# Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators

Peter Murray-Rust,[†] Henry S. Rzepa,*[,‡] Mark J. Williamson,[‡] and Egon L. Willighagen[§]

Unilever Centre for Molecular Informatics, University of Cambridge, Cambridge, UK,
Department of Chemistry, Imperial College London, London, UK SW7 2AY, and
Laboratory for Analytical Chemistry, University of Nijmegen,
Toernooiveld 1, NL-6525 ED, Nijmegen, The Netherlands

Received October 30, 2003

Examples of the use of the RSS 1.0 (RDF Site Summary) specification together with CML (Chemical Markup Language) to create a metadata based alerting service termed CMLRSS for molecular content are presented. CMLRSS can be viewed either using generic software or with modular opensource chemical viewers and editors enhanced with CMLRSS modules. We discuss the more automated use of CMLRSS as a component of a World Wide Molecular Matrix of semantically rich chemical information.

## INTRODUCTION

There is increasing recognition that the World Wide Web has vast untapped potential as an infrastructure for structured **data** interchange rather than just being a medium for delivering documents. This recognition underpins the Semantic Web,[1] Berners Lee's vision of its evolutionary future. Its construction will involve developing mechanisms which precisely and predictably associate data with descriptions of its meaning, context, and validity (whether it is fit for purpose). XML is now universally recognized as providing syntactic architectures for achieving this. XML is itself a specification for creating families and subfamilies of more specific markup languages. The best known of these is XHTML, which evolved from the original requirements of the Web to create documents which could be rendered readable for humans via the Web browser. In fact, XML was designed to serve an even more fundamental role for specifying data and data structures. Via a formalism known as namespacing, several XML languages can in turn be combined to create a compound document, and these components can be transformed into other appropriate forms by invoking other XML-based tools known as stylesheets. These can be appropriate either for presentation to a human for reading or for further processing (transformation) by empowered software agents according to defined algorithms. In recognition of the dual purposes that XML can serve, we have coined the term *datuments*[2,3] to describe these compound information objects.

As the structure of datuments and the number of components they may contain grows more complex and the datuments themselves become larger (possibly very much larger), methods for achieving higher order organization and aggregation become required. Metadata (data about data) provides a mechanism for providing concise descriptions of the type of content expected in the datument, enabling high level decisions about further processing or filtering to be made. What is required is a more finely grained elaboration of the MIME approach we used to achieve appropriate postprocessing of discrete data files on the first generation Web.[4]

At this stage, it is worthwhile noting an early experiment of ours in creating a complex environment of documents, chemical data, metadata, and processes applied to the collection, using the Web technologies available in 1995. The ECTOC electronic conferences[5] were designed to investigate innovative electronic metaphors for the conventional but expensive and time-consuming physical meetings which the scientific communities have evolved over many decades to promote cross fertilization of ideas among humans. Each of the four ECTOC conferences held during the period 1995−1998 contained about 100 posters and articles. These were an intertwingled mixture[6] of bit-mapped images, chemical data expressed in a variety of formats,[4] discussion forums and lists of titles, with associated provenance of authors, comments by participants, and clear time stamps. Part of this experiment was an attempt to create navigational aids to this diverse but inter-related information collection which would help participants to identify *chemical* subject matter of interest to them. This would in turn help identify similarities in this material which would promote serendipitous chemical discovery. While conventional navigational aids were presented, (tables of contents, subthemes, indices) we also introduced a novel metadata based mechanism using the Meta Content Framework (MCF) which had been developed by a small group within Apple Computer. MCF was used to provide metadescriptions of the various conference components and was presented to the human as a nonlinear visual navigation map of the conferences containing links between related components of the conferences based on this metadata. The MCF-based map and software to view it was included on the subsequent ECHET96 CDROM archives, although its use was not developed further at the time. A particular limitation was that chemical information such as "how many molecules are described in this article" still had to be (slowly) organized and then discovered by the human editor or reader.

* Corresponding author e-mail: h.rzepa@ic.ac.uk.
[†] University of Cambridge.
[‡] Imperial College London.
[§] University of Nijmegen.

10.1021/ci034244p CCC: $27.50 
Published on Web 01/20/2004

MCF itself underwent a number of evolutions after being abandoned by Apple in 1996, including adoption by Netscape for use in their own information portals under the name RSS. The ideas espoused by MCF were also adopted by the W3C for their Resource Description Framework (RDF), itself seen as an integral part of the Semantic Web noted in our introduction. These various concepts, along with a decision to recast the syntax into XML, merged around the year 2000 with the specification of a protocol now known as RSS (RDF Site Summary) 1.0. The background history to this evolution has been recently summarized elsewhere,[7] and this latter article also provides a concise description of various more formal metadata schemas that can be incorporated into RSS. These include the standard Dublin Core (DC) schema and PRISM[8] which provides an XML metadata vocabulary specifically for journal publishing.

With RSS now cast as an XML language, for the first time it becomes possible to consider how an entire collection of data, metadata, and information could be constructed using XML components (something not possible at the time of the ECTOC conferences) and which in turn could make use of the increasing array of standard (often opensource) software tools which have become available for processing XML. In an earlier article where we first introduced the ideas behind RSS,[9] we concluded by alluding to the prospects of such unification in the specific area of chemistry. In the present article, we provide explicit examples of the use of RSS to provide metainformation about three diverse chemical sites, including mechanisms for molecule discovery largely absent in conventional Web pages. We also show how the use of XML throughout greatly facilitates the development of authoring applications which make use of these concepts via reuse of standard components and tools.

