# Open Chemistry

***For Nature Horizons***

Peter Murray-Rust
*Unilever Centre for Molecular Sciences Informatics*
*Department of Chemistry*
*University Of Cambridge, CB2 1EW, UK*

*Note: This document is the first draft of an invited article (ca. 2007-09-15), provisionally to be published in January 2008. It is now being processed by the Nature editorial staff, and revision(s) are expected in early 2007-10. I am discussing what the copyright, visibility and re-use of various versions are, including the final "publisher's PDF". All images are deliberately free of restrictive copyright and I have asked that the article's copyright is CC-BY, or possibly with Sparc author's addendum. Pre-publication in Nature Precedings has been welcomed and agreed by all.*

*Note: The format of this article explores some new avenues. All references are to Open material on the web and there are no conventional citations. Where possible the reader is directed to Wikipedia; [Foo] implies http://en.wikipedia.org/wiki/Foo.*

*Note: this document is discussed on my blog at:*
*http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=650*

I am writing this article having just come back from the sixth annual UK All Hands meeting on [E-Science]. Several hundred delegates met to discuss how the Internet (and extensions such as the [Grid]) could support and change the way we do science. The message of one plenary lecture: "Digital Earth: The New Digital Commons" [1] by Timothy Foresman was starkly simple. Unless we act the higher organisms on the planet will die. This is not a conjecture; our data are sufficiently comprehensive and our simulation tools sufficiently good and cheap to make it a certain scientific fact. On the positive side, if we make every scrap of scientific research and knowledge fully public and develop and use collaborative tools across the planet we have a chance. The Internet may have come just in time.

Science is multidisciplinary. The [Keeling_curve] – the most beautiful and terrible graph of the twentieth century (Fig 1) links chemical bonding, spectroscopy, quantum mechanics, thermodynamics, fluid dynamics, meteorology, geology, astronomy, reactions, biochemistry, biology, oceanography, etc. directly to transport, economics, finance, politics, psychology and much more. It epitomises [eScience] which seeks to develop the tools, the content and the social science to support multidisciplinary collaborative science. How can we gather the data, formalize its representation, build the computational support, grow the community and share the results in forms that are appropriate to each reader? And who are the readers? Not just professional scientists, but children, senior citizens, lawmakers and funders in every country. And not only humans. To gather and manage this data and simulation we must make it accessible to machines since we cannot cope with the scale and complexity unaided.

