Visions of a Semantic Molecular Future

A Symposium and Hackfest celebrating the ideas of Peter Murray-Rust

January 15-17th 2011

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge

Unilever Cambridge
Centre For Molecular Science Informatics

UNIVERSITY OF CAMBRIDGE

This symposium addresses the creativity of the maturing Semantic Web to the unrealized potential of Molecular Science. The world is changing and we are in the middle of many revolutions: Cloud computing; the Semantic Web; the Fourth Paradigm (data-driven science); web democracy; weak AI; pervasive devices; citizen science; Open Knowledge. Technologies can develop in months to a level where individuals and small groups can change the world.

However science is hamstrung by archaic approaches to the publication, redistribution and re-use of information and much of the vision is (just) out of reach. Social, as well as technical, advances are required to realize the full potential. We've asked leading scientists to let their imagination explore the possible and show us how to get there.

This is a starting point for all of us – the potential of working with the virtual world of scientists and citizens, coordinated through organizations such as the Open Knowledge Foundation and continuing connection with the Cambridge academic community makes this one of the central points for my future.

The pages in this document represent vibrant communities of practice which are growing and are offered to the world as contributions to a semantic molecular future.

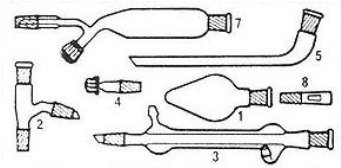
Peter Murray-Rust.

(Molecular backgrounds by Thomas Murray-Rust, Chem, 2000)

“Scientific AI”

An underlying theme of this symposium is that we have reached a stage where we can create “artificial intelligence” (weak AI) in chemical sciences. There are 3 main sub-themes:

- Semantics and interoperable components (software, data, ontologies)



- Open collaborative community

- Open Knowledge (Data, metadata, bibliography, citations, …)

AI will come from linked existing knowledge, physics calculations, search algorithms, machine-learning, enhanced perception and lightweight reasoning.

When these are all available then innovation takes off explosively. Our hackfest (Saturday and Sunday) will explore how existing tools and data can be combined within hours, *e.g.*

- CKAN metadata system and Open patents
- Bibliography and space-time data
- Molecules and Open maps

Why and what are semantic molecules?

Semantics gives “meaning” to information such that machines can make connections. Semantic Maps, timelines, molecules, publications can all be linked (mashups).

Mashups can be made rapidly; David Murray-Rust[1] made the Cambridge Molecules demo in less than a day by mapping molecular models onto a map of Cambridge. YOU can view it on most modern phones with GPS and a compass.

(i, above) OpenStreetMap (semantic) of the Chemistry Dept. and Pantonia;

(ii, right) Virtual molecule of morphine “in M-R’s garden”. We’ve hidden some more round Cambridge...

[1] (Churchill, 1996, Eng)

