The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World Wide Web Information Exchange

Henry S. Rzepa,*[,†] Peter Murray-Rust,[‡] and Benjamin J. Whitaker[§]

Department of Chemistry, Imperial College, London, SW7 2AY, Virtual School of Molecular Science,
Department of Pharmacy, University of Nottingham, Nottingham, and School of Chemistry,
University of Leeds, Leeds, LS2 9JT, U.K.

Received January 6, 1998

The proposal and subsequent global use of an Internet standard based on **chemical** primary Multipurpose Internet Mail Extensions (chemical MIME) media type is reviewed. Examples of the configuration of this standard for use with Internet-based electronic mail and World Wide Web clients are shown. The long-term objectives of the integration and interoperability of chemical information across the boundaries of Internet-based electronic journals, conferences, virtual courses, databases, modeling, and newly emerging information handling and modeling tools are set out. We believe that one way forward is by concentrating on more finely grained chemical information components, using generic tools based on XML (eXtensible markup language) and its support in chemistry via CML (chemical markup language).

INTRODUCTION

The development of Internet-based document and information delivery systems during the last four years has been rapid,[1] with a particular focus on the creation and delivery of chemically oriented World Wide Web (WWW)-based documents. This has in turn introduced to many concepts such as the use of structured and interlinked document collections specified by languages such as HTML (hypertext markup language). This review will focus on one aspect of this revolution, the chemical application of an Internet standard known as MIME (multipurpose internet mail extensions) to the WWW and electronic mail (e-mail) handling. We will discuss how a transparent integration of e-mail- and WWW-based exchanges of chemical information with chemical modeling and information handling tools can be achieved using this mechanism, and present some ideas for how we believe development beyond the MIME mechanism should proceed.

MULTIPURPOSE INTERNET MAIL EXTENSIONS (MIME)

Despite the attention given to the development of the WWW, e-mail arguably remains the more frequently used mechanism for electronic information exchange by scientists. E-mail is often regarded, however, as a temporary and informal communication medium, not well suited for the precisely defined exchange of structured information in a subject area such as chemistry. Its increasing adoption by chemists over the last 15 years or so has concentrated on the exchange of loosely structured messages based on ASCII text, which rarely if ever contain any explicit markup (chemical or otherwise) or easily machine-parsable semantics. In 1992, Borenstein and Freed recognized that the absence of a structuring mechanism was a serious deficiency in e-mail, and proposed a MIME protocol[2] for achieving this, which was subsequently adopted as an Internet standard by the Internet Engineering Task Force (IETF). MIME defined how a specified document could be associated with an e-mail message body, and how it should be handled upon receipt by a suitable client program.

The MIME protocol comprises two components. The first defines how binary computer files must be encoded to achieve so-called 7-bit transparency for compatibility with most text-based Internet mail routers (so-called base-64 encoding) and is not discussed further here. The second component defines a standard mechanism whereby computer files can be associated with an e-mail message via appropriate headers and delimiters, and allows the appropriate processing of such enclosures by mail handling programs in the possession of the e-mail recipient. Borenstein and Freed envisaged a multicomponent structure to an e-mail message, in which the first compoent would comprise the informal and unstructured message body, and subsequent components could include structured and well-defined data files that could be handled by programs other than the basic e-mail client. These components were to be known as media types, and in the original proposal, a number of such primary media types were defined, each sufficiently generic that default handling schemes could, at least in principle, be applied their content. Thus, it is clearly apparent that different processing and display mechanisms are required for the primary defined media types TEXT, IMAGE, AUDIO, and VIDEO. The APPLICATION media type has less well-defined boundaries, and tends to be used for the resolution of proprietary data types defined by the developers of software applications. A MULTIPART type was defined to allow the multicomponent collection to be created. Most recently, the MODEL primary type has been added to allow the processing of numerical and symbolic data for three and higher dimensional models.

* To whom correspondence should be addressed. E-mail: rzepa@ic.ac.uk.
[†] Department of Chemistry.
[‡] Virtual School of Molecular Science.
[§] School of Chemistry.

