Chimia 52 (1998) 653–657
© Neue Schweizerische Chemische Gesellschaft
ISSN 0009–4293

A History of Hyperactive Chemistry on the Web: From Text and Images to Objects, Models, and Molecular Components

Henry S. Rzepa*

Abstract. The chemical application of the Internet over the last decade is reviewed, and lessons from the experience set out. In particular, we emphasize that to create a high quality, indexed and structured chemical resource using this technology, new standards and methodologies have to be defined and adopted. One particular change, from the traditional use of text and illustrations in printed journals, to the deployment of object-based models built from, *e.g.*, molecular components is discussed in more detail.

Introduction: 1970–1993

A retrospective view of the development of the scientific and chemical Internet suggests that the chronology can be usefully divided into a classical period dating from around 1970 to 1993, and the modern post-1993 era. Chemists who were aware of and actively used the Internet prior to 1993 tended to have a background of theoretical or computational chemistry, or to have particular responsibilities relating to libraries or information science and retrieval. Although on-line systems such *CAS* were released for nonexpert use as early as 1983, 'point-to-point' nonstandard connection methods using proprietary programs were needed. The mid 1980s also saw a minority of chemists beginning to use simple text-based electronic mail facilities available *via* the local 'mainframe' computer and early experiments in the use of electronic journals based on full-text delivery of articles. Computational chemistry and molecular modeling techniques also began to make use of computer networks, and by the late 1980s, it had become common for people working in such areas to exchange computer files using point-to-point connections to remote resources such as national computer or database centers, although it was still rare to go beyond one's national boundaries. At that stage, it was still necessary to acquire considerable expertise in using a text editor for extracting and reformulating the information carried within these files into more useable form for local processing. This classical era was characterized by a lack of interoperability of the information flowing *via* the Internet, and by a host of incompatible methods used for retrieving any chemical content and meaning from the loosely structured text and other files [1].

The Changes since 1993

Several key infrastructures were introduced in 1993, which have dramatically changed the working model.

- Prior to 1993, use of the Internet had largely been session and file centered. By session is meant the user establishing a point-to-point connection with a remote information or computational resource *via* a simple screen-oriented terminal interface, in which a history of the session is recorded by the remote computer system, a history which serves to establish the context or state between the different transactions conducted by the user. Computer files were either text-based editable entities that served as either the input to or output from binary program files that allowed information to be processed. Any context that existed between such files was normally defined by a proprietary database-handling system, and rarely, it was possible for facile exchange of information between different databases other than by involving human intervention. Importantly, many of the session and file-based mechanisms had become highly parochial to particular subject disciplines. An area such as chemistry had developed its own proprietary on-line usage, and file types defined purely by the chemical programs that generated them or used them as inputs. This at a time, when it was beginning to be recognized that transcending subject boundaries was more important than ever.
- In 1993, the concept of a computer file started to transmute into that of a document. The term document implied that some sort of internal and predictable structure existed within the document, one that might even survive across subject disciplines, and also carried an implication that indexing of the document might be possible. A particular way of defining the structure of a document called hypertext-markup language (HTML) became standard (currently implemented at level 3.2, and with level 4.0 now agreed), and freely available programs to display such documents known as 'browsers' were introduced for the first time. In this scenario, the 'session' began to be replaced by a standard mechanism to define the context between two or more documents, known as the URL (Uniform Resource Locator) or more colloquially, the Internet hyperlink.
- Although in the five years since 1993 the URL has become commonplace not merely in science, but in everyday life, it is worth looking at a URL in a little more detail (http: //www.ch.ic.ac.uk/rzepa/chimia/index.html#introduction). The first component of this URL up to the # defines a resource, *i.e.*, a mechanism to obtain a document from a designated global location, although it could also be a processing resource such as a search query rather than a document. The string after the # specifies a document fragment, or more generally a definition of actions to be performed on the document or by the processing resource. The old-style ses-

**Correspondence*: Prof. Dr. H.S. Rzepa
Department of Chemistry
Imperial College of Science
Technology and Medicine
London, SW7 2AY, UK
E-Mail: rzepa@ic.ac.uk

sion in this model is thus replaced with a collection of hyperlinked documents and resources that together defined a structured set of resources.

