Chemistry for everyone

Peter Murray-Rust

Moves by chemists to help computers access the scientific literature have boosted the drive to make scientific information freely available to all.

Imagine a world in which all scientific information is instantly available — where anyone can get an answer to any question, no matter how abstruse or dependent on technical formalism. In such a world, science is published directly onto the Internet, the size of data sets doesn't matter, and machines can do dirty and everyday tasks, such as searching through millions of dense technical articles or calculating routine data. This is the emerging world of e-science or cyberscholarship.

e-Science seeks to develop the tools, content and social attitudes to support multidisciplinary, collaborative science. Its immediate aims are to find ways of sharing information in a form that is appropriate to all readers. This requires new methods for gathering and representing data, for improved computational support, and for growth of the online community. But who are the readers? They are not only professional scientists, but also children, senior citizens, lawmakers and funding bodies. And they are not only people — information must be accessible to machines, because humans won't be able to cope with the amount and complexity of the incoming data.

Cyberscholarship also embraces the revolution of 'web 2.0', in which humans and machines come together in unpredictable ways to create innovative knowledge resources. The resulting 'cyberlaboratory' — foreseen in the novels of William Gibson and others — is giving rise to virtual spaces where scientists share ideas and data, and where some of the traditional values of science are re-examined in emerging 'gift economies'.

Here I will show how chemistry, often thought of as a conservative discipline, is making important contributions to the nascent field of e-science. Indeed, the creativity shown by young chemists might transform the way that science is performed.

Open information

In the twentieth century, technical information was expensive, as it was gathered by humans, double-checked and then usually retyped. For example, the American Chemical Society collects bibliographic information and abstracts for all chemistry-related articles published worldwide. Its Chemical Abstracts Service (CAS) contains data on more than 27 million substances and is seen by chemists as the fundamental source of chemical information. But in the twenty-first century, this resource — along with all the other conventional sources of chemical information — is incompatible with the requirements of web 2.0. If chemists are to contribute to e-science, they must rethink their approach to the way information is accessed and presented.

Figure 1 | Virtual world. Second Life is a virtual world that allows people to interact using animated characters, or avatars. Here, people are discussing open science. Inset, an interactive molecule.

Even so, chemists — with almost no funding — have created some of the best aspects of web 2.0. The Nature Publishing Group has watched these developments carefully and encouraged them by providing scientific commentary blogs, discussion forums in the virtual-reality world of Second Life (Fig. 1), and, more recently, by co-sponsoring 'Foo camps'. These interdisciplinary brainstorming meetings exploit web 2.0 ideas to full advantage, and have helped to legitimize and encourage unconventional approaches to sharing information. Such fun initiatives may look trivial and are currently inefficient. But they emphasize the power of collaboration and show how verbal communication can be enhanced online. They are a serious part of the future of science.

But there is more to be done. A group of chemists, programmers and informaticians — myself included — have set up an informal, online community known as the Blue Obelisk to encourage openness in chemistry. The mantra is "open data, open source and open standards". The Blue Obelisk is a bottom-up movement, largely composed of young researchers inspired by web 2.0 and by the relative ease with which useful chemical software can be written. The emphasis is on open, interoperable software, reference data and algorithms, such as Jmol, which allows computer models of molecules to be displayed in various ways, and Open Babel, which interconverts the different electronic formats used to store chemical information. The Blue Obelisk provides a complete basic infrastructure for open-access chemistry, including a standard language for communication (chemical markup language, CML) and libraries of software applications for essential chemical functions (the Chemistry Development Kit, CDK).

The issue of open data is particularly problematic. Unlike astronomers, geoscientists and biologists, chemists have no global data-collection projects; their data are usually published in many different online journals and then collated by hand into CAS. In the era of real

©2008 Nature Publishing Group

paper, limited page counts ensured that most chemical data were never published, and so are effectively lost. Even now, most electronic documents still use visual representations of a printed page (such as PDF files), rather than machine-friendly formats that allow data to be shared across different information systems. Moreover, the default business model for chemical publishing is 'reader pays'. As a result, non-subscribers — that's most of the world — have no access to a large percentage of chemical data.