## IMPLEMENTATIONS OF RSS FOR CHEMICAL DATA SOURCES

We have previously described[9] the structure and use of a basic RSS document, noting how XML namespaces[10] allowed explicit chemical information and metadata to be added. Here we elaborate upon the topic of namespaces which we had introduced in the earlier article and then illustrate this usage via three deliberately diverse examples of how these concepts can be added to repositories of chemical data.

**Namespaces and RSS 1.0.** Namespaces are central to modern XML but not always widely deployed in some domains, including chemistry. Large documents (e.g. journal articles, regulatory submissions, patents, books, etc.) may contain material from many disciplines and be created by many authors. Moreover material may be copied or transcluded from other sources. It is unrealistic to expect a globally controlled vocabulary, and namespaces allow authors to create local information components and merge them without name collisions. Thus chemistry and XML both use the vocabulary "element" which would collide unless disambiguated.

Namespaces use URIs[10] for disambiguation. The creator of a namespace devises a unique URI, usually based on their domain name to provide uniqueness. Thus many XHTML documents start with the following syntax:

```
<html xmlns="http://www.w3.org/1999/xhtml">
```

This states that, by default, all elements and attributes in the document belong to the "http://www.w3.org/1999/xhtml" namespace, conventionally referred to as the "XHTML V1.0" namespace. The use of the HTTP protocol for namespaces is an unhappy and confusing syntax. It is not required and could be replaced by a URN (a naming, rather than addressing convention). We dispose of the following common myths:

1. "You have to be connected to the web to use namespaces". In fact no XML tool should try to resolve these as addresses and if it does it is an error.

2. "There is a web page with something useful at the URL address". It is not an address, and it is only coincidental if there is a page to which it resolves.

In practice many namespace designers do put some form of specification or help pages at the "URI address", but there is no consistency in content or syntax.

The namespace specification should be read carefully by designers of multinamespace documents (including RSS), but the following simple guide and example illustrating namespace use in an RSS 1.0 document[11] (Scheme 1) is sufficient for this article.

- The document may (but need not) start with an XML declaration (line 1).
- The document may (but need not) have a DTD reference: <!DOCTYPE foo SYSTEM "http://foo.org/bar.dtd">. In practice it is difficult to construct DTDs for multinamespace documents, and they are not normally DTD-validatable, so our approach does not use them. It is possible to create schemas which describe and validate, but content models and attributes are complex in RDF and schema-based validation is not always cost-effective.[12]
- Processing instructions (line 2) are not under namespace control. PIs are hints or instructions to processing software but do not affect the content of the document. The PI target ("xml-stylesheet") is used in user-agents (browsers) which support the W3C style guidelines. It means "if you are stylesheet aware, and if you support CSS stylesheets ("type" pseudoattribute) then retrieve the stylesheet at URL ("href") and apply it". In this example, the stylesheet specified ensures that if the RSS feed is displayed in a browser, the result is at least readable.
- There can be any number of namespaces in an XML document. For RSS there will normally be at least RSS, RDF, and DC. Most human-readable news feeds also include XHTML. In the example (Scheme 1) there are the following:

```
"http://www.w3.org/1999/02/22-rdf-syntax-ns#"
"http://purl.org/rss/1.0/"
"http://usefulinc.com/rss/manifest/"
"http://purl.org/dc/elements/1.1/"
"http://www.xml-cml.org/schema/cml2/core"
```

Namespaces do not need to be declared at the start of the document unless they are required in the rootElement. They can also be declared several times.

- All namespaces (except the default namespace, which in this example corresponds to the schema for RSS 1.0 itself) are associated with a prefix. This prefix is arbitrary and is required to be unique only within the document. The prefixes are determined by the xmlns pseudoattribute mechanism. Thus xmlns:dc="http://purl.org/dc/elements/1.1/" associates the prefix "dc" with the namespace "http://purl.org/dc/elements/1.1/". Any element (or attribute) whose name starts

