IUPAC PROJECT MEETINGS: EXTENSIBLE MARKUP LANGUAGE (XML) DATA DICTIONARIES AND CHEMICAL IDENTIFIER

NOVEMBER 12-14, 2003, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, (NIST), GAITHERSBURG, MARYLAND, USA

A report by Dr. Wendy A. Warr

Wendy Warr & Associates

June 2004

Dr Wendy A Warr
Wendy Warr & Associates, 6 Berwick Court
Holmes Chapel, Cheshire CW4 7HZ, England
Tel/fax +44 (0)1477 533837
wendy@warr.com http://www.warr.com

IUPAC PROJECT MEETINGS: EXTENSIBLE MARKUP LANGUAGE (XML) DATA DICTIONARIES AND CHEMICAL IDENTIFIER NOVEMBER 12-14, 2003, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, (NIST), GAITHERSBURG, MARYLAND, USA

Proceedings recorded by Wendy A. Warr

Introduction and Overview

Stephen Stein, NIST
steve.stein@nist.gov

IUPAC was formed in 1919 for international standardization in chemistry. The standardization of weights, measures, names and symbols is essential to scientific enterprise and the growth of international trade and commerce. NIST was formed even earlier with overlapping objectives, especially regarding “weights and measures”, so it is not surprising that there have been longstanding interactions between the two organizations. It is important to note that while NIST encourages the development and dissemination of standards, this has no force in law. NIST is not a regulatory agency and like IUPAC deals primarily with voluntary standards. Standards such as those discussed at this meeting are needed to ensure that the receiver of chemical information is able it interpret it as intended by the submitter. Stein welcomed everyone to the meeting, presented the agenda, and invited the participants to introduce themselves and outline their interests. About 50 people attended, on and off, over the three days. An attendance list is available, separate from this report.

IUPAC project 2002-022-1-024, on standard XML data dictionaries for chemistry, for which Stein chairs the Task Group, is due for completion in 2005. Other members of the Task Group are Kirill Degtyarenko, Jeremy Frey, François Gilardoni, Jiří Jirát, Robert Lancashire, Alan McNaught, Miloslav Nič, and Henry Rzepa. Stein introduced the big issues behind the project. Chemical information must be labeled for effective computer processing. Accepted labels for widely used chemical terms do not exist: each Extensible Markup Language (XML) implementation invents its own labels. The next stop could be Babel but IUPAC labels can provide some overlapping terms. Labels should be well defined, for example using IUPAC Glossaries such as the “Color Books” (defined later in this report). This is a basic data dissemination problem and XML provides the best current syntax. Earlier meetings on XML in Chemistry and a namespace for IUPAC XML projects are described in the talk by Davies below.

Why use XML? It is the currently accepted standard syntax, allows use of software developed for other purposes, can effectively represent complex data, and is extensible. It also enables integration of dictionaries: STMML (a markup language for scientific, technical and medical publishing, http://www.xml-cml.org/stmml/) provides an umbrella for namespaces such as CML, UnitML, ThermoML, and AniML (all covered later in this report). Increased access enables other benefits such as versions for users with visual impairment, and language-independent components. However, there are some arguments for *not* using XML. Data entry is primitive and is not yet used by authors. Progress depends on progress made by others, e.g., Microsoft and others for display. It is easy with XML to “roll your own” standard. Effective validation is cumbersome and XML is a moving target.

XML has two directions, vertical and horizontal. The former is a convenient way of representing one’s own information in a narrow domain under a single dictionary; the latter a way of handling data under multiple complex dictionaries, reusing dictionaries and data structures, and communicating with metadata. XML and chemistry should be divided according to “natural principles”; ordering and linking should be according to chemistry. The details of XML should not influence data dictionary development, for XML data representation is changing and primitive while chemistry is stable and deep.

