XML in chemistry

Tony Davies

External Professor, University of Glamorgan, UK
c/o Creon Lab Control AG, Europaallee 27–29, 50226 Frechen, Germany

Many organisations and companies delivering scientific information products have implemented or are looking at the use of the so-called eXtensible Mark-up Language or XML as a powerful alternative to conventional binary file storage and information exchange. XML can be regarded as an extension to the well-known HTML or Hyper Text Mark-up Language, which is the language which you will have most frequently encountered when viewing web pages on the Internet.

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, whether for commercial organisations, academic institutions or government bodies have been developing XML formats with similar content but differing data dictionaries and conventions so that they 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 standardisation in this field for the benefit of the community, the International Union of Pure and Applied Chemistry has decided to take an active role in helping to unify the various dictionaries and publicise their availability.

IUPAC's role

During the 2001 IUPAC General Assembly in Brisbane an *ad hoc* group outlined the dos and don'ts of a possible IUPAC role in helping XML in Chemistry and a timeline for further action. This list reflected the known strengths and weaknesses of IUPAC's organisation and procedures as well as the consequences of any actions taken by the world's standardisation body for chemistry.

The strategic importance of these decisions was reflected in the presentation of the chairman of the IUPAC Standing Committee on Printed and Electronic Publications—Wendy Warr—to the IUPAC Bureau in Brisbane[1] and the subsequent comments by IUPAC's secretary general Ted Becker in his article in *Chemistry International*.[2]

Some quotes explaining Extensible Markup Language (XML) and its parent, the Standard Generalised Markup Language (SGML)

"The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web." See http://www.w3.org/XML/

"The name emphasizes the key feature of the language as it will be seen by an HTML user—the ability to define your own tags and attributes, which, of course, HTML does not allow." [*XML—Questions & Answers* (by Jon Bosak, Sun Microsystems)] (http://www.isgmlug.org/n3-1/n3-1-18.htm)

for those who wish to know more see:

http://www.oasis-open.org/cover/xml.html#overview

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 standardisation process through IUPAC as long as their efforts remain in the public domain.
- Become formal members of W3C, OMG or other similar organisations, 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 (through IDCNS).
- Establish a method to identify and resolve problems in overlap of definitions (within IUPAC as well as with other scientific standards and other organisations).

Tony Davies at the XML in Chemistry meeting.

Timeline

One thing that was very clear from the Brisbane meeting was the urgency with which the issues raised had to be addressed. The IUPAC internal issues of identifying glossaries and project team members and contacts between divisions and standing committees needed to be completed by the end of December 2001. This was started directly upon return of the participants from Australia and completed on time. Professor Bobby Glen of the new Unilever Centre for Molecular Informatics at the University of Cambridge, UK, was approached and, as this type of initiative is of great interest to the fledgling centre, kindly agreed to host the follow-up meeting on 24–25 January 2002.

Some of the attendees: 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, Tony Davies. (Thanks to Robert Lancashire for the photos.)

IUPAC Nomenclature—"Colour Books"

Gold Book

Compendium of Chemical Terminology, 2nd Edition. Blackwell Science (1997) (ISBN 0-86542-6848)

Green Book

IUPAC Quantities, Units and Symbols in Physical Chemistry, 2nd Edition. Blackwell Scientific Publications, Oxford (1993)

Orange Book

Compendium of Analytical Nomenclature (definitive rules 1997), 3rd Edition. Blackwell Science, (1998)

Blue Book

A Guide to IUPAC Nomenclature of Organic Compounds (recommendations 1993). Blackwell Science (1994)

Purple Book

IUPAC Compendium of Macromolecular Nomenclature. Blackwell Scientific Publications, Oxford (1991)

Red Book

Nomenclature of Inorganic Chemistry II. Recommendations 2000. The Royal Society of Chemistry (2001)

White Book

IUBMB Biochemical Nomenclature and Related Documents, 2nd Edition, Portland Press, London (1992)

Silver Book

Compendium of Terminology and Nomenclature of Properties in Clinical Laboratory Sciences (recommendations 1995). Blackwell Science (1995)

Follow-up meeting

The invitations were duly issued to the IUPAC division and standing committee representatives who had responded to the original inquiry about possible interest. Delegates from outside IUPAC who are active in setting the direction in which the handling of chemical objects within their organisations were also invited to attend.

This historic meeting was organised by the IUPAC Committee on Printed and Electronic Publications and attended by delegates from three IUPAC Divisions (Analytical, Physical, Chemical Nomenclature) as well as the IUPAC JCAMP-DX Working Party and EPR/ESR Limited Term Task Group.