Extended Group Rapid Presentations

PM-R – The Group

Joe Townsend - CML and Chem4Word

Sam Adams - Chem# embargo management and repository

David Jessop - OSCAR and NLP

Daniel Lowe - OPSIN intelligent Name2Structure

Jens Thomas - Quixote community: semantic compchem

Noel O'Boyle - Blue Obelisk: collaborative Open software

Rufus Pollock - Open Knowledge Foundation

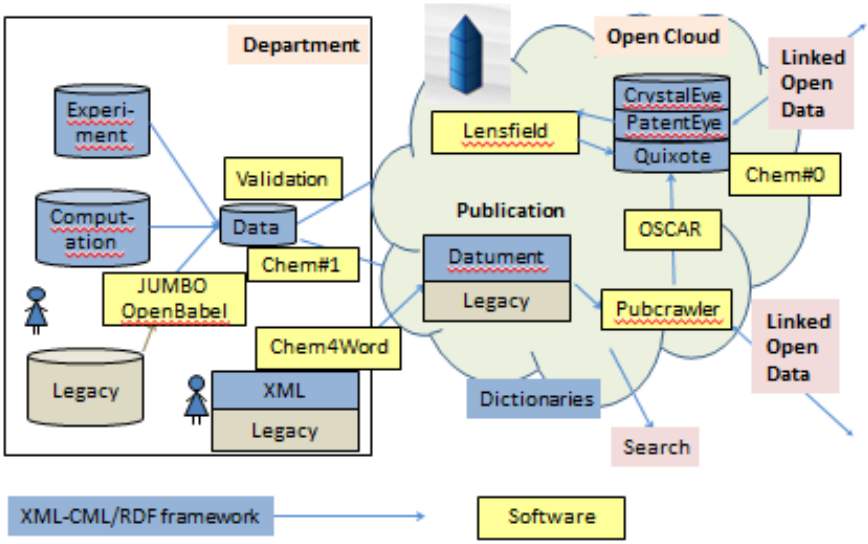
Ben O'Steen - Open Bibliography

Nick Barnes - Climate Code Foundation

Brian Brooks - Ami intelligent fume cupboard and lab

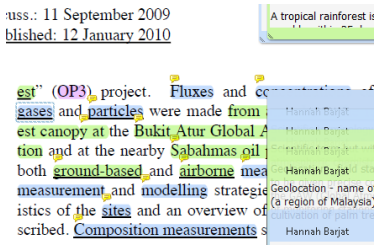
Infrastructure for Chemical Intelligence

Intelligence emerges as components connect to each other, and to humans. Here is part of our WWMM framework covering the creation, editing, publication and conversion of semantic chemical information. It interfaces with Blue Obelisk, Quixote and the wider Linked Open data cloud of Open human knowledge.



We can now start asking (chemistry) questions such as:

- *What vegetation is found near atmospheric terpenes?*
- *What hazardous solvents are used in pharma patents?*

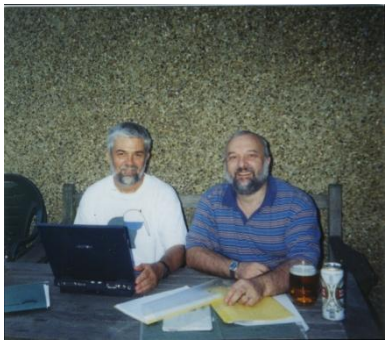


Machines should be able to answer some first-year chemical exam questions so “passing the domain-specific Turing test”.

CML – Chemical Markup Language

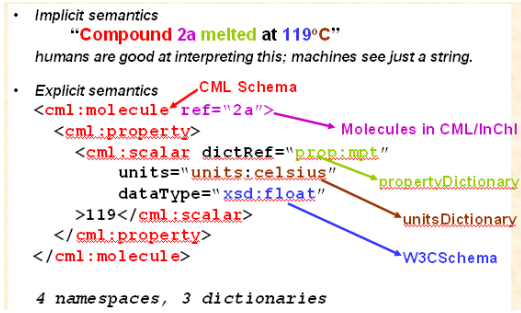
XML is a mainstream approach providing semantics for science, such as MathML, SBML/BIOPAX (biology), GML and KML (geo), SVG (graphics) and NLM-DTD, ODT and OOXML (documents). CML provides support for most chemistry, especially molecules, compounds, reactions, spectra, crystals and compchem.

CML (PM-R and Henry Rzepa) is 15 years old, the de facto XML, accepted by publishers, with > 1 million lines of Open Source support.

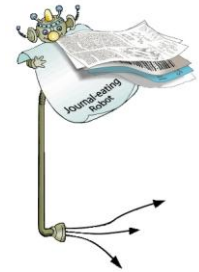


CML can be validated and built into authoring tools (Chem4Word, 250,000 downloads). The infrastructure includes legacy converters, dictionaries, Semantic Web and Linked Open Data.

The community creates meaning through dictionaries (*e.g.* prop:mpt points to a dictionary entry. The Quixote project is creating dictionaries for all major computational chemistry codes.



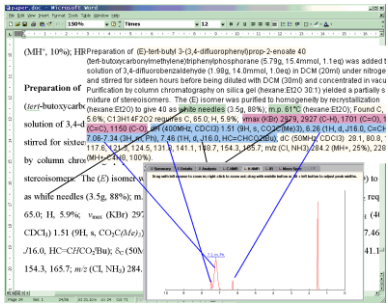
Hannah Barjat and we are creating similar dictionaries and markup for atmospheric chemistry.



OSCAR/OPSIN/ChemicalTagger

8-year project to develop Natural Language Processing to extract chemistry and other physical science from traditional publications.

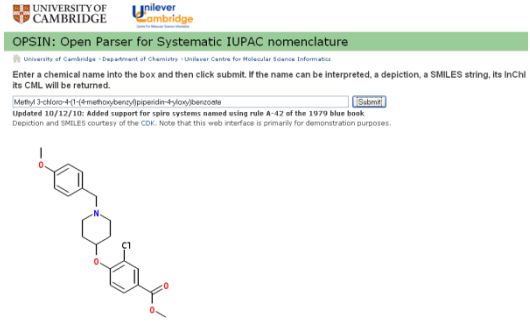
Mature modular *de facto* Open Source; community includes EBI, EPO, NIH, NaCTeM and Eur. Geoscience Union.

