Concatenated slides

Introduction

XML and the Launch of Chemical Markup Language

This talk consists of 26 HTML-based slides.

- Some slides contain GIFs as thumbnails. They are outlined (probably in blue). If you click on these a fullsized image will appear. To 'go back' you may have to click the right mouse button or or use some other way to 'go back in frame'
- Some slides contain technical bits (in red). Mainly examples of CML in action. These can be skipped and read later if necessary.
- The table of contents (an important XML concept) is slide001. It may be useful to clone a window with the TOC. You can also reach this page (and other pages) from the TOC.
- Each slide contains a NUMBER under the JUMBO icon (001-026) for navigation.
- There is NO Java during the lecture. To download the JUMBO browser, wait till after the lecture and follow the JUMBO icon
- JUMBO is being released in stages (Shakespeare at present). Chemistry will follow in about a week
- Technical JUMBO queries are best left to the post-lecture discussion list
- The Open Molecule Foundation is sponsoring a free JUMBO-based CDROM of CML.
- Many thanks to Chemweb and VEI.

toc

synopsis

Chemical Markup Language

- A platform- and convention-independent specification for information interchange in the molecular sciences
- Platform and application interoperability
- Compatibility with current W3C initiatives (XML)
- The Future of scientific publishing
- Addition of semantics and hypermedia to chemistry
- Tested in conjunction with working XML software
- Supported by the first XML browser,

  JUMBO

The Open Molecule Foundation is sponsoring a free demonstration CDROM for CML.

html

Simple, human-readable and authorable, easily learnt

Transmits text and graphics

Forgiving of errors;

Makes hypermedia (links) accessible to millions

Inspires 'memes' (spontaneous self-reproducing ideas)

Unsuited for machine2machine communication

Hyperlinks break easily and cannot be maintained

Tagset and syntax not extensible

Confusion of content and style/formatting

htmlmol

MIME provides type-stamping for files:

- text/html = HTML hypertext
- image/gif = Images (pixel maps)
- text/xml = an XML document

Browsers and other software make great use of these

- Helper applications can be set (e.g. application/postscript can call ghostscript)

Chemistry makes use of this through chemical/x-*

- chemical/x-pdb = PDB (Protein Data Bank) format [[Example]]
- chemical/x-mdl-molfile = MDL Mofile
- viewable with RasMOL or Chime (TM)

Problem: each type usually requires individual software

- Not always viewable on all platforms

Problem: few current 'standards' and little conformance

- Example: most 'PDB' files do not conform to PDB documentation
- Some file types are binary and platform-dependent

XML

The W3C language for structured documents on the WWW

- A complete solution for documents AND data

Format-free content + stylesheets

"SGML made simple"

- SGML is the most powerful document management language
- SGML is too complicated for the WWW

Major players such as SUN, Microsoft and Netscape

- Over 100 major players in the WWW Consortium, financing standards development

Targeted for browsers and servers

- All WWW software in 1998 will support XML

Supports structure, markup, and full hypermedia

- Addresses most current problems in WWW information infrastructure
- HTML is just an introduction; XML shows the real power of SGML and hypermedia

Customisable, extensible and designed for interoperability

- XML will run anywhere, anywhen and 'talk' to any other XML application

Simple to get started with; incredibly powerful

- If you know HTML you can start on XML tomorrow
- XML can provide much of the power of RDBs or OO systems

---

xdev

XML is part of the W3C effort which covers:

- Structured documents (XML)
- Hypermedia (links) (XLL)
- Stylesheets independent of content (XSL)
- Meta-data (data about documents and their structure) RDF
- Structured data (XML-data)
- Interoperability of domains (XML-name)
- Privacy, authenticity, etc.

XML is developed in an fast open collaborative process

- Editorial board WG (Working Group) meets (virtually) weekly

- 100 experts worldwide advise the WG
- A strictly managed process and timescale gives fast, high quality results
- Openness means that all particpants 'own' the result

Chemistry must learn from the XML process!

- CML is offered as a way to build on W3C/XML success
- The Chemical Markup Forum (CMF) offers its services to support this

---

whoxml

XML is exciting industries and activities such as:

- Banking (OFX)
- Metadata and Search engines (RDF)
- Push technology (CDF, Microsoft)
- Databases on the WWW (XML-data)
- Commerce (XML-EDI)
- Electronic drug submissions (CANDAs)
- Healthcare (HL7)
- Publishing

---

xmlbasis

XML provides structure and precise markup

- Examples of markup are given in later slides

Example of structure (technical, but visually important!):

Transformation of this example:

The fun of chemistry

by Molly Cool

Published by Elementary Press

Senior editor Nick L. Ion

---

structdoc

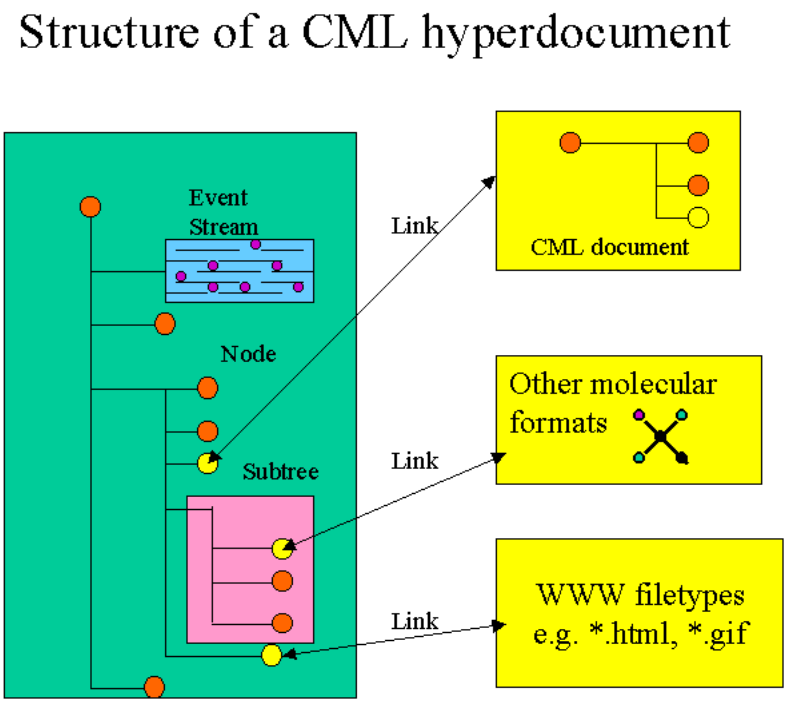
WARNING: Technical slide!

The structural power of CML

- CML documents can have a complex internal tree structure
- They can also link to other XML (and non-XML) documents.

(NOTE: ' [EventStream] ' corresponds to hypertext.)

[



]

Hypertext is an Event Stream

- A stream of character data with event flags to switch formatting on/off

[

<P>This is a <I>stream</I> where the <B>markup</B> switches formatting on and off. Hyperlinks can include material such as <A HREF="text.html">text</A> and <SRC IMG="image.gif">images. The XML format allows greater control of these as in <A HREF="mol.cml" XML-LINK="SIMPLE" SHOW="EMBED" ACTUATE="USER">'click to show molecule'.</P>

]

[

This is a *stream* where the **markup** switches formatting on and off. Hyperlinks can include material such as text and  images. The XML format allows greater control of these as in  (molecule has been clicked)

]

---

semantics

Tags alone have no meaning; semantics must be added

- <A> could mean Author, Anchor, Answer...
- Some tags (e.g. <MOLECULE>) may be human-readable, but are meaningless to computers

Semantics can be added through Java

- For <MOLECULE>, run jumbo.MOLECULE.class

Semantics can be added through hyperlinks to glossaries

- <ITEM TITLE="tryptophan" HREF="glossary.org/org/mols?tryptophan"/>

Semantics can be added through stylesheets

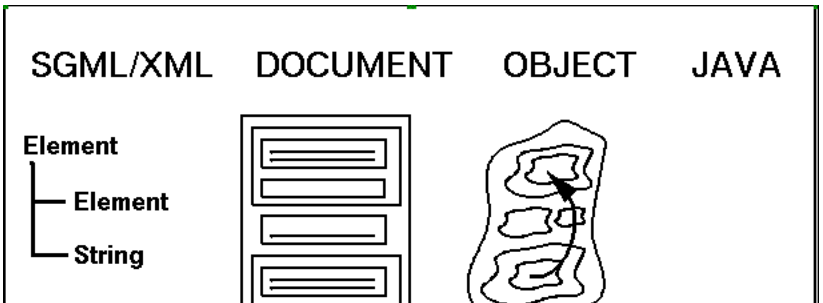
---

xmljava

"XML gives Java something to chew on"

- Jon Bosak (SUN) Microsystems; a driving force behind XML
- XML and Java complement each other perfectly.
- Java provides the technology to deliver XML anywhere in any form
- XML allows people to "touch and feel" Java objects
- Every XML element can have a corresponding Java class
- This philosophy is implemented in JUMBO

[



]