**Scheme 1.** RSS Code Generated for a CMLRSS Aggregation Feed (Shown Truncated)

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet href="http://www.w3.org/2000/08/w3c-synd/style.css" type="text/css"?>
<rdf:RDF
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns="http://purl.org/rss/1.0/"
         xmlns:mn="http://usefulinc.com/rss/manifest/"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:cml="http://www.xml-cml.org/schema/cml2/core">
 <channel rdf:about="http://www.ch.ic.ac.uk/rzepa/rss/">
   <title>Chemstock</title>
   <link>http://www.ch.ic.ac.uk/csdemo</link>
   <description>A chemical database based upon MySQL and PHP</description>
   <dc:publisher>Chemstock</dc:publisher>
   <dc:creator>rzepa@imperial.ac.uk</dc:creator>
   <image rdf:resource="http://www.ch.ic.ac.uk/logo.gif" />
   <items>
     <rdf:Seq>
       <rdf:li rdf:resource="http://www.ch.ic.ac.uk/csdemo/index.php?type=item&id=437" />
      </rdf:Seq>
   </items>
 </channel>
 <image rdf:about="http://www.ch.ic.ac.uk/logo.gif">
    <title>ChemStock,
       Imperial College London</title>
    <url>http://www.ch.ic.ac.uk/logo.gif</url>
    <link>http://www.ch.ic.ac.uk/csdemo</link>
    <dc:description>A chemical database based upon MySQL and PHP</dc:description>
 </image>
 <item rdf:about="http://www.ch.ic.ac.uk/csdemo/index.php?type=item&id=437">
   <link>http://www.ch.ic.ac.uk/csdemo/index.php?type=item&id=437</link>
   <title>Caffeine</title>
   <description>Caffeine</description>
   <dc:subject>Caffeine</dc:subject>
   <dc:date>2003-09-16T00:00:00-00:00</dc:date>
  <dc:creator>demo</dc:creator>
<cml:molecule xmlns:cml="http://www.xml-cml.org/schema/cml2/core"
 title="CAFFEINE">
<cml:metadataList title="generated automatically from Openbabel">
<cml:metadata name="dc:creator" content="OpenBabel version 1-100.1"/>
<cml:metadata name="dc:description" content="Conversion of legacy filetype to CML"/>
<cml:metadata name="dc:content"/>
<cml:metadata name="dc:rights" content="unknown"/>
<cml:metadata name="dc:type" content="chemistry"/>
<cml:metadata name="dc:contributor" content="unknown"/>
<cml:metadata name="dc:date" content="Wed Sep 17 15:03:57 bst 2003"/>
<cml:metadata name="cmlm:structure" content="yes"/>
</cml:metadataList>
<cml:atomArray atomID="a1 a2 a3 a4 etc"
elementType="C C N C O N C O N N C C C C H H H H H H H H H H"
formalCharge="0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0"
x3="0.310000 -0.355000 -1.534000 etc"
y3="0.025000 -0.629000 -0.232000 etc"
z3="0.266000 -0.740000 -1.149000 etc"/>
<cml:bondArray atomRefs1="a1 a1 a1 a2 a2 a3 a3 a4 a4 etc"
atomRefs2="a2 a7 a9 a3 a10 a4 etc"
order="2 1 1 1 1 1 2 etc"/>
</cml:molecule>
  </item>
</rdf:RDF>
```

with dc: has a namespace URI of xmlns:dc="http://purl.org/dc/elements/1.1/". Thus the element <dc:date>2003−09−16T00:00:00−00:00</dc:date> consists of the pair {localname+namespaceURI} of {"date"+"http://purl.org/dc/elements/1.1/"}. This is what is passed to an application program, and the actual prefix used is irrelevant. We emphasize this through the example: <dCore:subject xmlns:dCore=""http://purl.org/dc/elements/1.1/">Caffeine</dCore:subject> which is semantically identical to <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/">Caffeine</dc:subject> (or, since the dc prefix is in scope, to: <dc:subject>Caffeine</dc:subject>).