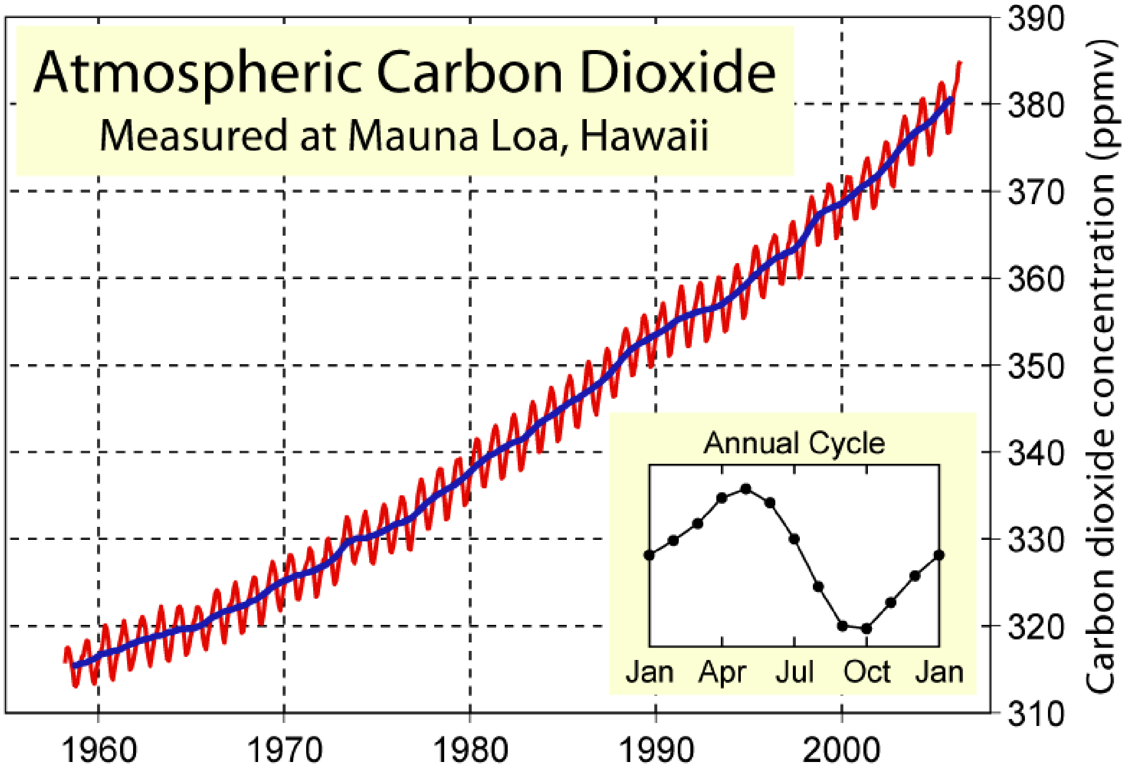


Figure 1. *http://en.wikipedia.org/wiki/Keeling_curve*

In this article I shall use chemistry as a microcosm of eScience, stressing implicitly that multidisciplinarity is essential. The fundamentals we shall encounter, which can be labelled as [Web_2.0] are:

- All information must be Open (Free). [Open_data]
- All information must have [semantics].
- Computing and storage are infinitely cheap and must be [Open_source]
- [Social_computing] is an essential component
- We work on a constantly shifting ground of technology.
- Simplicity is essential
- There is no centre

- Young people are fundamental

My style is first-personal; I have run a [blog] [3] for a year where I explore many of these issues. It seeds collaboration, helps extract information and acts as an advocate of Open information. The links in this article are all electronic and may occasionally point to ephemeral pages, including blogs and wikis – that is part of the constant change. I have provided many links to reference material but no citations, partly because many are inaccessible except to funded academics and partly because the emphasis on the monetary value of formal citations in the funder-academia-publisher complex underrates the value of Web 2.0 material. We need to develop mechanisms to give credit to new resources: blogs, software, data collections which are a core part of the future.

## ***Open Information***.

In the last century systematic information was expensive, as it was gathered by humans often checked and usually retyped. A pre-eminent example is the [Chemical_Abstracts] Division of the [American_Chemical_Society] with bibliographic information and abstracts for all articles in chemical journals worldwide, and chemistry-related articles from all scientific journals, patents, and other scientific publications. It contains information for over 27,000,000 substances and is seen as the fundamental source of chemical information in all mainstream chemistry departments. But, in the twenty-first century it – with all other mainstream providers of chemistry content, software, journals - ticks none of the points above. If chemistry is to contribute to eScience it must change every part of its approach to information.

Chemistry has, however, created some of the best aspects of Web 2.0, though with almost no funding. [Nature_Publishing_Group] has watched these carefully and encouraged them through its commentary blogs and most recently through the Nature/[Google]/[O%27Reilly_Media] [Foocamp]s which have greatly helped to legitimate and encourage unconventional approaches (Fig2.).

