The World-Wide Web as a chemical information tool†

Peter Murray-Rust,[a] Henry S. Rzepa[b] and Benjamin J. Whitaker[c]

[a] *Virtual School of Molecular Science, School of Pharmacy, University of Nottingham, Nottingham, UK NG7 2RD*
[b] *Department of Chemistry, Imperial College of Science, Technology and Medicine, London, UK SW7 2AY*
[c] *School of Chemistry, University of Leeds, Leeds, UK LS2 9JT*

CHEMICAL SOCIETY REVIEWS

The Internet-based information tool known as the World-Wide Web began to be widely used in 1993. This review will introduce the basic structures that characterise the Web and how standards and methods for expressing chemical content by this means have evolved during the last three years. We outline some of the challenges that face the chemical and related scientific communities in adopting this new medium as an information tool, with a focus on electronic journals and electronic conferences.

1 Introduction

Chemists were amongst the earliest users of computers for generating, acquiring or searching for data, and hence deriving information in order to create knowledge about the subject.[1] The established access model dating from the early 1980s is the use of proprietary on-line networks such as STN (Scientific and Technical Network). An alternative global information paradigm based on the Internet and known as the World-Wide Web was introduced in the early 1990s, and has subsequently gained widespread general use. This review will introduce some of the innovative and useful features that characterise the Web and in particular its chemical content, how it can be used to create new connections between related subject disciplines, and how it has the potential to influence the manner of future scientific collaboration.

† This article also appears in electronic form at http://chemistry.rsc.org/rsc/csr398.htm

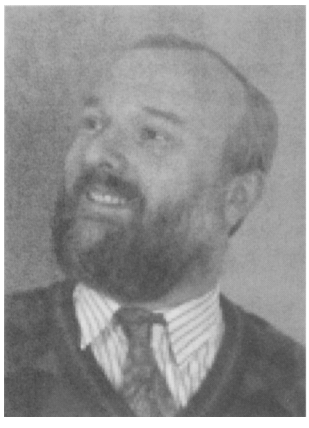
The review is not intended as a technical handbook. Nor in an area where the speed of change and growth is unusually high, can we attempt in print to be either comprehensive or entirely current. In keeping with the subject of this review, an electronic version of the article is available[2] for those readers who wish a starting point for their exploration of the chemical themes explored here, or who wish to investigate those aspects of this article which cannot be demonstrated in print.

2 The essential features of the World-Wide Web

The seamless integration of on-line chemical and biological data archives with other areas of scientific interaction such as journal publishing, conferences, electronic mail and discussions, books and other reference sources, laboratory systems, commercial catalogues, computer based research activities, and human resources could be described as something of a holy grail in science. The potential to achieve this ideal is thought to have come closer with the introduction in 1989 by Berners-Lee and colleagues[3] of the World-Wide Web system, and its subsequent rapid adoption from around 1994 onwards by sections of the scientific and chemical communities.

To help highlight why the Web differs radically from earlier infrastructures, the characteristics of the more traditional on-line information model are first summarised. Access is usually controlled by individual user accounts *via* a small number of remote server access points for users, and requires use of custom software which tends not to be inter-operable between different vendors. With a few exceptions, mechanisms for individuals or

Peter Murray-Rust was involved over a period of 14 years in setting up molecular graphics, computational chemistry, protein structure and object-based informatics facilities at GlaxoWellcome, where he also helped to develop the first Intranet in the pharmaceutical industry. As a visiting professor at Birkbeck College he developed the first virtual multimedia course on the WWW (Principles of Protein Structure) and is a member of the UK's BBSRC Bioinformatics Committee. He is now setting up the Virtual School of Molecular Sciences at Nottingham, where he has been involved in launching a virtual course in Bioinformatics Tools using Java. He is passionately devoted to collaborative ventures which include Collaborative BioMolecular Tools, Chemical Markup Language, Open Molecule Foundation and the Virtual Hyperglossary.



Henry S. Rzepa

Benjamin J. Whitaker

Henry S. Rzepa was born in London in 1950, and received both his BSc and PhD degrees from Imperial College, in 1971 and 1974. Following a post-doctoral period with Michael Dewar at the University of Texas, Austin, he returned to Imperial College, where he is now a Reader in Organic Chemistry. His research interests include the theoretical study of stereoelectronic effects, structural studies of unusual forms of hydrogen bonding such as π-facial interactions and those responsible for chiral recognition, and the development of the Internet as a chemical tool. He is the recipient of the 1995 Löschmidt Prize for physical organic chemistry.

Dr Benjamin Whitaker is a graduate in Chemical Physics from Sussex University, where he also obtained his DPhil, in 1981, for work in laser spectroscopy. His main research interests are in quantum resolved studies of elementary reactions using molecular beams and in the development of laser spectroscopy, particularly degenerate four wave mixing and cavity ring down spectroscopy. He is currently a Senior Lecturer at the University of Leeds.

local organisations to innovate with new components in a scaleable and integrable manner are not made available. In essence, the models have tended to be monolithic and centralised rather than collaborative and distributed. Local 'points-of-access' to such services have traditionally been located in libraries, and other specially designated areas. It is only in the last five years or so that the infrastructure has become available in most organisations for extending this access to offices, laboratories, meeting rooms, lecture theatres, the home, and indeed travel between these locations if desired.

2.1 The addressing scheme

The last five years have also seen the global adoption of a computer addressing scheme known as Internet Protocol (IP). This allows any Inter-networked device which supports the protocol to be uniquely addressed and located within seconds, irrespective of its geographical location. Originally capable of identifying some 10^8 devices, this number is already deemed inadequate and will shortly be replaced by a new standard (IPv6) which allows some 10^{38} addresses to be used. The World-Wide Web makes use of Internet Protocol *via* an identifier known as a URL (Uniform Resource Locator). This exquisitely named but often misunderstood label carries the essence to understanding the importance of the Web and so is explained in some detail here.

The use of the word 'resource' implies that not only documents, but other real or potential sources of data and information can be uniquely located on a global scale. In a chemical context, this could include database search engines and other computational algorithms and resources, instruments, people or scholarly articles. Even more importantly, quite finely grained data within a document might be addressed, as for example individual atoms in a set of molecular coordinates, or peaks in a spectral representation. A typical example of a URL might be the one associated with the article you are currently reading;

http://chemistry.rsc.org/rsc/csr398.htm#intro

The first part of the string indicates the use of Hyper-Text Transport Protocol (HTTP), a mechanism used to exchange information between the computer where the resource resides (the server) and the computer which the user is using to acquire the information (the client). HTTP has one relevant characteristic; it is said to be 'stateless'. Thus an HTTP transaction between a server and a client is entirely self contained, typically lasting perhaps a few seconds in duration, and there is no state or context maintained between any two HTTP transactions. This feature, which was originally chosen because of its excellent network characteristics, differs fundamentally from the type of context-rich environment associated with *e.g.* CAS on-line or other dedicated database systems. Instead on the Web, context is currently achieved predominantly *via* a device known as the hyperlink (*vide infra*), although there is some discussion that the stateless HTTP protocol might be replaced in future by a richer 'stateful' system.