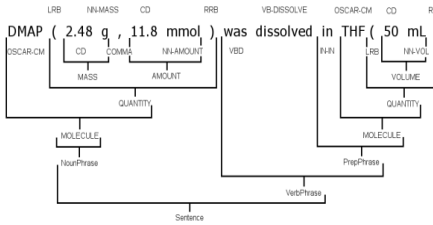


OSCAR (left) can process all PubMed abstracts and >500,000 reactions in patents.

OPSIN (right) >97% on IUPAC chemical names. [OPSIN speech recognition and mobile implementation.]

http://opsin.ch.cam.ac.uk/





ChemicalTagger tree (left). The most advanced chemical NLP software; it extracts context and sentence structure to help clarify meaning

These are examples of “Chinese Room” AI; they apply rules and often do better than humans. Searle would say they “think”. Would you?

Lensfield/Quixote

The Problem with computational chemistry…

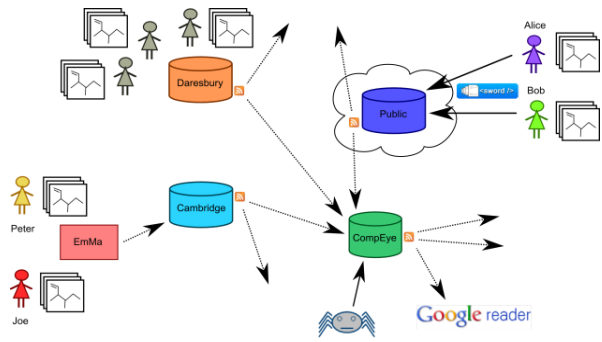
- no standard way to archive or search the data from compchem calculations; valuable data festers on disk.
- there isn't even a standard data format (despite the data being rigorously defined) so each computational chemistry code needs specialised tools to understand its output.
- lots of people have tried to solve this (seemingly trivial) problem, but no one can agree on a solution.

Quixote is…an internet-based, international community of scientists, passionate about open-source and open-data, looking to solve the problem in a bottom-up, pragmatic way at:

http://quixote.wikispot.org/

Our Solution...

A collection of modular, open source tools to convert data from legacy formats, organise and upload the data to local and remote servers, and provide the tools to search and share the data; both on scientists' own computers or on publicly accessible servers.



Blue Obelisk

An internet group dedicated to creating open resources for chemistry. Some of our software is the *de facto* approach in the specific subdomain.

Mantra: Open Data, Open Standards, Open Source

Cheminformatics Toolkits

RDKit
Open-Source Cheminformatics and Machine Learning

Open Babel

CDK

Cinfony

Indigo

http://www.blueobelisk.org/

Ami – a chemist's amanuensis, or, the intelligent fume cupboard

Jane: "Hello Ami!"
Ami: "Good moaning Jane"
J: What's the reaction on the left?" (On the screen, Ami shows the reaction summary, start time, name of the researcher running the reaction, and the safety information)
J: "I want to make an observation!"
A: "Fire away baby"
J: "Green, effervescing, no precipitate"
A: "Thank you" (Ami displays the observation and adds it to the experimental log)

What might life be like in a chemistry laboratory if the surroundings were more aware of what's going on?

Are there ways to help the chemist to access information, prepare their experiments, collect their data, and write-up their observations?

The Ami project is a short project investigating ways to improve the computing environment for chemists at the bench.

The project held a brainstorming session with Real Chemists to see what would help them in their lab activities.

The project then used off-the-shelf hardware and software to prototype possible solutions.