links

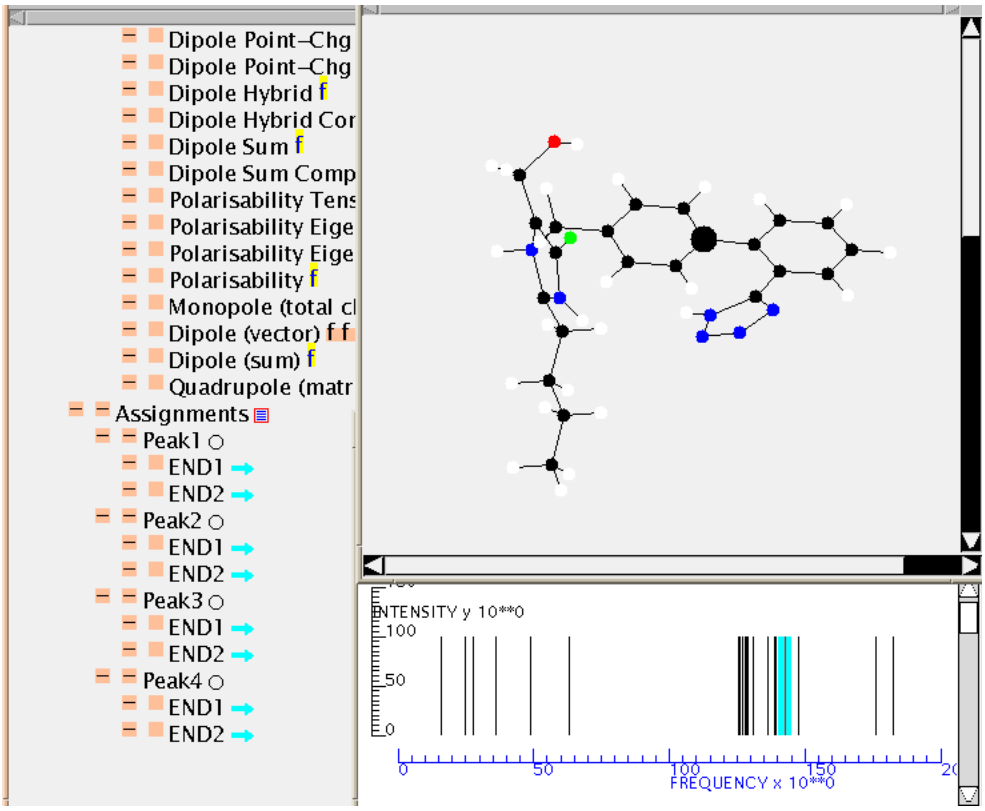
XML has extremely powerful hypermedia (XML-LINK)

- allows not only simple uni-directional links (e.g. <A HREF> in HTML)
- allows the creation of robust hyperdocuments (i.e. links won't break)

XML-LINKs

- can be multi-ended and bidirectional
- can represent different views (e.g. different graphs)
- can have several types of behaviour (e.g. automatic EMBEDding)
- can be typed (e.g. 'parentOf', 'pointerToFigure')
- can be in different documents

Spectral Assigments



WARNING: Quite technical

Reactions can be described by XML-LINKs!

In the esterification reaction
CH3CO2H + CH3OH = CH3CO2CH3
+ H2O

```
<!DOCTYPE CML><CML><P>In the
esterification reaction
<REACTION
XML-LINK="EXTENDED"
TITLE="Esterification">
<VAR XML-LINK="LOCATOR"
HREF="acetic.cml"
>
<VAR XML-LINK="LOCATOR"
HREF="methanol.cml"
ROLE="reactant"> <VAR
XML-LINK="LOCATOR"
HREF="ethylacetate.cml"
ROLE="product">
<VAR XML-LINK="LOCATOR"
HREF="water.cml"

</REACTION></P></CML>
```

---

search

Technical, but the facility is very exciting

TEI pointer searches in CML

---

tools

There are a very large number of high quality XML tools

XML will be in the major browsers, WWW database systems

For many applications you can get solutions off-the-shelf

For early adopters, see the XML-DEV list (slide 025)

For technical applications, especially molecules, JUMBO:

- is a generic XML browser
- although written to help develop CML, supports any DTD
- supports tree-structured display and editing
- supports XML-LINK
- supports TEI searches
- has over 300 Java classes
- maps XML elements onto Java classes

JUMBO 9801 has been released in alpha and is on the OMF CDROM

---

chembits

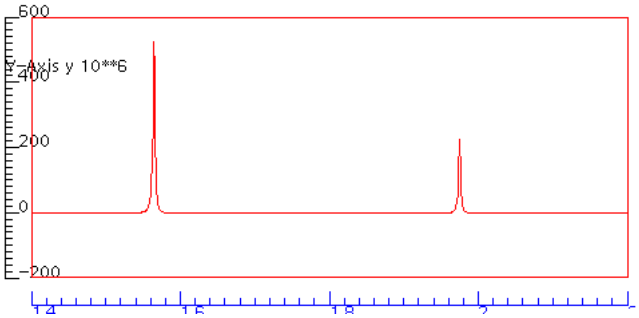
CML can support:

- Spectra and other instrumental output
- Crystallography
- Organic and inorganic molecules
- Physicochemical quantities (including units)
- MO calculations
- Macromolecules: Sequence protein, ligand and sequence
- Molecular Hyperglossaries : text and molecules
- Unidirectional hyperlinks and Multidirectional hyperlinks

Generic Markup

- <LIST>, <ITEM>, <ARRAY>, <TABLE>, <MATRIX>
- <PERSON>
- <BIB> (for citations)
- <UNITS>
- <INTEGER>, <STRING>, <DATE>, <URL>, <FLOAT>
- <FIGURE>