10.1021/ci9803233 CCC: $15.00 © 1998 American Chemical Society
Published on Web 08/14/1998

Table 1. Chemical MIME Media Types in Public Use During the Period 1994−1998

type	filename-extension	description	source of possible program or plug-in
chemical/x-cdx	cdx	ChemDraw eXchange file	*http://www.camsoft.com/plugins/*[a]
chemical/x-cif	cif	Crystallographic Interchange Format	*http://www.crystalmaker.co.uk/*
chemical/x-chem3d	cdx	Chem3D file	*http://www.camsoft.com/plugins/*[a]
chemical/x-cmdf	cmdf	CrystalMaker Data format	*http://www.crystalmaker.co.uk/*
chemical/x-cml	cml	Chemical Markup Language	*http://www.venus.co.uk/omf/cml/*
chemical/x-daylight-smiles	smi	Daylight SMILES	*http://www.daylight.com/* *http://www.synopsys.co.uk/*
chemical/x-csml	csml, csm	Chemical Style Markup Language	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-galactic-spc	spc	SPC format for spectral and chromatographicdata.	*http://www.galactic.com/galactic/Data/spcvue.htm*
chemical/x-gaussian-input	gau	Gaussian Input format	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-gaussian-cube	cub	Gaussian Cube (Wavefunction) format	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-isostar	istr, ist	IsoStar Library of intermolecular interactions	*http://www.ccdc.cam.ac.uk/*
chemical/x-jcamp-dx	jdx, dx	JCAMP Spectroscopic Data Exchange Format	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-kinemage	kin	Kinetic (Protein Structure) Images	*http://www.faseb.org/protein/kinemages/MageSoftware.html*
chemical/x-mdl-molfile	mol	MDL Molfile	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-mdl-rxnfile	rxn	MDL Reaction format	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-mdl-tgf	tgf	MDL Transportable Graphics Format	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-macmolecule	mcm	MacMolecule File Format	*http://www.molvent.com/*
chemical/x-macromodel-input	mmd, mmod	MacroModel Molecular Mechanics	*http://www.columbia.edu/cu/chemistry/mmod/mmod.html*[a] *http://www.camsoft.com/plugins/*[a]
chemical/x-mopac-input	mop	MOPAC Input format	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-pdb	pdb	Protein DataBank	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-xyz	xyz	Co-ordinate Animation format	*http://www.mdli.com/chemscape/chime/*[a]
chemical/x-vmd	vmd	Visual Molecular Dynamics	*http://www.ks.uiuc.edu/Research/vmd/*

[a] MIME type supported *via* a Browser plug-in.

The MIME protocol also defines a secondary media type header that allows the definition of more specific information on the expected content of a message attachment. For example, image/jpeg defines a bit-mapped image file in the specific standard format defined by the Joint Photographic Experts Group. The two-level mechanism also allows a separate name space to be defined for each primary media type.

In early 1994, we considered[3] how the MIME mechanism could be used to allow the exchange of standard (ratified or de facto) chemical data types using either e-mail mechanisms or the then emerging medium of the WWW. Although many of the so-called chemical legacy formats are not always fully documented and specified in the literature, and some such as the Brookhaven protein databank format have spawned a number of variants and mutations over the years, we nevertheless felt that the concept of "chemical" as a new primary MIME media type would have a number of distinct advantages. First, it was apparent that none of the original or subsequently proposed primary media types would allow any sensible component of default handling of implicit chemical information contained in a data file. Second, the MIME mechanism operates by assigning three or four letter filename extensions to the data files, and hence each primary type must operate within a closely regulated name space convention. By assigning a primary type chemical, this name space could be delegated to the community that defines the media type, rather than the less manageable Internet community as a whole. Finally, the adoption of chemical as a primary media type was seen as the first step in achieving a closer integration between the exchange of chemical information via document server systems such as the WWW and the exchange of the same data types using electronic mail mechanisms.