**Example 1. The ChemStock System.** ChemStock, described fully elsewhere,[13] is a simple data capture system designed using OpenSource software to maintain an inventory of molecules and selected properties. It is designed to answer questions such as "what are the latest chemical additions to the inventory?" or "what is the owner/location of a specific item?". At its heart is a user interface written using the PHP scripting system and populated by querying a MySQL database. Part of the data capture process involves contributors supplying molecular coordinate and atom connection table descriptors of a collection of molecules. Currently, two different formats for contributors to upload this information are implemented, the MDL Molfile format (which is not XML-compliant) and CML (Chemical Markup Language, which is XML based).[14] Because the Molfile cannot be used within the XML-based RSS format, it is necessary to preprocess this legacy format into CML using the OpenBabel converter.[15] As XML conformance in the chemical community increases, the need for this particular step, itself susceptible to potential data loss, will eliminate. The CML expression of the molecule is then stored in the MySQL database, along with other information such as the creation/modification date, the name of the author, and a text-based descriptor (i.e. name) of the substance. Other molecular properties could also be computed at this stage if needed, most prominently the IChI unique molecular identifier.[16]

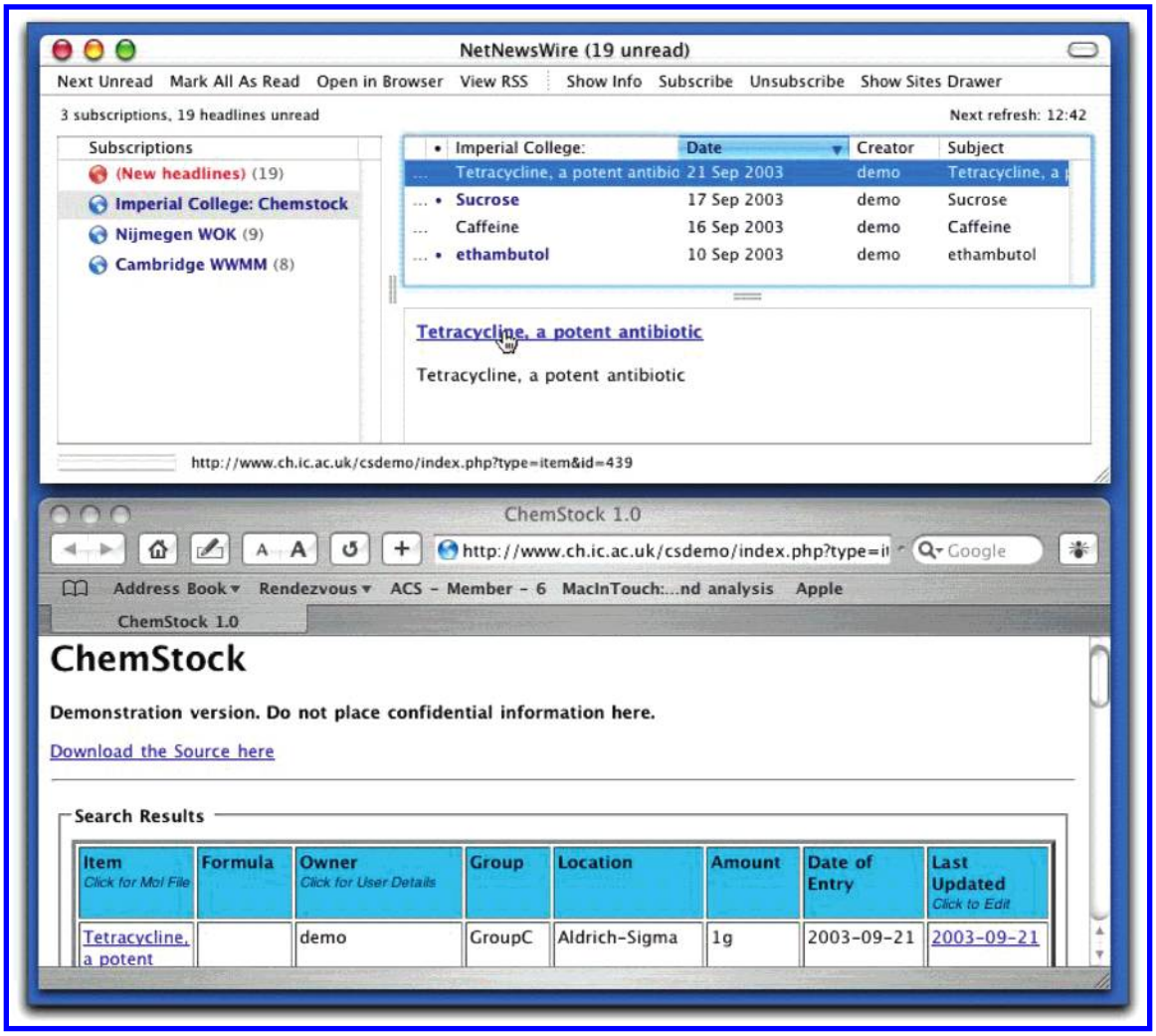


**Figure 1.** A generic RSS Viewer illustrating the RSS feed from http://www.ch.ic.ac.uk/csdemo/feed.php (top) and expansion of one item within a browser window (bottom).

To create a CMLRSS feed, the MySQL database must then be queried to retrieve the information and to format it according to the RSS 1.0 specification[11] using appropriate PHP tools.[17] The details of the changes required for the published ChemStock system are outlined in Appendix 1 in the Supporting Information. The resulting RSS document is shown in Scheme 1.

The subscription URL takes the form http://www.ch.ic.ac.uk/csdemo/feed.php?num=5. This default query retrieves the last five entries added by users to the ChemStock database, including any CML components, along with metadata appropriate for the DC schema such as the author, date, and description. An example of the RSS generated is shown in Scheme 1. If used within a generic RSS news reader (which does not support the CML namespace) the results take the form shown in Figure 1. Note particularly that the CML components are not displayed (since no handler for these is present or has been specified) but that the Dublin Core (DC) fields are displayed, and these can be used to sort the aggregated display. Selecting the link associated with any individual entry will display the ChemStock page.

**Example 2. The Dutch Dictionary on Organic Chemistry.** The Dutch Dictionary on Organic Chemistry (WOC, "Woordenboek Organische Chemie" in Dutch) is an 8-year old Web site about organic chemistry and mainly in Dutch.[18] It contains descriptions of terminology, named reactions, and compounds. The 10 most recently changed items in the dictionary have been made available as a CMLRSS feed at http://www.woc.sci.kun.nl/cgi-bin/rssfeed.rss. The content is similar to that of the ChemStock RSS feed, including CML metadata for molecular content. Though not available at this moment, it is planned that the named reactions will be available using the CMLReact namespace. Programming details are summarized in Appendix 2.