The component chemistry.rsc.org specifies the precise network device where the information resource will be found, being mapped onto one IP address of the type mentioned above. The final component /rsc/csr398.htm#intro specifies a hierarchical directory structure, a file name with an associated extension and a named 'anchor' (#intro) within the file. The anchor in this case refers to the introduction in this article, but it could just as easily refer to *e.g.* an individual atom in a molecular descriptor. The file name extension .htm indicates a document written in a descriptive language called HTML (Hyper-Text Markup Language).

2.2 Hyper-text markup language and indexing

Mention of HTML introduces another important theme, namely the use of a precisely structured formalism where the semantic content of the document is defined rather than the style in which it might be presented. This formalism is encapsulated in a set of guidelines known as SGML (Standard Generalised Markup Language). HTML is said to follow these guidelines in being specified in the form of a 'document type definition' or DTD. The HTML DTD has reached version 3.2, although strict adherence to the HTML specifications has been one of the less successful aspects of the practical implementation of the World-Wide Web.[4] The most basic structure of a document marked up in HTML is as follows:

<command>Text or other content acted upon</command>

where '<' and '>' are special characters describing an HTML command, and where <command> and </command> define the start and end of the content acted upon by the command. The choice of a markup language for the Web is significant. Because a distinction is made between the content of a document or resource and the manner in which it is displayed, this latter task is performed less by the author of the resource, and more by the client program used to display it. This program is often referred to as a 'browser', and the user of the program can specify, for example, the style, size, and colour of any displayed text depending of the device they are using to view the data, or in a chemical context how a molecule might be displayed.

Because of the focus on information content, documents written in HTML are relatively easy to index, at least in a non-chemical sense. Within two years of the increasing use of HTML to encode content on the World-Wide Web, software had been developed that allowed the global Internet to be indexed and hence searched. At the end of 1996, some 80 million documents had been so treated. However, because HTML is predominantly a generic descriptive language based around text, it does not define any specifically chemical descriptors. For this reason, other means of expressing this subject have hitherto been adopted. The earliest and still the most common solution was to use bit-mapped graphical images in so-called GIF (Graphical Interchange Format) or JPEG (Joint Photographic Experts Group) formats. Other than to a human being, these are largely 'content free' and almost un-indexable (*e.g.* Fig. 1; p. 4). Inevitably, as the World-Wide Web has become more commercially orientated with an increasing focus on style and 'look and feel', the use of bit-mapped images has proliferated, and this single aspect is causing concern as the world's networks become overloaded with their transmission.

2.3 Hyperlink connections between resources

The two unique aspects of the World-Wide Web, namely the ability to define where a resource is to be found using a URL and how to mark up the content of a document describing this resource using HTML, are combined using a device known as a hyperlink. The hyperlink can be resolved by the program used to display HTML coded content. Thus an HTML document can act as an index of content resident in other documents, whether on the same computer system, or elsewhere on the Internet, and can provide the context between different sources of information. In effect, the boundary of a structured collection of chemical or other scientific information need not be defined by the physical limitations of a sheet of paper, or printed book or journal. This has had profound consequences for how information is authored, presented to the user and subsequently indexed. Indeed it is because most of the information on the Web is extensively hyperlinked that searching and indexing the Internet has been possible. In the next section, we will analyse how productive searches can be accomplished.

3 Searching for information on the Internet

Unlike a service such as STN, there is no single unambiguous entry point for users wishing to make productive use of the World-Wide Web. Instead, there are a variety of mechanisms available, some of which are genuinely global in nature, others of which depend entirely on the local characteristics of the

user's environment, and many of which carry no cost to the user. Which one is used depends very much on the nature of the problem being addressed.

3.1 General global keyword searches

The Web is a highly decentralised model for information access. For example, in 1996 well over 1000 servers existed where the main focus was chemistry or molecular science. Inevitably perhaps, the quality of the information available from these sources is highly variable, sometimes evanescent, as is adherence to standards, whether generic or chemical. For this reason, the user is often recommended to start their search from their organisation's 'home page', where a local editor or 'webmaster' should have collected those resources that may have the most appropriate subject relevance to the user.

Because of the focus on hyperlinked contextual content rather than style in the original specification of the HTML, indexing the collective content on a global scale has proved a productive, and indeed commercially lucrative operation. Because HTML is not particularly appropriate for encoding specifically chemical information, one would use these global indices when searching for more specific concepts that are easily defined using simple keyword terms. Collected in Table 1 are some of the more popular indexed collections available in 1996, along with the results of a search based on using 'chemistry' or 'chemical' as keyword terms. Particularly noteworthy is the HotBot resource, which allows searches to be performed on files other than HTML which may be better suited for encoding chemical content (*e.g.* 3D coordinate files in the PDB format).

One striking result is that the results of each search and their computed relevance ranking can differ substantially. This is largely because the indexing of an HTML-based document is normally only performed on the text-based content, the interpretation of which can vary. For example, the content of the <TITLE> of a document is given a higher priority than the content of the <BODY> (<TITLE> and <BODY> being examples of HTML commands which can be declared). Unfortunately, not all authors of HTML documents always enter a title. Another aid to indexing a document is the so called <META> declaration, with which the author can specify key words for indexing, authorship, data of document creation *etc.* This too is rarely used by authors, and no chemical metadata standards for use in HTML yet exist. Chemical content in the BODY of a document is also inadequately handled. For example, there is no mechanism in HTML that can be adapted for identifying molecular formulae, compound registry numbers, *etc.* For higher quality chemical information retrieval, specific solutions have been adopted.

3.2 Global collections of chemical pointers

Creating the global indices referred to above is largely an automatic procedure, but as we have seen, this tends to function very unreliably for chemical content. Since late 1993, several sites have specialised in collecting and to a great or lesser extent evaluating on a manual basis high value chemical sites around the world (Table 2). Increasingly, these tasks are also being adopted by learned societies and publishers as part of more complete information services. In essence, such sites provide some element of peer review, and we feel that this approach will be increasingly adopted.

3.3 Chemical searches and database interfaces

The most innovative sites are those which provide the user with mechanisms to search for specifically chemical content (Table 3). There are some generic tools available *via* the HTML language for building an interactive interface into the displayed page. The use of such 'forms' allows the user to define text based search strings, and other search options, which can be passed to a program resident on the remote Web Server (*via* a so-called Common Gateway Interface or CGI). In turn, the remote program must cast any response into HTML so that it can be displayed back on the user's screen. Chemical structure input can be achieved by the use of text strings which can be entered into the search field (*e.g.* the Daylight SMILES descriptor, entry 4). Most recently direct molecule sketching and display schemes have been developed (entries 3–5). The mechanisms that allow this are discussed in the next section.