Sensors: RFID, barcodes, laser temperature, accelerometers, ultrasound detectors, gas detectors, vibration detectors

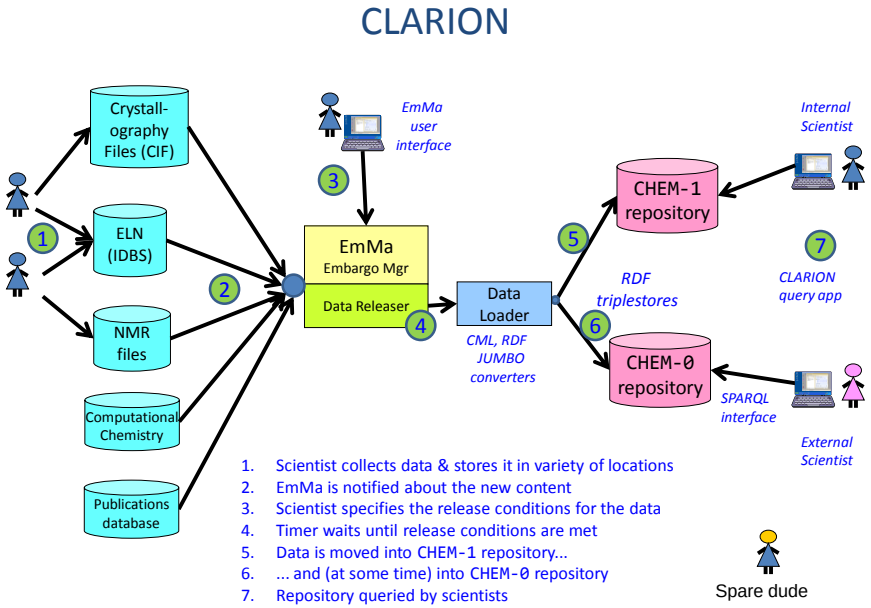
Comms: Mobile, Wii, gestures, infra-red, speech (bi-directional), video capture, Electronic Lab Notebook (eLN), visual projection

Chem#

Huge amounts of chemical data – crystal structures, spectra, experimental reports – are produced every day. The majority of this data is never published.

The JISC-CLARION project aims to address this problem through developing systems to automatically identify data as it is produced, and provide research groups with a simple mechanism for embargoing and publishing their data.

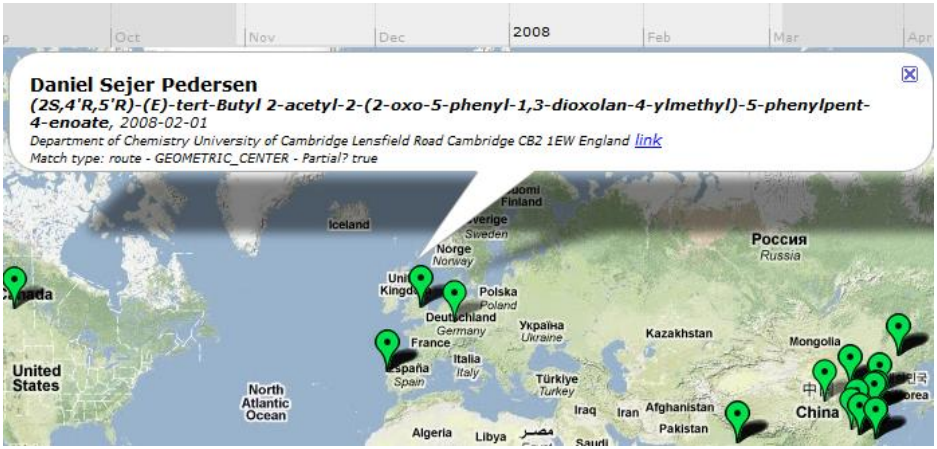
Chem# (*chem-pound)* is the repository tool we are developing to support the publication of semantically enriched linked chemical data. Chem# integrates CML and RDF making both human and machine readable and searchable representations of chemical data available.



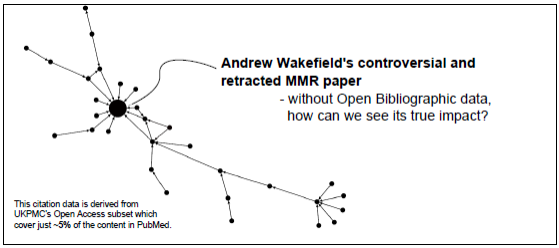
Open Bibliography

Bibliography is the essential infrastructure of scholarship. Key aspects: library collections / management of artifacts; and the scholarly record of (scientific) articles in journals.

In JISCOpenbib we've worked with the British Library to release the Brit. Nat. Bibliography under CC0 and converted it to RDF. With the Int. Union of Crystallography we've analysed Open Bibliographic Data from Acta Cryst:



Progress is rapid; we are extracting 20 million records from UK PubMedCentral. An exciting first result is David Shotton's graph of Bad Science where he tracks the legacy of retracted papers:



Realising the Semantic (Molecular) Future

PM-R's talk consists of two halves: Demos, some of which may have arisen from the Hackfest and will be short (1 minute presentations); and general principles for moving towards Open Scholarship.