CHEMICAL MIME TYPES

In the four years or more that have elapsed since the original proposal for chemical MIME types, their use via the WWW has become common. Listed in Table 1 are the chemical MIME types that as far as we are aware have actually been used to a greater or less extent during this period,[4] almost always in the context of the WWW rather than e-mail, together with suggestions for appropriate programs capable of processing and/or displaying the associated chemical content. In many cases, such programs can also serve as the starting basis for molecular modeling applications and database queries,[5] electronic journal[6] and conference browsing,[7] and numerous other applications. In this sense, chemical MIME has served as the infrastructure that has started to catalyze the development of a new generation of Internet-based chemistry tools.[1]

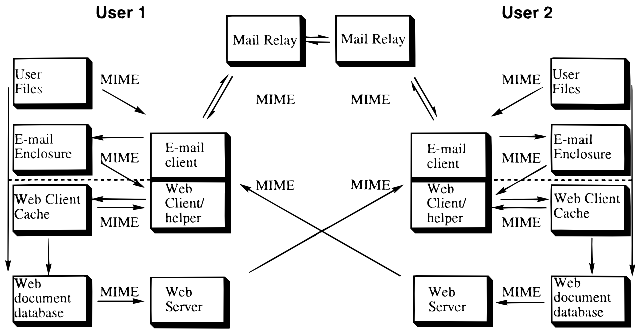
Chemical MIME types can be applied in three different contexts.

1. MIME types that have been configured for WWW (HTTP) document servers operating on an Internet-wide scale (i.e., associated with publically published documents). Such configuration is normally accomplished via a privileged account, and the use of standard types is essential so that different servers allow documents of the same type to be accessed by remote users in an identical manner. The precise manner in which any individual server is configured may differ, but a typical entry in a "mime.types" configuration file might appear as follows

```
chemical/x-mdl-molfile mol
```

This simply serves as an instruction to the server that any document associated with a filename extension.mol is issued

Scheme 1. Internet-based document and data flow, illustrating how MIME headers can be used to structure informaiton exchange and to integrate e-mail- and WWW-based mechanisms



upon request with a document header containing a specification of the MIME type as chemical/x-mdl-molfile.

2. n an Intranet environment (i.e., one associated with documents that are only accessible in a controlled private environment), it is common to define additional nonstandard MIME types for local use. The responsibility for coordinating the use of such private types lies entirely within the organization. This is to be contrasted with the use of public types, for which articles such as this serve to coordinate globally.

3. The configuration of client-side software for MIME is accomplished quite differently from that for servers. A number of the MIME types listed in Table 1 in fact derive from so-called "plug-ins"[8] that can be used to enhance the basic capability of a WWW browser and/or e-mail software, and that removes much of the burden of installation of the MIME mechanism from the user. An alternative is for the user to proactively specify that a designated "helper" program be used to resolve the chemical document. In some cases, such as the Netscape Communicator program, the same software package can be used for handling both WWW documents or e-mail messages, and the user's configuration for both is handled *via* a single plug-in installation process. For other programs, such as stand-alone e-mail clients, the user will have to do the configuration process explicitly.

APPLICATION OF CHEMICAL MIME USING E-MAIL- AND WWW-BASED CLIENT SOFTWARE

An overview of how MIME can be applied to the transport of specific chemical data types using the two principle Internet mechanisms of e-mail and the WWW is illustrated in Scheme 1. Four distinct data storage areas can be identified on any individual user's computer file system. These include the general user file area, an area specified by the user for receipt of e-mail attachments, a temporary area associated with the WWW client cache if specified by the user, and finally a WWW document collection area if the user has specified a personal WWW server or has access to a central WWW server. Chemical MIME at least in part provides one mechanism for achieving self-consistency in the handling of chemical files across these file areas. To illustrate this process, some specific examples of how MIME headers are added are shown next.

Examples of Chemical MIME Headers. A WWW client that makes a HTTP (hypertext transfer protocol) GET request to a WWW server configured to support chemical MIME types results in the following response;

```
GET /atp.pdb http/1.0
HTTP 200 Document follows
Date: Mon, 30 Mar 1998 13:54:40 GMT
Server: NCSA/1.5.2
Last-modified: Fri, 19 Aug 1994 15:46:58 GMT
Content-type: chemical/x-pdb
Content-length: 2916
```

The received MIME type is resolved via a suitable internal look-up table available to the WWW client, which maps the MIME types to an application program or plug-in capable of parsing, processing, and/or displaying the chemical data, in this example a simple PDB format file.