The entries in Table 3 emphasise how a context can be established between quite diverse datasets, ranging in this case from protein coordinates and sequences to small molecule collections. Establishing such connectivities is one of the most powerful aspects of the Web, and in areas such as metabolic pathways, ligand binding and other interdisciplinary subjects, one that is expected to grow significantly in the near future.

The WWW Chemical Structures[5] and the ChemFinder[6] databases are noteworthy because they are based in part on molecular information retrieved from the Internet itself. The ChemFinder database from CambridgeSoft is primarily a text-based index of a manually selected number of database sites containing CAS registry number, compound names, connectivity and physical data. The principal emphasis is on checking the compound names extracted from the indexed HTML pages and automatically correcting any errors. In contrast, the WWW Chemical Structures database was constructed from the results of an index robot designed for the task of finding 2D connection tables and 3D molecular coordinates on the Internet. In excess of 2200 such entries were retrieved from the Internet as it existed in 1995. The construction of these

Table 1 Collections of global index search engines

Search Engine	URL	Hits for 'Chemistry' or 'Chemical' (January 1997)
Lycos	http://www.lycos.com/	23 101
Alta Vista	http://www.altavista.digital.com/	About 600 000
HotBot	http://www.hotbot.com/	1 039 864
InfoSeek	http://ultra.//infoseek.com/	411 179
Yahoo Chemistry	http://www.yahoo.com/Science/Chemistry	231

Table 2 Some collections of global chemistry resources

Chemistry collection	URL
Virtual Chemistry Library Collection	http://www.chem.ucla.edu/chempointers.html
ChemDex at Sheffield	http://www.shef.ac.uk/~chem/chemdex
The RSC Collection	http://chemistry.rsc.org/rsc/
The ACS ChemCenter Collection	http://www.ChemCenter.org/
ChemWeb Collection	http://chemweb.com/

Table 3 A selection of Web-based chemical searching resources

Entry	Information Source	URL
1	The Brookhaven Protein Databank	http://pdb.pdb.bnl.gov/cgi-bin/browse
2	The Cn3D-Entrez System	http://www.ncbi.nlm.nih.gov/Structure/cn3d.html
3	MDLI	http://www.mdli.com/
4	Daylight Information Systems	http://www.daylight.com/
5	Tripos Sketch & Fetch	http://www.tripos.com/SandF.html
6	The WWW Chemical Structures Database[5]	http://schiele.organik.uni-erlangen.de/services/webmol.html
7	The ChemFinder System[6]	http://chemfinder.camsci.com/

databases involved a significant degree of manual error correction because the chemical content was not always created in a consistent and standard way by the original authors. That it was possible at all is because some standards do exist for this purpose.

4 Chemical standards and guidelines

If one compares the results of global keyword (Table 3) and molecular structure searches,[5] a ratio of one discrete structure was found for every 100+ Internet based documents which contain the word 'chemistry'. This is almost certainly because most molecular structures are currently represented on the Internet by bit-mapped GIF or JPEG images, from which the semantic molecular content can be extracted only with great difficulty, and which therefore cannot be automatically indexed in any chemical sense. The following example (Fig. 1) compares the characteristics of a representation of a helical carbon nanotube as an animated GIF file and as a PDB (Protein DataBank) file which has been rendered into visual form on the computer screen.

In order to explain the distinction between the two molecular representations, we have to explain the underlying standards and software technologies that have been introduced in the last three years.[7]

4.1 Chemical MIME standards

A large variety of file formats are currently utilized for storing electronic information. This is acceptable when running proprietary tools on dedicated hardware but begins to be problematic for information designed to be shared over a network comprising of many platforms and operating systems. In 1993, Borenstein and Freed[8] proposed a mechanism called MIME (Multipurpose Internet Mail Extensions) which would allow a variety of standard file formats to be exchanged over the Internet using electronic mail. It works by the simple device of adding a short header at the top of each datafile attached to a mail message which provides the information needed for the mail handler to process the contents. The header is hierarchical, with a primary designation designed to provide some measure of sensible 'default' handling of the content, and a sub-type which is more specific. So for example, an HTML file is indicated by text/html as a so-called media type.

The mechanism was rapidly adopted for use with World-Wide Web clients. When a user makes a selection through a hyperlink within a HTML document, the client browser 'posts' the request to the designated web server. Assuming the server accepts the request, it locates the appropriate file(s) and sends them to the client, with the relevant MIME header attached. When the browser receives the data, it reads the MIME type to determine what to do with it. For MIME types such as text/html or image/gif the browsers have been built in such a way that they can simply display the information in the browser window. For other MIME types, a local preference file is inspected to determine what (if any) local program (known as a helper application) can display the information, this program is then launched with the data file and the result displayed in a newly opened application window. The important aspect of this mechanism is that it achieves the delivery of semantic content to the user, who can specify the style in which it will be displayed *via* their choice of an appropriate application program.

In early 1994, we proposed[7] a coherent set of standard chemical file types and made a proposal for the introduction of what we called the Chemical MIME standard. A number of different chemical media types were originally proposed,[9] falling into several categories; (1) molecular coordinate and connectivity and sequence formats, (2) molecular modelling formats, (3) spectroscopic formats and (4) generic 'self-

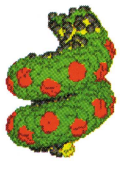
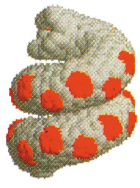
	Animated GIF	**Active 3D Coordinates in PDB form**
Visual Form		
Inter-operability of Information	No user Interaction other than copying of image	Structure can be annotated pentagonal rings heptagonal rings and saved for further analysis
Size of datafile	65,762 bytes	10,089 bytes
Advantages	No additional display software required	User manipulable data, indexable
Disadvantages	Information cannot be re-used or indexed	Requires additional software

Fig. 1 Comparison of molecular structures expressed in bit-mapped image and in PDB form. (If you are viewing the on-line version of this article, click on the image to get an active form of this table)

Table 4 A selection of proposed chemical MIME media types

chemical/x-pdb[a]	Brookhaven Protein database format
chemical/x-daylight-smiles	Daylight chemical connectivity format
chemical/x-mdl-molfile	Molecular and Reaction formats defined by MDLI
chemical/x-mdl-tgf	Transportable Graphic format defined by MDLI
chemical/x-c3d	3D Format defined by CambridgeSoft
chemical/x-embl-dl-nucleotide	SwissProt data format
chemical/x-ncbi-asn1	Protein Sequence Format
chemical/x-gcg8-sequence	Protein Sequence Format
chemical/x-kinemage	Protein Cartoon Format
chemical/x-csml	Chemical Structure Markup Language
chemical/x-mopac-input	Mopac Input Format
chemical/x-gaussian-input	Gaussian Input format
chemical/x-jcamp-dx	Standard for defining MS, IR and NMR spectra
chemical/x-cif	Crystallographic Interchange format
chemical/x-cxf	Chemical eXchange Format
chemical/x-cml	Chemical Markup Language

[a] The 'x' in *e g* x-pdb is included because this proposal has not yet completed the full ratification process [9]

defining' modern formats. A collection of representative types is shown in Table 4; a comprehensive and current list is available on-line.[9] The open architecture allows for new MIME types to be proposed and adopted without having to create new servers or clients. However, we envisage a small core of fundamental chemical MIME types passing through a formal ratification process, whilst new types of perhaps a more proprietary nature of interest to chemical software houses remains as 'x' types, although nevertheless registered *via* a central mechanism for coordinating and documenting these types.

