Chemical Terminology at Your Fingertips

by Miloslav Nic, Jiri Jirat, and Bedrich Kosata

The *IUPAC Compendium of Chemical Terminology*, commonly known as the **Gold Book**, is one of IUPAC's major contributions to communication among chemists. It collects terms and definitions from primary specifications and provides single-point access to them. With more than 6 500 entries, it is a real treasure for every chemist.

The first edition was published in 1987 and updated in 1997. Both editions are now out of print. Fortunately, the second edition is also available as a series of PDF files at <www.iupac.org/publications/compendium>. In addition, there is a searchable version via the Muscat interface at <www.chemsoc.org/goldbook>.

If users are looking for a reliable definition of a term, access to such information via the book or PDF indexes is sufficient. But very often their wishes are not so simple. The readers may be unsure under which heading the required information is hidden or they may have only a vague notion of what they are looking for.

A few years ago, the only strategy in such cases was to browse through a few books and rely on good luck in finding some clues. Nowadays, we have computers which can browse millions of pages and look up relevant information in a few seconds.

But nothing is perfect in this world. Computers are very fast, but they do not possess human intelligence. Humans can identify important points from context, they can see through irregular structures, inaccuracies, usage of uncommon words, and other pitfalls of text browsing. Computer programs rely on regularity and possibility to consult internal vocabularies and other sources, but they do not cope well with the unexpected. So far, no solution has been found that can compete with a trained human in understanding text.

Chemistry represents a particularly difficult area for computers to compete in. Chemists communicate with combination of texts, structural formulas, pictures, and equations, and so automatic understanding of chemical texts is especially demanding. The Muscat search of the Gold Book gives a realistic picture of what an advanced, generalized automatic process can provide. This search engine not only properly understands similar words, but it even clusters results that share some common features. It definitively offers an important improvement over standard full text searches, yet its capabilities are restricted. It does not understand chemical formulas and mathematics, it does not understand some clues any undergraduate student would pick up on, it may even run entirely astray if some textual similarity perplexes its algorithms. It is important to keep in mind that the search relies on generic analysis of text, not on chemical knowledge.

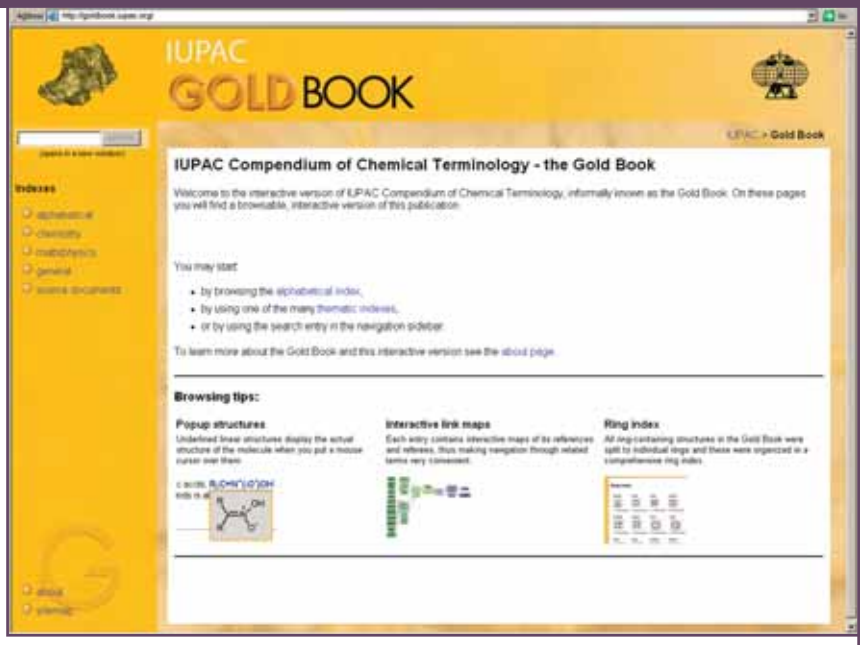
But the situation is not so bleak as it may sound from the preceding paragraphs. The art of communication with computers resides in their ability to organize text and other data into many small sections and annotate these parts in a way that makes it much easier to convey some sense to the computer.

