ChemComm

A universal approach to web-based chemistry using XML and CML

Peter Murray-Rust,[a] Henry S. Rzepa,[b] Michael Wright[b] and Stephen Zara[b]

[a] *School of Pharmaceutical Sciences, University of Nottingham, UK*
[b] *Department of Chemistry, Imperial College of Science, Technology and Medicine, London UK*

Received (in Cambridge, UK) 28th March 2000, Accepted 14th June 2000
Published on the Web 18th July 2000

We report the first fully operational system for managing complex chemical content entirely in interoperating XML-based markup languages.

The World-Wide-Web (WWW) was originally developed as a collaborative tool for scientists. In 1994 we proposed its use as a novel and widely applicable global model for expressing chemical information in an interlinked and re-usable form.[1] This model, based on HTML, has become widely used but suffers from a number of structural problems. Chemistry represents its message both *semantically* (*e.g.* as a machine-processable connection table) and graphically (*e.g. presentation* through human-readable arrows, boxes, diagrams). The existing use of HTML in chemistry emphasizes presentation, and provides no structured extension mechanisms. A presentational approach is too flexible to be reliably interpreted by machines, and cannot *validate* the integrity of the chemical *content*. Thus the chemical subscript and superscript conventions in *e.g.* $CH_4^{+\cdot}$ are only unambiguously interpretable by humans. This severely limits the re-usability of the rapidly growing amount of high-quality chemistry available on Web pages.

We therefore developed[2] our model to support both types of markup. Chemical Markup Language (CML) can carry molecules, crystallography and reactions in a formal manner. It forms the core of a comprehensive approach to publishing and communicating chemical information for both humans and machines. It uses the protocols of the WWW Consortium (W3C),[3] whose goal is to support such *interoperability* with the strategic aim of achieving a "semantic web" where automatic processing of information is possible. The core W3C approach is the *meta-language* XML (eXtensible Markup Language) which was formalised in 1998 and which encourages the creation of discipline-specific languages such as CML. The W3C have created a family of protocols supporting most aspects of managing Web-based information; the following are most relevant here: *XML Schemas* are a formal description of the language and can support arbitrary datatypes and or validate complex documents. *Scalable Vector Graphics (SVG)* provide a framework where presentation and content can be robustly combined. *XSLT (XSL stylesheets)* is a very powerful tool for transforming documents. *XML Query Language* and XLinking are under development. *XSL(FO)* provides high quality formatted output for any XML application.[3]

Much time is currently wasted on processing *legacy* unstructured and often binary documents and poor semantics leads to serious information loss. The 'plug-compatible' XML approach guarantees a document to be searchable, stylable, sortable, transformable, mergeable, transmittable and printable without extra cost. CML uses these developments and for the first time offers a universal platform- and application-independent infrastructure for chemical information. Ideally we see all future document and publishing systems being converted to XML and describe here an implementation of these concepts termed ChiMeraL.

Technical documents are often multidisciplinary and component-based and draw their components from several XML Schemas. Subdisciplines, *e.g.* MathML (for mathematics), CML, SVG, are identified within documents by discrete *namespaces*. To avoid *collision* their *tag names* are mapped onto globally unique URIs (Uniform Resource Identifier); http://www.xml-cml.org/ serves to uniquify any CML element. Components will often be glued together with XHTML (the XML version of HTML).[4] *Transformation/reformatting* through XSL stylesheets is a particularly important operation because it allows extraction of document sub-components such as molecules for *e.g.* redrawing in SVG (*vide infra*) or outputting in any desired format.

Besides transformation, various operations can be automatically performed, often without knowing the precise document structure. These include searching, whereby XML elements ('components') can be located by local or global context in a document or through content, re-use by fragmenting or combining XML documents, rendering/viewing in a browser window, high-quality printing, and data authentication services.[5]

'Chemical' information usually requires several other types of markup (text, numbers, tables, graphics, etc.). Such CML document components can come from many sources, either directly or after conversion from legacy formats. The sources include: instruments, databases, dictionaries and catalogues; hand-editing/authoring of chemical information; primary and secondary publications, and computational chemistry tools. To illustrate how the operation of transmitting, querying and processing compound chemical data within a Web browser can be combined, we have created a Web-based collection of XML tools termed ChiMeraL.[6] Its use involves five distinct stages (Fig. 1).