Early Web browsers could not directly display these data in the browser window, but by using the MIME headers of the type

chemical/x-pdb

were able to pass them to an external application program such as RasMol[10] for display. More recently this idea has been slightly reworked with the introduction of browser 'plug-ins'. These are platform specific application programs which can be added to Web clients such as Netscape Navigator (V 2+), Microsoft Internet Explorer (V 2+) or Apple Cyberdog (V 1.1+) to extend their internal support of MIME media types. Using the Chemscape Chime[11] or WebLab[12] plug-ins, chemical media types in the form of *e.g.* PDB or MDL Molfile coordinates can be directly embedded in the main window of the browser (Fig. 1). In this way interactive molecular models can be inserted into a chemical document in a seamless way.

The advantage of using chemical MIME as a simple descriptor of molecular information is twofold. First, at least for molecules containing fewer than around 500–1000 atoms, the data file describing the atomic coordinates is smaller than a bit-map image at reasonable resolution and colour depth of the molecular structure. Thus when the data are downloaded from the server to the client, less precious network bandwidth is used. For example the file size needed to describe a helical carbon nanotube as an animated GIF is over 6.5 times larger than the coordinate file (Fig. 1). Secondly, the quality of the information transferred is of a higher value. Once the atomic coordinates are known, the reader can manipulate the image in the browser to rotate the molecule, view it from a different perspective or change the manner in which the structure is presented, *e.g.* as a spacefilled model of a small molecule or a ribbon cartoon of the backbone structure of a protein.

Because the molecule is described in terms of atoms and their coordinates, even individual functional groups can be addressed. This was first achieved using a mechanism called Chemical Structure Markup Language (CSML),[7] which comprised small scripts held on the Web server and identified with the chemical MIME type

chemical/x-csml

This enables individual regions of molecular structures to be addressed from hyperlinks embedded in text files or graphical images. If you are reading the 'active' version of this article on the Web with the Chime molecular viewing plug-in installed, you can test this for yourself by clicking on the buttons in Fig. 1 labelled 'pentagonal' and 'hexagonal' rings respectively. This highlights the functionality which induces the carbon nanotube to adopt a helical conformation. CSML is now regarded as the precursor to the later development of a properly structured Chemical Markup Language (CML, see below) based on SGML guidelines.[13]

The availability of software supporting chemical MIME types has enabled the rapid development of many innovative applications, of which just a few examples[14] include medical biochemistry tutorials,[14a] correlating cross-peaks in 2D spectral data with atom pairs,[14b] the development of interactive workshops in which students can explore the structural properties of inhibitors specific to HIV protease,[14c] an annotation of the activity of taxol[14d] and a guided tour through the photosystem reaction centre highlighting the course of electron transport.[14e] The adoption of MIME standards can be viewed as the first key event in the development of chemical publication in the Internet by introducing a new paradigm for the publishing of a rich variety of chemical data.[15]

4.2 Markup and descriptive languages

The methods covered so far relate to the visualisation and interpretation of a relatively small range of molecular data. However, the variety of disciplines and techniques that chemistry covers is enormous, so it is not surprising that information exchange between different types of molecular data is difficult. The traditional approach has been either to try to standardise on a single format (*e.g.* chemical/x-pdb for proteins, and increasingly for small molecules also) or to write conversion programs such as BABEL.[16] Unfortunately the latter process always implies information loss: for example the MDLI Molfile format does not hold bibliographic information and in turn PDB files do not hold full connection tables.

A more serious problem is that electronic information decays. The formats used today may be (literally) indecipherable in five years time; many do not even have formally published standards but rely on word-of-mouth and guesswork. Even when manuals are available, it is often difficult to know whether two developers apply the same semantics to a given term. It is generally accepted that the best way to tackle these problems is through the use of markup languages and public discussion. Markup languages add so-called meta-information to a document allow it to be processed in a contextually rich manner.

4.2.1 Standard Generalised Markup Language

The most widely used approach is Standard Generalised Markup Language (SGML), which defines a procedure for describing in a well specified manner the type of content in a document, such that a suitable display or indexing program can

act appropriately on this content. As a simple example, an identified section heading might carry extra significance for indexing, it could be used to create a table of contents, and it might need to be displayed to the reader in a different font type, size, colour or alignment to regular text. The latter represent stylistic decisions defined by procedures specified by the reader, and not by procedures within the document. Sections of text are usually marked up or 'containerised' by placing the contents between two delimiters or tags. SGML is a standard procedural method which enables the definition of containers which specify the semantic content of the textual elements rather than their appearance. Each SGML document starts with a Document Type Definition (DTD), which lays out the rules for the markup of a document in terms of the elements it contains. This allows the markup to be designed with the specific data types likely to be encountered in mind, a feature which is particularly useful in specialised subject disciplines such as chemistry. As an example of descriptive markup in SGML parts of this paper might appear as;

```
<!DOCTYPE article PUBLIC "-//CSR/DTD article/EN">
<ARTICLE>
<TITLE>The World-Wide Web as a Chemical Information Tool</TITLE>
<ADDRESS>Department of Chemistry, . . .
</ADDRESS>
<AUTHOR>Peter Murray-Rust, Henry S. Rzepa and Benjamin J. Whitaker</AUTHOR>
<ABSTRACT> Chemists were amongst the earliest users of computers . . .
</ABSTRACT>
<SECTION>
<SECTIONHEADING>The Characteristic Features of The World-Wide Web
</SECTIONHEADING>
The information model . . .
</SECTION> . . .
</ARTICLE>
```

The first line contains a document type declaration identifying the DTD required for the document—in this case a (fictitious) declaration of a document in the style of an English language (EN) article for Chemical Society Reviews. This is followed by a start tag <ARTICLE>, identical to the tag used for the DTD declaration, which is called the base document element and which activates the DTD. The elements following the <ARTICLE> start tag are fairly self-explanatory. Finally the end of the document is identified with a closing end tag </ARTICLE>. Notice that only the logical structure of the document is important, for example the <ADDRESS> . . . </ADDRESS> container appears out of sequence from that in the printed version of this article. Textual elements may be nested so, for example, we could define a container <CITATION> . . . </CITATION> with sub elements <AUTHOR> . . . </AUTHOR>, <YEAR> . . . </YEAR>, <PAGES> . . . </PAGES>, and so on. SGML tags may be also qualified by attributes as in <CITATION type=Book> which might make certain sub-elements mandatory, *e.g.* <PUBLISHER>. Each SGML document, known as an SGML document entity, consists of three components, an SGML declaration defining the syntax and character set used within the document, the DTD defining the logical structure of the document, and the text itself.