[



]

Chemical Markup

- <MOL>, <FORMULA>
- <ATOM> and <ATOMS>
- <BOND> and <BONDS>
- <CRYST>, <SYMMETRY>
- <SEQUENCE>, <FEATURE>

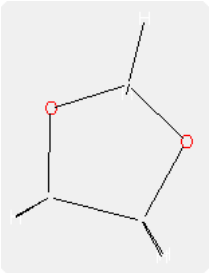
WARNING: Technical! (picture in next page)

Example:

---

mol

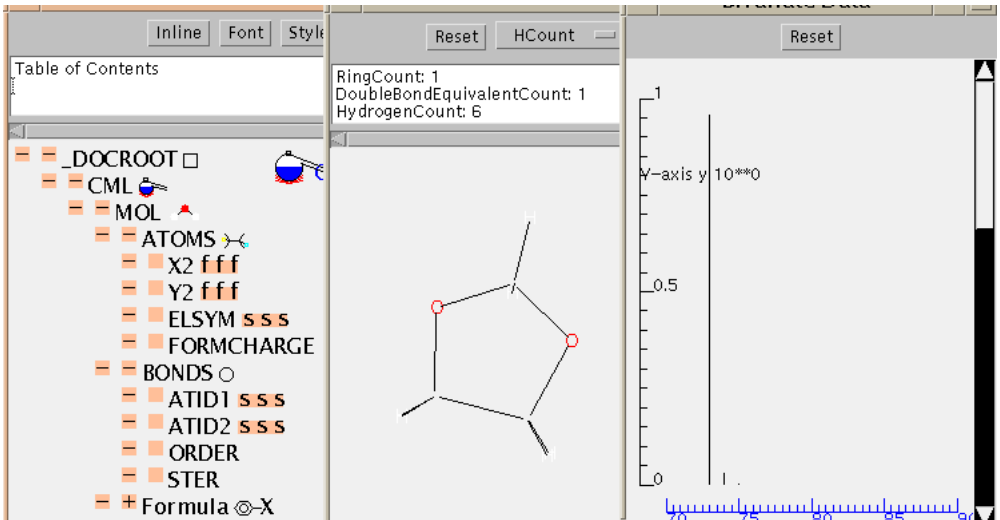
A 2-D diagram can be encoded in CML



Different views of the same molecule!

- The LHS is the TOC
- The RHS is the calculated isotopic distribution of the parent peak in MS
- The Centre shows several queries to the molecule

[



]

This is the CML representing it: (Technical)

---

cryst

JUMBO converts the following on-the-fly

- CCDC BIB/CON/DAT
- CIF (small molecule and mmCIF)
- PDB

The crystallographic data from the RSC Chem Comm:

```
└ #DOCROOT ? E **
  └ Mol At ? E mol1
    ├ Formula ? S Formula
    ├ Cryst At ? E Crystallographic data
    │ ├ Html At ? E Note
    │ │ └ P ? S These can be made compatible w...
    │ ├ Float At ? S G Units ACELL 11.105 (Angstrom)
    │ ├ Float At ? S G Units BCELL 18.297 (Angstrom)
    │ ├ Float At ? S G Units CCELL 20.027 (Angstrom)
    │ ├ Float At ? S G Units BETA 93.52 (Degree)
    │ ├ Float At ? S G Units cellvol 4061 (Angstrom+3)
    │ ├ String At ? S G spgp P21/n
    │ ├ Int At ? S G Z 4
    │ ├ String At ? S G color brown-green block
    │ ├ String At ? S G x-size 0.60
    │ ├ String At ? S G y-size 0.53
    │ ├ String At ? S G z-size 0.32
    │ ├ String At ? S G crystal system monoclinic
    │ ├ String At ? S G absorption coefficient 1.996
    │ ├ String At ? S G f000 2160
    │ ├ String At ? S G temperature 153
    │ ├ String At ? S G wavelength 0.71073
    │ ├ String At ? S G reflections 6009
    │ ├ String At ? S G min2theta 4
    │ ├ String At ? S G max2theta 50
    │ ├ String At ? S G R1 0.0481
    │ ├ String At ? S G wR2 0.1351
    │ └ String At ? S G GooF 1.124
    └ Html ? M Data were corrected  for Lorentz and□polarisa
```

- Note that many quantities have UNITS
- The CRYST object can calculate reciprocal cell, orthogonalisation matrix, etc.
- SYMMETRY can be managed