An e-mail client who makes an SMTP (simple mail transfer protocol) request to an e-mail relay will receive the following related set of headers;

```
Mime-Version: 1.0
Content-Type: multipart/mixed; boundary="===_-
1320854989==_===="
Date: Mon, 30 Mar 1998 15:18:23 +0100
To: recipient@somewhere
From: "Sender"
Subject: Illustration of chemical MIME headers
Status: O
- -====_-1320854989==_====
Content-Type: text/plain; charset="us-ascii"
This message contains a chemical attachment
- -====_-1320854988==_D====
Content-Type: chemical/x-pdb; name="ferrocene.pdb"
Content-Disposition: attachment; filename="ferrocene.pdb"
Content-Transfer-Encoding: base64
Q09NUE5EICAgIGZlcnJvY2VuZS5...
```

The e-mail program can be used to extract the appropriate component of the multipart message attachment (in this example, separated by the unique string 1320854989), decoding it if necessary from the base-64 scheme adopted to ensure 7-bit transparency of the file, and to save the file to the user's filebase in a segregated area identified for such attachments. If the user wishes to view the contents of the attachment, a mapping between the MIME types and a suitable application program can be achieved either via a specific look-up table associated with the e-mail client, or by invoking a WWW client to perform this task.

APPLICATION OF CHEMICAL MIME TO ADD VALUE TO DOCUMENT EXCHANGE

During the last four years, it has become increasingly commonplace to attach documents of various types to text e-mail messages via the implicit use of MIME protocols. The most common types of attached documents tend to be either bit-mapped images (MIME type image/gif or image/jpeg) or 7-bit encodings of binary word-processor documents (e.g., MIME type application/msword). The use of such nonchemical MIME types almost certainly means that any chemical information contained in the documents will inevitably degrade. Recovering chemical data from such formats into an active and reusable form first requires knowledge that the document actually does contain chemical information and second requires information about the likely structure of that information. It is precisely this missing information that the explicit use of the chemical MIME protocol will provide, via the primary and secondary types.

To highlight the advantages of this still infrequently used method of attachment, we include here specific details of how the mechanism can be used for three typical e-mail environments. These examples should also serve as prototypes for setting up other commonly used e-mail handling systems that support MIME.

Example 1. Chemical MIME Handling using the Unix Pine E-mail Client. This mechanism in fact constitutes the original Unix-based method developed by Borenstein and Freed[2] to test their MIME proposal. For outgoing e-mail messages, the chemical MIME headers are added according to a look-up table present on the users home directory called .mime.types. A typical entry is as follows

chemical/x-pdb pdb

For incoming e-mail messages, the association of a document MIME type with a program suitable for its resolution is accomplished by using a look-up table present on the users home directory called .mailcap

chemical/x-pdb; netscape %s

Example 2. Chemical MIME Handling Using the Eudora E-mail Client. Eudora is a stand-alone e-mail client available for Windows and MacOS operating systems. This program allows hyperlink-style resolution of an enclosed message attachment by a program designated by the recipient. Unlike a WWW client such as Netscape, where the chemical MIME types are simply defined on all three major platforms by adding an appropriate plug-in (Table 1), the configuration of Eudora both for sending and receiving chemical attachments is operating system dependent. On MacOS, a chemical MIME plug-in[10] is placed in the same folder as the Eudora application. To achieve the equivalent functionality on Windows 95/98/NT, the file Eudora.ini present in the application folder must have an entry of the following type added for each of the MIME types required

both=pdb,pdb,TEXT,chemical,x-pdb

When receiving e-mail messages that include a chemical MIME attachment, users will have to specify an appropriate program to resolve the attachment. This has to be done only once for each MIME type and can be, for example, by adding the filename extension appropriate for each type of MIME attachment via the Windows Registry file or by specifying this within the e-mail program.

Example 3. Chemical MIME Handling Using Netscape Communicator Illustrating Integration of WWW and E-mail Clients. Netscape Communicator (at the time of writing at version 4.05) represents, inter alia, an integrated WWW client (Navigator) and an e-mail client (Messenger). Configuration of chemical MIME types can be accomplished in two generic ways. The simplest way is via the Netscape plug-in mechanism. Several plug-ins offer support for chemical MIME types (Table 1), and their installation automatically configures both the WWW and e-mail client components of Netscape with the MIME types supported by the plug-in. This automatic mechanism can also be overridden by a user configuration option that will allow additionally defined or redefined chemical MIME types to be associated with other specific programs for processing any individual data type.