– A third vital component of the Internet infrastructure known as MIME (Multipurpose Internet Mail Extensions). This mechanism implemented a mechanism to unify the exchange of legacy files and documents that the previous 25 years of computer activity had helped define and create. MIME is a very simple two-level-hierarchical identifier of the content of such files (including documents as defined above).

MIME and Chemical MIME

Since MIME has been well described elsewhere [2], we will introduce only four commonly used MIME types to illustrate the themes in this article. These are

1) text/html
2) image/gif
3) chemical/x-mdl-molfile
4) model/vrml

The first of these is simply a text-based document, with the expectation that the internal structure of that document will be 'marked-up' using the HTML guidelines. Based simply on the globally accepted definition of version 2 of HTML, in combination with equally standard MIME declarations and the URL definition shown above, it had become possible by mid 1994 for a single person, with only limited resources at their disposal, to create a text-searchable index of most of the Internet. Such search facilities are now commonplace, although their utility for retrieving chemical information has been questioned [3]. I will return to this theme later in the article.

Finally, we will introduce chemical specifics into this discussion! The very earliest chemical content referenced within HTML-type documents was in fact *via* hyperlinks to images, most popularly defined by a format known as GIF (graphical interchange format) introduced around 1987. Chemical diagrams encoded in GIF (and a few other common formats) were easy to create, and from 1993 to the present, their use proliferated. I argue here that their use as carriers of interoperable chemical information has been an unmitigated disaster, and we will suffer the problems caused by their continued use for the foreseeable future. Why? Well, put simply, there is no standard way of identifying that their content is chemical. Only now are computer scientists starting to come up with efficient schemes for scanning the patterns in an image, creating databases of this content, and allowing search and retrieval schemes based on these patterns. The application of image recognition in chemistry remains nontrivial [4]. Perhaps a better approach has been to insert chemical information into 'hidden fields', which formats such as GIF and PNG (Portable Network Graphics) allow [5], and which would permit a reconstitution of the chemical information by suitable programs. Nevertheless, this approach is rarely adopted, and hence, attempts to index the chemical content of the Internet have accordingly encountered enormous difficulties [3]. An opportunity to insert chemical information in a more simple 'human readable' form was missed, however. The standard way of invoking an image from a document written in HTML 3.2 or earlier is as follows:

```
<IMG SRC="sildenafil.gif"
alt="C22H30N6O4S"
WIDTH="500" HEIGHT="490">
```

Firstly, note that the actual name of the GIF image suggests chemical content (to a chemist!), but such names are rarely chosen to be unambiguous or complete. The dimensions do not help at all. The so-called ALT field, however, provides an alternative text-based descriptor of the possible content of the image. In a chemical sense, no guidelines exist for how this field might be used. If present at all, it is normally entered as free-form text; formulae, or even better some form of atom connectivity such as a characteristic SMILES atom-connection string (*e.g.*, O=S(C1=C([H])C([H])=C(OCC)C(C(N2[H])=NC(C(CCC)=NN3C)=C3C2=O)=C1[H])(N4CCN(C)CC4)=O) have rarely been used.