**Example 3. The World Wide Molecular Matrix (WWMM).** The molecular matrix[19] is a bold and innovative attempt to create a global open repository of molecular information and associated properties using a grid-based peer-to-peer model for collaboration and dissemination. The adoption of XML syntax throughout ensures that the diverse molecular information held in the matrix can be aggregated in a fully extensible and interoperable manner and that it is exposed to other chemical and nonchemical disciplines that may wish to access it in a semantically rich manner. Another pervasive concept is that of adding value to existing information ("accretion"). The Cambridge node in the WWMM for example can process molecular information contributed by users and add to it via e.g. a full MOPAC-based[20] quantum mechanical computation of selected molecular properties. Other nodes on such a grid would compute other properties. The matrix gains content from the contributing user, the latter gains a valuable property calculation,

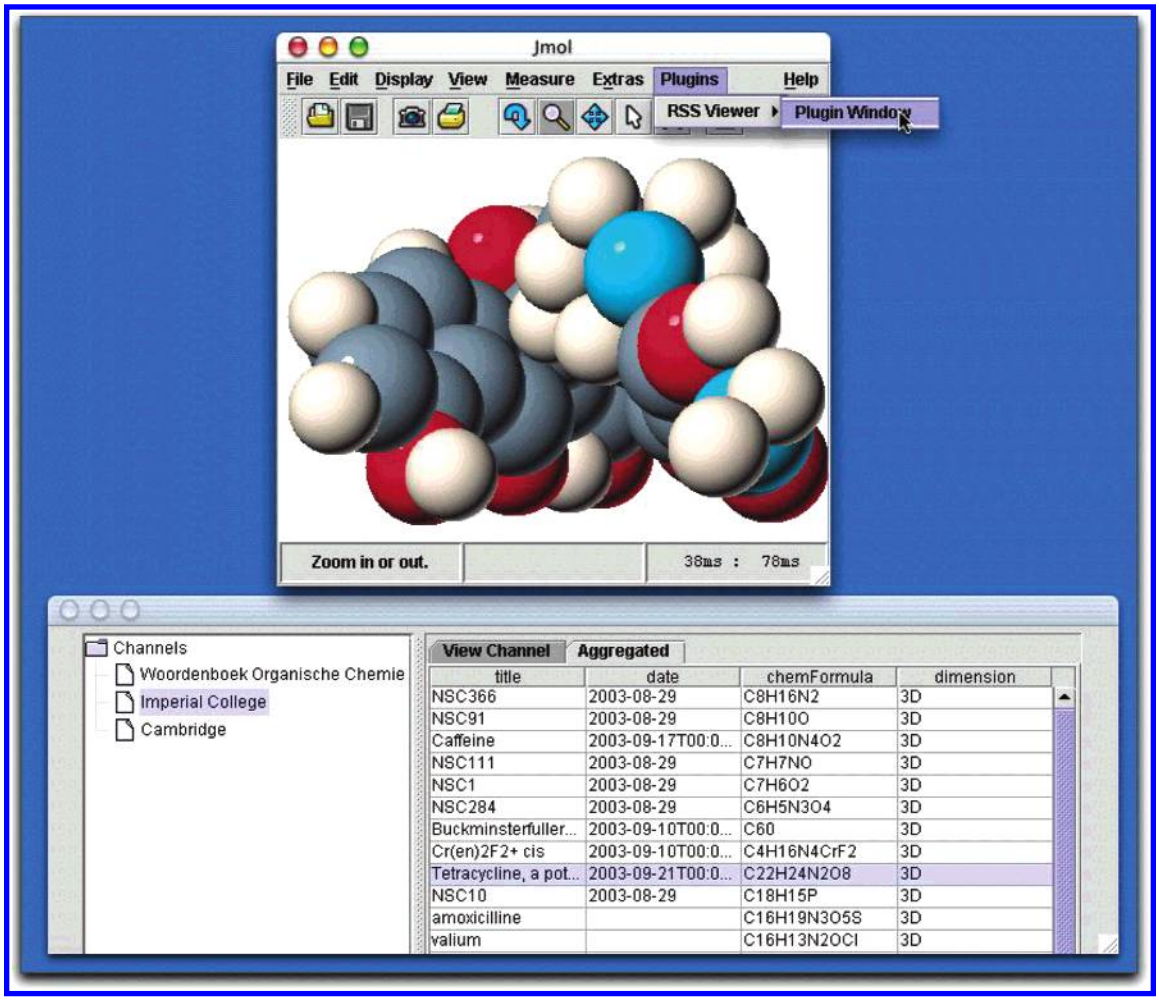


**Figure 2.** The Jmol 3D Molecule viewer showing the RSS plugin window. Three CMLRSS channels are shown sorted by date, with the Jmol window showing the most recent item. The formula is computed from the CML molecular information, and the dimension indicates what type of coordinates are available (1D, 2D, 3D, 2D+3D, fractional etc).

and the community gains from patterns that may emerge from the aggregation of this on a large scale. Within such an environment therefore, it becomes valuable to readily identify new entries, or newly computed properties for existing entries, and to filter and sort these according to specified criteria. CMLRSS provides a mechanism for achieving this. The RSS feed http://wwmm.ch.cam.ac.uk/Bob/rss contains the appropriate chemical metadata for selected entries to the WWMM which could form the basis for further interactions with the matrix.

## CHEMICAL POSTPROCESSING AND AGGREGATION OF RSS METADATA

RSS functionality is not limited to generic viewers but can also be incorporated into chemical application software. This has been done for the opensource programs Jmol[21] and JChemPaint[22] via a plugin module interface written as part of the CDK (Chemistry Development Kit).[23,24] This functionalized RSS reader is then rendered capable of parsing the XML and extracting both the DC and e.g. the CML namespaced components for display. If atom coordinates (of various dimensionality) are present in the CMLRSS feed, these are extracted and displayed within the Jmol (Figure 2) or JChemPaint (Figure 3) window when the item is selected.

Additionally, a call is made to the CDK toolkit[24] to compute (in this example) the molecular formula for display, although clearly a much wider range of computed properties could be included either via modular functionality of the program itself or a call to an appropriate Web service. If several CMLRSS feeds are defined in the Jmol or JChemPaint properties file, the aggregated molecule entries can sorted by the various fields such as title, date, or formula. An example of filtering the content by atom type is shown in Figure 4. More complex calls to the CDK toolkit could in principle provide other sorting mechanisms, such as by e.g. chemical substructure. If the molecule is associated with a published journal article, then appropriate Prism[8]-based metadata can link to this information.

