Review

www.molinf.com

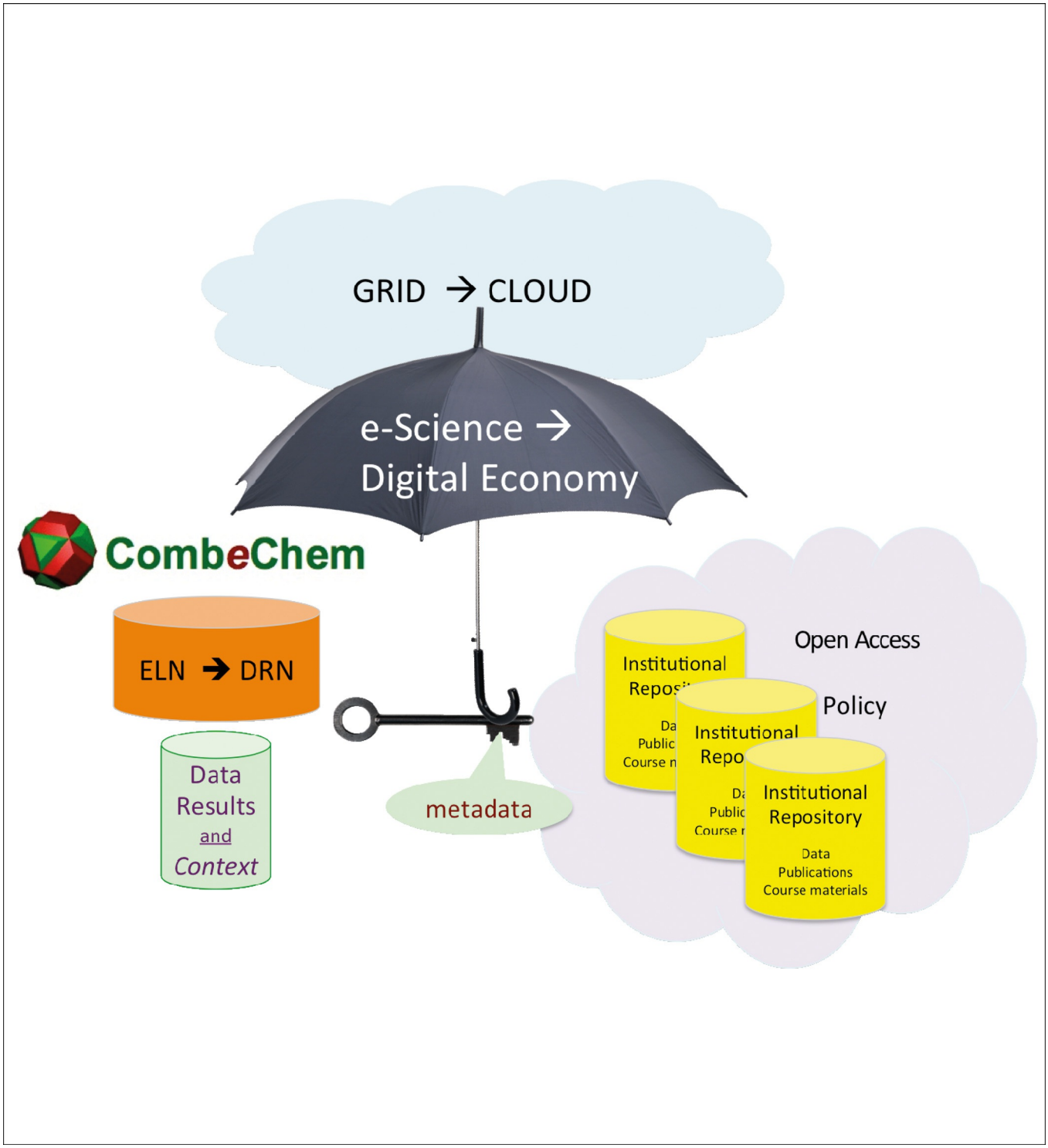
molecular informatics

Special Issue United Kingdom

DOI: 10.1002/minf.201500008

The Evolution of Digital Chemistry at Southampton

Colin Bird,[a] Simon J. Coles,[a] and Jeremy G. Frey*[a]



Abstract: In this paper we take a historical view of e-Science and e-Research developments within the Chemical Sciences at the University of Southampton, showing the development of several stages of the evolving data ecosystem as Chemistry moves into the digital age of the 21st Century. We cover our research on aspects of the representation of chemical information in the context of the world wide web (WWW) and its semantic enhancement (the Semantic Web) and illustrate this with the example of the representation of quantities and units within the Semantic Web. We explore the changing nature of laboratories as computing power becomes increasing powerful and pervasive and specifically look at the function and role of electronic or digital notebooks. Having focussed on the creation of chemical data and information *in context*, we finish the paper by following the use and reuse of this data as facilitated by the features provided by digital repositories and their importance in facilitating the exchange of chemical information touching on the issues of open and or intelligent access to the data.

Keywords: Chemical information · Digital · Semantic Web · Context · Modelling · Repositories

1 Introduction

To place our review of the current work of the University of Southampton Chemical Informatics group in context, it is useful to explain how we started to work in this area; a combination of a vision and funding opportunities.

In the early 1990's one of us (JGF), benefiting from increased computerization of his laboratory but still using the BBC Microcomputer,[1] started to be concerned with the storage of data with regard specifically to the reproducibility of experiments. Even on a small scale, in a pulsed laser experiment, the computer made it possible in principle to record the signal for each laser pulse, rather than as previously, an average integrated with a boxcar or similar gate. The advantage was that much better statistical calculations could be made, thresholds could be varied and the effect of changing them could be investigated. The disadvantages were much more data filling up disks and data being kept but without a good index, so its value could decay rapidly.

A vision, where the path between data and paper was traversable in either direction, the pathways would be maintained, providing permanent provenance, and context for data and the paper, was born, but could not be achieved with the technology then available. The World Wide Web was in its infancy, a Southampton system, Microcosm[2,3] was a possibility, but hyperlinks, data standards, in fact most of the necessary parts of a data ecosystem, were simply not available.

1.1 Navigation

A guide to the subsequent sections of this paper is appropriate. We start by outlining the e-Science origins and approach that we have taken to the evolution of chemical informatics in Southampton. In Section 2 we explain how the fact that, even in the just over 10 years that we have been pursuing the informatics agenda, the digital world has changed significantly, which has impacted on both the chemistry and the researchers creating, learning and using chemical information.

In Section 3 we discuss the representation of chemical information in a semantically meaningful manner within the worldwide digital infrastructure (Internet and the World Wide Web), summarising our published work and what it has led on to, together with an exemplar discussion on previously unpublished work on the representation of quantities and units on the Semantic Web.

In Section 4 we take a look at the role of electronic (or digital) laboratory notebooks (ELNs), setting our web-based ELN (primarily LabTrove) and semantic work (Smart Tea) within a general context of the evolution of scientific note taking and provide a description of our latest work on an "API-driven core" to support a mobile-first notebook system.

In Section 5 we consider the nature and functions of data repositories for chemical information from both a technical and a social perspective. In section 6 and the conclusion we look at how all these pieces join together to support the much more digitally enabled laboratory of the future, and consider how much of the original vision of CombeChem has been accomplished, surpassed or bypassed and how much is still to be achieved in the provision of tools and services to support the laboratory researcher in chemical sciences.

As a guide to the many Southampton projects an outline Gantt style chart is given in Figure 1.