Report on IUPAC and XML in Chemistry

Tony Davies, Creon Lab Control AG, Germany (Secretary, IUPAC Committee on Printed and Electronic Publications)
tony.davies@waters.com

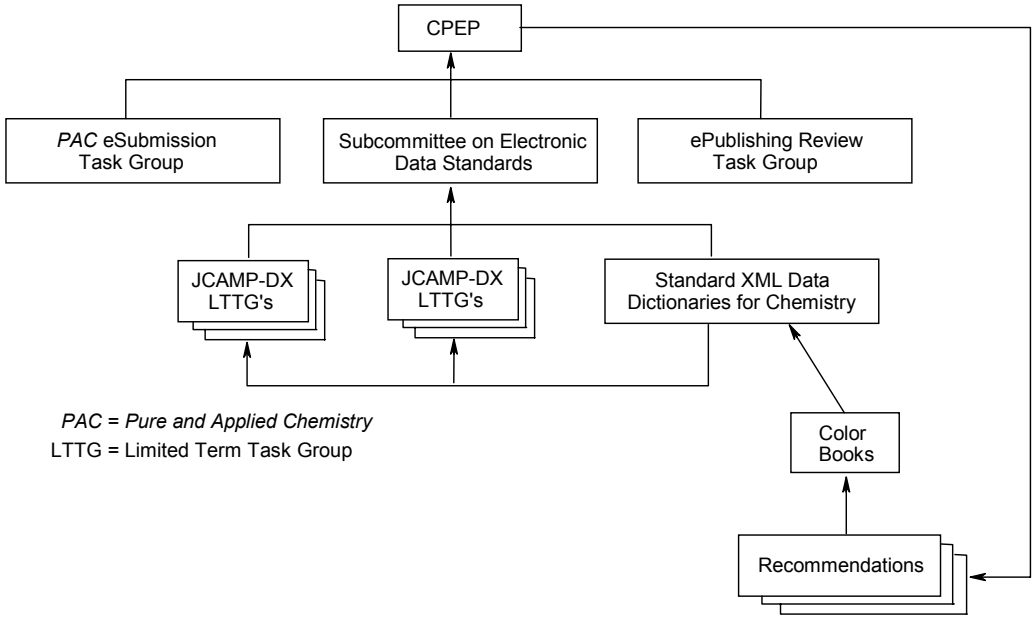
Davies outlined the history and nature of electronic data standards projects, and discussed coordination issues and the relationships between projects. IUPAC took over responsibility for the JCAMP-DX series of protocols from the Joint Committee on Atomic and Molecular Physical Data (JCAMP) in 1995. Protocols published to date are:

JCAMP-DX for IR. *Applied Spectroscopy* **1988**, *42(1)*, 151-162.
JCAMP-DX for Chemical Structures. *Applied Spectroscopy* **1991**, *45(1)*, 4-11.
JCAMP-DX for NMR. *Applied Spectroscopy* **1993**, *47(8)*, 1093-1099.
JCAMP-DX for Mass Spectrometry. *Applied Spectroscopy* **1994**, *48(12)*, 1545-1552.
JCAMP-DX v. 5.01 (IUPAC Recommendations 1999). *Pure Appl. Chem.* **1999**, *71(8),* 1549-1556.
JCAMP-DX for IMS (IUPAC Recommendations 2001). *Pure Appl. Chem.* **2001**, *73(11)*, 1765-1782.
JCAMP-DX NMR Pulse Sequences (IUPAC Recommendations 2001). *Pure Appl. Chem.* **2001**, *73(11)*, 1749–1764.

An IUPAC Working Party has had responsibility for the support and development of the JCAMP-DX scientific data standards series. Major developments of a generic nature are the responsibility of the Working Party as is the management of the technique-specific Limited Term Task Groups that develop new JCAMP-DX standards in different scientific disciplines.

What it is important to understand in all of these projects is that the key to success is the agreement of a data file terminology. Whether the current IT format is an ASCII flat file or XML is essentially irrelevant as in the long-term these formats will probably have to migrate to something better. What must migrate well into the future are carefully thought through data dictionaries at the core of the file.