But things are changing. The web is an almost infinite, comprehensive source of free data. Young scientists don't go to libraries and no longer look to traditional sources of information, but to search engines such as Google. They expect to be able to express their questions in natural language and to get instant answers. They have no time to learn proprietary systems with idiosyncratic approaches. For reference, one of the first places to look is Wikipedia.

Although relatively few chemists contribute to Wikipedia, the quality of its chemical content is high and increasing. Chemistry is an ideal subject for recording as factual information, and Wikipedia will soon be acknowledged as the primary reference for chemistry undergraduates. Proactivity is the key — if you find errors, correct them. And through the use of 'infoboxes' that contain searchable data, Wikipedia will evolve into a computer-searchable reference source that is more advanced than those provided by conventional commercial suppliers.

But many of the problems associated with capturing data are not technological but social. Most research institutions undervalue pure data, focusing instead on published papers as the hallmark of academic achievement. This is exacerbated by publishers, who generally do not require — and often oppose — the mounting of open data sets. Only 0.1% of the analytical spectra for the 20 million or so published compounds are openly available. But e-science and the demands of global problems are forcing this situation to change; data journals are starting to appear and will create markets for high-quality, citable data.

To encourage open data, my research group and others are exploring several ways of capturing data at almost zero cost. In the SPECTRa project, spectroscopic and crystallographic data are sent directly to open repositories. Again, the main barrier is social: many scientists wish to hide their data to prevent others from using them to their own advantage, showing them to be fallacious or making them unpatentable. Unfortunately, this leads to rapid data loss (often 80–99%). To avoid this, SPECTRa has surveyed how chemists actually work and proposes an 'embargo repository' that allows data to be released only after an appropriate period.

Another option is to make data publication a condition for all papers. For example, the International Union of Crystallography (IUCr) has campaigned over many years for the publication of raw crystallographic data and metadata (data about the data). As a result, more than 30% of all published crystallography data are openly available. CrystalEye is a web-based system that exploits this openness by scouring the Internet for crystallographic data and collecting them together in a searchable format (Fig. 2). Currently, CrystalEye has more than 100,000 entries gathered from daily visits to online journals. It uses Blue Obelisk software, such as Jmol and Open Babel, and the data are freely available for all users to use, reuse and redistribute.

The third strategy for cheap data capture is to extract data electronically from existing text. But there are problems with this. Text is usually only meaningful to people — there are few semantic flags that would allow machines to understand it. Furthermore, the full text of many documents is often not open access, and even if it were, many authors and publishers will not allow data to be extracted robotically from their work. Nevertheless, good progress has been made with natural-language processing software. For example, the OSCAR3 program, developed by my group, investigates how chemical information can be extracted from the text in PDF files.

But the single most crucial thing that chemists should do to simplify data capture is to abandon paper and create digital-only, semantic documents that can be understood by computers.

Metadata, semantics and ontology

The graph shown in Figure 3 (overleaf) is a good example of an object that is semantically void to a computer — humans can extract much meaning from it, but machines can understand nothing. To make it useful for e-science, we must represent the data as numbers in a standard form; use metadata to label the axes; and interpret the chemistry by mapping 'carbon dioxide' to a standard definition, such as that found in the open-access PubChem database (which lists more than 10 million compounds, each with its own universally agreed identifier). Adding such details to chemistry documents is simple and costs nothing, but is essential to allow a free flow of information.

It is more difficult to add ontological information — metadata that define concepts — to text. Figure 4 shows part of a paper in which chemical terms are recognized by the OSCAR3 language-recognition software. Ontological markers are used, so, for example, the word 'desilylation' is recognized as a chemical reaction, and the Greek letter 'α' is identified as a chemical prefix. Chemical names are also recognized; these can be linked to structural information in molecular databases (such as PubChem, ChEBI or the Gold Book), from which chemical structures can be produced and manipulated using open software such as CDK. The Royal Society of Chemistry uses this approach to enhance online publications in its award-winning Project Prospect, which builds in part on a collaboration with my group.