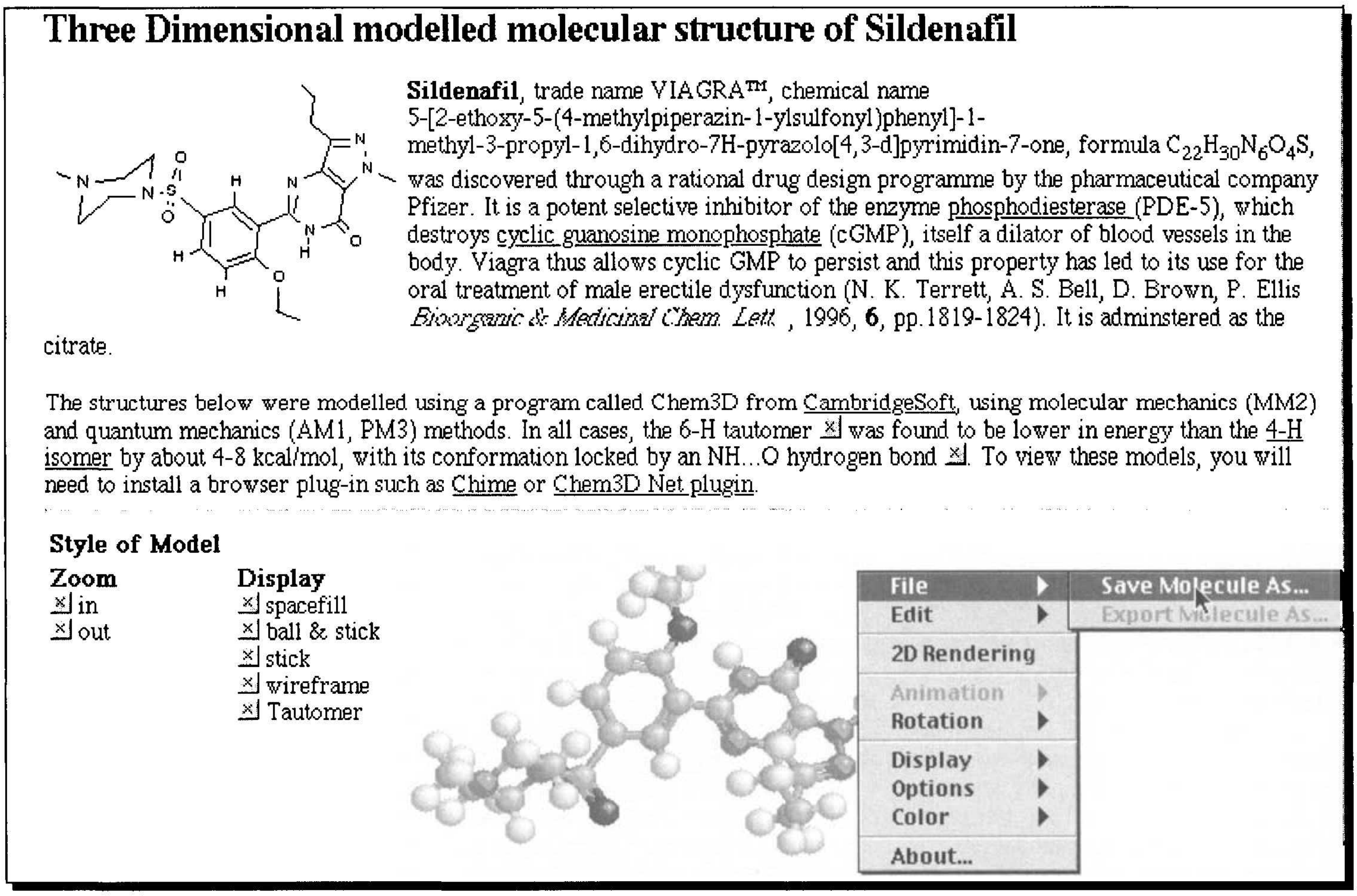
The image ALT field, however, does introduce one important new concept known as meta-data. This is simply a description of an information resource, with the term 'meta' deriving from the Greek word for change. Its purpose is to document the origins of, and/or track the change or use of, data.

The HTML 4.0 specification [6] replaces the image invocation with the syntax:

```
<OBJECT title = "Sildenafil (Viagra), Molecule-of-the
Month at Imperial College" data = "sildenafil.gif" type
= "image/gif" width = "500" height = "490"> 5-[2-ethoxy-
5-(4-methylpiperazin-1-ylsulfonyl)phenyl]-1-methyl-3-n-
propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one </
OBJECT>
```

Here, the title is a possible carrier of meta-information, whilst type refers to the media (MIME) type, in some ways another form of meta-data. The remaining text (in this case the chemical name) would only be displayed if for some reason the GIF image could not be shown. This field could serve a dual purpose in providing valuable text-based information as an alternative to attempting to recognize the image content.