In operation, the application of chemical MIME is almost entirely transparent to the user. Any chemical data set defined by the MIME types that is received by the WWW client Navigator will be displayed as either an in-lined model using an appropriate chemical plug-in or in an external window using a user-specified program. We note here that all incoming data files can also be saved in the Netscape client local disk cache, where in principle the chemical MIME labeling could be used to create a persistently stored chemical database using suitable software. A chemical attachment received by the e-mail client Messenger can be passed to the browser window for resolution with the MIME headers being processed internally between Messenger and Navigator, or externally via the file system and the filename extensions.

When Netscape Messenger is used to send a chemical e-mail attachment to an e-mail relay, the user selects the appropriate filename, and Messenger will insert the appropriate MIME headers by appropriately mapping the filename extensions. This mapping is, as before, defined either via the plug-in support or by explicit declaration in the configuration.

The Netscape implementation is the only one that works transparently across Unix, Windows, and MacOS client-based operating systems. One test of operating system transparency is for the test originator to attach a simple chemical coordinate file[8] to an e-mail message and to send this to a remote recipient. The entire process is then reversed by the recipient retrieving the received file from their e-mail attachments folder (Scheme 1) and sending it back to the original sender. The process will be regarded as successful if the test file received back is identical with the originally sent file, and can be suitably and automatically resolved by both parties via an appropriate three-dimensional (3D) coordinate display program or plug-in using either e-mail or WWW clients.

Example 4. Using Chemical MIME Handling to Add Value to an Internet Chemical Database. The fourth example illustrates how chemical MIME can be applied outside an e-mail environment. The application of chemical MIME in areas such as electronic conferences and journals has been amply documented elsewhere.[1] This example illustrates how MIME can be used to enhance a database of degradation schemes for atmospherically significant volatile organic compounds (VOCs).[11] In the Leeds Master Chemical Mechanism (MCM), simplified degradation schemes for 120 such VOCs have been constructed. The information contained in the database is, however, difficult to navigate because the degradation products and intermediates of one reaction may be reagents in an other. In the Leeds scheme, the organic component of the MCM contains in the range of 7000 reactions and 2500 chemical species. Such a complex web of information is best presented as a hypertext document, within which the chemical structures are represented using the SMILES notation (Table 1). The SMILES string can be interpreted by spawning an external viewer as a helper application (Table 1) using the chemical MIME mechanism, thus achieving a linkage between structural and kinetic information in the database. Furthermore, the chemical structural information contained within the database remains active and reusable in other contexts, for example substructure searching of other databases.

BEYOND CHEMICAL MIME: CHEMICAL INFORMATION COMPONENTS

Chemical MIME was an experiment in the sense of initiating activity and collecting data. In the spirit of the WWW, we had few preconceptions about what would happen, and our original choice of chemical MIME types was in part designed to stimulate development. A retrospective assessment is that the major use has been for distributing simple chemical information components, or what we have termed “legacy formats”. These derive from older databases or program input files, rather than for multicomponent documents such as envisaged in newer structured formats such as CIF, ASN.1, etc. Given the direction of the WWW community, this is entirely reasonable and has given an excellent platform for the next stage, which we outline here.

Most chemical information is a complex mixture of different components and disciplines. For example, a “compound data card” usually consists of

•a molecular connection table
•citations/references/authorship
•physicochemical properties
•(possibly) graphics, spectra
•links to other information.

Of these, only the first is specifically “chemical”, and the others are found in many other disciplines.

Starting in 1997, the W3C (World Wide Web Consortium)[12] has been developing a set of generic, discipline-independent protocols for the transmission of documents. These protocols allow for the description, encapsulation, and interoperability of “information objects” from specific disciplines. For example, in a cooperative international forum including support from the American Mathematical Society, a protocol for WWW-based transmission of “mathematics” has been developed (MathML). This protocol is specifically designed so that it will interoperate with existing (HTML) and emerging (XML) protocols. In a similar fashion, a W3C group is devising a protocol for meta data (RDF, or Resource Definition Framework, and DC, or Dublin Core) that will allow simple and complex descriptions of the role and content of documents. For example, the authorship, authenticity, location, ownership, and related attributes of a document can be described in RDF/DC framework. We believe that it should always be appropriate to examine these generic approaches before devising yet another proprietary format.