[a] *C. Bird, S. J. Coles, J. G. Frey*
Chemistry, University of Southampton
University of Southampton, Highfield, Southampton, SO17 1BJ, UK
phone: +44 (0)2380593209
**e-mail: j.g.frey@soton.ac.uk*

2 The UK e-Science Programme?

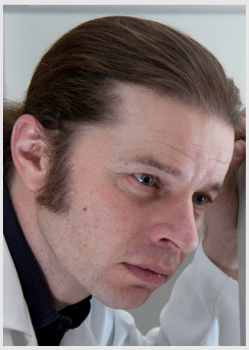
In early 2000s the UK initiated new interdisciplinary research funding: within the ICT programmes the Interdisciplinary Research Collaborations were complemented by the cross-research council e-Science programme The e-Science area was described as "research done through distributed global collaborations enabled by the Internet, using very large data collections, terascale computing resources and high performance visualisation."[4] In Southampton we put together a bid that involved chemists, statisticians and computer scientists, looking at the collection, analysis, and prediction of structure and function of different crystalline solid forms (polymorphs) taking advantage of the then relatively new high-throughput and combinatorial approaches to chemical discovery. The project, owing much to combinatorial approaches to experimental and computational polymorph studies, became known as the CombeChem project,[5] as an application of e-Science principles to high throughput and Combinatorial Chemistry (Combichem) The "visionary" UK effort led to *cyberinfrastructure* programmes in the USA and Australia".

The CombeChem project was unique amongst the initial e-Science projects in its emphasis on users and usability studies. Usability did subsequently become a significant area with in the e-Science programme. This initiated a considerable amount of chemical informatics research within Southampton. Particularly relevant for this paper are the research themes on chemical information and the Semantic Web[6–13] and the work on Electronic Laboratory Notebooks.[14,15]

The e-Science programme was not the only driver for the Southampton Chemical Information group. The EPSRC-funded Combinatorial Chemistry project led by Mark Bradley (which funded a new building at Southampton in 1998), offered huge potential in generating chemical structure and property data. The National Crystallography Service (NCS),[16] which moved to Southampton in 1998, provided not only a major capability in structural studies but also interactions with a diverse range of research groups around the country.

At about the same time the JISC (Joint Information Systems Committee, now Jisc) started a Data Management theme, which has continued in one form or another to address institutional issues with data management. The early work of the CombeChem project was adopted by members of UKOLN (United Kingdom Office for Library and Information Networking)[17] and led to our connection with reposi-

Having obtained his BSc and PhD in Chemistry at the University of Southampton, *Colin Bird* joined IBM UK Laboratories. After contributing to IBM's electrochromic display technology, he transferred to the IBM UK Scientific Centre to develop advanced image and visualisation applications. His work on content-based image retrieval led to a one-year secondment in 1999 back to the University of Southampton. On returning to IBM, he was involved in various aspects of information management, specialising in classification and metadata, and became an information architect. When he left IBM, he resumed his collaboration with Professor Jeremy Frey on e-Research projects, which began in 2000 as an industrial partner for the CombeChem project.



Associate Professor *Simon Coles* obtained his BSc from the University of Wales, Cardiff (1992) and continued on to complete a PhD (1997) and then a PDRA appointment with the Royal Institution, based at the Daresbury synchrotron. In 1998 Simon moved to Southampton to establish a new laboratory and manage the National Crystallography Service (NCS). He transferred to the School of Chemistry staff in July 2009 and became the NCS Director. Simon has diverse research interests ranging from structural systematics and high-resolution crystallography (charge density studies), through crystal growth and x-ray imaging to the development of new information technologies for recording, processing and sharing research data.

Jeremy Frey obtained his DPhil on experimental and theoretical aspects of van der Waals complexes, in the PCL, Oxford, followed by a NATO/SERC fellowship at the Lawrence Berkeley Laboratory. In 1984 he took up a lectureship at the University of Southampton, where he is now Professor of Physical Chemistry and head of the Computational Systems Chemistry Group. His experimental research probes molecular organization in environments from single molecules to liquid interfaces using laser spectroscopy from the IR to soft X-rays. In parallel he investigates how e-Science infrastructure can support scientific research with an emphasis on the way appropriate use of laboratory infrastructure can support the intelligent access to scientific data, with a focus, but not exclusively on chemical informatics. He is strongly committed to collaborative inter and multi-disciplinary research and is skilled in facilitating communication between diverse disciplines speaking different languages. He has successfully lead several large interdisciplinary collaborative RUCK research grants, from Basic Technology (Coherent Soft X-Ray imaging), e-Science (CombeChem) and most recently the Digital Economy Challenge area of IT as a Utility Network+, where he has successfully created a unique platform to facilitate collaboration across the social, science, engineering and design domains, working with all the research, commercial, third and governmental sectors.

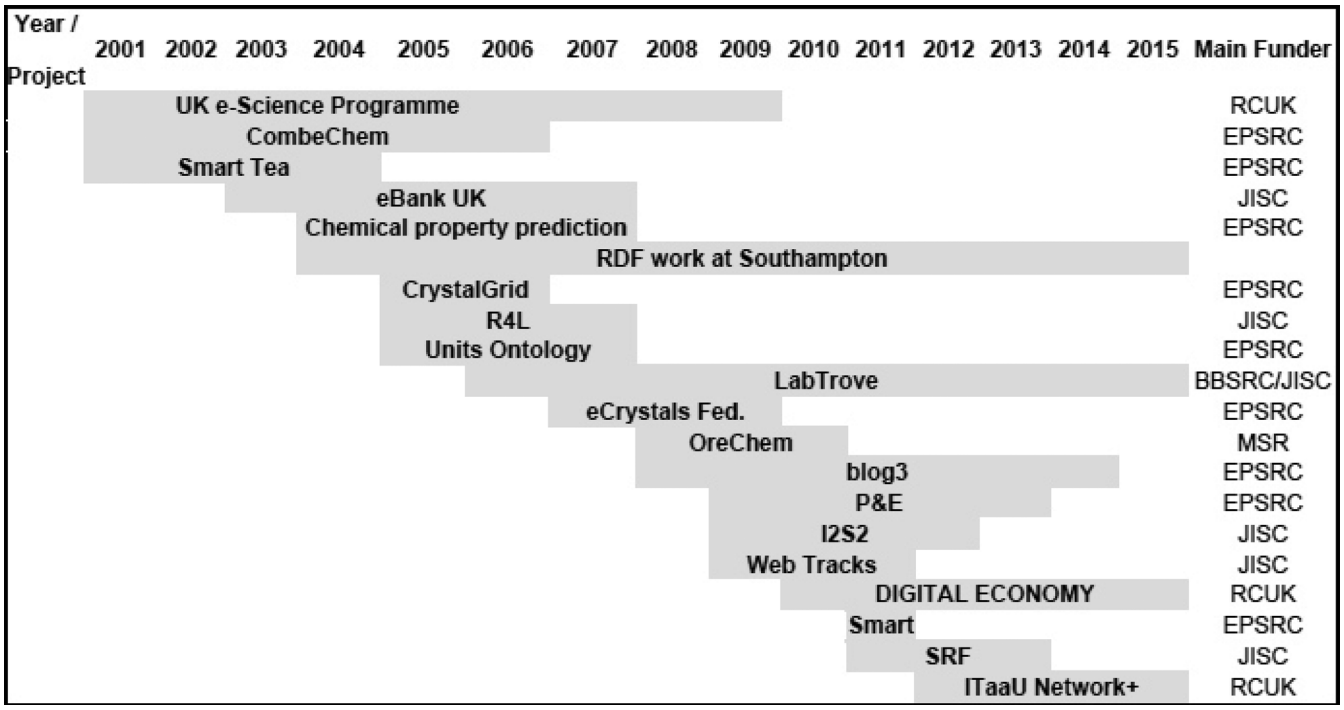


Figure 1. Chemistry related e-Science Projects based at the University of Southampton. RCUK: Research Councils UK; EPSRC: Engineering and Physical Sciences Research Council; JISC: (now Jisc); BBSRC: Biotechnology and Biological Sciences Research Council; MSR: Microsoft Research.

tories and in particular the eBank and eCrystals projects which initiated a whole stream of work using chemistry as an exemplar for data and repositories (discussed further in Section 6).