And this is how hypertext is integrated

- The hyperlinks (in blue) are active in the JUMBO browser

Data were corrected for Lorentz and polarisation effects and a semi-empirical absorption correction, based on [upsilon]-scans, was applied (T_{max}= 0.980, T_{min} = 0.558).The structure was solved by direct methods[ref14]and refined by full matrix least-squares on F^2, using all 5278 independent reflections (R_{int} = 0.0583). All the non-hydrogen atoms were refined with anisotropic atomic displacement parameters and hydrogen atoms bonded to carbon were inserted at

mo

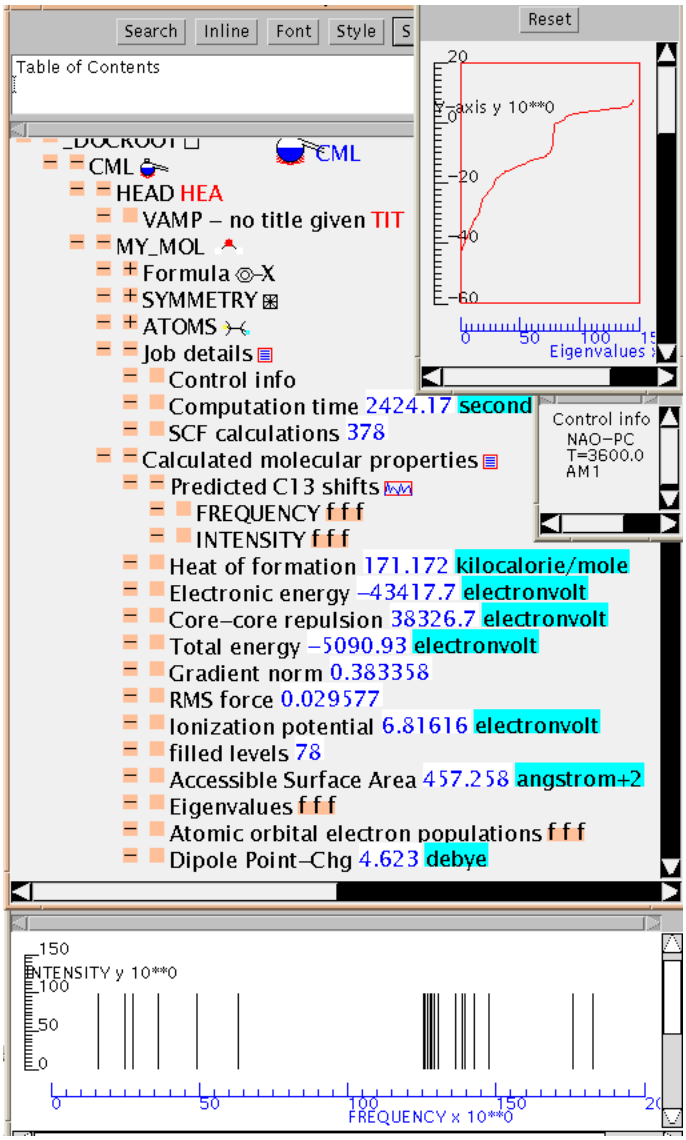
XML is very good for formatting ANY complex output

- Note the UNITS are marked up (and can be converted on-the-fly)
- Numbers can also be plotted graphically
- Much of this comes 'for free' with XML/CML/JUMBO

Three examples follow (VAMP, G94, MOPAC)

- VAMP

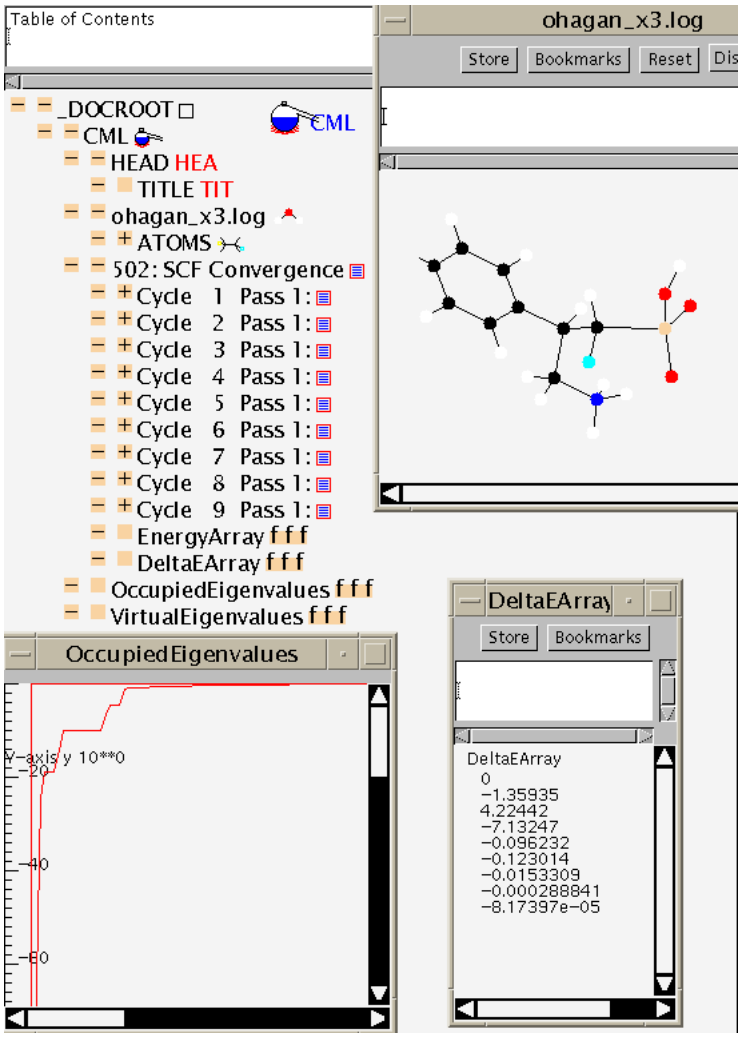
[



]

