

Internet chemical MIME

Antony N. Davies

ISAS, Institut für Spektrochemie und Angewandte Spektroskopie, Bunsen-Kirchhoff-Str.11, 44139 Dortmund, Germany.

Summary

This article covers the work being carried out to define a set of guidelines for Internet message bodies for files containing chemical related data. An Internet Draft was written and made available for comment and a discussion group initiated. Problems have arisen following discussions with the computer scientists comprising the "Internet Engineering Task Force" (IETF) who, although being the "de-facto" standards body for the Internet, did not feel themselves competent to assess the Draft. In order to make progress with this concept a IUPAC working party, chaired by Dr Henry Rzepa, has been set up to carry out an assessment of this work and report to the new IUPAC Committee on Printed and Electronic Publishing. Once the draft is accepted and published under the IUPAC banner it is hoped that the IETF will ratify the newly defined mail content "chemical".

Introduction

Any software package accessing data from the "Information Super-Highway" must understand a set of basic rules regarding the storage and transmittal of data. These rules are laid down in documents known as RFCs (strangely—Request for Comment!). RFC 1521 covers the so-called Multipurpose Internet Mail Extensions or MIME types for short. This RFC standardises the use of specific file extensions which can be expected to be found on the Internet and effectively reserves the use of these extensions for files of a particular data content. Anyone familiar with accessing the Internet with World Wide Web packages will have some experience of receiving HTML (Hyper Text Markup Language) documents. These consist of embedded codes which are interpreted by a web browser software package. Files of this type should carry the extension .HTM or .HTML. The coding should be invisible to the typical user but the information is vital to pro-

cessing such as database searching. An NMR spectroscopist may wish to rephase the data to his own standards before further processing.

ple responsible for writing the software packages and implementing these standards or to people having to plan for future oriented data storage, archival and retrieval.

Unfortunately, the original RFC 1521 is limited in the type of data it regards as being of interest to the users of the Internet. If you wish to make data content available as formatted text, or pictures, audio or video media types you are well catered for by RFC 1521 but chemical specific information must currently be forced into one of these definitions. An example would be a chemical structure with associated NMR and infrared spectral data. All three data sets would have to be transmitted as pictures if current MIME protocol was to be observed.

Chemical MIME

In February 1994 the original suggestion for an Internet primary chemical media type was suggested on the Computational Chemistry discussion list (CCL) and to chemistry and biology newsgroups. Using the input from these and other sources attempts have been underway to define a set of rules to ease the handling of scientific chemical related data over the Internet. A team consisting of Henry Rzepa at Imperial College London, Peter Murray-Rust at Glaxo in Stevenage and Ben Whitaker of Leeds University strayed into the realm of the computer scientists and designed to write an Internet Draft defining a so-called "Chemical Primary Content Type" for MIME. This would extend the allowed primary contents from purely IMAGE, AUDIO, VIDEO etc, to include CHEMICAL. The need for the new content-type is the belief that people handling chemical data have requirements from the data received well beyond just observing the data. An example for the use expected from received data would be an infrared spectroscopist converting his data from %Transmission into Absorbance then rescaling to the wavenumber range of interest to them before further process-

ing such as database searching. An NMR spectroscopist may wish to rephase the data to his own standards before further processing.

In the chemical community we already have one major advantage over Internet users from other fields in that we have for many years been developing and implementing standard formats for data transmittal, storage and retrieval. We have access to large data archives, and already use many "de-facto" standards for digital data transmission in our field.

Proposed chemical sub-types names

A list of the proposed chemical sub-type names is shown in Figure 1. This list is not supposed to be the total of all possible sub-types but to concentrate on a small number of existing formats across the field of chemical informa-

Primary/sub-type	Suggested file extension
chemical/cxf	cxf
chemical/mif	mif
chemical/pdb	pdb
chemical/cif	cif
chemical/mo1-mo1file	mo1
chemical/mo1-sd1	sd1
chemical/mo1-rxn	rxn
chemical/emb1-d1-nucleotide	emb, emb1
chemical/genbank	gen
chemical/nebi-asn1	asn
chemical/geg8-sequence	geg
chemical/daylight-smiles	smi
chemical/rosdal	ros
chemical/macromodel-input	mmd, mmod
chemical/mopac-input	mop
chemical/gaussian-input	gau
chemical/jcamp-dx	jdj
chemical/kinemage	kin

Figure 1. The proposed chemical sub-type names.

How MIME works

Anyone familiar with some of the finer details of Microsoft Windows could have come across similar functionality to that proposed here for Internet access software in the simple File Manager. If you view the icons alongside the filenames generated by the File Manager program those files which File Manager knows how to handle have an icon resembling a page of text, whereas those the File Manager does not recognise have an icon resembling a blank sheet of paper. It is possible using the option "associate" in the pull-down menu FILE to link unknown file formats to specific programs. An example would be to link the file extension .SPC or .SP to a third party spectra data handling program such as GRAMS/386 or SPECTACLE. Similarly a commercial Internet reader could spot the .MOL file extension and start ISIS-DRAW as would File Manager if ISIS-DRAW is installed on the system and linked to the .MOL file extension.

tion. To qualify for inclusion the format should be well defined in the scientific literature and in wide use.

In July 1995 the authors of the draft met in Stockholm with the IETF committee and after over two hours of discussion it was concluded with the IETF expressing two major concerns. One worry was that if the chemistry community were to be allowed to define their own primary data content then the committee would be overwhelmed by requests from other specialised user communities with their own definitions. The second problem which they foresaw, probably correctly, was their own inability to assess the proposal due to a lack of specialist knowledge in the field of chemical information.

The answer to their first problem is probably purely organisational and should be well within the resources of the IETF to solve. A solution for the second worry has been found by going to the International Union of Pure and Applied Chemists (IUPAC). The original work had already been presented to the CCDB database committee in December 1994 and so the logical choice was to return to this committee when seeking an internationally renowned standards body in chemistry to validate the proposal.

As of August 1995 the IUPAC database committee will join with the Publications Committee to form a new Committee on Printed and Electronic Publishing (CPEP). Two new working parties are to be formed reporting to the CPEP—one will continue the work of the Joint Committee on Atomic and Molecular Physical Data–Data Exchange, and the second headed by Henry Rzepa will work on Chemical MIME and produce a paper in the style of Pure and Applied Chemistry for the inaugural meeting of CPEP.

Further information

Henry Rzepa and the team now plan to produce a IUPAC Draft for the inaugural meeting of the CPEP in August. Anyone interested in contributing to this development can join the discussions on the chemical MIME initiative by sending an e-mail to listserv@ic.ac.uk with the one line message: `subscribe chemime yourname`. The discussions can be viewed on the Web at

<http://www.ch.ic.ac.uk/hypertext/chemime/>

and the latest copies of the relevant Internet Drafts and RFCs can be obtained from

<http://www.ch.ic.ac.uk/internet/>.

CIRCLE 025 FOR FURTHER INFORMATION