## DISCUSSION AND CONCLUSIONS

Scientific research generates large amounts of potentially valuable data. Most existing models for handling and disseminating such data adopt a variety of (often quite inadequate) approaches:[3,25]

1. The data are discarded at the completion of a project or archived on paper in a box file stored in a cupboard. The mere existence of such data is often forgotten.

2. The data are converted (by scanning or other means) to a PDF (Acrobat) file and submitted as Supporting Information along with the associated scientific publication. This material may be made available on a publishers Web site but possibly only for a limited period. It is unlikely to

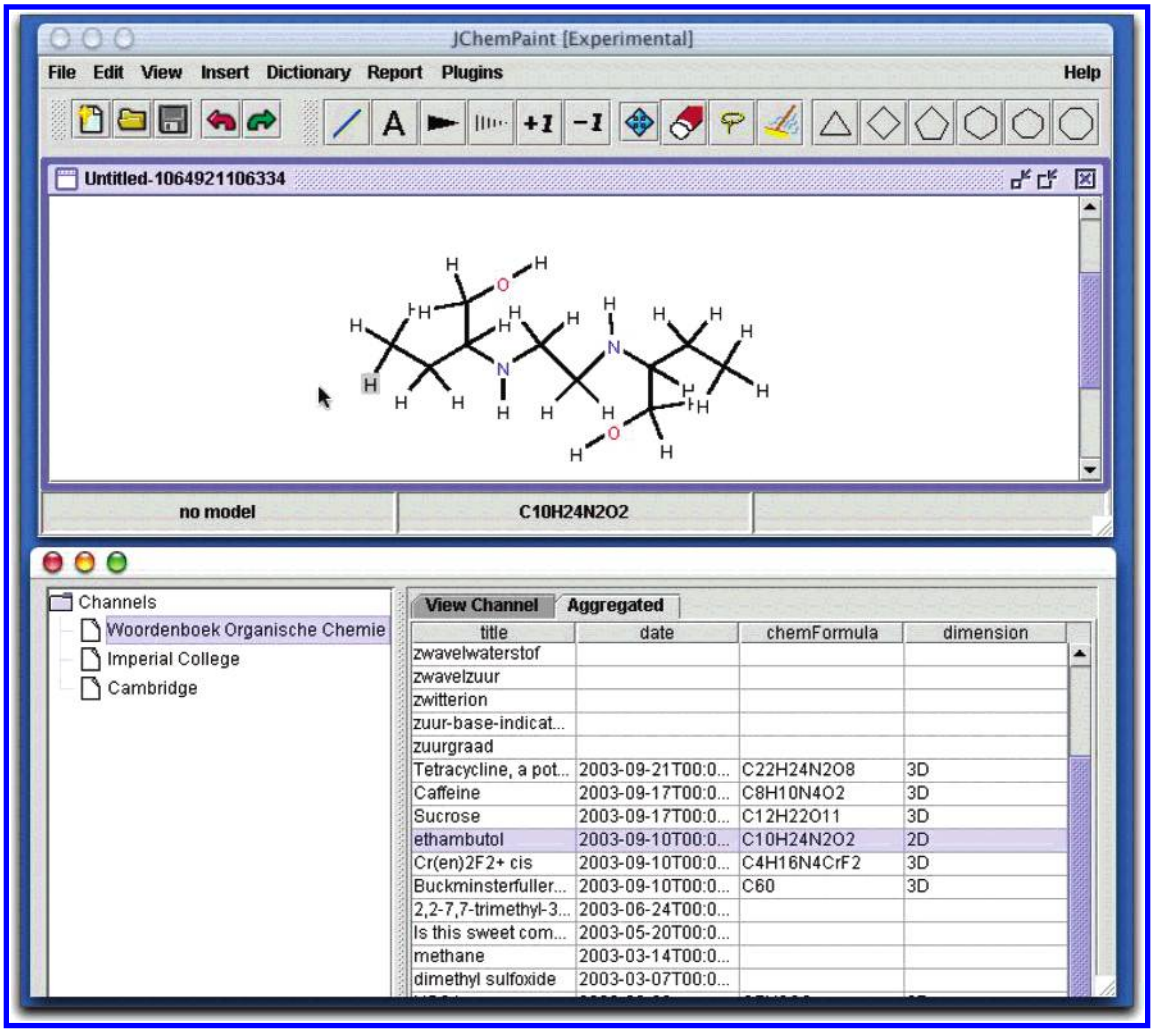


**Figure 3.** The JChemPaint 2D Molecule viewer/editor showing the RSS plugin window.

be indexed in any manner by the publisher and is therefore unlikely to be retrieved by any logical search procedures. Reuse of such data is only possible if a human (or good OCR system) rekeys it.

3. Data are saved in original (non XML) formats and made available via a publisher's (or author's) Web site in association with the article. It is potentially reusable by others if the particular formats (and any variations) are documented on the site or the prospective (human) user can "reverse engineer" the probable syntax and context and identify any software capable of handling it. These data too are unlikely to be indexed or searchable. Its existence is less likely to be advertised, and there may be many issues in its reuse, such as unambiguous knowledge of the particular scientific units used and ambiguity or multiple meanings for any terms describing the data.

4. Some forms of data, particularly those resulting from X-ray crystallography, may be submitted by the publisher to an agency or specialist for added-value processing such as validation and deposition into a formal database for subsequent searching and retrieval. Most such data are currently not "open" and hence available without a commercial license.

5. Ultimately, only a small proportion of scientific data will reach recognized definitive repositories such as Chemical Abstracts or Beilstein; here again its reuse is restricted to those who can afford it, and again proprietory software may be needed to process it.

It is quite likely that metadata about any stage in the above processes is also likely to be missing. There are no mechanisms for even identifying the existence of any data in categories 1 and 2, while metadata in category 3 may be restricted to information such as the date of deposition, the authors, and just possibly a hint (recognized only by a human) that it may include more specific information such as molecular connectivity information or molecular coordinates and their dimensionality (summarized as 1D, 2D, 3D, 5D (2D+3D), 3D/fractional, etc.). The provenance of the data may also be uncertain or inferred only by implicit association with (separately located) journal articles. The situation is not quite so dire in category 4, but even here it has been estimated that less than half of all determined crystal structures are actually deposited in any retrievable form. In addition, access to metadata (i.e. the existence of a molecular structure) is again restricted to those who have purchased access licenses and are using the dedicated software so provided. These data cannot easily be reconciled with other data about perhaps the same molecule (or indeed distinguished from data for an isomeric molecule) or with data resident in journal articles, etc. Some of the missing connections between the data and its provenance probably exist in collections in category 5 but again in a proprietary manner. Thus Chemical Abstracts and Beilstein can be seen as competitors and hence would have no (perceived) business case for providing mutual metadata about each other's