4.2.2 Chemical Markup Language

Chemical Markup Language (CML)[13] follows in the tradition of SGML. Whereas SGML has been primarily used for text and simple tables, CML allows its extension to numeric data and molecular systems. Syntax (the actual characters and their interpretation, the abstract structure of the document) is separated from semantics (the value put on words, context, attributes, *etc.*). Doing so gives chemical documents a much richer structure (*e.g.* for indexing, searching, *etc.*). Furthermore such documents can be linked to other documents rigorously. One important result is that terms in CML documents can be linked to publicly available glossaries. CML consists of three parts (in ascending hierarchy)

- HTML for the description of hypertext. The HTML used in CML strictly adheres to published standards.
- TecML (Technical Markup Language) which supports generic data in the STM field (*e.g.* numeric quantities with units, structured data, figures, bibliographies, other standards, *etc.*)
- MOL which is a specific DTD for molecular, atomic or crystallographic information.

There are quite general, so that markup might appear as:

```
<XVAR TITLE="Heat of Formation", DICTNAME="DeltaHF" CONVENTION="Thermochemistry" UNITS=kilocalorie/mole−1">12.34</XVAR>
or
<MOL TITLE="1,3,5-cyclohexatriene", DICTNAME="benzene" CONVENTION="http://www.some.where.org/terminology/organic" </MOL>
```

CML documents can have a very flexible structure which allows a large and finely grained body of chemical information to be encoded.

- Instrument output (*e.g.* spectra and crystallography).
- Program output (*e.g.* molecular orbital calculations).
- Database entries.
- Chemical journal articles, including molecules and reactions, numeric data and bibliography

In the case of a journal article for example, it is estimated that some 1000 separate chemically relevant data items might be identified *via* CML encoding. Extendibility comes through creating new document structures and adding terms to the glossaries, both of which can be done without redefining the format. An overview of the role of a language such as CML in the area of chemistry is shown in Fig. 2.

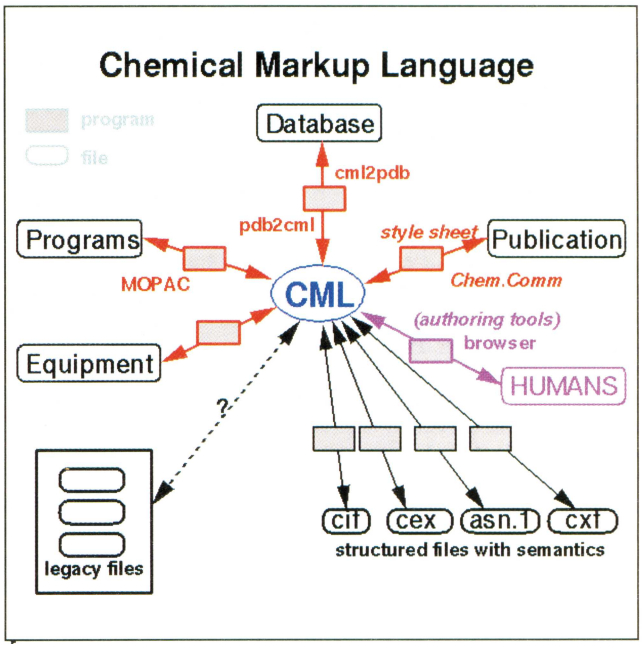


Fig. 2 A schematic showing the role of chemical markup language

4.2.3 Virtual Reality Modelling Language

In December 1994, a three-dimensional object description language called VRML or Virtual Reality Modelling Language was standardised.[17] VRML was designed to describe the relationships between sets of three-dimensional graphical objects such as spheres, cylinders and other primitives in much the same way that HTML and CML can be used to describe

textual and chemical entities. Support for VRML can be either using a separate browser program or more commonly using a browser plug-in and allows the user to navigate through complex 3D objects. A VRML encoded document fully supports the hyperlink concept *via* URLs. Thus, a molecule described using VRML can have hyperlinks associated with various atoms, or larger groups, and thus a bidirectional information flow between say an HTML and a VRML encoded document can be achieved, with each invoking the other as necessary. The most recent version of VRML introduced the concept of including scripted procedures for specifying actions on individual objects, and work is in progress to prototype a definition of chemical objects *via* a chemical extension of Open Inventor known as Molecular Inventor.[18]

A number of interesting chemical applications of VRML have been published in the last two years.[19] The earliest use was by Casher and coworkers to demonstrate how a so-called molecular collaboratory could be constructed.[19a] This involved the use of high-speed national computer networks combined with videoconferencing techniques to allow two or more groups of collaborators to simultaneously explore complex molecules. Associated properties such as computed molecular orbitals or electrostatic potentials are also well handled by this technique.[19b] Vollhardt and Brickmann[19c] have used VRML extensively to describe other complex molecular surfaces and properties including active site highlighting, and Robinson and Hardy[19d] have applied it *inter alia* to cell membrane model construction. The use of VRML is particularly appropriate where the context between a set of diverse three-dimensional properties needs to be analysed. For example, clustering analysis of intermolecular interactions derived from crystallographic information can be integrated with molecular diagrams, computed electronic properties and bibliographic information about individual molecules[19e] (Fig. 3).

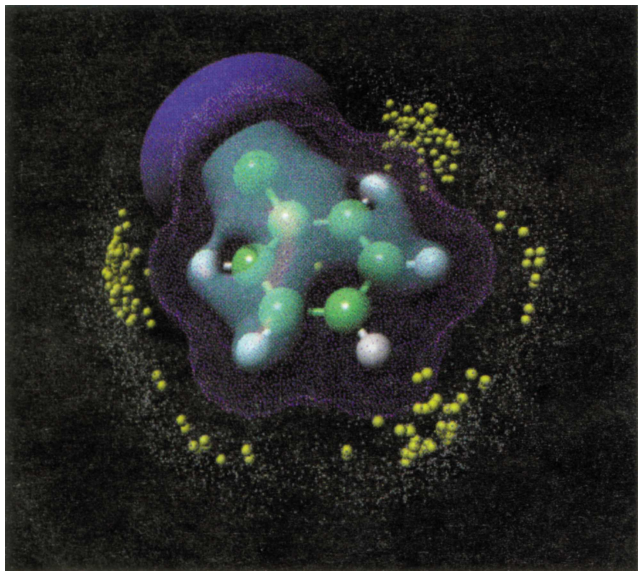


Fig. 3 A composite VRML scene showing a ball and stick molecular representation, with superimposed computed electrostatic potential difference map and various categories of intermolecular-contact 3D scatter diagrams derived from crystallographic data. Many of the objects are hyperlinked to other information. The example is available in both VRML 1 and VRML 2 from the active Web version of this article.