- Gaussian 94

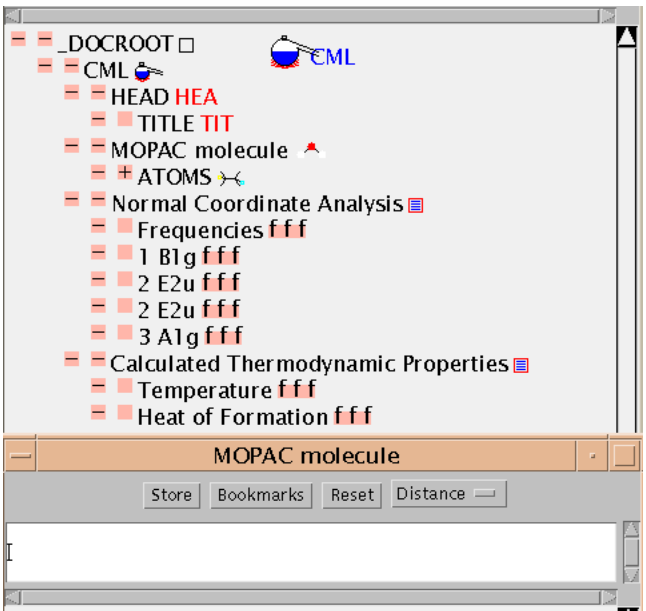
[



]

- MOPAC (shows molecular vibrations - JUMBO-MOL can display these)

[



]

protein

PDB files are not "flat" but highly structured

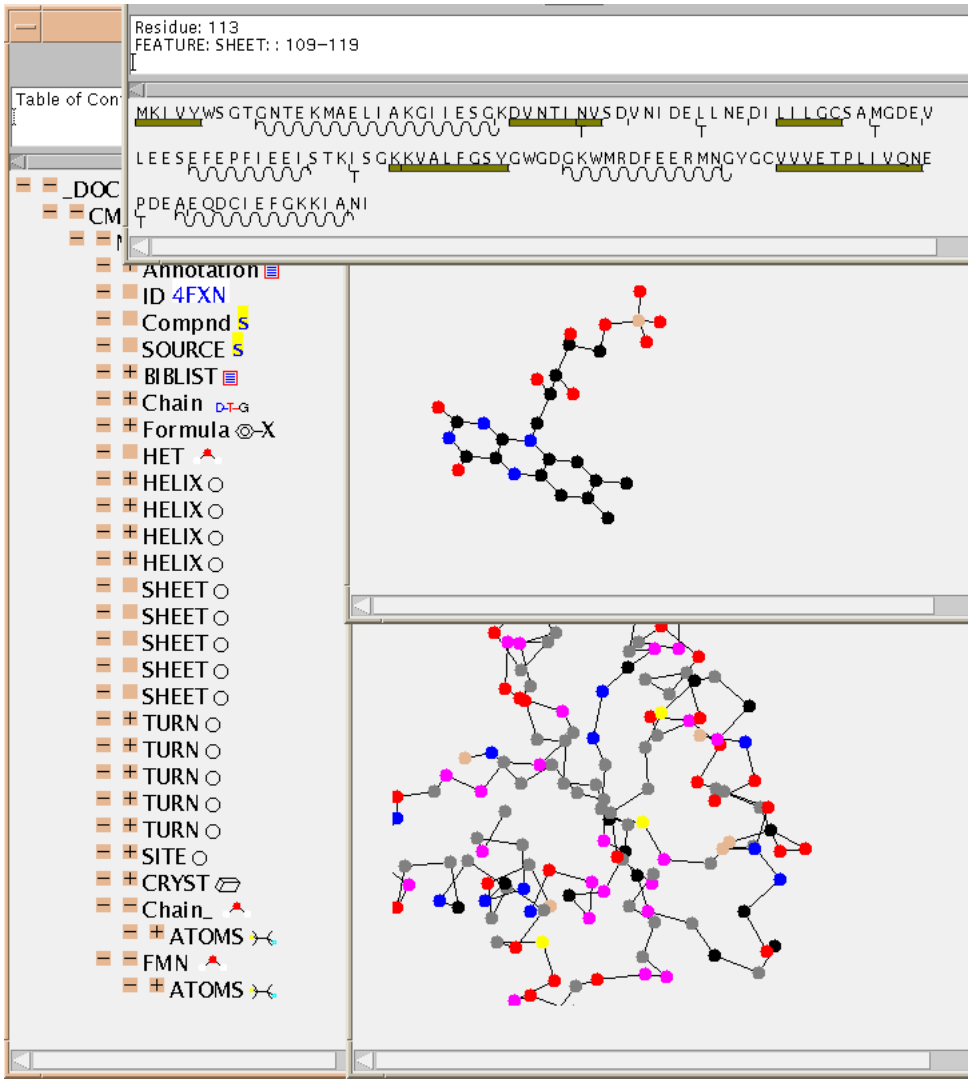
They contain many "objects" such as:

- <BIB>
- <FEATURE>
- <MOL>
- <SEQUENCE>

Here is a typical PDB file rendered in an early JUMBO

- NOTE: The TOC identifies many components
- The ligand and protein are separated
- The HELIX <FEATURE> is automatically mapped onto the <SEQUENCE>

[

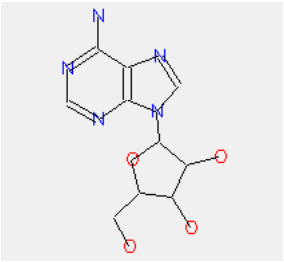


]

chemmath

Example of final display:

[



]

The hydrolysis of (I) can obey the
kinetics: (1)

dx/dt = -kx

<P>The hydrolysis of <A
HREF="mol.xml" XML-LINK=
"SIMPLE"> molecule(I)</A> can obey
the kinetics <A HREF="eqn.xml"
XML-LINK= "SIMPLE"
ACTUATE="AUTO"
SHOW="EMBED">

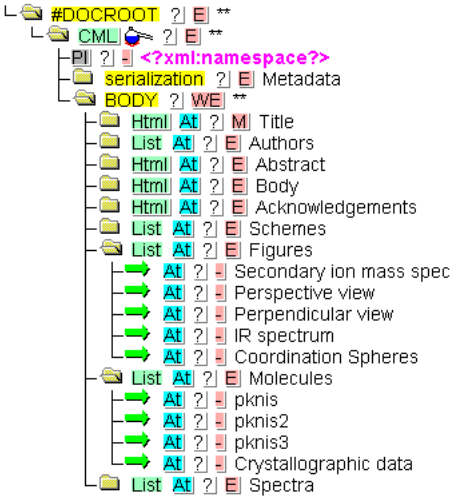
<!DOCTYPE MATHML><MATHML>
<EQN><EXPR><DIFF/>x<BVAR>t</BVAR></EXPR>
<EQ/><EXPR><MINUS/>k<TIMES/>x</EXPR></EQN></MATHML>

<!DOCTYPE CML><CML><MOL
TITLE="Methyl
ARRAY BUILTIN="ELSYM">C
/ARRAY></ATOMS><BONDS><ARRAY
BUILTIN="ATID1">1</ARRAY><ARRAY
BUILTIN="ATID2">2</ARRAY><ARRAY
BUILTIN="ORDER">1</ARRAY></MOL></CML>

chempub

Complete chemical publications are possible in CML

- This is an example from The Royal Society Of Chemistry's Chem. Comm.
- It appeared as CML on the CDROM of the latest ECTOC proceedings (thanks to RSC, Henry Rzepa and colleagues)
- This is the TOC (TableOfContents). Note that hypertext and data are integrated
- The green arrows are Hyperlinks (XLL) to other files. These files are both XML and non-XML. The latter are converted to XML on-the-fly.
- Note the metadata (in RDF format)



Some components of this paper

- Metadata
- Structure (Abstract, Body, Paragraphs, etc.)
- Schemes and Figures (with integrated captions/HTML)
- Molecules
- Crystallographic Data
- Spectra
- Citations

Molecular Hyperglossaries

[The hyperglossary concept]

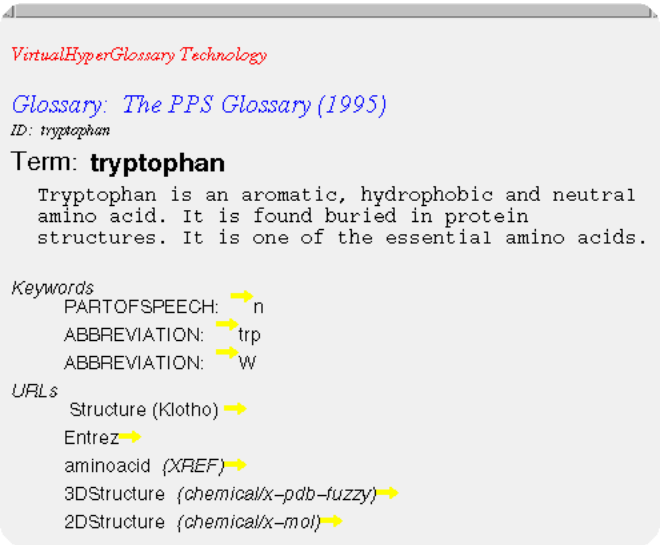
Semantic information can be added from glossaries