The approach we have taken in this paper is to outline the development of e-Science, Chemical Informatics, Data Science, and Data Management at Southampton by considering a matrix of three main themes:

- Representation of chemical information;
- The changing nature of laboratories and Electronic Laboratory Notebooks (ELNs);
- Digital repositories;

as they intersect with the topics of :

- Social vs. Technical or Human vs. Computer;
- Data & information integration;
- Sharing & collaboration (use/reuse).

It will become clear that a significant amount of our research in these areas has been as much concerned with the human and social aspects of researchers as it has with software and computational technology. Attitudes to the use and role of digital technology in chemistry are as diverse as the subject itself. Societal and funding drivers and commercial pressures do not necessarily align and have produced different views over the sub-disciplines and research linked to other major areas, such as the life sciences and bioinformatics. For these areas the drive is towards openness, whereas for medical research, or where business profits are involved, the trend has until very recently been quite the opposite.

A major concern with both funders and the wider community has been the willingness and ability to share information, both papers and data. Attitudes to this vary tremendously across the chemistry community, but technological considerations have also played a part. Barriers to efficient curation need to be overcome and reward for sharing clarified[18] and References therein.

3 The Evolving Digital World

From the outset of our e-Science research, the importance of the Web and Web services in providing chemists with access to chemical information was clear.[19] How this could be achieved was much less obvious. What was needed to provide versatile, easy to create and easy to use systems, and to ensure that the vast quantities of chemical information could and would flow into these systems, was much less obvious. That this needed to be done became increasingly clear. For Generation X (born early 1960s to early 1980s) and Generation Y (born early 1980s to early 2000s), studies of their views on access and availability of information, where searching on the Web was instinctive (if not in-

tuitive), showed the direction of travel and indicated what researchers of the future (as seen from the perspective of the early 2000's) would expect and demand.[20,21]

The rise of social networks and networking has changed the environment in which chemistry students and researchers operate. However, while many of these individuals use social networking tools in their everyday lives, the uptake and use within chemical research has been slow and patchy. In the early days limitations in the representation of chemistry on the Web could perhaps be blamed, but subsequently it is clear that entrenched attitudes towards sharing information are more significant barriers. As we will see in subsequent sections, as the technical problems are steadily solved the social problems remain and until and unless appropriate recognition goes with the sharing of information, little will change, even in the face of funders' demands.

This emphasis on the users' perspective fitted very well with the CombeChem project and the work on electronic laboratory notebooks, which had the aim of bringing people, materials and processes uniformly onto the Web, led by the need to provide a tool that researchers would actually use. Our focus integrated surprising well with the evolution of the e-Science programme as a whole towards the RCUK Digital Economy programme.[21] Moreover, the UK Government has set a service standard of Digital by Default:[22] the era of *digital by design* has arrived and "digital and mobile first" informs much of our current e-Science thinking.

The "always on" 3G and now 4G connectivity that smart phones have provided for many people, at least in the urban areas of the developed world – as well as wide scale Wi-Fi in educational environments – has changed expectations.

Work prior to the e-Science programme made great strides in the representation of chemical information by computers; the syntax of digital chemistry already had well defined standards.[23] In CML (Chemical Markup Language) there was even a standard that conformed to much more general computer science standards.[24,25] The aim of the CombeChem Project and subsequent research at Southampton has been to build upon these structures by adding semantics, bringing chemistry to and into the Semantic Web (and perhaps to an even more general concept of a Semiotic Web when disciplinary and cultural contexts apply[26]).

4 The Representation of Chemical Information

Firstly we briefly review the state of chemical information on the World Wide Web as it was and has developed over the last decade and a half. Much of this work was the inspiration of the chemical informatics groups at the University of Cambridge and Imperial College London, and is perhaps succinctly expressed in the existence of CML.[24,25] The initial scepticism about CML, and issues with the stability of the implementations of the associated tools, led to a delay in the adoption of CML as a standard for the interchange of chemical structures. This position has radically changed in the last 10 years and now almost all chemical tools can export structures as CML.[27] We therefore initiated work on mapping chemistry onto or into the Semantic Web and looked at the representation of chemical data and structures using RDF.

However, we also had a wider view in mind when mapping chemical data into and onto the Semantic Web. We wanted not simply to enable or facilitate data integration, a still vital task, but also to integrate data and process by using a common "language" to describe both, using the subject-predicate-object expressions of the Resource Description Framework (RDF) language. The importance of this viewpoint is particularly relevant in the following section, where we consider the development of electronic or digital research notebooks.

Our work focussed on the creation of a schema using RDFS to describe chemical data together with its provenance. This was our earliest contact with what was to become a major theme: the provenance of data and processes. The associated metadata ensured that the provenance was *as fully described as practically possible*, reflecting the need to be pragmatic rather than a purist to establish that the systems were acceptable.

The RDF schema (RDFS) and associated software (triple stores and interfaces) has been described in chemistry and computer science papers.[8–10] This was our first attempt to bring chemical data to the Semantic Web and met with all the issues about the nature of the object being described. This work directly contributed to the extensive research that generated the ChemAxiom ontology[28] and then subsequently the ChemInf ontology.[29] In principle these ontologies now allow most chemical attributes to be described in a consistent manner within the Semantic Web alongside biological interest.[30,31] Wider representations of people, places and topics are also feasible.[32]

During this period we also undertook research into information modelling techniques, including, but not confined to, machine-readable descriptions of quantities and units, and property prediction using QSPR models.[33] Some research into the semantic modelling of quantities and units had been conducted in the early days of the adoption of XML representations of data, but seemed to stall, as did several other semantic projects.

All practical scientific investigations rely on a well-understood framework of quantities and units. While it might seem that the issues would be well understood, closer investigation reveals that many potential pitfalls still exist. Our perspective on the representation of quantities and units within the Semantic Web framework was driven by two ideals. One of us (JGF) is involved in the IUPAC project on the terminology and symbols for physical chemistry (IUPAC Green Book[34]), so the computer representation of

this material was an obvious need. In a more pragmatic approach we considered that if text is marked up with semantics and, for example, the indication is that a particular passage is about a pressure, it should be possible to check that the quantities, symbols, units are consistent with the text. Historically, several high profile failures attributed to conversion issues, such as the Mars Climate Orbiter in 1999,[35] have led to a compelling general need for the semantics of conversions involving units to be clear.

To achieve that aim, the semantics of the quantities and units have to be represented in a schema form, for which we chose RDF. The flexibility and extensibility of RDF enable the capture of the many details required for chemical data.