Background of AI in Chemistry – it's still achievable

Cambridge virtual molecules (David M-R)

Semantic Atmospheric Chemistry (Hannah Barjat)

ChemicalTagger (Lezan Hawizy); text-mining

Speech recognition for chemistry (Sam Adams); OPSIN

Open Bibliography (Ben O'Steen) and Open Citations

CrystalEye (Nick England)

Launch of Open Bibliographic Principles (Adrian Pohl)

Some fundamentals for Semantic Science:

- Universities must jointly reclaim their scholarship
- Collaboration must be rewarded
- Strategies for creating revenue streams for Open content
- Recognition of the multi-project hacker-expert (vs. scientist)
- Community development of communal dictionaries/ontologies
- Author-side semantic authoring
- Communal domain-specific data-repositories and code
- Textual content universally data-minable
- Tables, graphs and other data must be Open
- Theses must be Open and Repositories coordinated

Open Scholarship

Open Source components, acting as the force of liberation;

Open Data for all data (graphs, images, tables, equations, chemistry);

Open Bibliography for discovery and identification; (Today we launch the Principles of Open Bibliographic Data.

Open Citations for the author's statements of referenced work;

Open Standards for representation and communication.

The tools to do this exist and are Libre; it is technically possible for authors to engage in Open Scholarship and to reclaim their central role in communicating semantic science.

Protocols for best practice for Open Scholarship (*e.g.* Panton Papers with BMC). Examples:

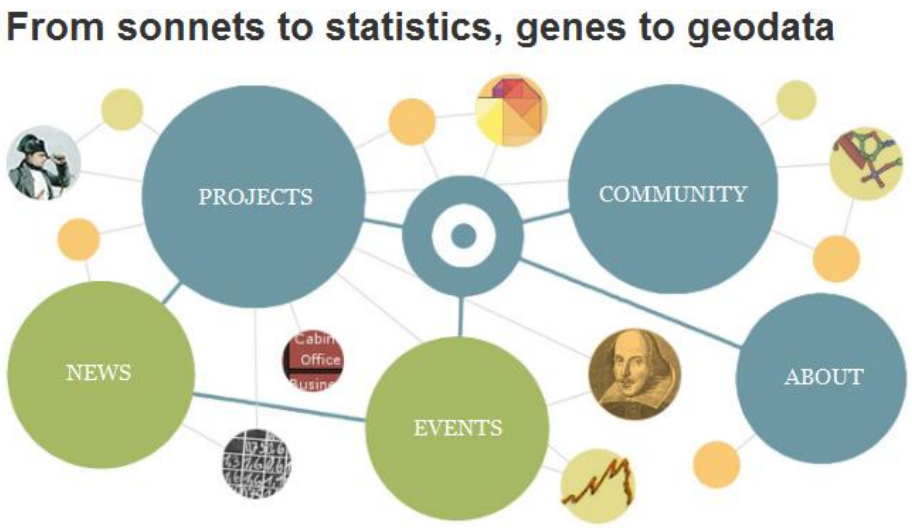
- **What is scientific data?**
- **Does Anyone Own Data?** -- http://okfnpad.org/PPRightsandOwn
- **Mining: Data-and text-mining from scholarly publications** http://okfnpad.org/PPDataTextMining
- **Why should *I* share *my* data?**
- **What is the best way to share my data?**
- **Discipline or Institutional repositories?**

This meeting launches a call for the community to adopt and engage these ideas.

Open Knowledge

Promoting Open Knowledge in a Digital Age

The Open (Knowledge) Definition (OD) sets out principles to define the “open” in open knowledge. The term knowledge is used broadly and it includes all forms of data, content such as music, films or books as well any other types of information.



http://okfn.org/

Panton Principles

OPEN KNOWLEDGE

OPEN DATA

Science is based on building on, reusing and openly criticising the published body of scientific knowledge.

For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open.

