

CombeChem: A Case Study in Provenance and Annotation using the Semantic Web

Jeremy Frey¹, David De Roure², Kieron Taylor¹, Jonathan Essex¹,
Hugo Mills² and Ed Zaluska²

¹ School of Chemistry, University of Southampton, UK
{j.g.frey, krt1, j.w.essex}@soton.ac.uk

² School of Electronics and Computer Science,
University of Southampton, UK
{dder, hrm, ejz}@ecs.soton.ac.uk

Abstract. The CombeChem e-Science project has demonstrated the advantages of using Semantic Web technology, in particular RDF and triplestores, to describe and link diverse and complex chemical information, covering the whole process of the generation of chemical knowledge from inception in the synthetic chemistry laboratory, through analysis of the materials made which generates physical measurements, computations based on this data to develop interpretations, and the subsequent dissemination of the knowledge gained. The project successfully adopted a strategy of capturing semantic annotations 'at source' and establishing schema and ontologies based closely on current operational practice in order to facilitate implementation and adoption. The resulting 'Semantic Data Grid' comprises around 45 million RDF triples across multiple stores.

1 Introduction

This paper reports on our experiences in building a Semantic Web infrastructure for chemical research as part of the CombeChem project funded by the U.K. e-Science programme. We set out to use the available Grid and semantic technology to support the entire chemical research sequence. This typically starts from an experiment producing data, which is then searched for relevant patterns, which lead to results, conclusions and publications and which in turn leads to further experiments. All progress depends on individual scientists building on the results already produced by others.

While in the past this process has served the science community well, it is now in danger of paralysis from the sheer quantity of data being produced [1]. We considered it essential that semantic support be introduced at every stage in the process to facilitate automated mechanisms for research support, providing an infrastructure where complex applications and services can be deployed with minimal manual intervention [2,3]. This level of automation is required to deal with the increasing rate at which scientific data is being generated and needs to be processed if the integration of the data and its transformation into information and knowledge is to keep pace with

the generation of the data. We describe our solution as a “Semantic Data Grid” as considered from a Grid perspective in [4].

Our architecture has been designed to enable effective capture of both data and metadata at the earliest opportunity during the scientific investigation. [5] Once captured the material is maintained and organized as it traverses the virtual organisation that represents the whole Chemistry community involved in converting a new piece of data into accepted chemical facts and knowledge. Our methodology has been to draw as far as possible on established chemistry practices and then augment them. Our principle is that by constructing schema and ontologies based on current operational practice we have a solution that is known to work, and we regard this as an essential first step to facilitate deployment and adoption.

In section 2 we explain our use of Semantic Web, leading us to the Semantic Data Grid. We then go on to explain our implementation and experiences in Section 3, illustrate the system by focusing on two parts – the “Smart Lab” and the large store of chemical descriptions. The conclusions and future work form Section 4.

2. Semantic Web Approach

To enter this semantically-described world we adopted a policy that we call “annotation at source”, recognising that the digital world necessary to implement our vision must start as soon as possible in the information chain. For example, the CombeChem project set out to support the relatively small-scale needs of everyday scientists in recording experiments. This resulted in an interest in the Electronic Laboratory Notebook (ELN) research area as part of the overall concept to provide effective semantic support for experimental and computational science. Here the “annotation” starts in the safety plan before the experiment has even begun.

We set out to provide comprehensive semantic support for the whole spectrum of chemistry research in as transparent a fashion as possible, adopting RDF to capture the semantic content and describe the scientific data. A fundamental part of this strategy was to study established practice in the field and to introduce as few changes to normal everyday working practice as possible. This was part of the objective of capturing as much metadata at source (i.e. as it is generated), completely automatically. We adopted the additional premise that it would be impossible to predict in advance the way that data would be accessed and used, hence flexibility of use was a fundamental objective. This led directly to the requirement that the information infrastructure to hold this data and metadata should be as general as possible. We adopted the InChI (a character string that uniquely describes an organic molecule based on its structure) as our shared identifier [6].