Recognizing that if chemical content on the Internet was to be rendered indexable, and hence retrievable, we proposed in early 1994 that where possible, generic images with chemical content should be replaced by a more explicit declaration. The so-called chemical MIME type was introduced, along with *ca.* 20 subtypes that represented a spectrum of chemical content carried by standard or *de facto* file types that the community had adopted over the preceeding 25 years, and for which tools for their generation and viewing were available to a greater or lesser extent. The molfile format documented comprehensively by MDL is a good example of this MIME type. It can contain either 2D or 3D molecule coordinates, and has an explicit declaration of the atom connectivity and bond types. Put simply, a reference to such a file from within an HTML document carries a strong assumption that an unambiguous declaration of a molecule might be expected. Other chemical MIME types defined other aspects of molecular connectivity, or specified analytical data carried in more or less standard formats which could be generated directly from analytical instruments. The support of this and *ca.* 10 other chemical MIME type *via* browser plug-in software such as Chime from MDL [7] or ChemDraw/Chem3D Net plug-in from CambridgeSoft Corporation [8], and *via* Java-based applets such as ChemSymphony [9] and analytical data interpreters [10] has ensured that the adoption of chemical MIME has gradually increased from 1994 onwards. A typical example of how this infrastructure could be used to deliver accurate and context-rich molecular information across a range of molecular sciences is shown in *Fig. 1*. This designation is used deliberately; if you are viewing this article in print, then you will inspect a figure (or illustration) of



Three Dimensional modelled molecular structure of Sildenafil

Sildenafil, trade name VIAGRA™, chemical name 5-[2-ethoxy-5-(4-methylpiperazin-1-ylsulfonyl)phenyl]-1-methyl-3-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one, formula $C_{22}H_{30}N_6O_4S$, was discovered through a rational drug design programme by the pharmaceutical company Pfizer. It is a potent selective inhibitor of the enzyme phosphodiesterase (PDE-5), which destroys cyclic guanosine monophosphate (cGMP), itself a dilator of blood vessels in the body. Viagra thus allows cyclic GMP to persist and this property has led to its use for the oral treatment of male erectile dysfunction (N. K. Terrett, A. S. Bell, D. Brown, P. Ellis *Bioorganic & Medicinal Chem. Lett.* , 1996, **6**, pp.1819-1824). It is adminstered as the citrate.

The structures below were modelled using a program called Chem3D from CambridgeSoft, using molecular mechanics (MM2) and quantum mechanics (AM1, PM3) methods. In all cases, the 6-H tautomer ☒ was found to be lower in energy than the 4-H isomer by about 4-8 kcal/mol, with its conformation locked by an NH...O hydrogen bond ☒. To view these models, you will need to install a browser plug-in such as Chime or Chem3D Net plugin.

Style of Model

Zoom
☒ in
☒ out

Display
☒ spacefill
☒ ball & stick
☒ stick
☒ wireframe
☒ Tautomer

Figure. *Viagra (Sildenafil) as an Internet model*

this concept. If you are viewing the article using the Internet, then you can inspect a model.

The model, as opposed to the figure, is invoked within HTML 3.2 as follows:

```
<embed src=viagra.mol
width=300 height=200
name=viagra
BGCOLOR="white" spinx=10
spinz=10 spiny=10
startspin=true
options3d=specular
display3D=ball&stick
alt="O=S(C1=
C([H])C([H])=C(OCC)C(C(N2[H])
=NC(C(CCC)=NN3C)=C3C2=O)=
C1[H])(N4CCN(C)CC4) = O">
```

The first line of attributes are generic, whilst the additional forms are specific to the chemical model, and to a large extent, to the software used to resolve this model on the computer screen. The ALT field in this instance is entirely nonstandard, but benign in the sense that it is ignored by the modelling software, and would serve only to provide meta-information.

