Recent Applications of Hyperactive Chemistry and the World-Wide-Web: Towards an Integrated Chemistry Information Environment

Henry S. Rzepa,* Omer Casher, Christopher Leach Department of Chemistry, Imperial College, London, SW7 2AY. and Peter Murray-Rust
Virtual School of Molecular Sciences

Article 6
New Initiatives in Chemical Education: An On-Line Symposium, June 3 to July 19, 1996

Abstract: We describe our model of a virtual chemistry information environment of inter-linked techniques, experimental and instrumental data, and information sources held at Imperial College and elsewhere. We have made use of a number of recent World-Wide Web technologies for incorporating active chemical information into documents, including chemical MIME standards, chemical structure markup language (CSML), Virtual Reality Modelling language (VRML), Java applets and Chemical Markup Language (CML). The overall objective is that the user will be able to follow a chemical "thread" to acquire an integrated portfolio of "hyperactive" molecular information, replacing the more traditional outcome of handwritten notes and photocopies of primary information found in more traditional libraries.

Information for browsing this article

Towards an Integrated Chemistry Information Environment: Introduction

1. Introduction

Although experiments in on-line provision of chemical information and journals can be traced back to the early 1970s, such services were introduced into most chemistry departments only in the mid 1980s, often via only a single low bandwidth "point of presence". The use of such services in taught courses has been much more variable, largely because significant, and in a teaching environment unmanageable, costs were associated with these services. During this period, the standard user interface was mostly restricted to either the 24 line by 80 character telnet "VT100" terminal mode, or Tektronix 4014 vector graphics mode. There was little integration between various information services from the point of view of query formulation or chemical structure definition. Equally variable was the quality of documentation and on-line help, too often depending purely on the user having access to printed material.

Other significant limitations were the lack of integration of information sources into other laboratory based excercises or molecular modelling themes, and on a wider scale of the incorporation of related projects being conducted in other teaching faculties around the world.

Recently, solutions to such problems have evolved in both a commercial and an educational context. The commercial model is an interesting one, in that it must of necessity evolve around robust and affordable charging models. For example, Current Science has recently launched the "BioMedNet" club, which offers an environment in which subscribers can browse electronic journals, perform keyword searches, and have access to other network resources in a self-consistent and "user-friendly" manner. This club makes use of standard software such as a World-Wide web client and an assumed Internet connectivity to provide access. A rather different, and very much more proprietary model is the "SciFinder" interface to the Chemical Abstracts database, representing the latest stage in the 15 year evolution of on-line services provided by this organisation. SciFinder in its current state of development is very much a closed turn-key client-single server system which does not appear to offer a viable model for the implementation of any local teaching resources. Moreover, the current cost of subscribing to such a service would represent a very significant increase in most library or teaching budgets, at a time when these budgets are under severe pressure to contract. The focus of SciFinder on the commercial sector means that this charging model fits with difficulty into any teaching environment.

A quite different open approach is rapidly evolving in many teaching institutions and is based on a client-multiple server model known as the World-Wide Web system which you are using to view this document. The Web originated in 1989 at the European Laboratory for Particle Physics (CERN) with the first definition of HTML or Hypertext-markup-language and a transport protocol called HTTP (Hypertext-Trasport-Protocol). The participation of the National Center for Supercomputing Applications (NCSA) in 1993 introduced a Web client called Mosaic, which allowed a combination of text and two dimensional images to be used to create a cohesive environment for describing various information services. The real technical innovation of the Web over earlier hypertext systems was the introduction of a global resource locator known as a URL (Uniform Resource Locator), which allows a section of text or a graphic to be seamlessly linked to other relevant documents or resources anywhere on the Internet.

Starting in September 1993, a significant chemical presence began to build up using these technologies [1]. The "critical mass" was probably achieved in 1995, when for the first time it became possible to devise an experiment in molecular information retrieval which could be completely integrated, not merely on a local but on a global scale, with other chemical resources [2]. In this article, we describe our own

implementation of an experiment in molecular information retrieval in a teaching environment which takes into account the increasing molecular richness and diversity of the Internet.