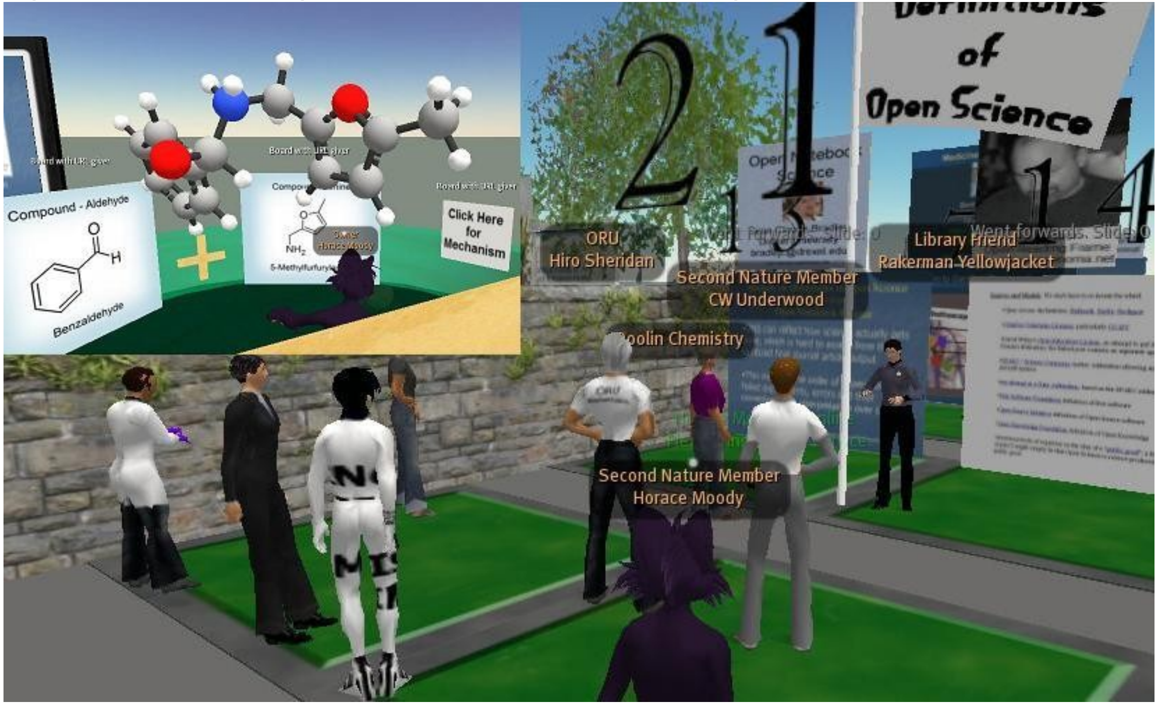


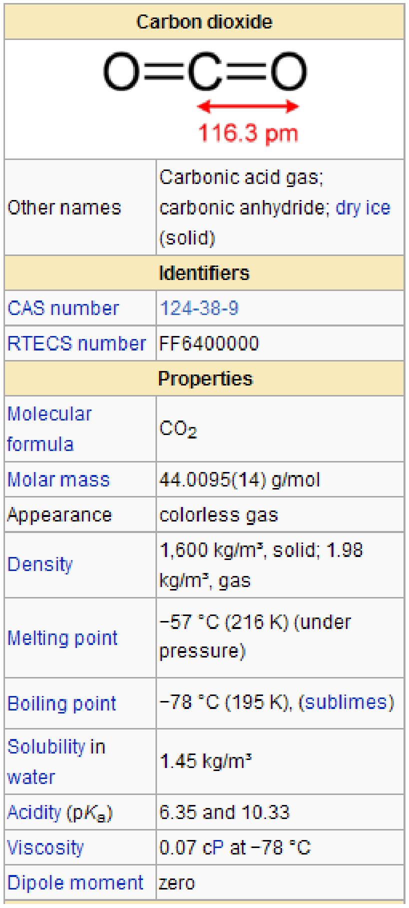
*Fig2  Joint activity between Jean-Claude Bradley and Nature on an island in [Second_life]. The group, including non-scientists, are discuss the definitions of [Open_science] with (inset)a valid interactive molecular object*

This is seriously part of the future of science. It's fun – currently inefficient – and will probably change dramatically but it emphasizes the collaboration and the ability to enhance verbal communication with active objects.

A number of us have formed an informal community – the Blue Obelisk [4] – to encourage the use of Open thinking, actions, objects and artefacts in chemistry. The mantra is "[Open_data], [Open_Standards] and [Open_Source]".  Mainstream chemistry has no tradition of Openness and electronic collaboration so this is a bottom-up movement, largely composed of young researchers inspired by [Web_2.0] and the relative ease of writing usable chemical tools. The primary emphasis is on Open interoperable software, reference data and algorithms such as [Jmol], followed by [Openbabel] and CDK [Chemistry_Development_Kit] and CML [Chemical_Markup_Language]).

Unlike global communities in astronomy, geoscience and biology, chemistry has no global data collection projects so data is primarily published through dispersed, heterogeneous conventional ePaper publications. Traditionally the data were then abstracted by hand. And in the era of real paper there were only limited pagecounts, so most was never published and or effectively lost for ever.  This mindset is still common, with emphasis on the visual form ("double column PDF") rather than machine-friendly [Xml]. Moreover the default business model for chemical publishing is "reader-pays" (Toll-access, TA). This means that much of the chemistry literature cannot be read by most of the world who therefore do not have access to effective chemical data.