What the W3C has also made clear is that there are generic operations that apply to any sort of document and that these are often not trivial. They include:

•authoring
•editing
•validation
•parsing
•rendering
•merging
•filtering
•searching
•transformation

To develop tools for each of these from scratch is expensive and error prone. The approach taken by the W3C is resulting in generic tools that apply to all document types, and we shall outline the general principles.

A document (which is not limited to text and can contain or consist of nontextual objects) is composed of smaller components. It contains sequence information (the order of the components can matter) and structure (one component can contain another). Thus, a book consists of chapters in sequence; the chapters can contain sections, which contain subsections. In a similar way, we can describe a labNoteBook as containing compoundDataCards and text. The compoundDataCards can contain molecules, spectra, etc. The other important concept is the identification of the components themselves through *markup*. The markup shows the limits of a component and gives a handle (tag) by which it can be identified or linked to semantic resources.

The W3C has now created a language (eXtensible markup language, XML)[13] for managing general structured documents on the WWW. XML is formally a very simple subset of SGML, but for those unfamiliar with SGML, it may help to think of it as an abstraction of the HTML approach. It is not a language, but a tool for creating one’s own language in such a way that it is compatible with existing and future WWW technology. At time of writing (April, 1998) it seems certain to be a ubiquitous component of all major manufacturers' WWW tools.

The XML effort is currently tackling the problem of how to manage documents with components from several domains, which has to address cases where the individual markup languages have been developed in ignorance of each other. An example would be a document containing text (HTML), maths (MathML), and chemistry (CML). The solution to this will initially be through the use of namespaces that guarantee that tags from the different domains will not clash.

It is clear, therefore, that a very important trend is toward multidomain documents using XML syntax. We believe that much information in molecular science is ideally suited to such an approach, and suggest that additional means of identifying chemical information will be required. In some cases this means could be the existing chemical MIME types (XML has a mechanism for encapsulating MIME), but will increasingly use other methods based on XML.

A survey of a number of technical publications suggests that there are a relatively small number of different abstractions of information components. For example, although the content of an image could vary widely, the basic technology to read and render it is domain independent. The commonest components (with existing technology) are:

•text (HTML)
•images (GIF, JPEG)
•structured graphics
•tables [*]
•graphs [*]
•numeric quantities with units [*]
•arrays and matrixes [*]
•citations (RDF, Dublin Core)
•links (XLL)
•mathematics (MathML)

For components labeled [*] there are existing solutions but they are fragmented and not yet developed for the WWW/XML. Note that many components that are apparently chemistry -specific (e.g. spectra, schemes) are simply concrete examples of these abstract types (e.g., a spectrum and an index of stock prices can use identical technology).

The challenge is semantic, how do we attach meaning to the labels or other components? In a similar way, reaction schemes can be described as molecular components embedded in a general semantic network.

A key feature of XML is the separation of syntax from semantics. Thus, XML guarantees that the document will be platform and software independent. However, when it "arrives at the reader" there must be a mechanism for adding semantic information; for example, what does MOL mean? In many cases this mechanism is supplied by stylesheets (e.g. formatting the document such that its meaning becomes clearer), but for technical data, additional processes are required. Tags and attributes can be linked to glossaries. Thus

```
<ITEM CONVENTION="mmCIF" TERM=
                         "_cell.length_a">23.4</ITEM>
```

can be linked to the International Union of Crystallography's mmCIF dictionary for macromolecules.[14] This relationship defines the quantity as the cell length (in Angstrom units by default). It will be a major advance if those with data dictionaries can make them available electronically. The Virtual Hyperglossary Technology[15] has been developed as a prototype to provide XML-based technology for such linking. Alternatively it is possible to link the document to actions provided by software. One such method is with Java classes, which can be specifically linked to given tags as in the JUMBO browser.[16]

The most challenging types of information to distribute are those that describe the relation of one quantity to another. Sometimes this can be done by containment (e.g., a spectrum and a molecule can be linked by being contained in a compoundDataCard). But for more complex mapping we require either semantic networks as in RDF or hypermedia as in XLL (XLink, the W3C'S protocol), which provide ways of connecting components in arbitrarily complex manners. Thus XLL could, for example, provide links from functional groups to peaks in a spectrum.