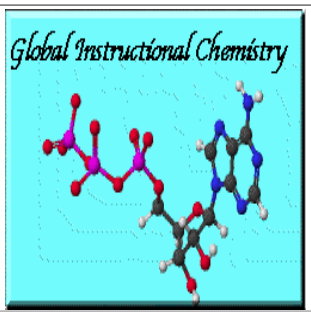
2. The Basic Chemical Web Technologies

The last two years has seen the introduction of a number of World Wide Web clients designed to display the contents of a HTML document written in hypertext markup language or HTML. Programs such as NCSA Mosaic, Microsoft Internet Explorer, Netscape Navigator or Apple Cyberdog allow the two dimensional page metaphor used in traditional chemistry texts to be translated to on-line form. However, this is no reason that we should continue to be bounded by two dimensions. From the outset, we considered it necessary to develop an infra-structure for linking documents written in HTML with chemically specific datafiles, which could be processed in an explicitly molecular manner by the user. In this section, we discuss various methods that we have evolved over the last two years for more closely integrating chemistry into the traditional "document".

2.1 Chemical MIME

Our first solution to this problem was to adopt a mechanism derived from mail handling programs called MIME or Multipurpose Internet Mail Extensions [3]. This mechanism is integrated in a generic manner into most HTML browsers. Our particular implementation of this was termed chemical MIME [4]. It enables a browser to pass on any documents of an explictly chemical nature to a program of the user's choice present on their computer. This means that HTML documents can contain hyperlinks to chemical data, which can then be displayed in a visual manner which a generic HTML browser is incapable of. A typical example is a hyperlink to a "pdb" file containing 3D molecular coordinates, which can be displayed using an external program such as RasMol or as an "in-lined" molecule using Chemscape Chime (Figure 1).

Figure 1. Adenosine Triphosphate displayed using Chemical MIME.



If you have installed Chemscape Chime as a "plug-in" to Netscape 2.0, this molecule should appear as a rotating image.	If you are using other WWW clients, and have configured the chemical MIME type as `chemical/x-pdb`, clicking on the thumbnail image will activate the molecule.

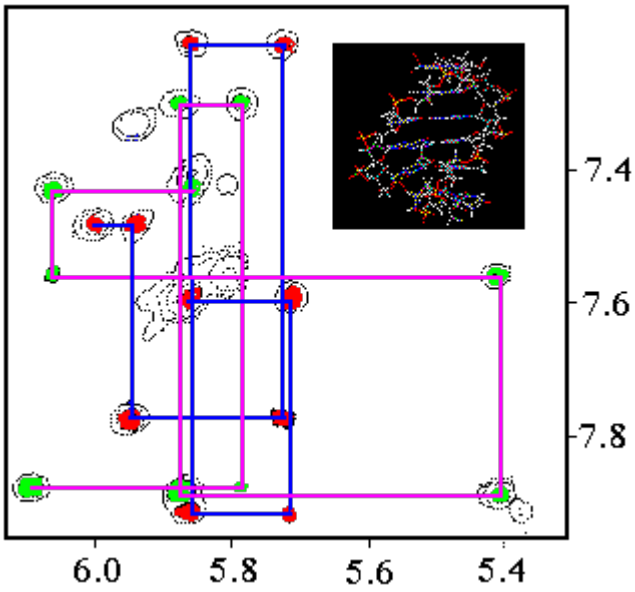
[Abstract |1: Introduction | 2: The Basic Chemical Web Technologies|MIME |CSML |VRML | Java | CML | The Virtual Chemistry Library | The Future | Acknowledgements and References | What's New]

2.2 Chemical Structure Markup Language.

Whilst such an approach is capable of adding a rich seam of chemical content to a document, there are specific limitations which soon become apparent. Because programs such as RasMol or derived plug-ins such as Chime cannot themselves resolve hyperlinks, the chemically specific document becomes something of a cul-de-sac into which further hyperlinks cannot easily be inserted. For example, one might wish to

have a hyperlink in the master HTML document which when invoked might highlight one specific atom or functional group in the molecule display. To accomplish this, it is necessary to establish further subsequent communication from the original HTML document to the chemical display window. Our original solution to this specific problem was to develop what we termed "CSML" or chemical-structure-markup-language [3b], achieving communication between the HTML browser and RasMol using a feature built into the Unix version of Rasmol, by which a script can communicate with a running RasMol process. By this means, we were able to associate peaks in a 2D NMR spectrum displayed in an HTML document with the individual protons responsible highlighted in a molecule display window.(Figure 2). The user could navigate around the spectrum using a device known as an "image-map", identifying individual proton pairs as they went. Subequently, the CSML mechanism has also been integrated into the Chemscape Chime plug-in, and applied to a "molecule-of-the-month" current awareness collection at Imperial College. Such annotation provides a powerful new teaching tool for use on the Web.