(1) Conventional HTML is used in conjunction with JavaScript in a browser to allow the user to select from an XML library containing CML components (Fig. 2). The library could be a server-based collection or could be generated by selective querying of a larger XML document held on an XML repository.

(2) The XML document is validated for integrity against a Schema. This is a major improvement over HTML, which has no validation for chemical or most other content. Validation can be server or browser-based.

(3) Selection of a suitable stylesheet for popular chemical editing and display programs, *e.g.* for 2D and 3D molecular co-ordinates and numerical information such as spectral/analytical data. The stylesheet is used to convert CML elements to legacy formats (*e.g.* MDL Molfile, Minnesota XYZ format, JCAMP DX format) for display using existing applets (*e.g.* JME, JMol, JSpec, Marvin, SDA), or it could be directed towards CML-compliant applets without the need for such legacy transformation. The XSL transform can again be server or browser based. The use of a stylesheet also provides the possibility of deriving new quantities from the original data such as *e.g.* aligning a range of pharmacophores or computing diverse molecular properties. The ChiMeraL demonstration (Fig. 2)† includes a range of XSL stylesheet fragments, XML example documents and a CML schema, together with utility programs which can generate CML documents.[6]

(4) The output from the stylesheet is displayed as a web page. The original CML components (*e.g.* < molecule > < /molecule >) are wrapped with suitable XHTML to allow appropriate display (Fig. 2).

(5) At this stage, the user could edit or add to the document

DOI: 10.1039/b002483j

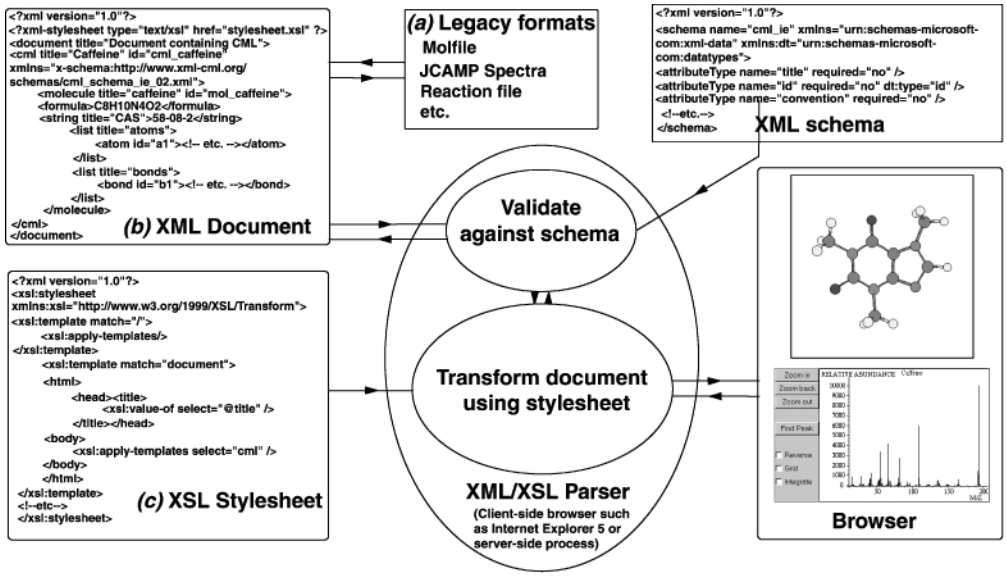

Fig. 1 Data flow between the principal components of ChiMeraL (a) The creation of a *structured document* in XML syntax from unstructured *legacy* formats (MOL, JCAMP, *etc.*). The validity of the process is checked by a CML-specific XML schema. (b) Illustrative CML document whose *root element* contains a *child molecule element*. In turn the *molecule* contains a *list* element, itself the parent of many *atom* elements (omitted for clarity); bonds are analogous. *Attributes* such as *convention* identify the usage of terms and values, while the *id* attribute identifies every element uniquely. (c) Illustrative XSL stylesheet, which transforms CML to XHTML. The *title* of the *molecule* is transformed to the XHTML document title, while the content of the *molecule* is transformed into various legacy formats (MOL, JME, *etc.*) for display by applets.

by selecting an appropriate stylesheet to invoke appropriate software tools, and with appropriate validation, create a new annotated XML document which could be returned to the original document repository. Multiple authorship of a document is possible since subsequent stylesheet transforms could *e.g.* extract either the original document, or any individual's additions to it.