Creating ontologies is laborious, so it is useful if the load can be spread. For example, the

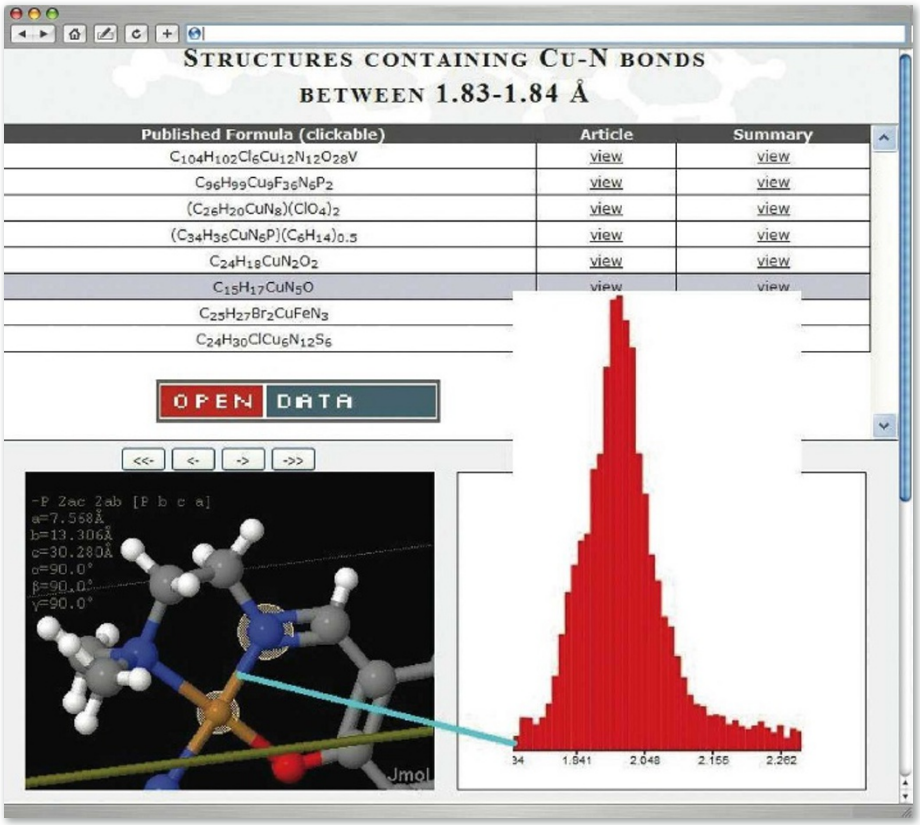

Figure 2 | Exploiting open data. CrystalEye is a free web application that gathers open-access crystallographic data and allows it to be searched and manipulated. The screenshot shows the results of a search for compounds with copper–nitrogen bonds. The graph plots the number of hits against the lengths of the bonds; compounds with the shortest bonds are listed in the table. The molecular structure of the compound highlighted in the table is also displayed.

©2008 Nature Publishing Group

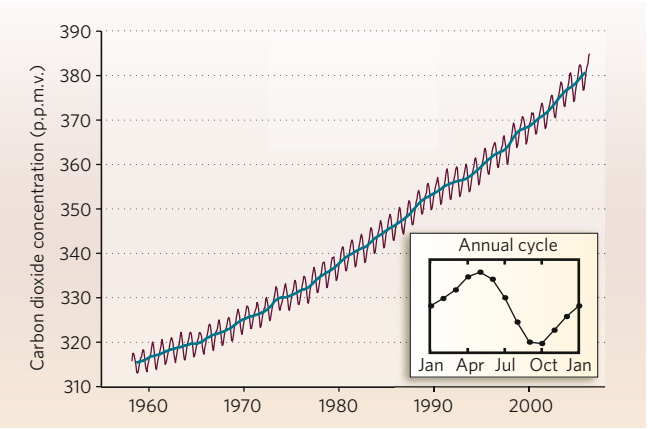

Figure 3 | Meaningless data? A Keeling curve plots the concentration of carbon dioxide in the atmosphere above Mauna Loa, Hawaii, since 1958. Although humans can obtain much information from this graph, it is meaningless to computers unless identifiers are added that allow the data to be interpreted electronically.