If we accept that XML and related standards (XLL, XSL stylesheets, RDF, DC, and DOM or document object model) will provide the generic capabilities, then the task facing chemistry becomes more clearly defined; that is, to provide extensible markup for chemistry-specific components and to analyze common relationships in chemistry.

To this end a prototype XML language has been developed and termed Chemical Markup Language.[17] This is provided as a starting point for the process of supporting chemistry in XML. Current elementTypes or tags have been deliberately kept simple:

- ATOM
- ATOMS (simplifies handling of large molecules)
- BOND
- BONDS
- ELECTRON
- FORMULA

These tags can be qualified with a small number of hard-coded attributes (e.g., ELSYM, X2, Y2, etc. for ATOM, ORDER, STER, etc. for BOND). Different conventions can be set with a CONVENTION attribute. In this way, most simple descriptions of molecules can be captured. The contents of these can be arbitrarily complex and could support, say, orbital components on atoms, quadrupoles, ^{13}C shifts, etc. Because much chemistry is solid state we include CRYST, and for macromolecules supply SEQUENCE and FEATURE. The key point is that XML has been developed in an open collaborative process (which has included many companies that compete vigorously in the markets). We believe that XML is appropriate for chemistry as well and offer CML as a starting point for such a process.

CONCLUSION

During the period 1970−1994, chemical applications of the Internet were largely based on a set of generic file and text transmission protocols, such as terminal (Telnet), file transfer (FTP), e-mail transfer (SMTP), and document handling systems (HTTP), methods where little explicit labeling and structuring of the chemical content was available and where little interoperability existed between either the chemical content itself or with nonchemical content in other disciplines.

During the period 1994−1998, mechanisms such as chemical MIME set the scene for a convergence both within the chemical community and with other scientific areas, bringing together applications such as electronic mail with database and modeling tools, electronic conferences, journals, books, and taught courses. In the future, more finely grained mechanisms such as XML/CML will undoubtedly enable further convergence to the point that the Internet will become a far better and more powerful "resource discovery" tool for the chemical and scientific communities.

ACKNOWLEDGMENT

Funding from the UK JISC e-Lib program for the CLIC electronic journal project, and the JISC JTAP program for the VChemLab project is gratefully acknowledged.

REFERENCES AND NOTES

(1) Rzepa, H. S.; Murray-Rust, P.; and Whitaker, B. J. *Chem. Soc. Rev.* **1997**, 1−10.

(2) Borenstein, N.; Freed, N. "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.

(3) Rzepa, H. S.; Whitaker, B. J.; Winter, M. J. *J. Chem. Soc., Chem. Commun.* **1994**, 1907; Rzepa, H. S. *Comput. Networks and ISDN Systems*, **1994**, *27*, 317−318; Rzepa, H. S. *Chem. Design Auto. News* **1994**, *9*, 1; Rzepa, H. S. In *The Internet: A Guide for Chemists*; Bachrach, S., Ed.; American Chemical Society: Washington, D.C., 1995; Winter, M. J.; Rzepa, H. S.; Whitaker, B. J.; *Chem. Br.* **1995**, 685; Davies, A. N. *Spectrosc. Eur.* **1996**, *8*, 42; Rzepa, H. S. *Sci. Progress* **1996**, *79*, 97; Whitaker, B. J.; Rzepa, H. S. *Proc. Int. Chem. Inf. Conf.* Collier, H., Ed.; **1995**, 62−71; Rzepa, H. S.; Casher, O.; Whitaker, B J. *Proc. Int. Chem. Inf. Conf.* Collier, H., Ed.; **1996**, 141−148; Rzepa, H. S.; Locke, W.; Murray-Rust, P.; Whitaker, B. J. In *Perspect. Protein Eng.* '96. Geisow, M. J., Ed.; **1996**, Paper no. 19; Rzepa, H. S.; Murray-Rust, P.; Whitaker, B. J. *Chem. Intl.* **1997**, *19*, 17.

(4) A description of the definitive list will be published elsewhere; Rzepa, H. S.; Murray-Rust, P.; Whitaker, B. J. *Pure Appl. Chem.*, in preparation. The latest information is available on-line at *http://www.ch.ic.ac.uk/chemime/*.