The failures we refer to above do not involve "unit conversions" as such: what is really occurring is the expression of a measure of some property in different units, which is a quantity conversion. For many quantities this is not an issue as the process is multiplicative. For example, we convert a length expressed in imperial units to metric units by simply scaling the unit. However, for temperature the conversion between °F and °C is not a simple scaling, as we have to adjust the origin.

Kieron Taylor's doctoral research at Southampton addressed the issues of quantity conversion as part of a programme of applying e-Science techniques to chemical property prediction.[33] Recently, during the preparation of this paper we became aware of the recent work by NASA and TopQuadrant, Inc.[36] that looks to have goals similar to those reported in Taylor's thesis.

In a paper presented to the 2006 UK e-Science All Hands Meeting, Taylor et al. set out the case for machine-readable unit descriptions and proposed an RDF schema for representing quantities, units, and unit conversions, where a quantity represents a dimension expressed in terms of a unit.[37] Subsequently, Taylor, Gibbins, and Frey extended this work to include values and developed a 'Units Ontology' using OWL (Web Ontology Language).[38] A value is represented by a resource with a rdf:value property bearing the numerical value, expressed as an XML Schema decimal datatype (xsd:decimal), while the unit and (optional) scaling factor prefix are indicated by the properties has-unit and has-prefix respectively. The 'Units Ontology' adopted seven base physical quantities, being the set comprising the ISO base quantities. They implemented the 'Units Ontology' in a prototype program, called "uniterator", that accepted RDF files containing numerical values and units, and requests to convert to other sets of units. The program reduces the input units to SI base units by expanding any derived SI quantities, and performing all necessary conversions to SI base units. The prototype had limitations, for example, being unable to identify the use of units out of context, but demonstrated nevertheless that it is feasible to make scientific units machine-understandable in a manageable manner consistent with best practice on the Semantic Web.

5 From Smart Tea to Blog^3

5.1 Electronic Laboratory Notebooks (ELNs)

The work on ELNs overlaps two of our main themes, that of the representation of chemical information and the changing nature of laboratories and is potentially a key provider of information to repositories. It plays heavily in both the areas of human and computer interaction and that of sharing of information (use and reuse). Three of our projects are of note here: Smart Tea, LabTrove and blog^3.

The initial (and indeed on-going) aim was to represent both materials and processes within the Semantic Web framework. Our work with the Smart Tea project was at – or perhaps beyond – the capabilities of the technology (both physical in terms of tablet computers and software in terms of the development of RDF). Accordingly, we concluded that it went beyond what users would accept at the time. That recognition led us to consider a more informal Web 2.0 approach based on a blog, which resulted in the LabTrove project.[14] LabTrove was followed by blog^3, which initiated a fully API-driven approach to the ELN blog. The principles developed for blog^3 are being migrated to the LabTrove system and will provide the basis for the future "mobile first" LabTrove-style ELN.

In this paper we will briefly describe the main principles of Smart Tea and LabTrove from the informatics and data handling points of view while referring the reader to recent papers for the details.

5.2 The Smart Tea Project

This is one area where our themes and topics interact. We were very interested in the potential of ELNs to capture synthetic chemistry. Our aim – somewhat optimistic for the time – was to represent both the chemical data and the process in a similar, common, methodology, and chose RDF as the means to provide this description. We realised that the UK COSHH[39] requirement to have a health and safety plan in place for each experiment undertaken meant that we could create a digital harness in advance of a synthetic procedure, requiring only relatively small changes to the processes followed by a synthetic chemist. This approach not only gave us *metadata in advance*, but also led directly to the concept of a plan and forward-looking provenance, which came to fruition in subsequent work on Planning and Enactment (P&E).[40,41]

The implementation of the desktop- and tablet-based Smart Tea system suffered from being too close to the "bleeding edge" of technology: the tablet systems and interface then available were not ideal, handwriting recognition was poor, and the RDF technology was underdeveloped. Specifically, the latter needed inventions like ORE (Object Reuse and Exchange),[42,43] RDF manifests, and better RDF databases (triple stores). Nevertheless we were able to trial the system and understand where user resist-

ance was an issue. We were not at the time able to carry out a sufficient number of experiments to create enough data to experiment with data integration and search. As we were doing this work before SPARQL (SPARQL Protocol and RDF Query Language)[44] was fully defined, our queries would in any event have been limited.

5.3 LabTrove

The complexities of providing the necessary detail of material and process, especially without the large scale open-access material databases now available (e.g., ChemSpider) suggested that we would be better bringing users into the digital world via a much gentler route. Most of the younger research students in the early 2000's were familiar with blogging (then the only widely used, or at least widely talked-about and viewed, social networking tool). This suggested that they would be comfortable with a blog-style approach that came with the necessary features for a laboratory notebook and enabled (but did not insist on) the use of metadata. The resulting LabTrove software[14] has now been used by several research groups,[44] of which we would highlight the open notebook science consortia run by Mat Todd.[45,46]

5.4 Blog^3

In this section we describe the blog^3 software in more detail, as this has not been published elsewhere. Blogging had moved on as a communication medium since we developed LabTrove as a blog-based system and more dynamic user interfaces had come into common use, for example, the adoption of the AJAX group of technologies[47] and the pervasive espousal of Facebook. It therefore became clear that we should adopt a more modern approach for the future to enable greater use of mobile technology and a wider variety of user interfaces and platforms. Consequently, we created an API-driven core for a digital laboratory notebook, which could be accessed from a browser, on either a desktop or a mobile device, and by agents running laboratory equipment. This core was implemented using the Ruby on Rails framework[48] as the blog^3 system, with role-based access control and semantic capabilities, by facilitating the representation of the content of the notebook using both a relational database and a triple store. Ruby on Rails is database-agnostic. For the demonstrator, we used MySQL. However, the system would work equally well with PostgreSQL, or any other supported relational database engine).

The blog^3 system uses the Ruby library to implement the OAUTH 2.0 protocol.[49] blog^3 authorises all actions that affect database resources, and access to private blogs requires the appropriate permission, thus giving users complete control over who can access their resources. All blogs are either public (visible to all users) or private (visible to selected users, using a white-list). blog^3 includes a user-configurable, role-based access control system, the default roles being: Reader, Author, and Editor. Default settings, with common roles, are provided for new blogs. However, the defaults can be overridden, and the blog owner can define new roles. blog^3 also ensures that users see only what they are allowed to access; users will not even be aware of the existence of a resource to which they do not have access.

The original blog^3 source code and documentation was held on RubyForge. As this repository has now closed we are in the process of relocating the source code and API Developer's Guide. Further information is available from reference.[50]

During the development of the blog^3 system, at one stage it marked up its posts as XHTML extended with RDFa attributes, thus enabling other applications to extract the semantics of the entry: posts were therefore simultaneously both human- and machine-readable. RDFa has not been as universally used as we had anticipated and the current version makes it as easy to obtain an RDF version of the post as it is to obtain the HTML. blog^3 uses the RDF.rb library for the Ruby programming language to convert relational database records into RDF on demand either by adding .rdf to the post URL or by content negotiation.[51] Similarly other formats are available. The RDF can then be imported into a triple store to facilitate graph-based queries. We had considered an automatic synchronization between the blog^3 relational database and a triplestore as is done with MyExperiment. However the problem with this approach, using one triplestore containing all the relationships pertinent to the entire blog^3 instance, is that the graphical queries would then have access to all the information in blog^3 and thus enable inferences using information to which a given user should not have access. For example, if two users are using the same synthetic procedure, one user might discover that the other user is synthesizing commercially sensitive substances.