However the Web is now seen as an infinite comprehensive source of free information, and young scientists no longer look to the traditional sources of information but to engines such as Google, [Yahoo], and [Microsoft] (GYM). They expect, rightly, to be able to express their requirements in natural language and to get answers instantly. They have no time to learn proprietary systems with idiosyncratic approaches. For reference the first place to look is Wikipedia (WP)



| Carbon dioxide | |
|---|---|
| Other names | Carbonic acid gas;<br>carbonic anhydride; dry ice (solid) |
| **Identifiers** | |
| CAS number | 124-38-9 |
| RTECS number | FF6400000 |
| **Properties** | |
| Molecular formula | $CO_2$ |
| Molar mass | 44.0095(14) g/mol |
| Appearance | colorless gas |
| Density | 1,600 $kg/m^3$, solid; 1.98 $kg/m^3$, gas |
| Melting point | −57 °C (216 K) (under pressure) |
| Boiling point | −78 °C (195 K), (sublimes) |
| Solubility in water | 1.45 $kg/m^3$ |
| Acidity ($pK_a$) | 6.35 and 10.33 |
| Viscosity | 0.07 cP at −78 °C |
| Dipole moment | zero |

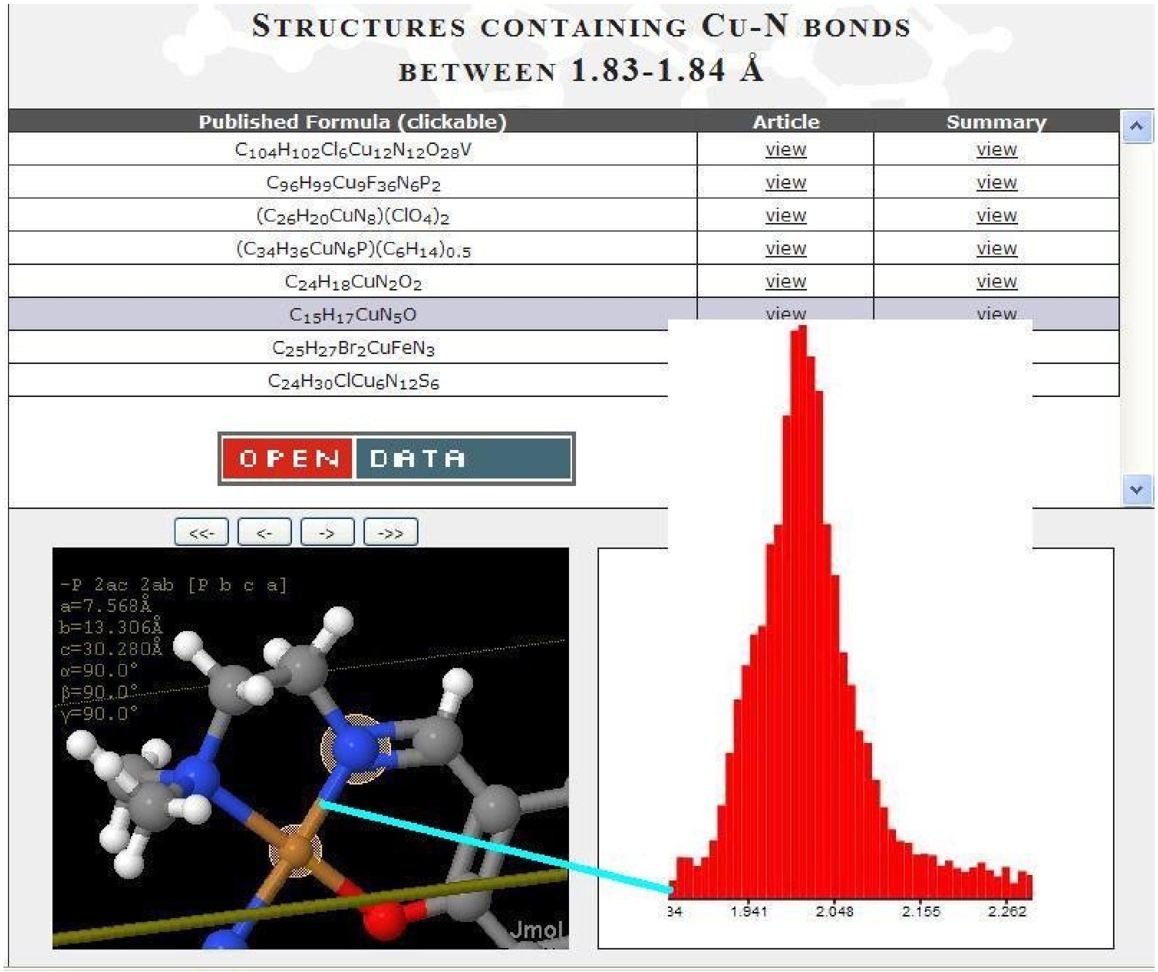
*Fig4. Part of the Wikipedia infobox for carbon dioxide.*