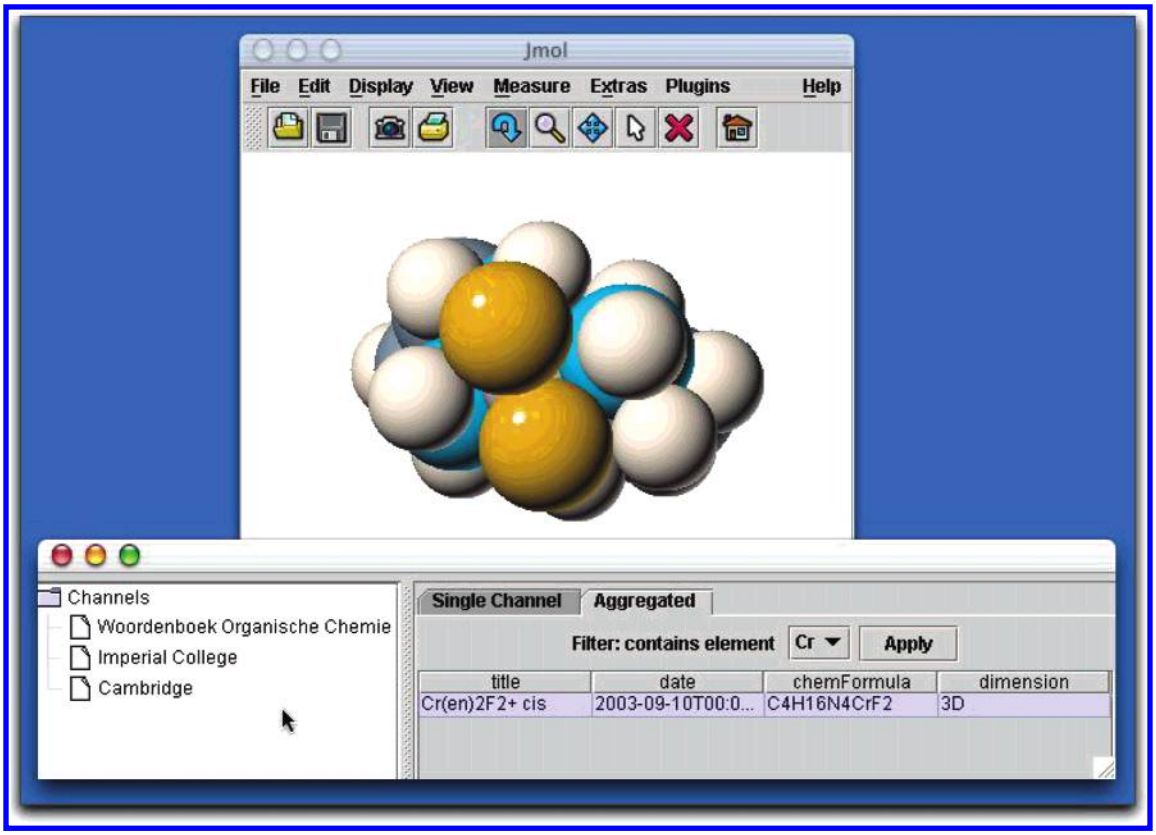


**Figure 4.** The Jmol 3D Molecule viewer showing the RSS plugin window with selection by atom type (in this example the element Cr).

holdings. Each of these databases holds extensive metadata about chemical substances, including for example enumeration of property lists (*plists*) for each substance. These *plists* however may not always overlap or inter-relate, and hence aggregation of the data is not possible.

Against this background, we introduce CMLRSS as an XML-based RSS carrier for chemical metadata. The three implementations described above allow the discrete capture of several types of metadata. Simple information such as date, author, text descriptions, etc. can be contained with the base DC schema. More finely grained chemical metadata (one could of course argue this to be an oxymoron!) is carried using the CMLCore schema. This enables capture of information such as the type of molecular coordinate available and also allows derived metadata such as a molecular formula to be algorithmically computed on the fly. The CMLRSS feeds described above include sufficient molecular information to allow humans (or software) to decide if the data is appropriate for the purpose they had in mind, including the possibility of filtering/sorting the information not just by e.g. molecular formula but also by substructure content or unique (IChI)[13] identifier.

The adoption of a unifying XML-based syntax has other particular benefits. We have illustrated this by writing a CMLRSS parser[26] using standard XML-compliant components which can be easily incorporated into a (2D) chemical editor (e.g. JChemPaint)[22] and a (3D) molecular viewer (Jmol).[21] This enables the user of either tool to subscribe to the appropriate CMLRSS feeds and hence to sort, select, and reuse the molecular data. The selection could be either by a human according to their own perceptions, or by software tools alone, but prearmed with particular chemical criteria. The current RSS plugin allows for filtering out news items that do not contain a specific element. By enfranchising chemical data sources of the types outlined above (particularly categories 1−4) with CMLRSS feeds, many of the problematic issues discussed above can be reduced if not entirely eliminated. This raises the intriguing prospect of what the role of aggregators such as 4 and 5 should indeed be. A case could indeed be made that the greater availability of interchangeable data and metadata from such sources, routinely accessible from standard chemical software, would greatly improve their business models. Certainly this prospect would move the chemical community a great deal closer to the realization of the chemical semantic Web.

**Supporting Information Available:** Details of the changes required for the published ChemStock system outlined in Appendix 1 and programming details summarized in Appendix 2. The copyright for the Supporting Information is held by the authors. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Berners-Lee, T.; Hendler, J. Publishing on the semantic Web. *Nature* **2001**, *410*, 1023−4.

(2) Murray-Rust, P.; Rzepa, H. S. Scientific publications in XML - towards a global knowledge base. *Data Sci.* **2002**, *1*, 84−98.