Chimeral is offered as OpenSource, *i.e.* available as source code which can be used for any purpose, but especially for the development of new ideas, tools and resources in this area. XML has benefitted greatly from this movement and Chimeral uses several OpenSource components (XML tools, JMol, JSpect).[7] OpenSource is an effective means of developing high-quality robust applications very rapidly and we encourage developers to increase the volume of WWW-based collaboration in chemistry.

We have described a method for completely transparent Internet-based transfer of chemical information from creator/author to reader/user. Applications of this method might include a document such as this journal article being automatically abstracted for molecular content, and stored or used for calculations or searching. One can envisage *e.g.* laboratory robots scanning CML-based reaction schemes for starting materials and ordering them from CML-based 'dot.coms' on the WWW, and recording this in electronic notebooks. Instruments could directly output universally processable spectra and data. Databases could accept chemistry in many traditional formats and re-offer them transparently. Computational chemistry and modelling programs could be used routinely for adding information content to molecules. Finally, we note that the use of XML allows other disciplines (bioscience, pharmaceutical, materials, patents) to include semantically rich chemistry in their information.

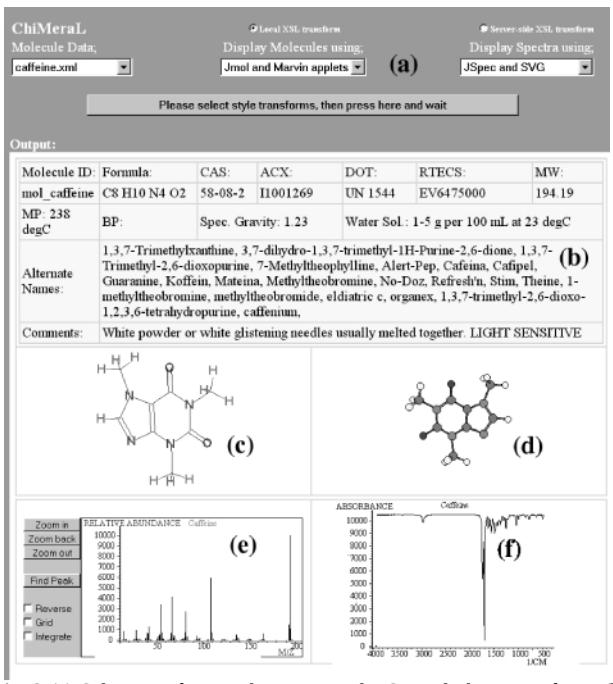

Fig. 2 (a) Selection of XML document and XSL stylesheet transform. (b) Property display using HTML table. (c) 2D Molecule display using Marvin applet. (d) 3D Molecule display using JMol applet. (e) Spectral display using JSpec applet. (f) Spectral display using SVG and Adobe plugin.

Notes and references

† This article expressed as XML is also available *via* the electronic supplementary information pages at http://www.rsc.org/suppdata/cc/b0/b002483j/

1 H. S. Rzepa, B. J. Whitaker and M. J. Winter, *J. Chem. Soc., Chem. Commun.*, 1994, 1907; 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; H. S. Rzepa, P. Murray-Rust and B. J. Whitaker, *Chem. Soc. Rev.*, 1997, 1; H. S. Rzepa, P. Murray-Rust and B. J. Whitaker, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 976.

2 The original concept was described in P. Murray-Rust, C. Leach and H. S. Rzepa, *Abs. Papers. Am . Chem. Soc.*, 1995, **210**, 40-COMP. For a formal description of CML version 1.0, see P. Murray-Rust and H. S. Rzepa, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 928.

3 World-Wide Web Consortium (W3C). See http://www.w3.org/

4 W3C Working draft: 'XHTML 1.0. The extensible HyperText markup language for a specification of the XHTML standard', http://www.w3.org/MarkUp/

5 W3C Working draft 'XML-Signature Syntax and Processing', http://www.w3.org/TR/xmldsig-core/

6 ChiMeraL can be viewed at http://www.ch.ic.ac.uk/chimeral/

7 V. Kiernan, *Chronicle Higher Education*, 1999, http://www.chronicle.com/free/v46/i11/11a05101.htm. See also D. Gezelter, http://www.openscience.org/