XML and BioDOM - Structured documents in molecular science

Adam Moore and Peter Murray-Rust

Virtual School of Molecular Science, University of Nottingham,
University Park, Nottingham, NG7 2RD,UK
Phone: +44-115-951-5087 Fax: +44-115-951-5110
paxamr@unix.ccc.nottingham.ac.uk Peter.Murray-Rust@nottingham.ac.uk

Abstract
Molecular information (sequences and molecular structures) are normally represented in legacy format ASCII files (e.g. SwissProt, PDB and MDL-molfile formats). Although there is a high level of content in these files the information structure is difficult to analyse and the data is not easy to re-use. For example, in PDB files most of the information apart from the ATOM records is discarded by many current tools (such as viewers).

There is an increasing demand for information to be format-independent for exchange protocols to be extensible and interoperable. This challenge has been addressed by the WorldWideWeb Consortium (W3C) who have developed a set of protocols for structuring documents and marking up data. These include:

XML: the eXtensible Markup Language; a metalanguage for creating Markup Languages
DOM: the (API for the) Document Object Model
XSL: eXtensible StyleSheet Language
XLINK/Xpointer: addressing systems for structured documents(SDs)
XQL: a query language for SDs
RDF: a metadata protocol

These allow users to create compound documents containing information from many sources. The XML technology supports many generic information types including: mathematics (MathML), vector graphics (SVG) and multimedia (SMIL), as well as allowing others to create their own languages (e.g. CML).

A key factor of the W3C approach is that SDs can be exchanged without loss of information. A large range of tools are now freely available to support generic operations (search/filtering/rendering/browsing, etc.) We have applied these to common document types in biomolecular science, particularly:

- protein sequence
- macromolecular structure
- 'small molecule' structures (using Chemical Markup Language, CML)

XML is an accessible discipline which allows users to create documents from well defined fragments/components. Thus a PDB 'file' is really a 'grove' of many components, which others can be added to or removed from. By adopting the XML approach we overcome the syntactic impedance between many biomolecular documents and data entries, making it easier to write automated processes for management.

References

http://www.vsms.nottingham.ac.uk/biodom
http://www.w3.org