Although there are relatively few Wikipedian chemists the quality of content is high and increasing. Chemistry is an ideal subject for recording factual information and WP will soon become acknowledged as the primary chemical reference for undergraduate study. If you find errors, don't moan but correct them. And, through the infoboxes (Fig 4) it will evolve into a semantic resource more advanced than conventional commercial providers.

We have the technology to capture research data (e.g. in theses, journal articles and patents) but run into social problems. Most research institutions undervalue data, concentrating on "full text" with its [citation] counts and [Impact_factor] as the hallmark of academic achievement. Many publishers do not require and often oppose the mounting of open data sets. Only 0.1% of the spectra relating to the 20+million published compounds are Openly available (in the Blue Obelisk volunteer NMRShiftDB[*]). But eScience and global problems are changing this; data journals are starting to appear and will create markets for quality and citability.

There are several ways of capturing data at near zero-cost which we are exploring, supported by national and international efforts in digital repositories and libraries:

- Link the data measurement and analysis directly to repositories. In the SPECTRa project [] spectroscopic and crystallographic data are sent directly into Open repositories. The main barrier is social: many scientists wish to "hide" their data in case others re-use it to their advantage, show it to be fallacious or make it unpatentable. Unfortunately this leads to rapid data loss (often 80-99%). The SPECTRa toolkit includes an "embargo repository" so that data can be exposed after an appropriate period. SPECTRa, like all software in this article, is Open, and designed for departments and institutions wishing to save data.
- Make data a condition of publication. The [International_Union_of_Crystallography] has campaigned over many years for data and metadata publication. Over 30% of all published crystallography is thus Openly available. Nick Day has built crystalEye [] with over 100,000 entries through daily visits to journals. Fig.* shows a search for Cu-N bonds (ca 10,000), analysis and interrogation. We use many Blue Obelisk components: Jmol (shown), CDK, JUMBO and OpenBabel. Data are released under the Open Knowledge Foundation licence [].



STRUCTURES CONTAINING CU-N BONDS BETWEEN 1.83-1.84 Å

| Published Formula (clickable) | Article | Summary |
|---|---|---|
| $C_{104}H_{102}Cl_6Cu_{12}N_{12}O_{28}V$ | view | view |
| $C_{96}H_{99}Cu_9F_{36}N_6P_2$ | view | view |
| $(C_{26}H_{20}CuN_8)(ClO_4)_2$ | view | view |
| $(C_{34}H_{36}CuN_6P)(C_6H_{14})_{0.5}$ | view | view |
| $C_{24}H_{18}CuN_2O_2$ | view | view |
| $C_{15}H_{17}CuN_5O$ | view | view |
| $C_{25}H_{27}Br_2CuFeN_3$ | | |
| $C_{24}H_{30}ClCu_6N_{12}S_6$ | | |

*Fig *. CrystalEye. A search for Cu-N bonds, with the shortest ones listed and displayed.*

- Extract data from "text". Most science is published as text but there are several problems. It is unstructured, with few semantic flags. The sources are often hidden and many "owners" refuse robotic extraction. Where possible, however, good progress can be made. Our SPECTRa-T project investigates how natural language processing (OSCAR3) can extract chemistry from free text in eTheses.
- Create born-digital semantic documents. This is one of the key messages that institutions should take from this article

