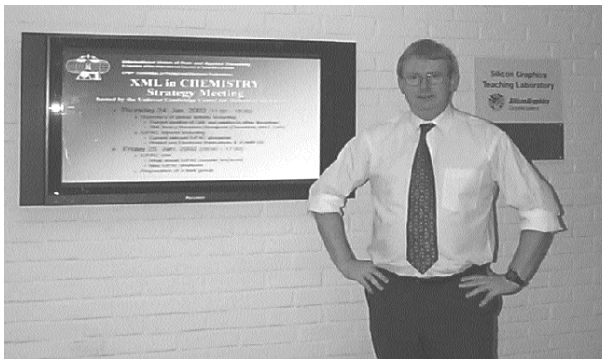

Antony Davies

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
```

of peer-reviewed papers. As McMahon explained, CIF was commissioned by IUCr following long-standing interest in the need for an open standard for data and information exchange. CIFs are divided into blocks, with each block consisting of individual labels or tags whose definition is stored elsewhere. Key points are that the semantic content is kept separate from the syntax of data representation, and that different dictionaries are used for different topic areas. McMahon concluded that one thing was abundantly clear from experience with CIF: "The design of a file format is an essential step, but it is only one component (and in many ways the least difficult) in the process of devising a feature-rich exchange mechanism. Far more difficult is the detailed definition of the tags that will be used within the file to ensure that applications attribute exactly the same meaning to the same item of information. The experience of the expert committees who undertake this work to extend CIF is that years of painstaking effort and discussion may be needed to define a few dozen tags, which are accepted across the community." As a contribution toward the establishment of content-rich XML applications in related areas of chemistry, the IUCr will make available its CIF-based definitions to the IUPAC groups working to establish XML-based applications. The scientific community said McMahon is looking forward to the day when effective chemical information exchange standards, widely accepted by the community, should complement and interoperate with CIF or its successors.

Peter Murray-Rust summarized other global activities surrounding the use of XML in science—see page 9 for a review of his work, co-authored with Henry Rzepa. At the meeting, Murray-Rust explained some of the benefits of using XML-based documents, including the ability to "validate" documents for correct or complete content, to create better electronically linked publications, and to significantly simplify information harvesting from such documents. According to Murray-Rust, for XML to function effectively for the sciences there needs to be agreement on the vocabularies or "ontologies" in use. He noted that the W3C expects that "domains" will create domain-specific tools and protocols for different subject areas such as chemistry. He also explained how the XML files differentiate between content, which has often been specified at different locations. Individual XML files may contain content from different ontologies such as a structure as defined by Chemical Markup Language (CML), a spectrum as defined by JCAMP-DX or SPECTROML, and a mathematical relationship as defined by MathML. This can be regarded as a powerful bonus, but again poses the question about reliability of the links the content needs to be put. This is currently leading to situations where "<element> carbon" might need to be handled differently, such as "<cml:element> carbon". The key is in the

. . . for XML to function effectively for the sciences there needs to be agreement on the vocabularies or "ontologies" in use.

The same file may be transferred from diffractometer to computational workstation to molecular graphics software, with each program in the chain importing and adding data. Authors using text editors or more complex editorial tools to create a full commentary and discussion of the structure may further extend the same file. Consequently the journals of the IUCr require all supplementary files recording crystal structure data to be in CIF format, and two of its journals will only accept papers submitted in this format. Such submissions are not only accepted and transformed by typesetting software into formatted research publications, but their embedded data are extracted and subjected to a battery of analytical and diagnostic calculations that provide referees with an objective assessment of the quality and consistency of the reported results.

The consequence of adopting such a standard is that data exchange becomes more efficient, computation is facilitated, transcription errors are removed from the publication process, and the quality of published data tends to improve. Overall, publication of structural reports journals becomes more efficient, onward transmission of the results to databases is also simplified, and readers may see any published crystal structure in three dimensions (and interact with the structure, generating stereo pairs, packing plots, and hydrogen bond networks ad libitum with the appropriate browser plug-ins or helper applications).

CIF has a somewhat different and rather simpler structure than XML. This is largely because it was developed at a time when SGML, the precursor of XML, was expensive and unwieldy to work with. Nevertheless, it is clear that automatic transformation between CIF and suitably devised XML formats is entirely feasible. Since its earliest days the CIF community has worked with pioneers in the chemical information field to work towards interoperability with emerging chemical information standards.

Brian McMahon <bm@iucr.org> is research and development officer at the International Union of Crystallography in Chester, United Kingdom.

explanation of the data dictionary associated with the defined name space "cml."

Namespaces do not have to be registered and so it is simple for any group or company to define their own version of "element." For example, although they could quite correctly claim to be using XML for data storage and transfer, the files generated would be as limited to their own internal applications as if they were using 17-bit binary encoded files. One way in which IUPAC could play a significant role in furthering XML for chemistry explained Murray-Rust is by ensuring that dictionaries are future safe and don't vanish from the Internet when a particular professor retires or a software or publishing house is bought out or goes bankrupt.

Goodman and McMahon agreed that IUPAC needed to identify the customers who would benefit from XML projects.

Jonathan Goodman, of the Unilever Centre, presented an amusing view from an academic and educational standpoint. His group has developed several databases that could lend themselves to being made available in an XML format. But, Goodman asked, what would be the immediate benefit? Quite simply, there would be none he stated. Should IUPAC take a clear lead in laying down guidelines on the presentation of chemical information in XML then it would be worthwhile to take this additional step as then other chemists and projects would be able to access and use the information more easily.

To conclude, Goodman said "there is a long way to go before XML is used routinely to improve and enhance chemical communication. However, XML friendly structures are already in place, and this should mean that a lot of data can easily be moved to this marked-up language. If an XML-based standard is accepted, then this process could be very rapid and data could be shared and reused much more easily than is now possible."

This supported the views of McMahon, who had commented that to generate an XML file from CIF would be a simple enough task, but questioned whether this would be "good" XML and "fit for purpose." Goodman and McMahon agreed that IUPAC needed to identify the customers who would benefit from XML projects. This includes clearly identifying stakeholders who will make the effort to implement whatever is developed.

Other presentations dealt with XML from various information providers' standpoints. Bill Town from ChemWeb and Sandy Lawson from MDL Information Systems pointed out the difficulties in achieving the uptake of technical developments in large organizations. Efforts have been made across the publishing industry to establish electronic submission and presentation of published papers, but authors still are unhappy about changing their habits. A general discussion was also held on the lack of decent authoring tools.

Kirk Schwall summarized the views of the Chemical Abstracts Service (CAS). According to Schwall, CAS has a collection of highly integrated data that have been organized using SGML since 1994. Since 1997, XML has been used for some data that have required frequent updating and interchangeability. Both the document and authority data collection concepts at CAS have XML as

Some of the attendees at the IUPAC Strategic Meeting on XML in Chemistry: (from left to right) Robert Lancashire, Bill Town, Jonathan Goodman, Sandy Lawson, Peter Murray-Rust, Kirk Schwall, Brian McMahon, Alan McNaught, Gary Mallard, Steve Stein, David Moore, Steve Heller, Bobby Glen, Kirill Degtyarenko, Richard Cammack, Peter Lampen, and Tony Davies.