web tools growing. The 'hype' maximum in the adoption curve has been reached, and we can now hope for real solid progress. While the semantic tools and data collections may not yet be competitive with well-established solutions to current problems, it is in the reduced cost of instituting solutions to new problems that the versatility of Semantic Web-enabled data and resources will make their mark.

The Semantic Web will not solve all the problems of communication and collaboration. The communication of computer to computer may be radically eased by the adoption of Semantic Web technologies but here will still be a problem at the more basic level of understanding how the material presented 'on the screen' to researchers is understood by those researchers, especially when they may have a widely varying experience, with different cultural backgrounds. Simply consider how many of the conventions for placement of objects depend on assumptions of language flowing from left to right; the semiotics of the communication are important. To handle this level of user engagement with user interfaces requires even more care and attention to detail, but has the benefit of bringing wider audiences together [30].

Acknowledgements

I would like to acknowledge the support and collaboration of my colleagues in Chemistry, Physics and Computer Science, and IBM

and Microsoft, without whom my exploration of the Semantic Web and e-Science in general would not have been possible. The support of the EPSRC and JISC in providing funding for the research is gratefully acknowledged.

References

1 Berners-Lee, Tim (1999). *Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor* (Hardcover). Harper San Francisco. By Michael Dertouzos (Foreword), Mark Fischetti (Collaborator) ISBN-13: 978-0752820903

2 Berners-Lee, T. *et al.* (May 2001) The Semantic Web. In *Scientific American Magazine*.

3 Feigenbaum, L. *et al.* (December 2007) The Semantic Web in action. In *Scientific American Magazine*.

4 Stephen, P.G. (2005) Ontologies and semantic data integration. *Drug Discov. Today* 10, 1001–1007

5 Williams, A.J. (2008) Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* 13, 502–506

6 Curcin, V. *et al.* (2005) Web services in the life sciences. *Drug Discov. Today* 10, 865–871

7 Tetko, I.V. (2005) Computing chemistry on the web. *Drug Discov. Today* 10, 1497–1500

8 Murray-Rust, P. *et al.* (2001) Development of Chemical Markup Language (CML) as a system for handling complex chemical content. *New J. Chem.* 618–634

9 Casher, O. and Rzepa, H.S. (2006) SemanticEye: a semantic web application to rationalize and enhance chemical electronic publishing. *J. Chem. Inf. Model* 46, 2396–2411

10 Coles, S.J. *et al.* (2005) Enhancement of the chemical semantic web through INChIfication. *Org. Biomol. Chem.* 3, 1832–1834

11 McNaught, A. and Frey, J. (2006) Tools for the trade, pp.12–15, Chemistry International, 28 (http://www.iupac.org/publications/ci/2006/2806/4_tools.html)

12 Taylor, K.R. *et al.* (2006) Bringing chemical data onto the semantic web. *J. Chem. Inf. Model* 46, 939–952 (doi:10.1021/ci050378m)

13 Bolelli, L. *et al.* (2007) ChemXSeer: a chemistry web portal for scientific literature and datasets. *Open Repositories Conference*, San Antonio, Texas, 2007 In: http://openrepositories.org/2007/program/files/6/bolelli.pdf

14 Frey, J.G. *et al.* (2002). Position Paper: Publication at Source: Scientific Communication from a Publication Web to a Data Grid. Euroweb 2002 the Web and the GRID: from e-Science to e-Business, British Computer Society. http://ewic.bcs.org/conferences/2002/euroweb/session3/paper3.htm.

15 Frey, J. (2008) Curation laboratory experimental data as part of the overall data lifecycle. *Int. J. Digital Curation* 3

16 Frey J. et al. (2006) CombeChem: a case study in provenance and annotation using the Semantic Web, revised selected paper, lecture notes in computer science, vol. 4145 (Information Systems and Applications, incl. Internet/Web, and HCI) (Moreau, Luc and Foster, Ian, eds.), XI, 288, ISBN: 3-540-46302-X, http://www.springer.com/uk/home/generic/search/results?SGWID=3-40109-22-173681711-0

17 Frey, J.G. *et al.* (2004) Less is more: lightweight ontologies and user interfaces for Smart Labs. In *Proceedings of the UK e-Science All Hands Meeting*, Nottingham, p. 8, EPSRC ISBN-1-90 4425-21-6 In: http://www.allhands.org.uk/proceedings/papers/187.pdf

18 Hughes, G. *et al.* (2004) The Semantic Smart Laboratory: a system for supporting the chemical eScientist. *Org. Biomol. Chem.* 2, 3284–3293 (doi:10.1039/b410075a)

19 Frey, J.G. *et al.* (2008) The laboratory blog-book: how a laboratory blog notebook has developed to support, and in turn has been influenced by, experimental laboratory practice. *4th International Conference on e-Social Science* In: http://www.ncess.ac.uk/events/conference/programme/workshop1/?ref=/programme/fri/4cfrey.htm

20 De Roure, D. et al. The Design and Realisation of the Virtual Research Environment for Social Sharing of Workflows, Future Generation Computer Systems (in press) (Corrected Proof, Available online 5 July 2008, ISSN 0167-739X, doi:10.1016/j.future.2008.06.010) (http://www.sciencedirect.com/science/article/B6V06-4SX9FTN-4/2/e44404603ec05e03f8add717d5069d25)

21 Myers, J.D. *et al.* (2005) A collaborative informatics infrastructure for multi-scale science. *Cluster Computing—J. Networks Software Tools Appl.* 8, 243–253

22 Talbott, T. *et al.* (2005) Adapting the Electronic Laboratory Notebook for the semantic era. *International Symposium on Collaborative Technologies and Systems, Proceedings (CTS 2005)* pp. 136–143

23 Hunter J. and Cheung, K. (2007) Provenance explorer—a graphical interface for constructing scientific publication packages from provenance trails, Special Issue—Digital Libraries and eScience. Int. J. Digital Libr. 7

24 Horrocks, I. *et al.* (2004) SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission (Available at http://www.w3.org/Submission/SWRL/)

25 Coles, S.J. *et al.* (2006) An eScience environment for service crystallography—from submission to dissemination. *J. Chem. Inf. Model.* 46, 1006–1016 (doi:10.1021/ci050362w)

26 Frenklach, M. *et al.* (2004) Collaborative data processing in developing predictive models of complex reaction systems. *Int. J. Chem. Kinet.* 36, 57–66

27 Gao, Y. *et al.* (2006) SWAN: a distributed knowledge infrastructure for Alzheimer disease research. *J. Web Semantics* 4, 222–228 (doi:10.1016/j.websem.2006.05.006)

28 Clark, T. and Kinoshita, J. (2007) Alzforum and SWAN: the present and future of scientific Web communities. *Briefings in Bioinform* 8, 163–171 (doi:10.1093/bib/bbm012). PMID: 17510163

29 Ciccarese, P. *et al.* (October 2008) The SWAN biomedical discourse ontology. *J. Biomed. Inform.* 41, 739–751 (Epub 2008 May 4. PMID: 18583197)

30 Frey, J. (2008) The Semiotic Web: don't forget the user amongst all the semantics. In *Proceedings of the First International Workshop on Understanding Web Evolution (WebEvolve2008)*, 22 April 2008, Beijing, China. ISBN-978 085432885 7 pp. 44–46. In: http://journal.webscience.org/38/