Providing a way to download the information to which a user has access, as RDF, allows that user to create a triplestore with only the information they should have access to. We believe this is simpler than trying to place access controls on the inferences in a single triplestore.

The system is "RESTful", with each entity-type in blog^3 consisting of two parts: a machine-readable API with CRUD (create, read, update and destroy) and other operations, like "publish" for posts, using a variety of content types, including XML, JSON and RDF) and a human-readable HTML interface, accessed through the Web browser. Clients communicate with the machine-readable APIs using REST. blog^3 provides Atom feeds (related to the traditional RSS) for posts in a blog, comments in a blog, and comments for each individual post. Chemists can enrich the content of their posts by embedding depictions of chemical structures using CML. As the system recognises a "chemical formula" (SMILES, InChI, CML) it can automatically extract the content and resolve the chemical structure to a record in the

ChemSpider database.[52] All the information about the chemical entity is embedded in a post copying this transfers all this information, not just the visible figure. Equation objects are treated similarly. The various plugins to facilitate blog[3] were implemented via the TinyMCE editor.[53]

6 Digital Repositories and Chemistry

The succession of projects related to digital repositories evolved from the CombeChem project, with particular focus on data provenance and the reuse of primary data not only for research but also for teaching and learning. Beginning in 2003, Jisc funded the eBank UK programme to investigate a range of issues associated with the management of research data: discover, access, use and reuse, provenance, and metadata schema for datasets. Jisc support continued in subsequent years, particularly with the series of Programmes: Digital Repositories,[54] Repositories and Preservation,[55] and Managing Research Data.[56]

We describe the projects to which Southampton University were leading contributors in more detail in the next section, following which we examine the knowledge gained with regard to the data lifecycle, the challenges and requirements associated with managing data repositories, and the roles and responsibilities of researchers and the wider community.

Digital repositories are a vital element in the research data lifecycle: the processes and workflows tend to be cyclical: archiving, accessing, and using both primary and derived data. Eventually such data could be used for teaching and learning.[57] The eBank project investigated "*linking from primary data to other research outputs within the scholarly knowledge cycle*".[58] One of the challenges identified during the eBank project was that, while institutions were developing their document repositories, there was "*little evidence that institutions are examining the curation and preservation of primary data within their Faculties, Schools and Departments*."

The culture of e-Science rests upon collaboration, sharing, and interoperation, features that have flourished over the ten years of this review. Nevertheless, at around the mid-point, the final report of the eCrystals Federation, which originated from the eBank Project, noted, inter alia: "*Advocacy programmes will be essential to assist with populating the data repositories, since there is no established culture of sharing data within the chemistry domain.*"

Metadata is a fundamental part of the data lifecycle and is essential for effective sharing and interoperability: "For a repository to be interoperable with other repositories, via an integrated research infrastructure, and to enable a harvesting process by third party services, it must publish its metadata according to a strictly controlled schema".[58]

In reviewing the Southampton contributions to the field of Repositories, we begin with an overview of four projects to which we were leading contributors, then identify a selection of specific findings regarding the data lifecycle, the challenges and requirements, and roles and responsibilities.

6.1 Overview of Southampton Projects

The first three projects that we describe were associated with and/or evolved from the Jisc-funded eBank UK project, which was "an interdisciplinary project that ran between 2003 and 2007, across 3 funding phases".[59] Jisc, supported a wide range of projects related to digital repositories, under programmes such as: Semantic Grid and Autonomic Computing; Digital Repositories; Repositories and Preservation.

The University of Southampton is the primary location for the National Crystallography Service (NCS), which is currently one of the mid-range facilities funded by the Engineering and Physical Sciences Research Council (EPSRC).[16] The NCS has been closely associated with digital repositories, most notably as a partner in eBank, which resulted in leading the eCrystals federation. Under the auspices of the CrystalGrid Network, also funded by the EPSRC,[60] a group of small molecule crystallographers held a workshop to begin developing a life cycle model for crystallographic data. This meeting led to a second workshop that reported findings in the following topic areas:[61]

- Standards-based infrastructure for archival of raw data, processed data and results
- Operational and archival data formats and metadata schema
- Community interest and involvement in the process of developing standards and building infrastructure for data management
- The findings discussed in the workshop report influenced the future research into data repositories at Southampton.

6.2 R4L

The Repository for the Laboratory project (R4L) ran from May 2005 until May 2007, specifically to investigate the requirements of a repository for experimental, laboratory-based science, focusing on support for chemical analysis.

"The laboratory repository is a separate entity from the institutional repository not out of architectural necessity, but in order to emphasise a difference in purpose and to ensure the development of appropriate policies".[62]

The exemplar repository developed by the R4L team was "capable of ingesting, storing, managing and presenting a cross section of some ten different types of data holding arising from different analytical techniques. The ingest processes have been carefully designed, following detailed analysis of laboratory workflows, in order to ensure complete capture of the raw, derived and descriptive data and

thus provide a full provenance trail and support a comprehensive preservation process."

The requirements gathering and preliminary analysis phases of the project identified issues arising from the wide range of file formats that would need to be accommodated. The general lack of standards led to a change in the aims of the project, thus shifting the emphasis to producing a proof-of-concept demonstrator supporting a limited range of file formats. The overall intention remained "*to enable the scientific reporting process to keep up with the speed of modern scientific analysis and to improve the accuracy, quality and reusability of scientific reports*."

Although the primary output of the project was the exemplar repository, R4L also conducted a survey of chemists' use of IT tools and methods in the course of their research.

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

6.3 eCrystals

The eCrystals Federation project ran from November 2007 until January 2009, as a continuation of the eBank UK work. The key issues were explored at a consultation workshop held in October 2006, to which prospective partners and stakeholders were invited.[58] The aspirations of the workshop were:

- Develop a widespread understanding for the role of data repositories in scientific research, learning and dissemination.
- Scope an initial set of minimal requirements for a data repository to underpin the chemistry publication and dissemination processes.
- Bring to light and probe issues surrounding interoperability, preservation, harvesting and aggregation in the data repository environment.
- Produce an initial set of recommendations on schema design for construction of data repositories and data capture at the instrument level.

Discussion groups considered: mechanisms for data capture; federation and interoperability; and the requirements of the wide range of parties with a potential interest in data repositories. The specific interests and concerns of nine stakeholders are presented in full in the comprehensive report: "Scaling Up: Towards a Federation of Crystallography Data Repositories".[58] The report also includes a full analysis of the then state of institutional repositories in the following areas: policy and practice; laboratory workflows; standards and interoperability; metadata schema; semantic issues; data citation; identifiers and linking; federation architecture; rights and licensing; quality; and curation.

The eCrystals project established a federation of institutional crystallography data repositories, comprising a small international group of partner sites and other linked data repositories. The project investigated aggregation issues arising from harvesting metadata from repositories in an international environment, thereby enabling interoperation with subject archives in other countries and with other third party harvesters. The project also made recommendations regarding good practice for preservation in institutional data repositories and evaluated sustainable models for partnership.[63]

An eCrystals institutional repository holds raw, derived, and results data from a crystallography experiment, together with chemical and bibliographic metadata Record metadata complies with Dublin Core standards,[58] which is extended to include some necessary elements not in the core that are required to describe properties of datasets.