There now exist technologies based on XML (eXtensible Markup Language) that are very good at such partitioning and further processing of the marked information. Research into the efficient markup of chemical text and in finding the optimal structure of information is in its infancy, and the XML Gold Book represents a milestone in this development.

The Gold Book includes chemical formulas, mathematical symbols, units, and other features. Its structuring and further transformation has required enormous effort over five years, but the work finally paid off. Relevant information is now captured in a way that makes it accessible using standard software techniques.

It happens too often that the usefulness of excellent books is spoiled by mediocre indexing. The new version of the Gold Book does not suffer such a fate. The fact that all information contained in the book is meticulously marked in XML enabled the creation of many indexes that are difficult to find in similar publications.

Some indexes extract chemical meaning so that entries can be selected based on the compounds they



contain (index of structures, chemical formulas, ring index), while others summarize information about physical constants, units, and quantities. Other indexes list images and offer selections of acronyms and abbreviations. All indexes are generated automatically from the available source text, so there is no need for manual intervention. Any upgrades and additions automatically appear in the indexes. One InChI identifier was generated for each compound to make chemical information accessible for search engines and data mining.

The XML version's "link maps," which are also generated automatically, are very useful as well. Every entry is accompanied by a map that graphically displays the relation of the entry to other terms and definitions. These maps often reveal relationships that are difficult to decipher by other means.

Full text incremental searching is also available. This search does not rely on an Internet connection and can be used directly from the CD-ROM.

The Gold Book is a very useful resource on its own, but its usefulness does not end here. Concurrently, software has been developed that enables automatic incorporation of Gold Book data into other resources. One of the features of this software is its ability to recognize Gold Book entries in various texts from independent sources. The software automatically links these entries to the Gold Book, so that in the near future anyone reading materials from IUPAC literally will be a click away from finding proper definitions of terms he/she does not understand.

As with any major undertaking, the XML Gold Book required the cooperation of many people. The activity started as part of the IUPAC project "Standard XML Data Dictionaries for Chemistry" (2002-022-1-024) under the leadership of Steve Stein from NIST. Miloslav Nic, Jiri Jirat, and Bedrich Kosata from the Laboratory of Informatics and Chemistry at ICT Prague did the XML work and other programming. Jiri Znamenacek from ICT Press wrote the search engine and also implemented the graphic design created by Ladislav Hovorka. Eva Dibuszova, the head of ICT Press, provided valuable editorial advices. Cheryl Wurzbacher, the production editor for *Pure and Applied Chemistry*, proofed very thoroughly the XML version against the original printed version of the book, which resulted in the correction of many errors and mistakes that sneaked in during the initial XML conversion. Aubrey Jenkins continues to update the Gold Book with new entries and to correct the old ones. Alan McNaught, Steve Heller, Leslie Glasser, and Jack Lorimer were also instrumental.

As mentioned at the beginning of the article, the new version of the Gold Book is a very important step in improving the availability of IUPAC materials to chemists and the general public. Further developments will transform IUPAC materials from a collection of independent sources to an integrated information resource. Many articles have been recently published about Semantic web both in the scientific and popular press. With this new development, IUPAC again will be a pioneer in a new territory, showing how to manage and distribute complex scientific information.

Miloslav Nic <Miloslav.Nic@vscht.cz> is the head of the Laboratory of Informatics and Chemistry at ICT Prague and the coordinator of the B.Sc. and M.Sc. study programs "Informatics and Chemistry" and "Applied Informatics in Chemistry." He is an invited observer on the IUPAC Committee on Printed and Electronic Publication. Jiri Jirat <Jiri.Jirat@vscht.cz> is a lecturer at ICT Prague (chemical informatics, XML technologies) and a researcher at the Laboratory of Informatics and Chemistry at ICT Prague. Bedrich Kosata <Bedrich.Kosata@vscht.cz> is a lecturer at ICT Prague (chemoinformatics, programming) and a researcher at the Laboratory of Informatics and Chemistry at ICT Prague. He is the author of the open source molecular editor BKChem <http://bkchem.zirael.org/>.

http://goldbook.iupac.org