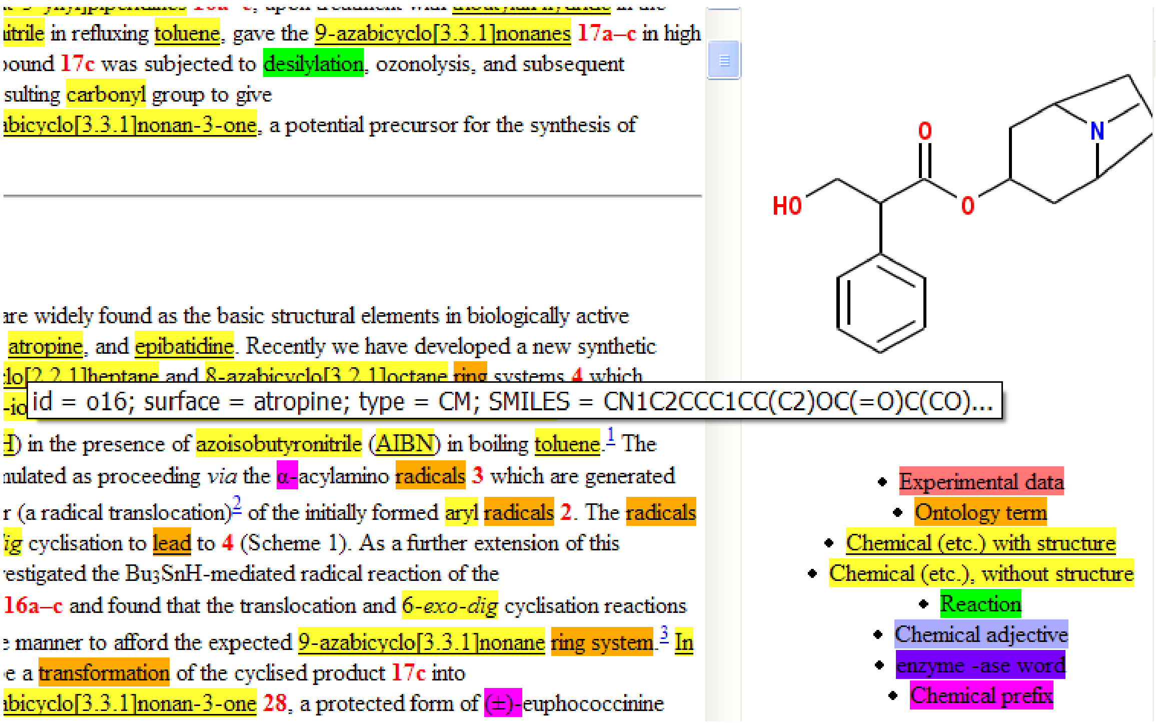
## *Metadata, semantics and ontology*

Fig.1 – a JPEG image - is a good example of a semantically void object. Although a human can extract much meaning, a machine can get nothing. To make it useful for eScience we have to do a lot of work:

- Represent the data as numbers in a standard form such as XML or NETCDF.
- Add metadata to each axis. Date (use [Xml_schema]) and concentration (use [ChemicalMarkupLanguage] (CML)).
- Interpret the chemistry. “Carbon dioxide” maps to [Carbon_dioxide], or the Open [Pubchem] collection of over 10 million compounds with universally agreed identifiers (here CID 280) and [InChI] (InChI=1/CO2/c2-1-3). Universal, non-proprietary, identifiers are essential for semantics.
- Add geospatial and temporal coordinates

Adding identifiers to all theses, articles and patents is simple, free and essential.

The next step is to add ontological information. Here is a snippet of text from a [Royal_Society_of_Chemistry] publication with chemical terms recognized by OSCAR3:



itrile in refluxing toluene, gave the 9-azabicyclo[3.3.1]nonanes 17a–c in high
ound 17c was subjected to desilylation, ozonolysis, and subsequent
sulting carbonyl group to give
ibicyclo[3.3.1]nonan-3-one, a potential precursor for the synthesis of

are widely found as the basic structural elements in biologically active
atropine, and epibatidine. Recently we have developed a new synthetic
lo[2.2.1]heptane and 8-azabicyclo[3.2.1]octane ring systems 4 which
-io

id = o16; surface = atropine; type = CM; SMILES = CN1C2CCC1CC(C2)OC(=O)C(CO)...

I) in the presence of azoisobutyronitrile (AIBN) in boiling toluene.[1] The
nulated as proceeding *via* the α-acylamino radicals 3 which are generated
r (a radical translocation)[2] of the initially formed aryl radicals 2. The radicals
ig cyclisation to lead to 4 (Scheme 1). As a further extension of this
estigated the $Bu_3SnH$-mediated radical reaction of the
16a–c and found that the translocation and 6-*exo-dig* cyclisation reactions
manner to afford the expected 9-azabicyclo[3.3.1]nonane ring system.[3] In
e a transformation of the cyclised product 17c into
ibicyclo[3.3.1]nonan-3-one 28, a protected form of (±)-euphococcinine

- Experimental data
- Ontology term
- Chemical (etc.) with structure
- Chemical (etc.), without structure
- Reaction
- Chemical adjective
- enzyme -ase word
- Chemical prefix

*Fig 5. Automatic markup of chemistry in OSCAR3 (Peter Corbett and others).*

Pubchem provides the connection table for atropine and CDK computes the diagram. Other ontological markup comes from [ChEBI] and [Gold_book] (also used in SPECTRa-T). The RSC has adapted this in its recent prize-winning Project Prospect.