The recommended future method of invoking a model within HTML 4.0 is:

```
<object data="viagra.mol" width="300" height="200"
id="viagra" title="Viagra" style="spinx: 10;
display3D: ball&stick" type="chemical/x-mdl-molfile">
<OBJECT title="Sildenafil (Viagra), C22H30N6O4S"
data="sildenafil.gif" type="image/gif" width="500"
height="490"> </OBJECT> 5-[2-ethoxy-5-(4-methylpiper-
azin-1-ylsulfonyl)phenyl]-1-methyl-3-n-propyl-1,6-
dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one</object>
```

This defines a cascading order in which attempts will be made by the browser to display either the chemical/x-mdl-molfile, the image/gif or the text field objects. Within the model shown above are also objects which serve to identify small molecular components of the molecule object, such as the region of tautomerism, or a key hydrogen bond, *i.e.*, the objects themselves can have relationships to each other.

Even by 1998, only perhaps a few percent of all molecular content on chemical web pages was identified using MIME types. These chemical MIME types are also largely seen now as a legacy from the days of proprietary formats and programs. Often, these formats lack modern mechanisms for defining internal structures and have to be considered as a single component (a 'blob'). One would not expect to easily identify smaller well-defined components within the document such as molecular subcomponents. Such legacy files also did not provide any definition of a standard mechanism for specifying meta-data. Frequently, it might even prove impossible to identify which version or flavor of file one was dealing with (*i.e.*, tracking the change in use of data with time). Before turning to molecular components and meta-data, one further important MIME type should be discussed.

The Transition from Images to Models

The adoption of chemical MIME types has allowed the chemical content of the Internet to be carried not by the poorly structured figures, schemes, or illustrations, but by what we call the MODEL [11] and by much more formally structured document types. Thus, a molecule *e.g.*, can be clearly and unambiguously specified not by an unindexable image, but by a model represented for the user's use by appropriate software within a browser window. Most importantly, the chemical model is available for the user to reuse in whatever context they wish. Such chemical models form one basis for what we term the third generation of electronic journals, where the journal is best considered as a tool for placing into context and manipulating the models it delivers [12–14]. This contrasts with the much more passive illustrations offered by print, or print-like classical electronic journals.

Alongside chemical models, the MIME-type model/vrml allows a more generic modelling functionality, known as Virtual Reality Modelling Language (VRML) [15]. This type of model serves to integrate molecular models with complex 3D data representations, molecular surfaces, animations, and most interestingly, processing functionality within the model. Such models can have so-called script nodes associated with model components based on defined algorithms. For example, two molecule components of a VRML model could be associated with a defined force field that could dynamically compute the energy of interaction of the two components during an attempt to dock one model with another. Such composite scenes also allow a much richer integration of chemical with nonchemical models, and these work particularly well at the boundaries between chemistry and other disciplines. Virtual Reality models have been recently reviewed from both a general [5] and a chemical perspective [11].

In one sense, the problems alluded to above in identifying the chemical content of two-dimensional images are also inherent in three-dimensional models. Whilst chemical models generated from formats such as the MDL molfile can be readily indexed and search for, much still needs to be done in chemically indexing the more generic VRML models. The equivalent of the image ALT field or HTML document meta-data could to be found in the so-called VRML viewpoints, but as with image ALT declarations, no standards in their use are employed, and such viewpoints are not yet indexed routinely by any index and search services. Interesting progress has been made in the reidentification of chemical content from VRML models [16], and progress is expected to be rapid in this area in the future.

Meta-Data, Molecular Components, and Modularity

Meta-data is a description of an information resource, and may be used for a variety of purposes, such as identifying whether the resource meets a particular information need, evaluating the quality or fitness for, *e.g.*, defined chemical application of the resource, or tracking the characteristics of a resource for subsequent maintenance or usage over time.

A meta-data record consists of a small set of attributes, or elements, necessary to describe the resource in question. These include basic attributes of a document such as its title, date, description, creator (author), format (MIME type), subject (keywords), and relation (of the resource to other resources, normally *via* a URL declaration). The meta-data is normally contained in a particular component of the document known as the header. Its implementation within any particular type of document can vary; that for HTML documents has recently been standardized [17], and proposals for implementation within images and perhaps even VRML types might be expected in the future. One example of using such declaration to enable a resource to be evaluated for a particular need was the inclusion of the following (highly nonstandard!) meta-data header in each of the articles comprising the ECTOC-3 electronic conference proceedings [18].