4.2.4 Object-oriented languages: Java

Many of the relatively old chemical descriptors enshrined in the chemical MIME types discussed above require specialist programs for their display. The availability of programs such as RasMol[10] or developed plug-ins such as Chime or WebLab played an important early role in the chemical development of the Web, but there are several aspects of this concept that are undesirable. First, such programs or plug-ins are platform and operating system specific. Secondly, exchanging chemical information between two or more different helper programs or plug-ins is even more problematic, since no well established standards exist for doing so. Chemically descriptive languages such as CML have a highly hierarchical and object oriented structure and it would be desirable to map such data objects to well-defined software objects that can act appropriately upon the data. One object-oriented language which allows such connections to be made is called Java.[20] It is in fact possible to map on a 1 to 1 basis the semantic content in say a CML document to corresponding Java components, called classes.[13]

The Java language allows the creation of a stand-alone application, but more interestingly of a network-portable version called an applet. An applet is downloaded to the client upon request and executed locally within a Web browser. Unlike browser plug-ins, Java code is a meta-language and hence platform independent, requiring only a Java-compliant local environment. This can be a Web client such as Netscape or it can be a feature of the operating system itself. This eliminates the often complex step of a user having to install a new program on their local computer system. If a document written in a language such as CML is being viewed, only those Java applets that relate to the document content need be acquired. If the reader wishes to view an infrared spectrum for example, only the Java objects that can read and display such data need be acquired.

Java applets are relatively new technology. Their immediate appeal has resulted in an explosion of interest and development, and in late 1996 there were already a number of excellent applets freely available on the Web,[21] including sketching applications which allow structures to be defined in database queries (Table 3). An organisation called the Open Molecule Foundation[22] has been established to serve as one focal point for identifying, resourcing and coordinating development in this area.

5 Applications to chemical conferencing

In this final section, we concentrate on two areas of scientific collaboration and dissemination which have been significantly influenced by the technologies and infrastructures described above. There are of course many other areas which space does not permit us to discuss, including the impact of the Web on teaching practice, the development of commercial models in this environment and aspects such as copyright, privacy and data protection, peer-review and authentication.

In November 1994, the first entirely Web-based chemical conferences, one on chemometrics and the other on computational chemistry were organised (Table 5). Their description as a 'conference' seems to have persisted, although it seems best to regard such events as representing something of a new genre for scientific exchange. The novel aspect of these conferences was their integrated use of various Internet-based technologies such as thematic electronic mail, conference articles presented as HTML documents, and the use of a 'MOO' (Multi-user Object Oriented) text-based interactive discussion environment. ECCC-1[23] was the first to adopt the use of chemical MIME to present 3D molecular models to the reader, and indeed was the first such conference where the proceedings were subsequently published on CD-ROM and recognised by Chemical Abstracts. This model was also followed in 1995 by the ECTOC-1 conference on organic chemistry,[24] which introduced keynote articles, full text-based keyword searching and an integrated molecular hyperglossary which could be searched by chemical sub-structure. Participants could also look at the access statistics to the conference to find out which were the most popular articles, and submit to a conference photographic collage. The second conference in this series (ECHET96) further developed the genre by allowing authors to edit their own contributions interactively *via* a Web interface, and by the use of molecular

Table 5 A selection of electronic conferences held on the World-wide Web from 1994–1996

Conference	URL	Published proceedings
InCIN'94. Chemometrics	http://www emsl pnl gov.2080/docs/incinc/homepage html	Various Journals
ECCC-1. Computational Chemistry	No longer available	S M Bachrach, W Hase, D B Boyd, H S Rzepa and S K Gray, (ed), *Proceedings of the First Electronic Computational Chemistry Conference*, (ARInternet Corp., Landover, Md), 1995
ECTOC-1· Organic Chemistry	http://www.ch.ic ac uk//ectoc//	H S. Rzepa, C Leach and J. M Goodman, *Electronic Conference on Trends in Organic Chemistry*, Royal Society of Chemistry, ISBN 0 85404 899 5, 1996
ECG-1: Glycoscience	http://bellatrix.pcl.ox.ac uk/egc/Welcome.html	B H Wilson, B. J. Hardy, *Tr Glycoscience Glycotechnology*, 1996, **8**, 301 Proceedings in *Glycoconjugate J*
TMMec: Molecular Modelling	http://uqbar.nciferf.gov/agora/welcome.html	Various Journals
ECCC-2 Computational Chemistry	http //hackberry chem niu edu/ECCC2/	*Theochem*
ECHET96· Heterocyclic Chemistry	http://www.ch ic.ac uk/ectoc/echet96/	H S Rzepa, J Snyder and C. Leach (ed), *Electronic Conference on Heterocyclic Chemistry*, Royal Society of Chemistry, 1997, to be published
e-MGMS Molecular Modelling and Graphics	http://bellatrix.pcl.ox.ac.uk/mgms/Welcome.html	*J Mol Graphics*

similarity clustering,[25] offering suggestions to participants on papers that might be related to the one they were reading. Due possibly to their relative novelty, these early conferences accumulated impressive access statistics, typically involving in excess of ten thousand visitors in a 1–2 month interval.

Events such as the Electronic Glycoscience conference (ECG-1) evolved a slightly different metaphor, in which the various themes of the conference were mapped onto an imaginary location analogous to a real conference venue. Participants had to register (at that time for no fee) before they could browse, and interactive MOOs were used extensively for real time discussions of individual contributions and other themes. Unlike previous e-conferences, the ECG-1 proceedings were subsequently published in a conventional printed journal, and this trend has been continued with subsequent events by the same organisers (ECG-2, e-MGMS).

The preceding examples were of conferences conducted exclusively on the Internet. However, such electronic enhancement has had a valuable role as an adjunct to physical events, and since 1995, many international scientific conferences have offered a Web component.[26] This can range from making available before the start an electronic book of abstracts, an electronic registration desk at which participants could supply their personal details, and obtain travel and tourist information.

In late 1996, it seems likely that several models for electronic conferences will continue to evolve. All will continue to offer new forms of 'added value' functionality, some will be purely evanescent, with authors having the option of subsequently submitting their contribution to more conventional journals, others seem likely to operate on a commercial basis with registration fees being levied. Perhaps the ultimate future of