Figure 2. The Partial 2D NOESY Proton NMR Spectrum of the oligonucleotide CGCGTTTTCGCG illustrating the Application of CSML



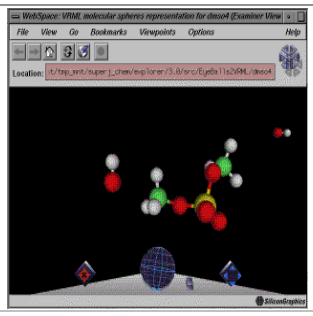
This represents a "clickable map", locally resolvable if you use Netscape V2, or remotely if you use other browsers. Activate the molecule first, and then spectral cross peaks to "annotate"the RasMol view by highlighting selected protons

c1-c1 c1-g2 g2-g2 g2-c3 c3-c3 c3-g4 g4-g4 g4-t5 t5-t5 t5-t6	t7-t7 t7-t8 t8-t8 t8-c9 c9-c9 c9-g10 g10-g10 g10-c11 c11-c11 c11-g12	
Because "clickable maps" cannot link to "in-lined" molecule displays, the list above represents an alternative way of achieving the same result. If you have Chemscape Chime installed, clicking on these links will highlight individual pairs of protons.		The Chemscape Chime display associated with the list of proton contacts on the left

2.3 Virtual Reality Modelling Language

The information display mechanisms described thus far represent essentially a one-directional communication between a hyperlinked document and a molecular visualiser. There is no capability for reversing the direction, from a "marked-up" molecule to HTML or other documents. Two recent developments offer solutions to this problem. During 1995, a three dimensional object description language called VRML or Virtual Reality Modelling Language was introduced. If HTML is thought of as a language used to choreograph the two dimensional ASCII character set, then VRML would correspond to a similar description of a set of three dimensional objects such as spheres, cylinders and other primitive graphical objects. A VRML browser can display these objects in 3D space, and the user can navigate around in this space. Unlike a custom display program such as RasMol, VRML browsers also fully support 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 document can now be achieved, with each invoking the other. As with Rasmol, the VRML scene can be rendered in either a separate window, or as an in-line image using a "plug-in".

Figure 3. Dimethyl sulfate encoded in VRML, containing embedded hyperlinks associated with individual atoms and bonds illustrating the hydrolysis of this species.

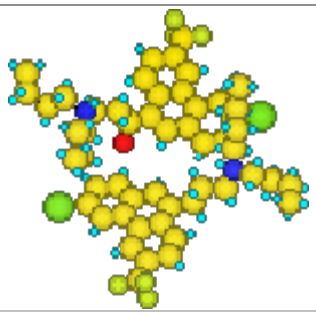


2.4 Java Applets

Currently, VRML in version 1.0 supports no chemical semantics, i.e. bonds and atoms are not explicitly identified as such, and hence the way they are displayed cannot be changed. Java is a programming language which allows the molecular display code (e.g. RasMol), the display data (e.g. a pdb file) and the hyperlink communication to be built into a single file, or "applet". The applet window is in-lined into the main body of the HTML document. Furthermore, two or more Java applets can establish mutual communication, such that a 2D NMR spectrum can be associated with a 3D rotatable model of the corresponding molecule, with appropriate atoms again highlighted. Thus Java allows small compact applets to be written by users for a specific task. In this, it does not necessarily supercede a specialised display program such as Rasmol, and all three mechanisms outlined above have their particular roles to play in the creating of a rich chemical environment for the user.