- <A HREF="www.some.glossary.org/org/mols?tryptophan">tryptophan</A>
- This can be added to any element in a CML document
- UNITS can linked:
- <ITEM TITLE="pressure" UNITS="glossary.org/units?pascal">23.27</ITEM>

A typical entry in a molecular hyperglossary

- The textual entry (uses ISO12620 terms) gives a definition, etc.
- links to other resources

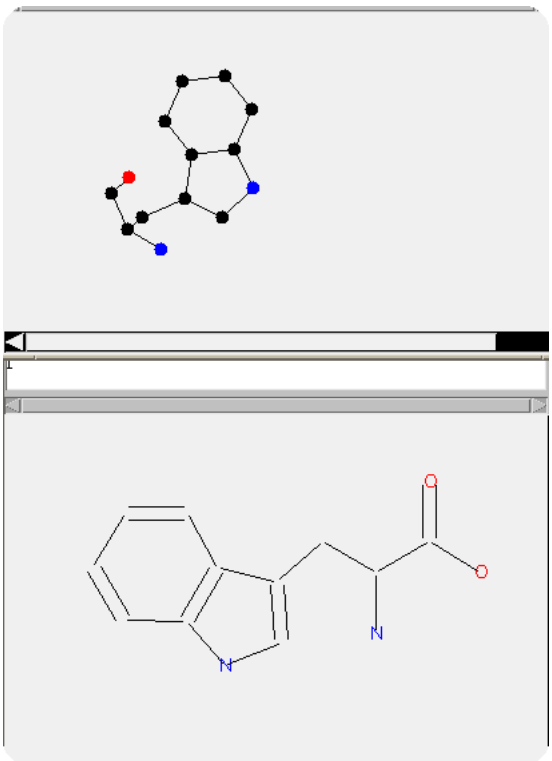
[



]

- The molecular information is both 2D and 3D
- Because it's in XML it can be searched, extracted for calculations, etc.

[



]

XML/CML and Intranets

XML is set to become the language of intranets

XML has major attractions in production environments

E-journal publication, e.g. complete Chem. Comm.

Database entry input/output

Data capture and instrument control

Program control and output postprocessing

Future-proofing of data in legacy systems

Representation of large scale documentation (e.g. regulatory)

Precise delivery of critical information (patents, safety, etc.)

Non-textual approaches (e.g. compound data cards)

The future of CML

- CML is a starting point, not a finished product.
- CML can support many philosophies and allow interconversion between them
- CML is supported by the Open Molecule Foundation, a consortium for interoperability.
- The OMF is sponsoring a free run-anywhere CDROM showing the potential of
- CML will interoperate with MathML and XML metadata standards
- CML can be developed Virtually (like XML)
- CML is not bound to any software platform
- JUMBO and JUMBO-MOL are freely available for collaborative development

The molecular community is certain to need XML soon

- Support communal development
- Make your requirements known. XML/CML can probably address many of them
- Get involved with prototypes; learn the power of XML
- Develop interfaces to XML-based systems

The OMF is sponsoring a free CDROM with CML/JUMBO

Join the OMF/CML effort!

Thanks and acknowledgements

Many thanks to the following

- Henry Rzepa for constant encouragement, promoting CML, especially at ACS 1997. And his long-suffering colleagues.
- The Open Molecule Foundation for support for CML and for the CDROM
- Colleagues in the VSMS and Nottingham
- A huge number of virtual friends in the XMl community. Many have contributed ideas, given webspace and mirrors, etc. Especially Jon Bosak who gave the first public demo of Jumbo
- Venus Internet for continued support and web pages
- The Royal Society of Chemistry for invitations into the CLIC project and allowing me to use their material for markup experiments
- Adam Precious (MDIS) and Moni Pangali (SUN Microsystems) for their vision, support and promotion of CML
- Chemweb and VEI for giving me this opportunity

URLs and other resources

Some Resources

- XML-FAQ. ([http://www.ucc.ie/xml/]) run by Peter Flynn.
- XML ([http://www.sil.org/sgml/xml.html]) maintained by Robin Cover and updated almost daily.
- XML-DEV. ([http://www.lists.ic.ac.uk/hypermail/xml-dev]) Henry Rzepa and I run a discussion list for anyone interested in developing XML applications. Highlights are identified at XML-Jewels ([http://www.vsms.nottingham.ac.uk/vsms/xml/jewels.htm])
- CML ([http://www.venus.co.uk/omf/cml/]) and [http://www.vsms.nottingham.ac.uk/vsms/java/cml] Chemical Markup Language, shortly to be circulated on CDROM.
- VHG ([http://www.venus.co.uk/vhg]) glossaries and terminology in XML, especially created for hyperlinked semantics.

- JUMBO ([http://www.vsms.nottingham.ac.uk/java/jumbo/])
- and feel free to contact me at [peter@ursus.demon.co.uk] or [peter.murray-rust@nottingham.ac.uk]
- [http://www.ch.ic.ac.uk/omf]