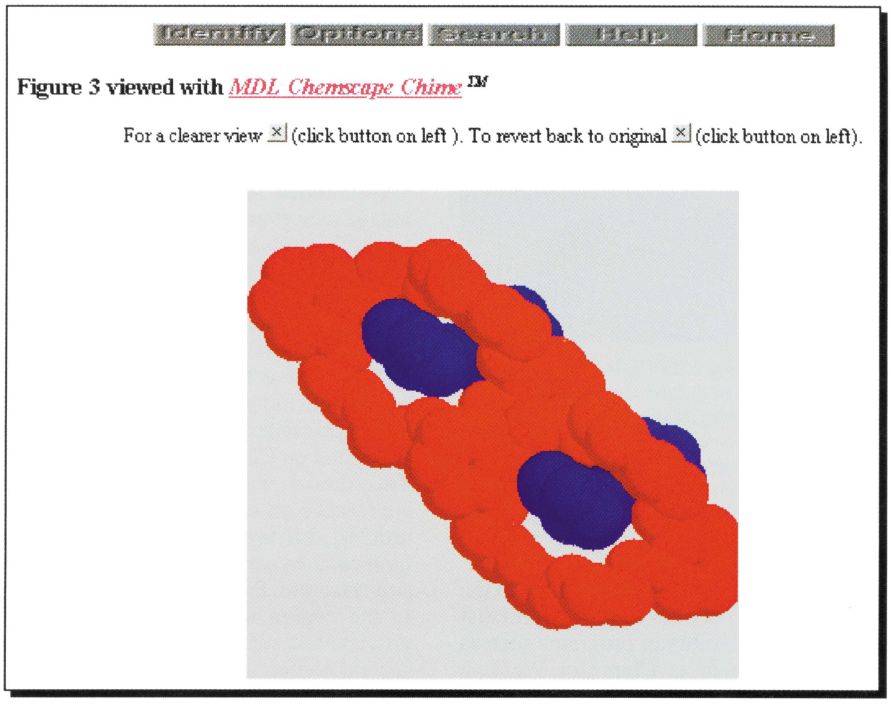


Fig. 4 Enhanced 3D representation as a component of an electronic article

such events is as part of a more complete information club in which the distinction between journals, conferences and informal discussion becomes progressively more blurred.

6 Innovation in chemical electronic journals

There are many issues to consider when applying an electronic medium to a scientific journal, from the perspective of the publisher, the author and the reader alike. The opportunities for all are obvious. From the publishers' point of view, to name but a few, there are opportunities to speed up the process of publication, to create 'value added' products and to monitor document access and usage by readers.

From the author's and the reader's perspective, the attraction lies in the novel ways in which data can be presented and interacted with. More fundamentally, the Web enables access to a wide variety of information on-demand, conveniently and efficiently. The drawbacks of the electronic medium can in the main be put down to lack of precedent. The technology is relatively new and is developing so rapidly that there has simply not been time to legislate for problems such as copyright, privacy and data protection. However, the obvious promise that this medium offers has led to a number of projects experimenting with the possibilities of electronic publication. These can be broadly divided into two types, those which derive from printed journals and which strive for electronic delivery with a degree of faithfulness to the printed form, and those where the prime focus is to enhance the journal concept without the constraint of requiring a precise printed equivalent.[23] Whilst experiments in the former category have been conducted for more than twenty years, the latter is a relatively new phenomenon. Here, we will focus on one such project of this type, the CLIC consortium initiative[15] to establish an electronic version of *Chemical Communications*, published by The Royal Society of Chemistry.

Although mainly used by structural and organic chemists *Chem. Commun.* is a general chemistry journal which aims to disseminate rapidly internationally important research results. Each issue also contains a longer feature article highlighting the work of a leading group. Whilst developing a strategy for the implementation of an electronic publication, several issues have arisen.

- Document definition and archiving—the electronic processing and indexing of submitted manuscripts
- Document delivery—the method by which documents are provided to the user.
- 'Added value'—exploiting to the greatest possible extent the features of the Web, providing information which, by its nature, could not be conveyed by the printed page.

The solution adopted in the first two areas by the RSC, along with many other publishing houses, has been to make use of SGML-derived technologies. This is a complex technical area, and a detailed discussion of this is beyond the scope of the present review.[28] Instead we will focus on some of the 'added value' aspects that the use of the World-Wide Web enables. Examples of added value could be to embed interactive 3D molecular viewers, using VRML or a mechanism based on MIME types to automatically launch external applications when a piece of embedded data is accessed by the user. NMR spectral data, reaction schemes and equations can also be treated similarly. We have called this concept 'hyperactive molecules' and demonstrated it in a number of applications.[7]

Indeed, one can envisage any object in an article becoming an interactive feature, *e.g.* equations could be dynamically linked to graphing or symbolic algebra tools. Other possible features of an electronic journal include searching and retrieving data on-the-fly, indexing, hyperglossaries, keyword searching and forward referencing. All but the latter are self evident—forward referencing is the concept that a document in the archive can be dynamically updated whenever a subsequent document refers back to it so generating a hypertextual link 'forward in time'.

Examples of what can be achieved[29] are the enhanced versions of feature articles taken from *Chem. Commun.* The paper by Stoddart and coworkers[30] describes recent advances in the area of the molecular design of interlocking molecules. In the electronic version figures in the paper version depicting molecular structures become animated three-dimensional models using the Chemscape Chime plug-in. Fig. 4 depicts this feature of the article as it appears on the readers screen. Other features in the demonstration article include hypertextual markup of spectral information and extensive cross-referencing.

Projects such as CLIC demonstrate that true electronic chemistry journals are in the very early stages of their development cycle, and that the very nature of many journals may change radically over the next five years or so.

7 Conclusions

Inevitably, in a very rapidly changing subject, this review can only present a very selective snapshot of how use was being made of the Internet and the World-Wide Web in late 1996. For this reason, we have tried to focus on the basic infrastructures that have been built up over a three-year period, the aspects of chemistry that have been enhanced as a consequence and our manifesto for how we feel the area should develop in the future. Many challenges, of course, remain to be solved, not least of which are the sociological ones in enfranchising and persuading the global chemical community to participate.

8 Acknowledgements

The authors acknowledge the following sources of the funding of projects in the areas described above; the Joint Information Systems Committee (JISC) e-Lib and JTAP programs, British Telecom and GlaxoWellcome. We are also grateful to our coworkers, in particular O. Casher, C. Leach, C. S. Page, C. Hildyard, S. Green, D. Riddick, J. Griffiths and D. James.

Glossary

Term	Brief description
Bookmark	A user pointer to a previously visited World-Wide Web document
Browser	A Client program for viewing World-Wide Web documents
CERN	European Laboratory for Particle Physics
CGI	Common Gateway Interface for scripts and other external programs
CML	Chemical Markup Language
DTD	Document Type Description
FORM	A graphical interface for interacting with WWW services
GIF	Graphical Interchange format (Compuserve)
Helper	An independent program invoked from a WWW browser
HTML	Hypertext-markup-language
HTTP	Hypertext-transport-protocol
Hyperglossary	A glossary of terms created and accessed via the WWW
Hyperlink	A pointer within a document to another network object
Inline	An image inserted within text and displayed by a WWW browser
IP	Internet Protocol, currently Version 4, but migrating to Version 6
Java	An object-orientated language with characteristics suited for the Internet
Javascript	A scripting language with some characteristics related to Java
JPEG	Joint Photographic Experts Group image format
MIME	Multipurpose Internet Mail Extension, used as two-level descriptor for data files
PHP	Personal home page on the World-Wide Web

Plug-in	A small program which enhances the capability of some Browsers
Server	A program providing access to a database of WWW documents
SGML	Standard Generalised Markup Language
URL	Uniform Resource Locator, used within HTML to located hyperlinks
VRML	Virtual Reality Modelling Language, a 3D scene descriptive language
Web	World-Wide Web
Webmaster	The administrator of a World-Wide Web server
WWW	World-Wide Web

9 References

1 H. S. Rzepa, *Science Progr.*, 1996, **79**, 97.

2 The on-line version of this article is available from http://chemistry.rsc.org/rsc/csr398.htm. For an early review of the chemical aspects of the World-Wide Web, see H. S. Rzepa, in *The Internet: A Guide for Chemists*, ed. S. Bachrach, Am. Chem. Soc., 1995.