The meeting started with a welcoming address by Bobby Glen, briefly explaining the background to the new Unilever Centre at Cambridge as well as providing a useful overview of the type of projects running at the centre.

Activities in scientific unions

Alan McNaught, Tony Davies and Robert Lancashire brought the meeting up-to-date as to IUPAC's intentions, current activities surrounding IUPAC glossaries and the JCAMP-DX file formats. The current state-of-the-art internally within IUPAC workflow was also presented. From currently eight IUPAC divisions there exist seven glossaries, with the IDCNS standing committee having the job of trying to make them self-consistent. Jeremy Frey pointed out that he had had difficulty during the creation of the Green Book with IUPAC bodies having different interpretations of the same entry in the data dictionary. Steve Heller pointed out that although nm was recognised widely as being nanometres in the scientific community, there is a significant body of opinion which would claim that the letters obviously referred to nautical miles!

Brian McMahon had been able to attend at short notice and travelled down from Chester representing the International Union of Crystallographers. This group is of special interest because of their rules on the deposition of crystallographic data with the publication of peer-reviewed papers. This has involved the development of a standard format for such depositions. Brian presented the status of the work being undertaken by the International Union of Crystallography and outlined the Crystallographic Information file format CIF. CIF files are divided into blocks with each block consisting of individual labels or tags whose definition is stored elsewhere. One of the key points here is to note that the semantic content is kept separate from the "syntax of data representation". There also different dictionaries for different topic areas. Private name spaces are allowed but in contrast to the JCAMP-DX standards, they are registered.

Activities in other bodies

Peter Murray-Rust summarised other global activities surrounding the use of XML in science. Peter explained some of the rationale behind the use of XML-based documents including emphasising the benefits of using an XML approach. These benefits include the ability to "validate" documents for correct or complete content, the ability to create better electronically linked publications as well as the added advantage of making information harvesting from such documents significantly simpler than is currently the case. For XML to function there needs to be agreement on the vocabularies or "ontologies" in use. Peter noted that the W3C expects that "Domains" will create domain-specific tools and protocols such as for the world of chemistry.

Peter 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 CML, a spectrum as defined by JCAMP-DX or SPECML and a mathematical relationship as defined by MATML. This can be regarded as a powerful bonus but again poses the question about reliability of the links within the files to the explanations of the use to which the content needs to be put. This is currently leading to situations where:

<element> carbon

might need to be handled differently to

<cml:element> carbon

the key being in the 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". 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 of the areas where IUPAC could play a significant role is in 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.

Jonathan Goodman, Unilever Centre, presented an amusing view from an academic and educational standpoint. His group have developed several databases which could lend themselves to being made available in an XML format. However, and here is an important point, what would be the immediate benefit? Quite simply, there would be no immediate benefit. 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 made available much more easily than is currently the case. This supported the views of Brian McMahon who had commented that to generate an XML file from a CIF file would be a simple enough task but would this be "good" XML and "fit-for-purpose". They agreed that IUPAC needed to identify the customers who would benefit from XML projects. This includes identifying 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 organisations. Moves have been made across the publishing industry to get electronic submission and presentation of the 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 presented some views from the Chemical Abstracts Service standpoint. The vast complexity of their operation meant that they were forced to handle about every possible mode of information delivery with only a tiny minority of their information suppliers delivering content in an XML format. Even when it is available, it is not used as the tags are stripped before being re-generated at the end of the document handling process. CAS does have an extensive thesaurus but this is not publicly available. It was agreed that there is a need for CAS and IUPAC to discuss common ontologies.

Gary Mallard from the US National Institute of Standards summarised activities within NIST. A quick search of nist.gov had located over 6300 documents with XML content on their web site. Gary was, however, quick to point out some of the drawbacks highlighted by problems associated with files being essentially uninterpretable if the explanations of the individual labels used are not open and freely available. Having created a nice presentation of the various efforts underway, a problem had arisen when several of the reference web sites for the ontologies turned out no longer to exist.

Conclusions

This very successful meeting concluded by drawing up a list of volunteers and appointing Steve Stein of NIST to chair a new limited term task group reporting to the IUPAC committee on printed and electronic publications. This task group will take the action forward within IUPAC in the coming years.

The next presentations on this topic can be expected at the CAS/IUPAC Conference on Chemical Identifiers and XML for Chemistry at the Pfahl Executive Education and Conference Center and The Blackwell at The Ohio State University, Columbus, Ohio, USA, 1 July 2002, http://www.iupac.org/symposia/conferences/CIandXML_jul02/.

References

1. http://www.iupac.org/news/archives/2001/41_council_minutes.pdf.
2. http://www.iupac.org/publications/ci/2001/september/CI0109.pdf.