The so-called chemical prototype attempts to define a single molecule that best represents the overall molecular content of document. In principle, this would allow automated analysis of the document to provide an indication to a user of whether the chemical content is close to their interests, or conversely whether the document represents a 'dissimilar' contribution in a scan of molecular diversity. Because agreement on meta-data types and syntax within HTML documents has only recently been formalized as the Dublin Core standard, [17] the five-year evolution of the Web has assimilated few of these guidelines. Almost no HTML documents contain any significant meta-data declarations, and of those that do, even less are chemically useful. As an example, the on-line version of this article contains some Dublin Core declarations. A project to define a discreet set of standard chemical meta-data declarations such as coordinates, substance, computation/simulation, biological activity, safety, synthesis, characterization, instrumentation, physicochemical data, reaction data, and crystallography (provisionally christened Dublin Chem) is under way [19].

The chemical document illustrated in *Fig. 1* illustrated how discrete molecular components could be identified within a larger molecular model. This is a very specific example of a general problem in chemistry as a whole. The underlying documents used to create the Model utilized a combination of HTML to define the text and links between objects, MDL molfiles to carry small molecule connectivity and 3D coordinates, CSML (Chemical Structure Markup Language) to define molecular fragments, and Brookhaven PDB files to carry macromolecular information. In the summer of 1995, a project was started to define a chemical equivalent of HTML that would serve to provide a single self-consistent syntactic framework to replace these *ad hoc* methods with modular components. The latest version of CML (Chemical Markup Language [20]) follows a set of guidelines known as XML (Extensible Markup Language) [21].

Conclusions

Prior to the introduction of Internet-based document delivery systems such as the World-Wide Web, a significant proportion of on-line chemical information was held in the form of proprietary databases, almost invariably requiring user authentication and custom software for access, but also offering high and consistent quality. Since 1993, *ca.* 2 million Internet-accessible documents have been indexed based on their title or the appearance of the words chemistry or chemical in the text. A significant proportion of these might be expected to contain links to images with further, but undetectable chemical content, and the overall quality of these documents is highly variable. The number of documents with strongly typed molecular information, *i.e.*, apparently *ca.* 7200 files in PDB format, 1400 MDL molfiles, 1200 XYZ, and 130 J-CAMP spectral files, comprise only *ca.* 0.5% of the total supposed chemical documents. Of these 2 million chemical documents, it is supposed that few carry meta-information, alternative chemical fields, or links to models, a supposition because the current generation of search engines are not programmed to search for such information. There is no doubt that the quality of the retrievable chemical information on the Internet would rise substantially if more use was made of the types of mechanisms described above.

Another approach to creating chemical documents involves generating them dynamically from closed databases in response to a suitably constituted request or search query. For example, the ChemFinder site [3] contains a large amount of chemical information, but this presented to the user only in the form of a 'just-in-time' document, and hence this content is not reflected in the statistics quoted above. It may well be that the majority of useful Internet-based chemical information will become available only *via* such controlled molecule or journal databases, created by a large number of authors, but controlled by a small number of publishers, a situation which reflects of course the current situation in printed publishing. The Internet does offer an alternative paradigm, in promoting the use of chemical models and other high value and modular chemical data in an open manner, and one where entirely new 'added value' models of chemical resource discovery and 'knowledge capture for compounds' could be created using information management techniques. If this second scenario is to come about, then the creators of this information will have to make it happen. If we adopt the same models that lead to the creation of an over-abundance of paper-based information graveyards, then it will probably not happen. The future is in our hands.

Received: September 11, 1998

[1] H.S. Rzepa, *Science Progress* **1996**, *79*, 97; H.S. Rzepa, P. Murray-Rust, B.J. Whitaker, *Chem. Soc. Revs.* **1997**, 1.

[2] P. Murray-Rust, H.S. Rzepa, B. Whitaker, *J. Chem. Inf. Comput. Sci.* **1998**, September issue.

[3] J.S. Brecher, *Chimia* **1998**, *52*, 658; W.-D. Ihlenfeldt, *Chimia* **1998**, *52*, 649; See also S.J. Clarke, P. Willett, *ASLIB Proceedings* **1997**, *49*, 184.