3 T. J. Berners-Lee, R. Cailliau, J.-F. Groff and B. Pollermann, CERN, *Electronic Networking: Research, Applications and Policy*, Meckler Publishing, Westport, CT, USA, 1992, vol. 2, p. 52.

4 D. Raggett, J. Lam and I. Alexander, *HTML 3—Electronic Publishing on the World-Wide Web*, Addison Wesley, 1996. A specification for HTML3.2 can be found at http://www.w3.org/pub/WWW/MarkUp/Wilbur/

5 W. D. Ihlenfeldt, 1996, *The WWW Chemical Structures Database*, http://schiele.organik.uni-erlangen.de/services/webmol.html

6 J. Brecher, 1996, *Chemical errors found on WWW sites*, http://www.camsoft;com/chemfinder/errorsfound;html;

7 (*a*) H. S. Rzepa, B. J. Whitaker and M. J. Winter, *J. Chem. Soc., Chem. Commun.*, 1994, 1907; (*b*) O. Casher, G. Chandramohan, M. Hargreaves, C. Leach, P. Murray-Rust, R. Sayle, H. S. Rzepa and B. J. Whitaker, *J. Chem. Soc., Perkin Trans 2*, 1995, 7.

8 N. Borenstein and N. Freed, *MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies*, Internet RFC 1521, Bellcore, Innosoft, September 1993.

9 H. S. Rzepa, P. Murray-Rust and B. J. Whitaker, *Pure Appl. Chem.*, 1997, in preparation. For the background to this project, together with a current list of the MIME types, see http://www.ch.ic.ac.uk/chemime/

10 R. A. Sayle and E. J. Milnerwhite, *Tr. Biochem. Sci.*, 1995, **20**, 374.

11 Available from MDL Information systems, http://www.mdli.com/

12 Available from Molecular Simulations Inc., http://www.msi.com/

13 P. Murray-Rust, Proc. Nimes Conf. 1995. The main CML site is at http://www.venus.co.uk/omf/cml/

14 For a general collection of examples, see http://www.mdli.com/chemscape/chime/sampidx.html For specific examples, see (*a*) J. R. Grant, G. M. Helmkamp, Jr. A. B. Rawitch, S. H. Goodman and G. L. Kern, Kansas University, http://www.kumc.edu/research/medicine/biochemistry/bioc800/start.html; (*b*) H. S. Rzepa, Imperial College, http://www.ch.ic.ac.uk/motm/; (*c*) S. Green and B. J. Whitaker, University of Leeds, http://chem.leed-s.ac.uk/Project/Teaching/cadd.html; (*d*) K. Harrison, University of Oxford, http://www.chem.ox.ac.uk/mom/taxol/taxol.html; (*e*) P. Nixon, H. S. Rzepa and C. Leach, Imperial College, http://www.ch.ic.ac.uk/hyperactive/1prc.html

15 D. James, B. J. Whitaker, C. Hildyard, H. S. Rzepa, O. Casher, J. M. Goodman, D. Riddick and P. Murray-Rust, *New Rev. Information Networking*, 1995, 61. Information about the project can be found at http://chemcomm.clic.ac.uk/clic/

16 W. P. Walters, M. T. Stahl, A. V. Shah and D. P. Dolata, *Abstr. Papers Am. Chem. Soc.*, 1994, **207**, 36-Cinf, 37-Cinf.

17 M. Pesce, *VRML Browsing and Building Cyberspace*, New Riders ISBN 1-56205-498-8, 1995. For a current bibliography, see http://www.sdsc.edu/vrml/books.html

18 M. Benzel, http://www.sgi.com/ChemBio/MolecularInventor; J. Krieger, *Chem. Eng. News*, 1996, **74**, 10.

19 (*a*) O. Casher and H. S. Rzepa, *Comp. Graphics*, 1995, **29**, 52; O. Casher and H. S. Rzepa, *J. Mol. Graphics*, 1995, **13**, 268; (*b*) G. A. Suñer, O. Casher and H. S. Rzepa, *Electronic Conference on Trends in Organic Chemistry*, ed. H. S. Rzepa, J. M. Goodman and C. Leach, Royal Society of Chemistry, ISBN 0-85404-899-5, 1996; (*c*) H. Vollhardt, C. Henn, G. Moeckel, M. Teschner and J. Brickmann, *J. Mol. Graphics*, 1995, **13**, 368; (*d*) B. J. Hardy and A. Robinson, *Theochem.*, 1996, **368**, 111; (*e*) O. Casher, C. Leach, C. S. Page and H. S. Rzepa, *ibid.*, 1996, **368**, 49. See http://www.elsevier.nl:80/section/chemical/theochem/eccc2.htm for electronic versions of the latter two articles.

20 K. Arnold and J. Gosling, *The Java(tm) Programming Language*, ISBN 0-201-63455-4, Addison-Wesley, 1996; M. A. Hamilton, *Computer*, 1996, **29**, 31; E. Yourdon, *Computer*, 1996, **29**, 25.

21 For a comprehensive keyword searchable site, see http://www.gamelan.com/

22 For further details of the Open Molecule Foundation, see http://www.ch.ic.ac.uk/omf/

23 S. M. Bachrach, *J. Chem. Inf. Comp. Sci.*, 1995, **35**, 431.

24 H. S. Rzepa, *Tr. Analyt. Chem.*, 1995, **14**, 240.

25 G. M. Downs and P. Willett, *Rev. Comp. Chem.*, 1996, **7**, 1.

26 For a typical example, see 14th International Symposium on Gas Kinetics held at Leeds during September 7–12, 1996, http://www.chem.leeds.ac.uk/conferences/.

27 H. S. Rzepa, *Tr. Analyt. Chem.*, 1995, **14**, 464.

28 C. Hildyard and B. J. Whitaker, *Proceedings of the On-line meeting*, December 1996, London.

29 These articles were prepared by O. Casher, J. Griffiths and D. Riddick. See http://chemcomm.clic.ac.uk/cccenha.htm

30 P. T. Glink, C. Schiavo, J. F. Stoddart and D. J. Williams, *Chem. Commun.*, 1996, 1483. See http://chemcomm.clic.ac.uk/c08349_1.htm

Received, 10th October 1996
Accepted, 3rd November 1996