Because Java is highly customisable, and also secure, several other issues come to the fore which the community will need to solve. Firstly, is the recognition that two or more Java Applets may need to intercommunicate. To achieve this, chemical standards will have to be created to allow this to happen easily and seamlessly. Secondly, some mechanism for indexing the action and content of a Java applet will need to be created. Such issues also apply to the VRML concepts outlined above. We envisage the major thrust of such work coming from the commercial software developers, but perhaps with an impartial standards body set up to attempt to control the evolution.

Figure 4. A Molecule Rendered using a Java Applet



If your WWW client is "Java-aware", the image you see should be rotatable. If your client does not support Java, a simple static image will be present.

2.5 Chemical Markup Language

The technologies covered thus far relate to the visualisation and interpretation of molecular coordinate data, with spectral data represented as simple bit-mapped images. However, the variety of disciplines and techniques that chemistry covers is enormous, so it's not surprising that information exchange between different types of molecular datafile is difficult.

It is generally accepted that the best way to tackle these problems is through the use of markup languages. You are reading a markup language (HTML) at the moment! Markup languages add meta- information to a document to tell the recipient more about it. In this spirit, we have started to develop what is termed Chemical Markup language [5].

CML consists of three parts (in ascending hierarchy):

- HTML (strict), for the description of hypertext.
- XML (Xperimental Markup Language), which supports generic data in the STM field (e.g. numeric quantities with units, structured data, figures, bibliographies, other standards, etc.)
- MOL (a specific DTD for molecular, atomic or crystallographic information).

These are quite general, so that markup might appear as

```
<X.VAR TITLE="Heat of Evaporation" REL="glossary"
HREF="/chem/theor?deltahevap"
UNITS="kilocalorie/mole">34.12</X.VAR>
```

The most important result of this is that a very large body of current chemical information can be encoded with CML. CML documents can have a very flexible structure and have already been used to describe precisely:

- Instrument output (e.g spectra and
- crystallography). Program output (e.g.
- molecular orbital calculations). Database entries.
- Publications (management of whole papers is already tractable).

In the future, we expect mechanisms such as this to achive a closer intergation of virtual chemistry libraries

Towards an Integrated Chemistry Information Environment: Discussion

3.1 A collection of On-line Chemical Information Sources

A virtual chemistry information environment must to a significant extent still depend on a variety of technologies, both old and new. Here we describe a model that was first used in an undergraduate environment in October 1994. The library consists of an integrated series of information sources, in which the information is presented to the user in a variety of methods, ranging from simple text, text in the form of HTML with integrated hyperlinks, "forms" where the user can provide feedback, 2D images, 3D molecular datasets to be displayed as embedded images, 3D "scenes" in VRML with embedded hyperlinks to further information, and documents containing Java "applets" where chemical semantics can be coded into the document. Clearly, this concept is still at a very early stage, and most of the information sources currently available implement very little of this technology. The following section represents a snapshot of what is available in early 1996, and can be expected to evolve very rapidly indeed as more people develop specific content.

Each section in this virtual library is associated with an icon, as found in Table 1. A single mouse click on this icon will initiate connection to the data source or technique described. To obtain a fully-functioning library, you will have to set the Helper configuration first.

Table 1. The Virtual Chemistry Library

Start Point	Information Source	Program Invoked (if any)	Information about the Supplier
	Starting the Electronic Notebook	Claris Works	Claris Works
	Search the Chemistry Virtual Library		Search the World using Lycos
	A guided tour of the World-Wide Web Information system		The World-Wide Web Organisation
	The BIDS Science Citation database	Telnet	Bath Information Delivery System
	The CAS On-line system	Telnet or SciFinder	Chemical Abstracts
	The "WebSpirs" bibliographic database system		Silver Platter
	Safety information/chemical availability		The Fisher Catalogue
	The World Drug Index and Savant		Daylight Information systems
	The ChemFinder system		CambridgeSoft

			
	The Crossfire system	Beilstein Commander	Beilstein
	The ISIS/Base reaction database	ISIS/Draw and Base	MDL Information Systems
	The Cambridge Crystal Structure database	X-Window Server	CCDC
	The Virtual 3D Library	VRML and/or Chime plug-in	The vchemlib project
	Electronic Conferences and Journals		The CLIC Project

3.2 Navigating the Library

The visitor to the virtual library enters by a "door" provided by a World-Wide Web client such as Netscape, Mosaic, Internet Explorer, etc. They collect the "shopping trolley" by opening up a "notebook" such as Claris Works, in which they will collect bibliographic and graphical information as they proceed with their search.