Creating ontologies is laborious, and should be collaborative – a good example is IUCR's [Crystallographic_Information_File] system (used in CrystalEye). We are extracting ontologies directly from corpora, with social computing where possible.

We also add semantics to numeric data (measurement or chemical computation). Several of the main computational codes for quantum mechanics, molecular mechanics and crystal computation have been "CMLized" using FoX []. The semi-structured CML vocabulary is microformatted, with local dialects for each program. To extract the structure and relationships, Andrew Walkingshaw has created Golem, based partly on the map-reduce computational paradigm. With OSCAR and Golem we can trawl existing chemistry for hidden structure and meaning.

Ontologies really become powerful with large amounts of semantic datuments (data + documents) and [Resource_Description_Framework] "triple" relationships, such as:

```
pubchem:CID280 pubchem:name "carbon dioxide" .
pubchem:CID280 cml:boilingPoint "195"^^xsd:float .
cml:boilingPoint cml:units cml:Kelvin .
```

Every object (other than strings) has a namespace (e.g.
**`xmlns:cml="http://www.xml-cml.org/schema"`**)
so that all triples are globally unique. Several organizations and companies have espoused this vision of the semantic web and are building stores that can scale to gigatriples and beyond.

The WWW2007 meeting finally convinced me that RDF was stable, highlighted by [DBPedia]. This project takes Wikpedia as-is and extracts triples from the categories and infoboxes. Recall that no authority has orchestrated this, and that the volunteers had no idea that their work would be used like this. Chemistry is not yet well extracted, but the potential can be shown by real queries such as :
Soccer player […] number 11 from club with stadium with >40000 seats born in a country with more than 10M inhabitants

This is simply the result of linking 5 triples, but the result is amazing.  Try it! This methodology is ideal for chemistry when we create RDF versions of out ontologies and CML-RDF datuments.

## *Social Computing and Collaborative Science.*

This millennium has seen the explosion of social computing where humans and information systems have great [Intertwingularity]. We do not know what the precise features are – why one system grows and another does not - but it is essential for the global eScientific challenges. It is an ecosystem where there are many expriements and the fittest (at a given time) survive and flourish. The most valuable include:

- wikis - structured information with history and hyperlinks
- blogs – ephemeral personal communications with wide exposure. Discovery mechanisms include [Trackback]s and [Blogroll]s.
- virtual collaborative environments.
- [Recommender_system]s and [Linked_data], possibly generating a meritocracy.

Unfortunately these are currently limited to text and images, with no common interfaces for adding scientific material such as equations, chemistry, code and visualization. There is a pressing need for a common system of scientific tools in this area, including plugins for browsers.

Even so the chemical [Blogosphere] has been spectacularly successful. At least 100 bloggers find roles and produce consistently interesting content. This ranges from laboratory chat, to the perils of postdoccing, accounts of actual experiments with photographs, gels and spectra. Some blogs are effectively personal review journals with high-quality commentaries on chemical articles, while others report on domains of interest such as drug discovery, patents, and company business. An important group of technology blogs, emphasizes software and data, mainly Open. A recent development is the Blue Obelisk [Greasemonkey], a browser plugin alerting readers to unseen features in the pages they are viewing. It can highlight any publication mentioned in the blogosphere or any paper with a structure in crystaleye. None of this involves consent from the publisher. It therefore gives mechanisms for alternative community review of the literature, such as quality assessments independent of [Institute_for_Scientific_Information]. Egon Willighagen's meta-blog, "chemical blogosphere" [*] reviews all blogs and shows how semantic chemistry based can be automatically extracted .