IUCr's crystallographic information file (CIF) system is the result of a 15-year collaborative effort. This universal text-file format allows many different computer programs to interpret crystallographic data in order to share and visualize molecular structures.

Ontology can also be added to numeric data. Similarly, computer code can be converted into a 'universal' language — for example, several of the codes often used for quantum mechanics, molecular mechanics and crystallographic calculations have been converted into CML. The semi-structured CML vocabulary can then be reformatted in different ways to allow different programs to use the codes, using cross-platform toolkits such as FoX software (which allows FORTRAN programs to produce output in modern formats such as CML or XML).

To extract structure and relationships from heterogeneously computed data, software such as Golem has been developed in our group. Golem spots patterns that describe how data are expressed in chemistry documents (whether written by humans or machines), and uses these patterns to extract and correlate data from document repositories. This could reveal meaning that was not explicit in the original versions.

Ontologies are powerful when dealing with large amounts of text and data, as they can exploit 'triple' relationships. Triples are statements that come in three parts: subject, predicate and object (the predicate defines the characteristics of the subject, and expresses a relationship between the subject and the object). A simple example is: "The car has the colour red." Here, the car is the subject and the predicate "has the colour" describes the relationship of the car to the object, which is red. Triples can describe almost any concept and can be described in standard formats that are recognized by machines. For example:

pubchem:CID280 pubchem:name "carbon dioxide"

Translated into human terms, this means "the chemical defined as CID280 in the PubChem database has the name carbon dioxide". Triples can each be given a unique location (similar to a URL on the web) and saved in triple stores. If semantic flags in electronic documents link to triples, then any computer can extract information from those documents. Used in combination with each other, triples allow machines to deduce a deeper 'understanding' of information. Several organizations and companies espouse this vision of the semantic web, and are building stores that can host gigatriples of information.

This approach for data mining has come of age, as highlighted by the DBpedia project, which extracts triples from the categories and infoboxes in Wikipedia. This allows sophisticated questions to be asked of Wikipedia, by linking together information spread across the entire resource. Remarkably, DBpedia arose spontaneously — no authority orchestrated it, and the volunteers who wrote entries for Wikipedia had no idea that their work would be used in this way. Although DBpedia is currently poor at extracting chemistry information, its potential can be shown by real queries such as:

Soccer player […] number 11 from club with stadium with >40,000 seats born in a country with >10 million inhabitants

The question simply links five triples, but amazingly it returns a short, precise list of names fulfilling the criteria. We have now created triples for more than 1,000 chemical compounds in Wikipedia and shown that similar queries can be used to mine these data. This search method will be ideal for finding chemistry information once the appropriate ontologies are created.

Social computing and collaboration

This decade has seen an explosion in social computing, in which humans and information systems have become greatly interconnected. Such connectivity is essential to meet the needs of e-science. The most valuable elements of the social computing 'ecosystem' include wikis, blogs, virtual collaborative environments and recommender systems, which suggest links of interest to readers based on their previous choices, creating a meritocracy of information. Unfortunately, all these innovations are currently limited to text and images, and lack interfaces for adding scientific material such as equations, chemistry, molecular visualization or computer code. There is a pressing need for a standard system of scientific tools in this area, including plug-ins for browsers.

Even so, the chemical 'blogosphere' has been spectacularly successful. At least 100 bloggers produce consistently interesting content, ranging from laboratory chat to accounts of actual experiments with photographs, gels and spectra. Some blogs are personal review journals, with commentaries on chemical articles; others report on drug discovery, patents and business. There is even a 'meta-blog' that reviews the other chemistry blogs, and which has pioneered semantic links for the automatic extraction of chemistry information from these resources.

A crucial group of technology blogs focuses on software and data, most of which is open access. One development is Blue Obelisk's 'greasemonkey', a browser plug-in that alerts readers to unseen features in the pages they are viewing. It can highlight any publication mentioned in the blogosphere, or any paper that has a structure in CrystalEye. This provides an alternative mechanism for reviewing the literature, and allows chemists to assess the quality of papers, independent of impact factors. None of this requires consent from publishers.