The visitor starts with simple "one dimensional" keyword searches to retrieve simple bibliographic information. Modern implementations of such keyword search systems use the "forms" HTML interface, which offers a self-contained user interface. The Lycos search system offers perhaps the most comprehensive general catalog of the Internet, and allows two or more keywords to be specified using boolean-like logic. Lycos indexing is performed by an automatic search robot, which is programmed to look specifically at the content of HTML documents. In particular, only the content of documents enclosed by <html> and </html> tags is indexed, which would exclude detailed indexing of any document with chemical content, such as for example a 3D coordinate file. A more recent global search index is Alta Vista introduces some novel features deriving from the structure of the Web itself. For example, one can search for documents which "cite" a particular URL (Uniform resource locator), much in the same way that the Science citation index can be used to follow a "thread" of chemical information.

Other more chemically specific bibliographic searches are offered to the visitor, including the Science Citation Index itself via a UK wide national licensing scheme referred to as "BIDS", demonstration files on CAS On-line system, and more specific databases such as the Fisher catalogue, and "samplers" of the WebSpirs system from Silver Platter. The results of such searches are normally presented as text-based documents. There is little the user can do with such information other than copy-paste the text into a word processor.

To achieve keyword retrieval in a chemical context, one has currently to enter a turn-key system designed to perform this task. Examples include the Daylight system, which allows retrival of structural information using SMILES strings as the search term, and the CambridgeSoft ChemFinder system, which allows a variety of chemical and keyword search terms to be specified. Both these services offer small "sampler" databases via a "forms" interface for the user to experiment with. Here, the results of a search are presented in not only a graphical form, generated in real-time by background programs and scripts, but one also has the opportunity of acquiring 2D or 3D coordinate and connectivity data using the MIME mechanism referred to previously. This allows the user to open a separate molecular window using local programs, and if necessary to save the information on their local disk. Another MIME implementation involves acquiring reaction or 3D query definitions from a remote site, and using these to search a locally implemented

database (or one operating a local client/remote server system). We have implemented this with the MDLI ISIS/Base system for searching for synthetic transformations of penicillins using a search definition saved in the "TGF" format, for the Beilstein Crossfire system for searching for molecular properties and for the Cambridge crystal structure database.

These types of search will need various local tools and programs to help with the searching, and these more specialised programs will be invoked from various hyperlinks via MIME definitions. In addition a "telnet" terminal emulator for bibliographic searchs which do not support the "forms" interface is required.

Table 2. MIME Definitions for Activating Chemical Search Programs

MIME Type	Program Activated
application/x-claris	Claris Works
chemical/x-pdb	RasMol or Chime
chemical/x-mdl-tgf	ISIS Draw/Base

One way of avoiding the need for a plethora of additional "helper" programs is to provide the service via what are called "cgi-bin" (common gateway interface) programs, held on a central server. Such a route has been strongly advoated by Weininger, and illustrated via the Daylight CGI interface. Such a service was also a component of the ECTOC conference for creating a hyperglossary of molecular coordinates and other information. Such a hyperglossary enables auxillary information to be stored to complement conferences and electronic journals. It can hold chemical information, whether it be textual, 2D or even 3D. Once entered into the hyperglossary, the information is readily accessable by the rest of the community in a structured and indexed form. The electronic nature of this system allows the information contained to be indexed for easy searching and the complexity of the searches can range from simple text keyword to the complicated substructure search engine, depending on the sophistication of the server holding the information. The hyperglossary within ECTOC was a collection of molecules that were discussed during the conference. Another example of a hyperglossary in action is the one that supplemented the Internet course on The Principles of Protein Structure to hold a collection of terms on proteins. Both of these hyperglossaries allowed users to add their own contributions.

Future developments in this area include presenting the molecular data as a hyperlinked set of coordinate files encoded in the VRML format, with additional functionality provided by Java encoded applets and scripts illustrate specific topics in chemistry.