The eCrystals Federation model demonstrated the interoperability that is essential for such distributed systems. eCrystals also provided a superb platform for developing and demonstrating the tools and services that could operate not only on individual repositories but also over a federated network of data repositories.[63]

As a conclusion to the project and to begin embedding this approach into the crystallographic community a satellite workshop was held at the IUCrXXI congress (August 2008, Osaka, Japan).[64] This workshop discussed new routes to data publication and started a discussion theme in the community that is still going on to this day.[65]

6.4 KeepIt

The KeepIt project ran from April 2009 until September 2010, with the overall aim of improving the long-term preservation of digital repository content. The project manager, a preservation specialist worked with the managers of four existing repositories, one being eCrystals. The project enabled those managers to formulate practical and achievable preservation plans, using existing and newly developed preservation tools and services with the support of training and advice.

The preliminary phase of the project identified a range of preservation requirements that informed the design and development of a training course, which comprised five modules that are described in the publication "Preserving

repository content: practical tools for repository managers":[66]

- Module 1: Organizational issues, audit, selection and appraisal
- Module 2: Institutional and lifecycle preservation costs
- Module 3: Primer on preservation workflow, formats and characterisation
- Module 4: Putting storage, format management and preservation planning in the repository
- Module 5: Trust, of the repository and of the tools and services it chooses

The managers of the exemplar repositories "*applied at least one of the tools to their own repositories*" and as a result of the course revisited their preservation objectives. Furthermore, the managers and repository staff improved their understanding of the implications of organizing and administering their repository content.

"Additionally, the project helped managers to raise awareness (of the repository as well as digital preservation) among repository users, colleagues and managers and provide tangible evidence to contributors and senior managers that repositories indeed take seriously their responsibility to ensure secure preservation of the content entrusted to them".[66]

From the eCrystals perspective, a very important aspect of the involvement with the KeepIt project was the registration of the CIF and CML file formats, thus enabling preservation services to recognise and understand repository content automatically.

6.5 WebTracks

The WebTracks project[67] ran from August 2010 until November 2011, addressing inter-repository communication rather than repositories themselves. The outcome was a specification for an application-layer protocol, InterCom, enabling communication between digital data repositories of any type.[68]

The project evolved from the earlier CLADDIER[69] and StoreLink projects, in recognition of the benefits of exchanging citation information between repositories, thereby enabling researchers to trace links between data and publications[70] and produced a prototype communicating between partner repositories. The specification describes the InterCom protocol as "*more flexible than StoreLink as it does not specify a fixed format for the metadata ontology and it allows the metadata properties to be defined per link.*"

6.6 Data Lifecycle Issues

The eBank project report noted in 2008:[58]

"Whilst there is a growing body of work relating to institutional policy associated with document repositories, there is as yet, little evidence that institutions are examining the curation and preservation of primary data within their Faculties, Schools and Departments."

The project also identified a number of indicators pertinent to the future implementation of federated repositories for crystallographic data, two of which were particularly relevant to the data lifecycle:[63]

"It is clear that preservation and curation issues will have to be addressed politically by both institutions and the community."

"Advocacy programmes will be essential to assist with populating the data repositories, since there is no established culture of sharing data within the chemistry domain."

Such issues were clearly recognised during the KeepIt project:[66]

"…the eCrystals team knew that whilst it is relatively easy to set up a new repository, it is in populating it with older data that the costs really mount up."

Preservation and curation are fundamentally dependent on metadata and on its capture at the earliest possible stage in the data lifecycle. The importance of curation at source is stressed in our recent paper about data curation issues in the chemical sciences (DCICS). The eBank and eCrystals projects clearly recognised the importance of metadata for interoperability:[58]

"For a repository to be interoperable with other repositories, via an integrated research infrastructure, and to enable a harvesting process by third party services, it must publish its metadata according to a strictly controlled schema."

6.7 Challenges and Requirements

The projects described in this Repositories Section have each identified challenges and requirements in a range of areas.

6.7.1 Standards

The second workshop run the auspices of the CrystalGrid Network came to a number of conclusions with regard to standards:[61]

"These findings relate to archival standards, practices and policies; archival formats and infrastructure; and community organization and mobilization to address the challenges of data management in the crystallography community."

The workshop also concluded that further research was required with regard to standards for referring to data sets from publications and also that standards were required to enable instrument vendors to offer raw data in open formats as well as in their own proprietary formats.

In considering policy and practice, the eBank project observed:[58]

"The RIN Data Stewardship Principles provide an appropriate framework into which institutional data policies can be positioned, however data policies need to be developed

locally and reflect organisational requirements and repository maturity."

With regard to standards for security, the CrystalGrid workshop noted:

"Other important issues include standards for managing identity, privacy and access security at the file, sample and archive level, and mechanisms for creating and using handles used to refer to data sets and components of data sets across the entire crystallographic community."

6.7.2 Services

The need for interoperability introduces its own challenges and requirements, in that file formats must be recognisable, so that preservation services can understand the content of a repository automatically. File formats therefore need to be incorporated into an approved registry, such as PRONOM.[71] Of equivalent importance is the compliance of metadata with strictly controlled standards, thereby enabling harvesting by third party services.

6.7.3 Community and Infrastructure

The CrystalGrid workshop called for "*broad community involvement*" in the process of developing the necessary data management standards, practices, and infrastructure.[61] Another project in which Southampton was a partner, "Infrastructure for Integration in Structural Sciences (I2S2)" made a collection of findings that included:[72]

"A robust data management infrastructure which supports each researcher in capturing, storing, managing and working with all the data generated during an experiment.

Where crystallography data repositories already exist, there is a requirement to develop them into a robust service incorporating curation and preservation functions.

The potential of data for reuse and repurposing could be maximised if standard data formats and encoding schemes, such as XML and RDF, are widely used."

The community also needs to underwrite clarity of ownership of data. The eBank project asserted that "there is a need to categorise roles such as that of 'creator', and to allocate public responsibility for creation of a record".[58]

6.7.4 Costing

Justifications for making and sustaining major investments in repositories and data curation will depend greatly on cost benefit analyses. Jisc funded two studies to understand the long-term preservation costs for research data: the first study reported in 2008; the second in 2010, having conducted a costs data survey and performed data preservation cost modelling exercises for four of the organisations contributing to the study, one being the eCrystals repository.[73]

The study "… identified and analysed collections of long-lived research data and information on associated preservation costs and benefits and provides a larger body of material and evidence against which existing and future research data preservation cost modelling exercises can be tested and validated."

It is widely accepted that preservation and curation are much better performed at the time of the experiment. Curation delayed is less reliable and it is significantly more expensive to recreate data at a later time. If the sample no longer exists, the costs may become prohibitive.

6.7.5 Roles and Responsibilities

A key responsibility of a data repository is for the quality of the data it holds. The Cambridge Crystallographic Data Centre (CCDC)[74] was a stakeholder in the eBank project and acknowledged issues not only with acquiring data for the Crystal Structure Database (CSD) but also with ensuring the accuracy of that data.