There are many new approaches to social computing and data sharing, of which the dynamic, interactive world of Second Life is well known. The Blue Obelisk community has invested in some virtual real estate and collaborated with *Nature* in a new generation of interactive arenas and intelligent objects. Encouraged by iPhones and multi-touch screens, we expect the next few years to revolutionize the way that humans interact with information. This will inevitably find its way into everyday scientific practice.

Another critical aspect of collaborative science is the open notebook, which records experiments on the Internet as they happen. When coupled with semantic documents, this generates globally visible, machine-readable information. It challenges the current ethos that chemists may not disclose their work before it has been formally published. Open notebooks are especially suited to computational and simulation processes.

My group, in collaboration with researchers at Imperial College, London, has recently mounted a system that routinely predicts analytical spectra for new compounds presented in publications. Such a system could act as a robot reviewer to judge whether published data seem reasonable. Predicted spectra would be published as soon as they are calculated, so the whole world can comment on the method and individual data (both for the experimental data and the predictive software). A publisher could then be approached to give the final seal of approval to experimental data.

Processing power and data storage

Currently, most of the information typically published for an organic compound can be stored using just 1 megabyte of data. With about 1 million new compounds discovered every year, this amounts to a paltry annual output of just 1 terabyte — less than a single day's calculation output for some astronomical or geophysical laboratories.

©2008 Nature Publishing Group

Clearly there are gaps in chemistry data. As an example, high-quality molecular modelling data are not available for most compounds, although in the majority of cases such information can be calculated in just one or two days. The process of modelling chemical structures is well suited to high-throughput, simultaneous computation. Only about 5,000 machines would be needed to calculate fundamental data for the world's annual complement of new compounds. Several groups, including my own, have therefore taken over spare computer capacity — such as university teaching machines at night — to fill this data gap.

Of course, such initiatives create problems of data storage. Fortunately, many companies are now supporting open systems and data — at the 2007 Science Foo Camp, for example, Google offered to host open scientific data free of charge. In the emerging arena of community systems and content, data and software will be free, and openness will be seen as a big commercial advantage.

Simple technology and decentralization

The web is not just a triumph of technology — it is equally as dependent on human input. If the Internet is to develop successfully, systems are needed to make human involvement as easy as possible. The current basic protocols for web interactions (such as SOAP) are heavily engineered approaches that are unnecessarily complex, so many information technologists are adopting a new style of software architecture known as REST, which is much simpler.

Most REST applications are based on features in the hypertext transfer protocol (HTTP, the standard rules used to transfer information on the web), and the uniform resource identifier (URI) framework, which gives all information an 'address' in a common format. A key advantage is that the interfaces to REST applications are simple and uniform, whereas older systems often required specific implementations and tools to allow different software to 'talk' to each other. As a result, REST allows users to focus on their data, rather than having to second-guess how any particular set of users might want to use it. Frameworks and tools that support REST tend to embrace this simplicity throughout their design — REST systems are easier and quicker to use. For scientific data, the combination of triples with REST is replacing traditional portal services.

The next few years will see a shifting balance between data and computation held locally and centrally. Decentralization is often the key to high-throughput data processing, for example by farming out tasks to any idling machines in a network. A crucial strategy is seen in the MapReduce system pioneered by Google. In this approach, users send data sets to a hub, which distributes the data across hundreds of thousands of computers for processing. After reduction of the outputs, the results are returned to the user. But because the service relies on a central provider, there is a danger that openness could be destroyed. The traditional approach, in which chemists store and process their data locally, will still have value.