3.3 Application of these Techniques at Imperial College

Starting in October 1994, students in an advanced organic chemistry laboratory have had access to this integrated chemical information environment as part of the "techniques" component of their course. In the laboratory they will have converted a pencillin derivative to a cephalosporin, and purified this using chromatography. They are asked to perform a rounded literature search on this synthetic conversion and purification, and asked to find any information on these or related compounds that relates to safety, 3D structure etc.

During the course, the students have to select 8 techniques out of a menu of 19 available. The "IT" technique has proved very popular, with the majority opting to do it. Typically, the students will spend about 1 week of laboratory time learning the various techniques, and many continue to use the techniques during subsequent research projects and indeed Ph.D. programs.

The comment made most often by students is of the confusion brought about by using various user-interfaces for defining chemical structures and keyword searches, and the difficulty of transporting information across the divides introduced by the use of different programs. In 1996 we anticipate that some of these concerns will slowly be addressed as the suppliers of chemical information move to integrate their

own user-interfaces with those enabled by the Web. At a time of rapidly evolving Web technology, applying standards is inevitably difficult, but at least a small glimmer that this is possible is beginning to emerge. A fuller maturity of this system may however not be seen for several years yet.

3.4 The Future

The virtual chemistry library as currently available has many tantalising hints of what might be possible, but clearly much still remains to be achieved. Thus the library as implemented in April 1996 still requires computers endowed with prodigious amounts of memory which must support the many different program windows that the user might need to open during the course of a session. The custom programs listed in Table 2 are currently capable of little inter-communication with each other beyond simple cut-paste operations on text. Thus the user is still faced with a variety of user interfaces, both for bibliographic keyword searches, and for sub-structure drawing. Some standards do exist. For example, a search query in one system can produce results in the form of a SMILES string that can be used to initiate further searches in other systems. Both HTML and VRML define ways of encoding information which can be cross-referenced or hyperlinked to achieve a coherent theme. Java may provide a transparent interface to incorporate heuristic algorithms and other functionality which can operate on molecular information, spectra, etc and deliver it in customised form to the user.

Gradually, we envisage that modularity in the software components will be accomplished. Perhaps the highest priority now must be to create a structure editing interface as a "plug-in" (or an OpenDoc "part" for the Cyberdog WWW client) which would enable a much more seamless interface to be constructed. The recent announcement by Tripos of a Java applet called Sketch and Fetch appears to be the first such product which seamlessly intergrates into a Web page, although the extent of its modularity is not yet know. Both Java and VRML are now integrated into this environment, so the grand synthesis is gradually coming together.

Given the richness of the tools available now or in the near future, the prospect for developing exciting new methods of presenting molecular sciences looks very good. Over the next year or two, many of these mechanisms will mature into a product that will offer a sea change in the way molecular "information technology" is both used and taught.

Acknowlegements

We thank the many people who, over the Internet or in person, have helped this theme of "Internet Chemistry" to come together. We are also very grateful for financial support from GlaxoWellcome and for a studentship (to CL), to the JISC Electronic Libraries Program, and to British Telecom.

References and Citations

1. The Internet: A Guide for Chemists (Ed S. M. Bachrach), ACS publications, 1996.
2. For review articles, see B. M. Tissue, Distributing and Retrieving Chemical Information Using the World-Wide Web, Tr. Anal. Chemistry;(28 July, 1995); Y. Wolman ; Chemical education on the Internet, Tr. Anal. Chemistry (19 February 1996); G. Wiggins, Use Of The Internet In Teaching ChemicalInformation Courses, Paper 09, New Initiatives in Chemical Education.
3. N. Borenstein, and N. Freed, "MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, Bellcore, Innosoft, September 1993.
4. P. Murray-Rust, H. S. Rzepa and B. J. Whitaker, IETF Internet Draft, May-October 1995. See http://www.ch.ic.ac.uk/chemime/; 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. For a review of the history of the development of chemical MIME, see A. Davies, European Spectroscopy News, 1996, **8**(1), 42.
5. P. Murray-Rust. See http://www.venus.co.uk/OMF/cml06f/newintro/chem.html for details.