At the time of the projects described in this section, data citation was perceived to be a potential issue, in part because data and publications tended to be managed differently in institutional repositories and data archives. This situation has been alleviated by the establishment of DataCite in 2009 as an international body that supports the assignment of persistent Digital Object Identifiers (DOIs) to enable the preservation, citation, discovery, and reuse of data.[75]

Institutional repositories undoubtedly have a role in disseminating experiment data as well as the publications that refer to those results. In 2008, Southampton crystallographers co-organised a workshop under the auspices of the International Union of Crystallography to initiate a debate about the effective and efficient dissemination of the ever-increasing volume of crystallographic raw and results data.[76]

That workshop also considered whether Open Access would assist in discharging the dissemination responsibility. There was a "general consensus that some of these new technologies and approaches can help, especially in the case where there is never going to be any associated journal article." Overall, there was recognition that Open Access "could play a role in several different models."

7 Conclusions

In this review of the evolution of digital chemistry at the University of Southampton we summarize for the first time our work on the interaction of chemical representation on the Semantic Web with the trajectory of the development in Electronic Laboratory Notebooks and link this to the summary of the wider data management projects we have undertaken. We show how the attitudes of researchers to recording and sharing information have been as important as the technical developments that have occurred over the last decade. While we report significant achievements, we

also indicate how much work there is still to be done to bring chemists and chemistry laboratories fully in the digital world.

Acknowledgements

Our views have been formed both as practicing chemists and information researchers and been brought into focus over the last decade with work funded by the RCUK e-Science programme (EPSRC Grant GR/R67729, EP/C008863, EP/E502997, EP/G026238, BBSRC BB/D00652X), the EPSRC National Crystallography Service (Tender RCUK/D/EPSRC/Facilities/XRC/10), the HEFCE and JISC Data Management Programme and the University Modernisation (UMF), and most recently the RCUK Digital Economy Theme as part of the IT as a Utility Network+ funding (EPSRC EP/K003569). These views could not have been honed without considerable interaction with our colleagues in *Chemistry, Computer Science, and Statistics in Southampton* and *the e-Research South Consortium* (EPSRC EP/F05811X), especially UKOLN, OeRC, STFC and the DCC, and our professional society and industrial colleagues at the *Royal Society of Chemistry, Microsoft Research* (*MSR*) and *IBM*.

Conflict of Interest

None declared.

References

[1] *The BBC Microcomputer and me, 30 years down the line*; http://www.bbc.co.uk/news/technology-15969065; accessed January **2015**.
[2] H. C. Davis, W. Hall, I. Heath, G. J. Hill, R. J. Wilkins, *MICROCOSM: An Open Hypermedia Environment for Information Integration*, Monograph (Technical Report) **1992**; http://eprints.soton.ac.uk/250713/; accessed January **2015**.
[3] *Microcosm*; http://www.cs.cf.ac.uk/Dave/ISE_Multimedia/node325.html; accessed January **2015**.
[4] *RCUK Review of e-Science 2009*; http://www.epsrc.ac.uk/newsevents/pubs/rcuk-review-of-e-science-2009-building-a-uk-foundation-for-the-transformative-enhancement-of-research-and-innovation/; accessed January **2015**.
[5] *The CombeChem Project*; http://www.combechem.org/index.php, EPSRC grant GR/R67729/01; accessed January **2015**.
[6] J. G. Frey, D. De Roure, L. Carr, 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* **2002**; http://eprints.ecs.soton.ac.uk/7852/1/index.html; accessed January **2015**.
[7] J. G. Frey, G. V. Hughes, H. R. Mills, M. C. Schraefel, G. M. Smith, D. De Roure, Less is More: Lightweight Ontologies and User Interfaces for Smart Labs, in *Proc. UK e-Science All Hands Meeting, Nottingham 2003*, EPSRC **2004**; http://www.allhands.org.uk/proceedings/papers/187.pdf; accessed January **2015**.
[8] K. R. Taylor, R. J. Gledhill, J. W. Essex, J. G. Frey, S. W. Harris, D. C. De Roure, *J. Chem. Inf. Model.* **2006**, *46*, 939–952.
[9] J. Frey, D. De Roure, K. Taylor, J. Essex, H. Mills, E. Zaluska, CombeChem: A Case Study in Provenance and Annotation Using the Semantic Web, in *IPAW* (Eds: L. Moreau, I. Foster), Springer, Heidelberg, **2006**, LNCS 4145, 270–277.
[10] K. Taylor, J. W. Essex, J. G. Frey, H. R. Mills, G. Hughes, E. J. Zaluska, *J. Web Semantics* **2006**, 4; http://eprints.ecs.soton.ac.uk/12505/; accessed January **2015**.
[11] J. G. Frey, *Drug Discov Today* **2009**, DOI:10.1016/j.drudis.2009.03.007.
[12] M. Borkum, C. Lagoze, J. Frey, S. Coles, A Semantic eScience Platform for Chemistry, *IEEE 6th Int. Conf. on e-Science* **2010**, 316–323.
[13] J. G. Frey, C. L. Bird, Cheminformatics and the semantic web: adding value with linked data and enhanced provenance, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, DOI: 10.1002/wcms.1127.
[14] A. J. Milsted, J. R. Hale, J. G. Frey, C. Neylon, *PLoS ONE* **2013**, *8*(7), e67460; DOI: 10.1371/journal.pone.0067460.
[15] C. L. Bird, C. Willoughby, J. G. Frey, *Chem. Soc. Rev.* **2013**, DOI: 10.1039/C3CS60122F.
[16] M. B. Hursthouse, S. J. Coles, *Crystallogr. Rev.* **2014**, *20(2)*, 117–154.
[17] *UKOLN Archive*; http://www.ukoln.ac.uk/; accessed January **2015**.
[18] C. L. Bird, C. Willoughby, S. J. Coles, J. G. Frey, *Inform. Stand. Quarterly*, **2013**, *25(3)*, 4–12.
[19] J. G. Frey, C. L. Bird, *Expert Opinion Drug Discov.* **2011**, *6 (9)*, 885–895.
[20] Jisc and British Library, Researchers of Tomorrow: the Research Behaviour of Generation Y doctoral students, *Report* **2012**.
[21] Research Councils UK Digital Economy Theme; http://www.rcuk.ac.uk/research/xrcprogrammes/digital/; accessed January **2015**.
[22] *Digital by Default Service Standard*; https://www.gov.uk/service-manual/digital-by-default; accessed January **2015**.
[23] H. S. Rzepa, P. Murray-Rust, B. J. Whitaker, *J. Chem. Inf. Comput. Sci.*, **1998**, *38 (6)*, 976–982.
[24] P. Murray-Rust, H. S. Rzepa, *J. Chem. Inf. Comput. Sci.*, **1999**, *39(6)*, 928–942.
[25] P. Murray-Rust, H. S. Rzepa, *J. Cheminform.* **2011**, *3*, 44.
[26] J. G. Frey, in *Proc. 1st Int. Workshop on Understanding Web Evolution (WebEvolve2008)*, **2008**, pp. 5–7.
[27] *Chemical Markup Language*; http://en.wikipedia.org/wiki/Chemical_Markup_Language; accessed January **2015**.
[28] N. Adams, E. O. Cannon, P. Murray-Rust, in *Int. Conf. on Biomedical Ontology*, Vol. 1, Nature Publishing Group, **2009**, p. 2.
[29] J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, M. Dumontier, *PLoS ONE*. **2011**, *6*; e25513.
[30] E. L. Willighagen, A. Waagmeester, O. Spjuth, P. Ansell, A. J. Williams, V. Tkachenko, J. Hastings, B. Chen, D. J. Wild, *J. Cheminform.* **2013**, *5*.
[31] *ChEMBL documentation*; http://www.ebi.ac.uk/rdf/documentation/chembl; accessed January **2015**.
[32] For example see: http://Schema.org, http://www.foaf-project.org/.
[33] K. R. Taylor, Modernisation of computational chemistry and cheminformatics with e-Science techniques: Applications to chemical property prediction, *PhD Thesis*, University of Southampton, **2007**.
[34] *Quantities, Units and Symbols in Physical Chemistry*, 3rd ed., International Union of Pure and Applied Chemistry; http://www.iupac.org/fileadmin/user_upload/publications/e-resources/ONLINE-IUPAC-GB3–2ndPrinting-Online-Sep2012.pdf; accessed January **2015**.