In summary our design approach adopted four principles:

1. Grounding in established operational practice – our starting point is to study chemists at work;
2. Capturing a rich set of associations between all types of things, expressed pervasively in RDF and hence explicitly addressing the sharing of identifiers;
3. Metadata capture should be automated as far as possible – our goal is augmentation not disruption;

4. Information will be reused in both anticipated and unanticipated ways.

Other approaches to the problem include the World Wide Molecular Matrix [7], while the Collaboratory for Multi-Scale Chemical Science [8] is developing an informatics-based approach to synthesising multi-scale information to create knowledge.

3. Implementation

3.1 Smart Lab

The back-end of the Smart Lab [9] system consists of a database which stores the details of the experiments, and presents a query interface for interrogating the data-store. The data for the experiment is stored in an RDF triple store based on Jena [10], and is accessed by the front-end applications through a SOAP-based query interface.

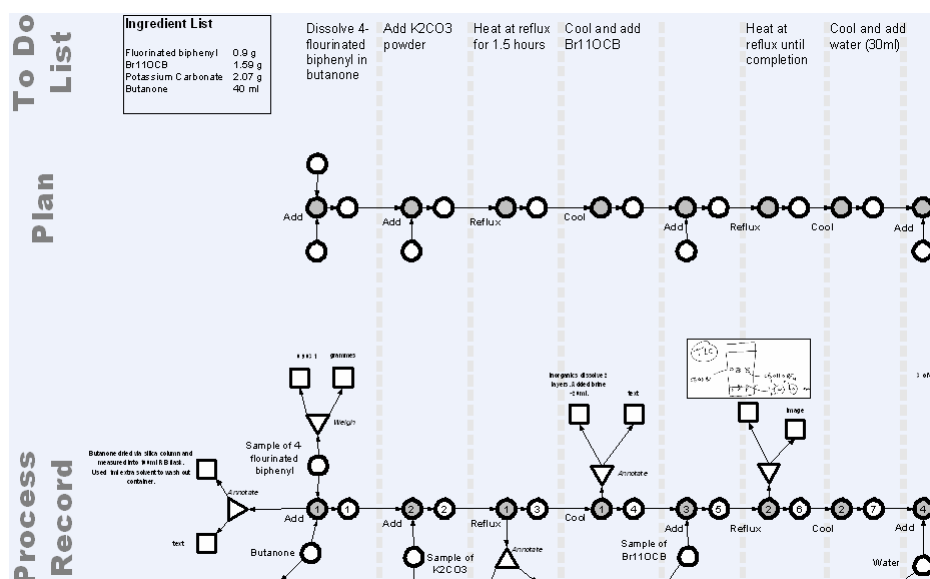


Fig. 1. Fragments of the RDF graph, showing the distinction between Plan and Record.

We have developed three primary applications: a planning tool, which is used to set up the plan and ingredients for the experiment; a weigh-station/liquid-measure application, used for recording the quantities of ingredients actually used, as an example of a measurement device; and a "bench" application, used for making notes and annotations on the plan while performing the experiment. The methodological approach is described in [11]. The latter two applications we have implemented on a Tablet PC, to be carried around in the laboratory. The current prototype planner ap-

plication is implemented as a set of dynamic, form-based web pages. The "smart lab" system is modular. For instance, other measurement devices, such as a digital camera or a formatter for adding mass spectrograph recordings, can also be added to the system in the same way as the weigh-station application.

The functional core of the SmartLab software is the libtea library. This library provides abstraction of the low-level RDF data structures kept in the triple store, and queried and served by the ModelServer. The programming interface (API) for libtea abstracts away some of the details of the underlying RDF structures from the programmer, and presents a simpler interface for the "standard" structures and properties encoded in the base SmartLab ontology. It does not, however, explicitly hide the RDF from the programmer, so if it is necessary to add new features or structures to the schema, it is possible to do so in a flexible manner without having to add support directly to libtea. The libtea API presents a set of objects to the programmer, each object representing a different concept within the RDF structure, and encapsulating a subgraph of one or more triples.