Jean-Claude Bradley has developed [Open_notebook_science], the practice of recording experiments on the Internet as they are done. When coupled with semantic documents (e.g. in CML) this is true eScience, globally visible and machine-readable. It challenges the current ethos where chemists may not expose their work before it has been formally "published" – a luxury we cannot afford in the face of global challenges. It is especially suited to computational and simulation processes. While writing this article I have explored with the Blue Obelisk how we may build systems which routinely compute predicted spectra in publications. The calibrated system will then act as a robot reviewer to judge whether published data may be “incorrect”. Spectra will be published as soon as they are calculated, so that the whole world can comment on the methodology and individual data. Only later will a publisher be approached to give the seal of approval.

## *Computing and Storage*

Chemistry is well suited for high-throughput embarrassingly parallel computation. For each of 1 million new organic compounds per year we can hold 1 megabyte of published information and the whole of our annual output - 1 terabyte - is less than a single day’s astronomical or geophysical calculation in some laboratories. High quality structures and energies can be computed for these in a day or two for each. We only need – say – 5000 machines to compute this basic data for the whole world’s chemistry. We and many others have taken over spare capacity (e.g. teaching machines at night) using the experience of the Condor system in the eScience program.

Many companies are now espousing Open systems and data. At scifoo2007 Google offered to host much Open scientific data for free. Craig Mundie of Microsoft says “the next breakthroughs in science and engineering will come from harnessing the power of software and data … the software industry can play a key role in developing tools that automate these data management tasks” [1] and Tony Hey has moved to MS after the success of the UK eScience program. In the emerging arena of Community systems and content, data and software will be free and Openness is being seen as a major commercial advantage.

## *Simple Technology with no centre*

The success of the web has been, and continues to be, the  mixing of humans and computers. Heavyweight approaches such as [SOAP], [WSDL], [BPEL], and relational databases are being replaced in many areas by [REST], a philosophy as much as a technology. It is based on the simplicity of HTTP, addresses all resources as URIs, and can be implemented in hours. It is often coupled with simple scripting languages (Python, Javascript, Ruby), perhaps in AJAX applications. Servers are true servants, and do not dominate the clients as in current quasi-centralised systems. We and others are exploring lightweight repositories and servers based on REST.

An important paradigm is [MapReduce] a system of distributed computing pioneered by Google.  In this the program is taken to the data and the results, after reduction, returned to the user. This suits operations over large databases but has the danger of centralising the ownership and destroying Openness. The current alternative, where scientists like to keep data close to themselves, suffers from the problems of software distribution.

A concern is that Web 2.0 does not support strong typing that physical science requires. Browsers assume only text, images and hyperlinks;  the human will correct errors. The Blue Obelisk has therefore congregated round Bioclipse, a prize-winning application developed by Ola Spjuth from IBM's Open [Eclipse] Java framework. This integrates Jmol, JUMBO, CDK, JSpecview, OpenBabel, etc. into a single framework.

## *Conclusions*

The world is changing much faster than established chemistry; it must adapt rapidly or fracture. Closed publications, tool-access databases and binary software are being swept away by the new philosophies and technologies. Young scientists do not read or use closed systems, and are increasingly frustrated by out-of-date approaches. The technology is now in their hands – several of our systems were pioneered by undergraduates. A single, compelling, blog voice can seed a large community and ideas and technologies spread almost instantaneously. They demand Open, high-quality, integrated, semantic systems.

To add chemical information to our planet-saving engine we must make everything Open as rapidly as possible. We need to be working alongside current publishers and learned societies. But we also need new social protocols as the current ones aren't working. So here are some suggestions, based on the spirit of the blogosphere:

- support young people from an early age. They are already shaping the future through interactive collaborative systems
- use the global challenges – carbon dioxide, disease, ageing to drive our information systems
- reach out to unconventional communities

And, perhaps most importantly, redesign the information economy so that reward is given for making information Open, rather than selling it. If we can create a 30 B USD carbon trading market (possibly expanding to 1 trillion) we could do the same for scientific information. It requires government action, but it could work. Let's sell Chemical information credits, rather than journal subscriptions.

But you don't have to take my word for it. Ask the blogosphere.

Thanks to come

Limited references to come

Digital Earth:
http://www.allhands.org.uk/2007/proceedings/proceedings/introduction.pdf,
http://www.isde5.org/

Chemical Blogspace (from which all other chemical links can be discovered)
http://blueobelisk.sourceforge.net/wiki/index.php/Main_Pagehttp://cb.openmolecules.net/

[1] http://www.technologyreview.com/read_article.aspx?id=16461&ch=infotech