(3) Murray-Rust, P.; Rzepa, H. S. The Next Big Thing: From Hypermedia to Datuments. *J. Digital Information* **2004**, in press. For a reprint, see http://www.ch.ic.ac.uk/rzepa/jodi/.

(4) Rzepa, H. S.; Murray-Rust, P.; Whitaker, B. J. The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World Wide Web information exchange. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 976−982. See also http://www.ch.ic.ac.uk/chemime/.

(5) Proceedings of the Electronic Conference on Trends in Heterocyclic Chemistry (ECHET96), Rzepa, H. S., Snyder, J., Leach, C., Eds.; ISBN 0-85404-894-4; CD ROM Version, The Royal Society of Chemistry, 1997. See also http://www.ch.ic.ac.uk/ectoc/.

(6) The term intertwingling was coined for this process by Ted Nelson in his own far sighted vision encapsulated in the Xanadu project, now reincarnated as Zig-Zag: http://xanadu.com/zigzag/.
(7) Hammond, A. Why Choose RSS 1.0. http://www.xml.com/pub/a/2003/07/23/rssone.html.
(8) See http://www.prismstandard.org/.
(9) Murray-Rust, P.; Rzepa, H. S. Towards the Chemical Semantic Web. An introduction to RSS. *Internet J. Chem.* **2003**, *6*, article 4.
(10) For formal recommendations, see http://www.w3.org/TR/REC-xml-names/.
(11) RDF Site Summary (RDF); http://www.w3.org/2000/08/w3c-synd/.
(12) For an RSS validation service, see http://feedvalidator.org/.
(13) Rzepa, H. S.; Williamson, M. J. Chemstock: A Web-based Chemical Inventory system built from OpenSource Software Components. *Internet J. Chem.* **2002**, article 6. For a reprint, see http://www.ch.ic.ac.uk/csdemo/article/.
(14) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the Worldwide Web. Part 4. CML Schema. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757−772 and references therein. See also http://cml.sourceforge.net/.
(15) See http://openbabel.sourceforge.net/.
(16) Linstrom, P. J.; Tchekhovskoi, D. V. Transitioning to a structure based identification system. *Abstr. Papers, 226th ACS National Meeting*, New York, United States, September 7−11, 2003, CINF-085. See also http://www.iupac.org/projects/2000/2000-025-1-800.html.
(17) For information about RSSWriter see http://usefulinc.com/rss/rsswriter/.
(18) Willighagen, E. L.; Josten, G.; Fleuren, M. http://www.woc.sci.kun.nl/.
(19) World Wide Molecular Matrix (WWMM), http://wwmm.ch.cam.ac.uk/.
(20) Stewart, J. J. P. MOPAC: a semiempirical molecular orbital program. *J. Comput. Aided Mol. Design* **1990**, *4*, 1−105. See also http://www.cachesoftware.com/mopac/.
(21) For information about Jmol, see http://jmol.sourceforge.net/.
(22) Krause, S.; Willighagen, E.; Steinbeck, C. JChemPaint - using the collaborative forces of the Internet to develop a free editor of 2D chemical structures. *Molecules* **2000**, *5(1)*, 93−98. See also http://jchempaint.sourceforge.net/.
(23) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(24) Willighagen, E. L. Processing CML conventions in Java. *Internet J. Chem.* **2001**, *4*, article no. 4.
(25) Rzepa, H. S.; Murray-Rust, P. A New Publishing Paradigm: STM Articles as part of the Semantic Web. *Learned Publishing* **2001**, *14*, 177.
(26) The relevant module (rssviewer.jar) is available via the CDK site: http://cdk.sourceforge.net/.

CI034244P