[35] *Mars Climate Orbiter*; http://mars.jpl.nasa.gov/msp98/orbiter/; accessed January **2015**.
[36] *QUDT – Quantities, Units, Dimensions and Data Types Ontologies*; http://www.qudt.org/; accessed January **2015**.
[37] K. R. Taylor, E. Zaluska, J. G. Frey, Semantic Units for Scientific Data Exchange, in *UK e-Science* All Hands *Meeting*, Nottingham **2006**; http://www.allhands.org.uk/2006/proceedings/papers/614.pdf; accessed January **2015**.
[38] K. Taylor, N. Gibbins, J. Frey, *An Ontology for Scientific Units*, Private Communication, **2008**.
[39] *Control of Substances Hazardous to Health (COSHH)*, Health and Safety Executive; http://www.hse.gov.uk/coshh/; accessed January **2015**.
[40] M. Borkum, C. Lagoze, J. Frey, S. Coles, A Semantic e-Science Platform for Chemistry, *IEEE 6th Int. Conf on e-Science*, **2010**, 316–323.
[41] J. G. Frey, M. I. Borkum, C. Lagoze, S. J. Coles, *Using the oreChem Experiments Ontology: Planning and Enacting Chemistry*, American Chemical Society 240th Meeting, **2010**, abstract available from: http://bulletin.acscinf.org/node/224; accessed January **2015**.
[42] *Open Archives Initiative: Object Reuse and Exchange*; http://www.openarchives.org/ore/; accessed January 2015.
[43] *SPARQL Query Language for RDF*, W3C Recommendation **2008**; http://www.w3.org/TR/rdf-sparql-query/; accessed January **2015**.
[44] K. A. Badiola, et al. (32 authors), *Chem. Sci. Adv. art*. **2014**, DOI: 10.1039/C4SC02128B.
[45] K. A. Badiola, D. H. Quan, J. A. Triccas, M. H. Todd, *PLoS ONE* **2014**, DOI: 10.1371/journal.pone.0111782.
[46] Open Source Malaria, Looking for New Medicines, http://opensourcemalaria.org; accessed January 2015.
[47] J. J. Garrett, *Ajax: A New Approach to Web Applications*; http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications/; accessed January **2015**.
[48] *Ruby on Rails*; http://rubyonrails.org/http://rubyonrails.org/; accessed January **2015**.
[49] *Web Authorization Protocol* (*oauth*); http://datatracker.ietf.org/wg/oauth/documents/http://datatracker.ietf.org/wg/oauth/documents/; accessed January **2015**.
[50] M. I. Borkum and J. G. Frey, *Private Communication*.
[51] *Content Negotiation*; http://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html; accessed January **2015**.
[52] *ChemSpider, Search and Share Chemistry*; http://www.chemspider.com/; accessed January **2015**.
[53] *TinyMCE: JavaScript WYSIWYG Editor*; http://www.tinymce.com/; accessed January 2015,
[54] *Digital repositories*, Jisc; http://www.tinymce.com/; accessed January **2015**.
[55] *Repository and Preservation*, Jisc; http://www.jisc.ac.uk/rd/projects/repository-and-preservation; accessed January **2015**.
[56] *Managing research data*, Jisc, http://www.jisc.ac.uk/rd/projects/managing-research-data; accessed January **2015**.
[57] S. J. Coles, G. C. Conole, J. G. Frey, E. Lyon, Integrating research data and pedagogy: Experiences of the eBank-UK project demonstrator, *Report prepared for the eBank-UK project*.
[58] L. Lyon, S. Coles, M. Duke, T. Koch, Scaling Up: Towards a Federation of Crystallography Data Repositories, *eBank Phase 3 report*, **2008**.
[59] eBank UK; http://www.ukoln.ac.uk/projects/ebank-uk/; accessed January **2015**.
[60] *The Crystal Grid Network, Engineering and Physical Sciences Research Council (EPSRC)*, http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/C007271/1; accessed January **2015**.
[61] S. J. Coles, J. G. Frey, D. DeRoure, M. Hursthouse, *The CrystalGrid Collaboratory Foundation Workshop, Southampton, 13–17 September*, 2004: a selection of presentations, **2004**, http://eprints.soton.ac.uk/9777/; accessed January **2015**; The CrystalGrid Collaboratory, http://www.crystalgrid.org/; accessed January **2015**.
[62] S. Coles, L. Carr, J. Frey, *R4L: The Repository for the Laboratory, Project final report to the JISC*, **2007**.
[63] S. Coles, *JISC Final Report – eCrystals Federation*, Project final report to the JISC, **2009**.
[64] *IUCrXXI Congress Workshop, New Routes to Crystallographic Data Publication*, **2008**, http://wiki.ecrystals.chem.soton.ac.uk/index.php/IUCrXXI_Congress_Workshop; accessed January **2015**.
[65] *Communicating Crystallography, Chemical Crystallography Group Autumn Meeting* **2014**, http://ccg.crystallography.org.uk/documents/AM_2014_programme.pdf; accessed January 2015.
[66] M. Pickton, D. Morris, S. Meece, S. Coles, S. Hitchcock, Preserving Repository Content: Practical Tools for Repository Managers, *J. Digital Information*, **2011**, *12*, 2; available from: https://journals.tdl.org/jodi/index.php/jodi/article/view/1767/0; accessed January **2015**.
[67] *WebTracks, Web-scale link tracking for research data and publications*, JISC, 2012; http://webtracks.jiscinvolve.org/wp/; accessed January 2015.
[68] S. Crompton, B. Matthews, J. Casson, A. Shaon, M. Borkum, *Intercom: A protocol for link notification, Project specification for the JISC*, **2011**.
[69] *CLADDIER Project*, Jisc, **2007**, http://webarchive.nationalarchives.gov.uk/20140702233839/ttp://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2005/claddier.aspxh; accessed January 2015.
[70] B. Matthews, M. Borkum, S. Coles, A. Duncan, P. Hunter, C. Jones, C. Neylon, A protocol for exchanging scientific citations, *Private communication (S. Coles)*.
[71] The National Archives, *PRONOM Registry*; http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx; accessed January **2015**.
[72] M. Patel and project partners, *Infrastructure for Integration in Structural Sciences (I2S2)*, Project final report to the JISC, **2011**.
[73] N. Beagrie, B. Lavoie, M. Woollard, *Keeping Research Data Safe 2*, Study final report to the JISC, **2011**.
[74] Cambridge Crystallographic Data Centre (CCDC); http://www.ccdc.cam.ac.uk/pages/Home.aspx; accessed January **2015**.
[75] *DataCite*, http://www.datacite.org/; accessed January **2015**.
[76] S. J. Coles, *New Routes to Crystallographic Data Publication*, **2008**, Workshop report – private communication.

Received: January 29, 2015
Accepted: April 16, 2015
Published online: July 20, 2015