3.2 Triple Store

The digital record from the Smart Lab data then feeds into the scientific data processing. The creation of original data is accompanied by information about the experimental conditions in which it is created. There then follows a chain of processing such as aggregation of experimental data, selection of a particular data subset, statistical analysis, or modelling and simulation. The handling of this information may include explicit annotation of a diagram or editing of a digital image. All manipulation of the data through the chain of processing is effectively an annotation upon it and the provenance is explicit. The annotations are required to be machine processable, and useful for both their anticipated purpose and interoperable to facilitate subsequent unanticipated reuse. This is achieved by RDF being used through the system. At the time of writing there are 45 million RDF triples in the Combechem triplestore. The current target is about ten times this number, representing a very substantial Semantic Web deployment.

Figure 2 shows the schema for the CombeChem data grid, based around chemical properties. Objects are marked as ellipses, the arrows show how predicates link objects together, and rectangles are literal values.

We evaluated several triplestores and adopted 3store [12] because it has good scaling properties. Additionally it is easily batch or perl scriptable, supports RDFS, and it can use RDBMS tools for maintenance of data (e.g. backups and migration) as all application state is held in the database. 3store supports the SPARQL query language.

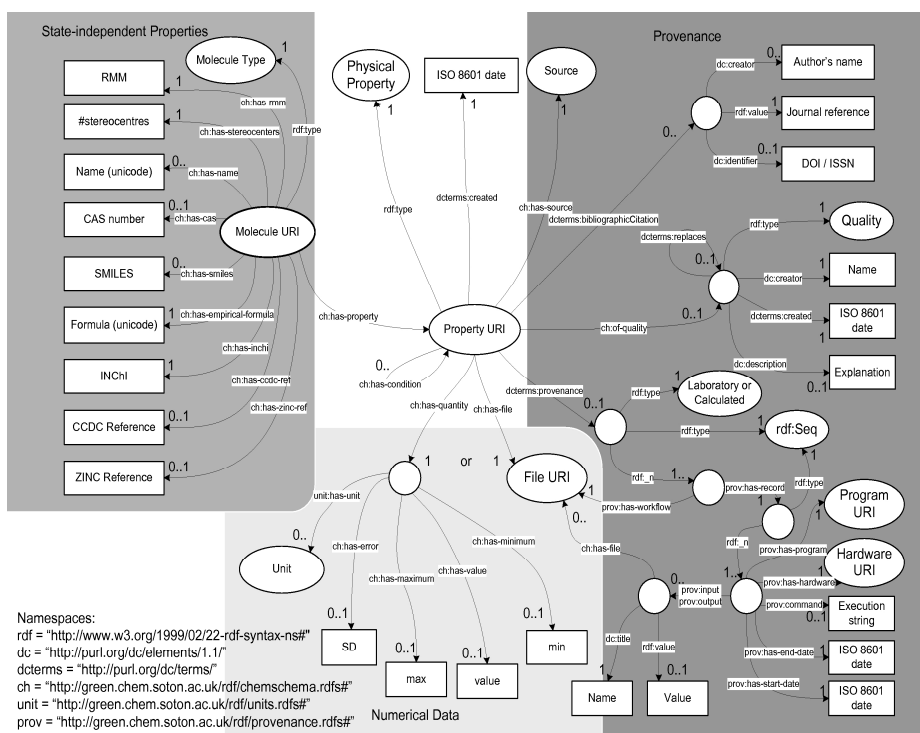


Fig. 2. The schema for the CombeChem Semantic Data Grid

We have succeeded in feeding approximately 80 million triples into the triplestore. Queries remain responsive, but data import performance has begun to degrade, i.e. reassertions of the RDF schema are taking a long time. Write performance on large stores is known to be a challenging issue and we can envisage how a single large triplestore with frequent insertions would be unable to cope with potential demand. 80 million triples equates to a reasonably-sized chemical dataset, but could easily be doubled or trebled when populating with computed properties. Hence we are now contemplating alternative ways of partitioning and maintaining the triples across multiple stores. Progress is also now being made in linking the RDF structures for the molecular properties and those describing the experiments from the ELN, tied together by the molecules URI.