[4] A. Simon, A.P. Johnson, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 109.

[5] For further details, see W.D. Ihlenfeldt (http://www2.ccc.uni-erlangen.de/services/gif.html).

[6] D. Raggett, A. Le Hors, I. Jacobs (http://www.w3.org/TR/REC-html40/).

[7] B.A. Van Vliet, T.M. Maffett, *Abs. Papers Am. Chem. Soc.* **1997**, *214*, 177-COMP (see also http://www.mdli.com/chemscape/chime/).

[8] A. Hinchliffe, *Elec. J. Theor. Chem.* **1997**, *2*, 215 (see also http://www.camsoft. com/plugins/chem3D.html).

[9] P. Tebbutt, A. Hodgkin, A. Krassavine, *Abs. Papers Am. Chem. Soc.* **1997**, *214*, 58-CINF (see also http://www.cherwell. com/chemsymphony/).

[10] Viewers are available for the JCAMP standard; A.N. Davies, *Analyst* **1994**, *19*, 539; http://www.isas-dortmund.de/projects/jcamp/, and the SPC analytical chemistry format (http://www.galactic.com/). See *e.g.*, a viewer available as a Java applet (http://wwW.ch.ic.ac.uk/java/applets/jspec/spectra/).

[11] O. Casher, C. Leach, C.S. Page, H.S. Rzepa, *Chem. Brit.* **1998**, in press. For earlier articles, see O. Casher, G. Chandramohan, M. Hargreaves, C. Leach, P. Murray-Rust, R. Sayle, H.S. Rzepa, B.J. Whitaker, *J. Chem. Soc., Perkin Trans. 2* **1995**, 7; C. Conesa, H.S. Rzepa, *ibid.* **1998**, 857.

[12] CLIC Project: D. James, B.J. Whitaker, C. Hildyard, H.S. Rzepa, O. Casher, J.M. Goodman, D. Riddick, P. Murray-Rust, *New. Rev. Information Networking* **1996**, 61; H.S. Rzepa, B.J. Whitaker, J. Goodman, O. Casher, D. Riddick, J. Griffiths, D. James, C. Hildyard, *Abs. Papers Am. Chem. Soc.* **1997**, 214, 45-COMP. For examples of articles enhanced as part of the CLIC project, see http://www.rsc.org/is/journals/current/chemcomm/cccenha.htm

[13] *Journal of Molecular Modelling*: T.R. Clark, *Advanced Materials* **1996**, *8*, 787; H.H.K. Roth, T.R. Clark, *Abs. Papers Am. Chem. Soc.* **1997**, *214*, 43-COMP.

[14] *Internet Journal of Chemistry*: S.M. Bachrach, D.C. Burleigh, A. Krassavine, *Issues in Sci. Tech. Librarianship* **1998**, No. 17.

[15] W.D. Ihlenfeldt, *J. Mol. Modeling* **1997**, *3*, 386.

[16] M. Leipold, W.D. Ihlenfeldt, personal communication: http://www2.organik.uni-erlangen.de/dissertationen/

[17] S. Weibel, *Bull. Am. Soc. Inf. Sci.* **1997**, *24*, 9. For a userguide, see http://128.253.70.110/DC5/UserGuide4.html

[18] H.S. Rzepa, C. Leach, Eds., *ECTOC-3*, Royal Society of Chemistry, 1998, ISBN (CD-ROM) 0-85404-889-8.

[19] P. Murray-Rust, H.S. Rzepa, in preparation.

[20] The CML project was first described by P. Murray-Rust, C. Leach, H.S. Rzepa, *Abs. Papers Am. Chem. Soc.* **1995**, *210*, 40-COMP; P. Murray-Rust, H.S. Rzepa, *ibid.* **1997**, *214*, 23-COMP.

[21] P. Murray-Rust, 'Chemical Markup Language, A simple introduction to structured documents', in 'XML, Principles, Tools and Techniques', Ed. D. Connolly, O'Reilly, 1997, pp. 135–149.