XML in Chemistry

by Antony N. Davies

Extensible Mark-up Language (XML) is a powerful alternative to conventional binary file storage and information exchange. As many scientific organizations and companies delivering scientific products have implemented or are looking at the use of XML, IUPAC decided to review and evaluate what could and should be its role in advancing the use of XML in chemistry. In January this year, the IUPAC Committee on Printed and Electronic Publications (CPEP) organized a two day Strategic Meeting to assess the Union's position and options. Hosted by the Unilever Cambridge Centre for Molecular Informatics in the University of Cambridge Department of Chemistry, delegates from all interested IUPAC Divisions gathered together with key players in the field.

XML can be regarded as an extension to the well known HTML or Hyper Text Mark-up Language, which is the language most frequently encountered when viewing web pages. XML is considered to be the universal format for structured documents and data on the Web.[1] As with a conventional Web page, it isn't the use of XML itself that is interesting or even particularly novel, but the content stored within the XML files. In chemistry and associated technical fields, various groups—commercial organizations, academic institutions, and government bodies—have been developing XML formats independent of each other. These formats have similar content but differing data dictionaries and conventions. This means they are not compatible with each other and, what is far worse, resources are being deployed to address problems already solved by other groups. In order to support standardization in this field for the benefit of the community, IUPAC has decided to actively explore ways in which it can help to unify the various dictionaries and publicize their availability.

It isn't the use of XML itself that is interesting or even particularly novel, but the content stored within the XML files.

IUPAC'S Role and Timeline

During the 2001 IUPAC General Assembly in Brisbane, an ad hoc group outlined the dos and don'ts [see box] of a possible IUPAC role in advancing the use of XML in chemistry and developed a timeline for further action. The strategic importance of these decisions was reflected in the presentation of Wendy Warr—CPEP chairman—to the IUPAC Council[2] and the subsequent comments by IUPAC's secretary general Ted Becker in his article in CI.[3]

It was very clear from the Brisbane meeting that there was an urgent need to address the issues that were raised there. Hence, by the end of December 2001 the issues of identifying glossaries, project team members, and contacts between divisions and standing committees had been addressed. By then, Professor Bobby Glen of the new

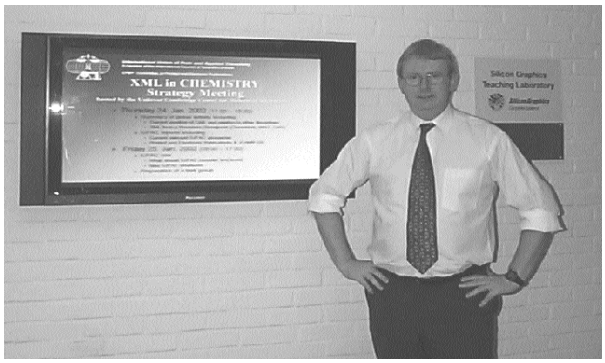
DOS AND DON'TS

IUPAC should not:

- Commence activities better left to the computer scientists
- Re-invent the wheel—the current activities at various locations should be invited to contribute to a standardization process through IUPAC as long as their efforts remain in the public domain
- Become formal members of World Wide Web Consortium (W3C), Object Management Group (OMG) or other similar organizations, however they should be informed of IUPAC activities in this area and we should continue to monitor their work.

IUPAC should:

- Establish "ownership" of the definition of standard terms in chemistry to be used in digital communications through formal IUPAC recommendations.
- Generate a glossary of standard terms in chemistry for use in applications involved in digital communications such as scientific data exchange or electronic publishing.
- Locate potential interested parties within IUPAC who "own" glossaries of terms or who are in the process of creating them
- Establish a method to identify and resolve problems in overlap of definitions (within IUPAC as well as with other scientific standards and other organizations)



Antony Davies

Unilever Centre for Molecular Informatics at the University of Cambridge, United Kingdom, agreed to host a follow-up meeting from 24-25 January 2002, as this type of initiative is of great interest to the fledgling center. Those invited to attend included IUPAC division and standing committee representatives and delegates from outside IUPAC who are active in establishing guideline for handling of chemical objects within their organizations. The IUPAC Analytical Chemistry Division was represented by its president David Moore; the Physical and Biophysical Chemistry Division represented by Jeremy Frey; and the new Chemical Nomenclature and Structure Representation Division, represented also by its president, Alan McNaught. In addition, I represented the IUPAC JCAMP-DX Working Party.

Meeting Overview

The meeting started with a welcoming address by Bobby Glen, who briefly explained the background of the Unilever Centre and provided a useful overview of the type of projects underway at the center.

Alan McNaught, Robert Lancashire, and I discussed IUPAC's intentions, current activities involving IUPAC glossaries, and the status of the JCAMP-DX file formats. Currently, within the eight IUPAC divisions there exist seven glossaries that are supervised by the Interdivisional Committee on Terminology, Nomenclature, and Symbols, which is responsible for ensuring conformity with existing IUPAC recommendations and consistency within and between each volume. These compendia, known as the IUPAC color books, cover chemical terminology, quantities, units, and symbols in physical chemistry, inorganic, organic, macromolecular, and analytical nomenclature, as well as the terminology and nomenclature of clinical laboratory sciences.[4]

Jeremy Frey pointed out that one difficulty encountered during the revision of the "green book" (which covers quantities, units, and symbols in physical chemistry) was the accommodation of different definitions, which originated from different fields of chemistry, for single entries in the data dictionary. Steve Heller offered an even broader example of the problem: although nm is widely recognized as nanometers in the scientific community, there is a significant body of opinion that feels that the letters obviously refer to nautical miles!

The International Union of Crystallographers (IUCr), represented at the meeting by Brian McMahon, has a very special interest in mark-up language because it has developed a standard format—the Crystallographic Information File (CIF)—for the deposition, storage, and distribution of crystallographic data with the publication

Crystallographic Information File (CIF)

by Brian McMahon

Commissioned by the International Union of Crystallography (IUCr), CIF consists of a very rich set of descriptors, allowing a file to contain raw and processed experimental data, a detailed experimental log, information about subsequent structure solution and refinement cycles, and a complete description of crystal and chemical structure and connectivity. A small excerpt from the standard example file for submissions to Acta Crystallographica Section C is presented here; the complete file can be viewed at ftp://ftp.iucr.org/pub/example.cif.

```
data_99107abs
_chemical_name_systematic
; 3-Benzo[b]thien-2-yl-5,6-dihydro-1,4,2-oxathiazine 4-oxide
;
_chemical_name_common          ?
_chemical_formula_iupac        'C11  H9  N  02  S2'
_chemical_formula_moiety       'C11  H9  N  02  S2'
_chemical_formula_sum          'C11  H9  N  02  S2'
_chemical_formula_weight       251.31
_chemical_compound_source      'synthesized by the authors,
see text'
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
S4      S 0.32163(7) 0.45232(6) 0.52011(3) 0.04532(13) Uani
S11     S 0.39642(7) 0.67998(6) 0.29598(2) 0.04215(12) Uani
O1      O -0.00302(17) 0.67538(16) 0.47124(8) 0.0470(3) Uani
O4      O 0.2601(2) 0.28588(16) 0.50279(10) 0.0700(5) Uani
H5A     H 0.1284 0.4834 0.6221 0.060 Uiso
H5B     H 0.1861 0.6537 0.5908 0.060 Uiso
```