Figure 3 illustrates one interface to this data, developed in the eBank project [13]. The information contained within an entry in this archive is all the underlying data generated during the course of a structure determination from a single crystal x-ray diffraction experiment. An individual entry consists of three parts: core bibliographic data, such as authors, affiliation and a number of chemical identifiers; data collection parameters that allow the reader to assess at a glance certain aspects of the crystallographic dataset; files available for download (visualisations of the raw data, the raw

data itself, experimental conditions, outputs from stages of the structure determination, the final structural result and the validation report of the derived structure).

4. Conclusions and Future Work

RDF has been shown to be an effective method for capturing highly detailed chemical data, allowing it to be indexed in a persistent triplestore such that it can be searched and data-mined in useful ways. The triplestore has now reached a viable state with further addition of chemical properties as an ongoing process. We are now beginning to develop automated calculations using the many available structures and to store the results alongside all the details of the computations that produced them. Beyond that we can achieve high-throughput data processing and begin to develop new models based on those computations.

2-Benzyl-4',6'-bis(trifluoromethyl)-6'-hydroxy-2',3',7',7a'-tetrahydro-6'H-spiro-(furan-3,7'-(furo(3,2-c)pyran))

S. J. Coles, J. M. Mellor, A. H. El-Sagheer, E. E. -D. M. Salem and R. N. Metwally.

University of Southampton

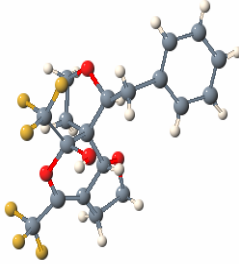
C₁₉H₁₈F₆O₄

CCDC Code: WOSSUS
ICHI Code: C19H18F6O4,20-17(21,22)13-12-6H2-8H2-28-15H(12)16(18(26H,29-13)19(23,24)25)7H2-9H2-27-14H(16)10H2-11-4H-2H-1H-3H-5H-11
([google for ichi](#))

Compound Class: Organic
Keywords: additions; Grignard reagents; trifluoromethylketones
Creation Date: 13 September 1999
Deposited By: [Susanne L. Huth](#)
Deposited On: 30 July 2004

Data collection parameters

Chemical formula	C19 H18 F6 O4
------------------	---------------



Available Files

Final Result	
99sot029_data/99sot029.CIF	14k
99sot029_data/99sot029.cml	7k

Fig. 3. The eCrystals interface.

The ELN proves to be an excellent model system for the meeting of the semantic chemical grid descriptions of materials and services with the pervasive environment needed to capture the information within the source laboratory. We are currently conducting more investigations in the use of the smart lab systems in an active synthetic organic chemistry laboratory looking at the use and re-use of information captured using our systems. The studies will soon be extended to investigate the pervasive aspects of the grid looking at the use of handheld systems (e.g. PDA, tablets) systems vs. distributed computers in different positions within the laboratory.

More work is required to build up the chemical ontology to the level comparable with the XML structure provided by CML. In the INCHI we have a computable URI for organic molecules, but this leaves large areas of compounds (inorganic, mixtures, materials etc.) without adequate URIs of this form. We have seen in the need to provide an RDF structure for units that the process of describing a piece of scientific data, with all the necessary descriptions and provenance, propagates requirements out in an extensive net, meeting though with other domains, where we can link up with other semantic descriptions. An area which demands rapid attention because of the importance of unhindered and accurate information flow between different knowledge domains, is the need to integrate the chemical ontology with for example the LSI identifiers is a part of the whole processes of linking Bio- and Chemical Informatics, to aid for example drug modelling (sample and model selection in QSAR) and on to the larger spatial scale of the environment, all of which are major purposes of our current investigations.

We have also commenced an exploration of capturing scientific discourse within the Data Grid, fully linked in following the publication@source approach. This includes materials from meetings and videoconferences, and is achieved through the use of meeting support tools which capture semantic annotation following a similar approach to the smart lab. [14]

The importance of provenance cannot be overstated – not just with respect to understanding where information has come from, but in understanding how to interpret it. An item of information is rendered almost useless if the details of the provenance are not known (in practice this leads to experiments being unnecessarily repeated).