By open data in science we mean that it is freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. **To this end data related to published science should be explicitly placed in the public domain.**

Formally, we recommend adopting and acting on the following principles:

1. *When publishing data make an explicit and robust statement of your wishes.*
2. *Use a recognized waiver or license that is appropriate for data.*
3. *If you want your data to be effectively used and added to by others it should be open as defined by the Open Knowledge/Data Definition – in particular non-commercial and other restrictive clauses should not be used.*
4. *Explicit dedication of data underlying published science into the public domain via PDDL or CCZero is strongly recommended...*

Peter Murray-Rust
University of Cambridge (UK)

Cameron Neylon
STFC (UK)

Rufus Pollock
Open Knowledge Foundation and University of Cambridge (UK)

John Wilbanks
Science Commons (USA)

PM-R Group

Adam Thorn (PDRA, 2010), **Alan Tonge** (Project Manager, 2005), **Andrew Walkingshaw** (PDRA, 2005), **Ben O'Steen** (PDRA, 2010), **Brian Brooks** (PDRA, 2009), **Charlotte Bolton** (Coordinator, 2010), **Chris Waudby** (Summer student, 2003), **Dan Hagon** (Summer student, 2008), **Daniel Lowe** (PhD (RCG), 2008), **David Bebb** (Summer student, 2008), **David Jessop** (PhD & PDRA, 2006), **Diana Stewart** (PDRA, 2007), **Ed Cannon** (PDRA, 2007), **Egon Willighagen** (PDRA, 2004, 2010), **Erica Wise** (Summer student, 2003), **Fraser Norton** (Summer student, 2002), **Gemma Holliday** (PhD (JBOM), 2003), **Hannah Barjat** (PDRA, 2010), **James Bell** (Summer student, 2004), **Jason Lee** (Summer student, 2008), **Jim Downing** (Development Manager, 2005), **Joe Townsend** (PhD & PDRA, 2002), **John Aspden** (Developer, 2009), **Jürgen Harter** (PDRA, 2001), **Justin Davies** (Summer student, 2005), **Lee Harper** (Summer student, 2005), **Lezan Hawizy** (PDRA, 2008), **Matt Smith** (Summer student & PDRA, 2010), **Nick Day** (PhD & PDRA, 2005), **Nick England** (PhD & PDRA, 2005), **Nico Adams** (PDRA, 2005), **Peter Corbett** (PDRA, 2005), **Peter Matthews** (Summer student & PDRA, 2010), **Ramin Gorashi** (Summer student, 2005), **Richard Moore** (Summer student, 2006), **Rufus Pollock** (Collaborator, 2010), **Sam Adams** (Summer student & PDRA, 2003), **Shaoming Chen** (Summer student, 2010), **Simon Tyrrell** (PDRA, 2003), **Vanessa de Souza** (Summer student, 2003), **Volker Thome** (PDRA, 2005), **Yong Zhang** (PDRA, 2001)

Thanks.

There are hundreds of people and organizations who have contributed to the projects in which I and my group have been involved – I can't name you all. Special thanks to those beliving in me and helping me get started: the Department of Chemistry, the Unilever Centre, Unilever Research and Bobby Glen for appointing me late in my career path. My colleagues Jonathan, John, Peter and Andreas for many collaborative ventures in research and education. Susan and Emma. Churchill College for so much support. Tim Dickens, Charlotte and the Chemistry COs. Personal support from Roger Leach, Natraj, Jerry Winter, Dominic Tildesley, Ian Stott and others in Unilever. Henry for his unswerving support. Tony Hey, Andy Parker and others in eScience for a great opportunity to develop ideas in the context of an exciting national program. To Martin Dove and many colleagues in the eMinerals and MaterialsGrid projects (Mark Calleja, Toby White, Richard Bruin) who believed in CML and helped make it a reality for computational chemistry. To Ann Copestake, Simone Teufel and Ted Briscoe who helped lay the basis of our Natural Language Processing in chemistry. Markus Kraft for believing. Peter Morgan. Many in JISC and collaborating projects (Brian Mathews, Jeremy Frey, Simon Coles) and now David Shotton. To Tony Hey, Alex Wade, Lee Dirks and others in Microsoft Research for a concerted and personal approach to making semantics a reality through Chem4Word and OREChem. Dan Zaharevitz. To IUCr, RSC, Nature, BMC, PLoS. Peter Sefton in Toowoomba and many other Australians in our thoughts. To the many unfunded people who have helped build a communal semantic resource including the Blue Obelisk (Egon Willighagen, Geoff Hutchison, Noel O'Boyle, Bob Hansen), Quixote (Jens Thomas, Marcus Hanwell, Jorge Estrada, Pablo Echenique). An enormous debt to The Open Knowledge Foundation; Rufus Pollock, Jonathan Gray, Jo Walsh and many more.

Funding bodies and collaborators:

Engineering and Physical Sciences Research Council