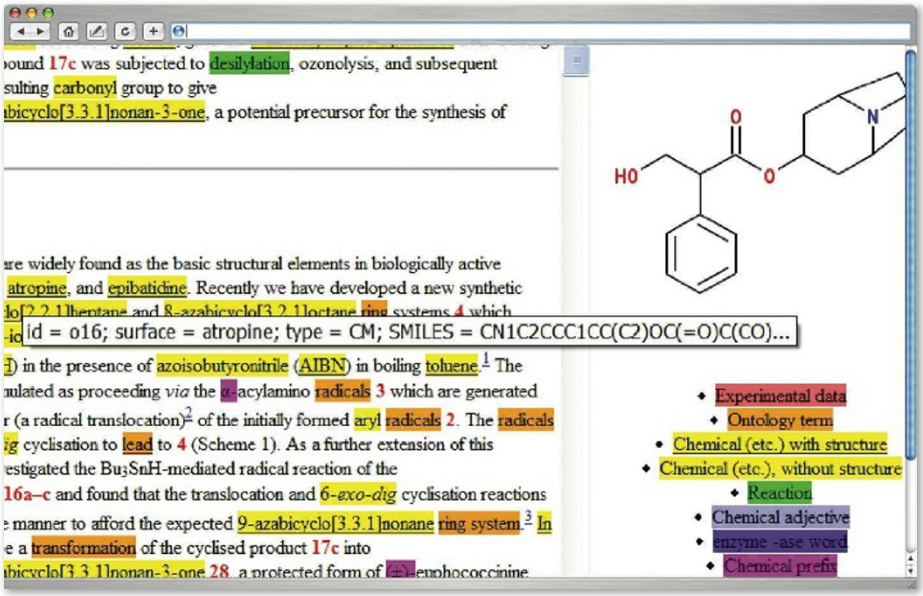

Figure 4 | Text for computers. OSCAR3 is a free web application that can extract chemical information from text in natural language. In this screenshot, the highlighted text has been recognized and categorized according to the colour key (bottom right). By selecting the word 'atropine' in the text, structural information for the compound is retrieved from the web and converted into a manipulable structure (top right) by another open-source algorithm.

A concern for scientists is that web 2.0 is based on human vocabulary, which can be ambiguous. For example, chemists would recognize 'CO' as the chemical formula for carbon monoxide, but a computer would confuse it with the abbreviation for Colorado; browsers currently rely on humans to distinguish between such things. The Blue Obelisk has therefore developed a browser with chemical 'intelligence', known as Bioclipse, based on an open-source development framework known as Eclipse. Bioclipse can recognize molecules, reactions, proteins and their sequences, spectra and crystallographic data, and fire up specific applications to handle each of them. When semantically rich data become common on the web, and discipline-specific search engines evolve, cyberscience will truly have arrived.

Conclusions

The world is changing rapidly, and the chemistry establishment must adapt quickly or fracture. Closed publications, binary software and toll-access databases are being swept away by the emerging philosophies and technologies. Many young scientists do not read or use closed systems, and are increasingly frustrated by out-of-date approaches. Perhaps for the first time in history, the technology for change is in their hands — indeed, several of Blue Obelisk's systems were pioneered by undergraduates. As new ideas and technologies arise, the blogosphere spreads them almost instantaneously. And the message from the blogosphere is clear: the next generation of chemists needs open, integrated, semantic systems.

If chemical information is to address worldwide problems, it must be made open as rapidly as possible. This will involve working alongside the publishers that currently produce most of the scientific literature. 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. We must support young people; they are already shaping the future through interactive collaborative systems. We must use global challenges — such as climate change, disease and the ageing population — as spurs to drive the evolution of our information systems. And we should reach out to unconventional communities for their ideas.

But perhaps most importantly, the information economy must be redesigned so that rewards are given for making information open. If we can create a US$30-billion carbon-trading market to help deal with carbon dioxide emissions, why can't we sell chemical-information credits, rather than journal subscriptions? This would require government action, but it could be made to work. There is broad support for such a move. But you don't have to take my word for it. Ask the blogosphere. ■

Peter Murray-Rust is at the Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK.
e-mail: pm286@cam.ac.uk

FURTHER READING AND ONLINE RESOURCES

Hundreds of people have contributed to open chemistry and they are best acknowledged by following the links from the following web pages. Most projects also have Wikipedia entries.

- http://wwmm.ch.cam.ac.uk
- http://blueobelisk.sourceforge.net/wiki/index.php/Main_Page
- http://cb.openmolecules.net
- http://www.okfn.org
- http://en.wikipedia.org/wiki/List_of_chemistry_topics

©2008 Nature Publishing Group