Our principle of pragmatism has brought us a long way – this is a significant piece of the Semantic Web. In a sense, we are now ready to begin! We have benefited from flexible associations and from the sharing of identifiers, and from graph queries and chaining in the triplestore. Much of the power of the Semantic Web that comes from ontologies is yet to be explored. Also at this level we see opportunities for use of rules as these solutions emerge within the Semantic Web stack.

Acknowledgements

The work in this paper was partially supported by the UK e-Science CombeChem project (GR/R67729/01), the Advanced Knowledge Technologies IRC (GR/N15764/01), CoAKTinG (GR/R85143/01) and Semantic Media (EP/C010078/1). eBank is funded by JISC.

References

1. Hey and Trefethen, Cyberinfrastructure for e-Science, *Science* 2005 308: 817-821.
2. C. A. Goble, D. De Roure, N. R. Shadbolt, and A. A. A. Fernandes, "Enhancing Services and Applications with Knowledge and Semantics," in *The Grid 2: Blueprint for a New Computing Infrastructure*, I. Foster and C. Kesselman, Eds.: Morgan-Kaufmann, 2004, pp. 431-458.
3. De Roure, D. Jennings, N.R. Shadbolt, N.R. The Semantic Grid: Past, Present, and Future, *Proceedings of the IEEE*, Volume 93, Issue 3, March 2005, Pages 669-681.
4. Taylor, K., Gledhill, R., Essex, J.W., Frey, J.G., Harris, S.W. and De Roure, D. "A Semantic Datagrid for Combinatorial Chemistry", *Proceedings of IEEE Grid Computing Workshop at SC05, IEEE, Seattle, WA. November 2005.*
5. J.G.Frey, M.Bradley, J.W.Essex, M.B.Hursthouse, S.M.Lewis, M.M.Luck, L.Moreau, D.De Roure, M.Surridge and A.Welsh, 'Combinatorial Chemistry and the Grid', published in 'Grid Computing: Making the Global Infrastructure a Reality', edited by F.Berman, G.Fox and T.Hey, Wiley
6. InChI International Chemical Identifier <http://www.iupac.org/inchi/>
7. P. Murray-Rust, "The World Wide Molecular Matrix - a peer-to-peer XML repository for molecules and properties," presented at EuroWeb2002, Oxford, UK, 2002.
8. Collaboratory for Multi-Scale Chemical Science (CMCS) <http://cmcs.org/>
9. G.Hughes, H.Mills, D.De Roure, J.G.Frey, L.Moreau, m.c.schraefel, G.Smith and E.Zaluska, 'The semantic smart laboratory: A system for supporting the chemical e-Scientist', *Org. Biomol. Chem.* Vol. 2, No. 22, pp3284-3293, 2004.
10. Jena – A Semantic Web Framework for Java, <http://jena.sourceforge.net/>
11. m.c.schraefel, G.Hughes, H.Mills, G.Smith, T.Payne and J.Frey, 'Breaking the Book: Translating the Chemistry Lab Book to a Pervasive Computing Environment', published in *Proceedings of the Conference on Human Factors (CHI)*, 2004.
12. Harris, S and Gibbins, N.3store: Efficient Bulk RDF Storage. In *Proceedings of the First International Workshop on Practical and Scalable Semantic Web Systems (PSSS2003)*, Sanibel Island, Florida, USA.
13. M. Duke, M. Day, R. Heery, L.A. Carr and S.J. Coles, "Enhancing access to research data: the challenge of crystallography". *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, Denver, CO, USA. Pages: 46 – 55, 2005
14. Bachler, M. S., Buckingham Shum, S. J., De Roure, D. C., Michaelides, D. T. and Page, K. R. Ontological Mediation of Meeting Structure: Argumentation, Annotation, and Navigation. In *Proceedings of 1st International Workshop on Hypermedia and the Semantic Web (HTSW2003)*, Nottingham, 2003.