(5) Rzepa, H. S. "Internet-based Computational Chemistry Tools", In *Encyclopaedia of Computational Chemistry*; Wiley: New York, in press.

(6) An example of the use of chemical MIME to integrate a variety of chemical data types into the body of an electronic journal is the CLIC Electronic Journal Project; James, D.; Whitaker, B. J.; Hildyard, C.; Rzepa, H. S.; Casher, O.; Goodman, J. M.; Riddick, D.; Murray-Rust, P.;, *New. Rev. Information Networking* **1996**, 61. For the project itself, see *http://www.rsc.org/is/journals/current/chemcomm/cccenha.htm* or

the original site at *http://chemcomm.clic.ac.uk/*. For details of how a "chemically enhanced" article was prepared, see Casher, O.; Rzepa, H. S. In *Proc. E. Conf. Trends in Organomet. Chem.: ECTOC-3*. Rzepa, H. S., Leach,C., Eds.; Royal Society of Chemistry, 1998; ISBN (CD-ROM) 0−85404−889−8.

(7) For examples of the application of chemical MIME to electronic conferencing, see *ECTOC-1* Leach, C., Rzepa, H. S., Eds.; Royal Society of Chemistry, 1996. Also ECHET96, 1997; ECTOC-3, 1998, and ECHET98, 1998. The conferences are on-line at *http://www.ch.ic.ac.uk/ectoc/*.

(8) Maffett, T.; van Vliet, B. MDL Information Systems, 1996. For further details of Chime, see *http://www.mdli.com/chemscape/chime/*. This plug-in derives from the Rasmol program written by R. Sayle and applied within the chemical MIME project as described in Casher, O.; Chandramohan, G.; Hargreaves, M.; Leach, C.; Murray-Rust, P.; Sayle, R.; Rzepa, H. S.; Whitaker, B. J. *J. Chem. Soc., Perkin Trans 2*. **1995**, 7.

(9) A simple test molecule is available at *http://www.ch.ic.ac.uk/rzepa/jcics/molecule.pdb*. A site for testing an extended set of chemical MIME types is available at *http://www-dsed.llnl.go*υ*/documents/tests/chem.html*.

(10) Rzepa, H. S., available as *http://www.ch.ic.ac.uk/rzepa/jcics/chemical10.hqx*.

(11) Pilling, M. J.; Saunders, S.; Jenkin, M.; Derwent, D. "Tropospheric Chemistry"; *http://www.chem.leeds.ac.uk/Atmospheric/MCM/main.html*.

(12) For details of all W3C (World Wide Web Consortium) recommendations, proposed recommendations, working drafts and notes, see *http://www.w3.org/TR/*.

(13) Murray-Rust, P.;, "Chemical Markup Language, A simple introduction to structured documents" In *XML, Principles, Tools and Techniques*; Connolly, D., Ed.; O'Reilly: 1997; pp 135−149.

(14) For specifications see Bourne, P. E.; Berman, H. M.; McMahon, B.; Watenpaugh, K. D.; Westbrook, J.; Fitzgerald, P. M. D. *Methods Enzymol*. **1997**, *277*, 571−590. See also *http://www.iucr.org* and *http://www.iucr.org/cif/mm/index.html*.

(15) Murray-Rust, P.; West, L. *ASLIB Managing Information*, **1997**, *4*, 36−39. See also *http://www.*υ*hg.org.uk/*.

(16) Murray-Rust, P.; In *Proc. E. Conf. Trends in Organomet. Chem.: ECTOC-3*; Rzepa, H. S., Leach, C., Eds.; Royal Society of Chemistry, 1998. ISBN (CD-ROM) 0−85404−889−8. A fully working version of JUMBO is included on this CD-ROM. See also ref 13, pp 197−207.

(17) The CML project was first described in Murray-Rust, P.; Leach, C.; Rzepa, H. S. *Abstracts of Papers, Am. Chem. Soc.* **1995**, *210*, pp 40-COMP (*http://www.ch.ic.ac.uk/cml/*) and in Murray-Rust, P.; Rzepa, H. S. *Abstracts of Papers, Am. Chem. Soc.* **1997**, *214*, pp 23-COMP. Details of CML itself are available in refs 13 and 16. Further details will be published in a forthcoming paper.

CI9803233