In 2002 the Committee on Printed and Electronic Publications (CPEP) established the Subcommittee for Electronic Data Standards to coordinate the various technique-specific limited term task groups, and provide expert oversight in the different fledgling XML projects. CPEP also reviews new project proposals to ensure that they have a representative group of end users (in order to prove a *need* for the project) and an agreement of key commercial vendors to implement and support the standard. An electronic data standard without agreement to implement is a waste of time and money. The projects must produce IUPAC Recommendations that include an agreed terminology for describing the data (the so-called data dictionaries) as well as guidelines for their implementation. Davies produced a diagram of CPEP coordination:



Data dictionaries in infrared spectroscopy, mass spectrometry, nuclear magnetic resonance and ion mobility spectrometry have been agreed. Draft data dictionaries in electron paramagnetic resonance/electron spin resonance spectrometry (EPR/ESR), near infrared (NIR) spectrometry and chemometrics, chromatography, liquid chromatography/mass spectrometry (LC/MS), and multi-dimensional nuclear magnetic resonance (NMR) are being drafted. Proposed task groups will provide data dictionaries in thermodynamics and catalysis, the XML Limited Term Task Groups having been proposed by IUPAC Divisions. CPEP has reviewed these projects as strictly as the usual JCAMP-DX task group proposals and intends to treat them in a similar manner. It is proposed to create the Group Identity as is currently the case with the JCAMP-DX Task Groups shown through the Web site http://www.jcamp.org.

In the past, IUPAC IT projects have suffered from a lack of formal coordination. Unfortunately, work carried out by a particular divisional commission has often found no resonance in the wider world; although the content of the work has been excellent, it has died with changes in the commission membership. Additional problems arise with the selection of particular technologies where the failure of IUPAC to ensure that such projects were “future safe” has led to databases being created that no longer run on current technology.

IT projects must include safeguards to ensure that the technology selected can migrate to whatever becomes the *de facto* standard in the future. This has been achieved in the past in the JCAMP-DX spectroscopy standards by rigorously avoiding any hardware- or software-specific solutions or “quick fixes”. This makes the work harder in the short-term but guarantees longevity. Operationally CPEP has also included at least one of the core working party members, with experience in the technique, in each of the task groups to ensure that the specialist knowledge and working practices is always available and that the guidelines are followed during all stages of the development of the new standards.

Davies next discussed the IUPAC XML in chemistry initiative (see *Chemistry International* **2002**, 74(4), 3-8. http://www.iupac.org/publications/ci/2002/2404/XML.html). The initiative started with a one-day meeting at the IUPAC General Assembly in Brisbane in 2001. This was followed by a consultation exercise with all divisions and then an open meeting in Cambridge, UK in January 2002, for all those who had expressed interest. The Cambridge meeting came to the conclusion that IUPAC should establish “ownership” of the XML data dictionaries being developed for chemical nomenclature around the world and resulted in the “Standard XML Data Dictionaries for

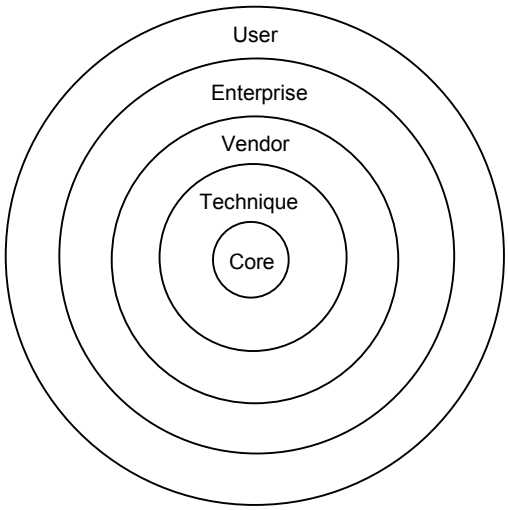
Chemistry" project being drawn up which is now being chaired by Steve Stein and reporting to CPEP through the data standards subcommittee.

There is a multiplicity of XML terminology: STMML, CML, GAML, MatML, ThermoML, UnitsML, SpectroML, AtoML, BSML, ToxML, and CatML, together with some obsolete MLs such as GEML, GeneML, BioML, and PhysicsML. The LTTG overseeing standard XML data dictionaries for chemistry aims to translate existing IUPAC standard terminologies and related information to data dictionaries in XML format; to establish a strategy for future IUPAC involvement in applications of XML for chemistry; and to enable IUPAC to serve as the *principal, authoritative source* of basic chemical terminology for electronic communications. [At this point Jeremy Frey interjected, pointing out that it is important for the *computer* to be able to do the look-up.]

XML has many advantages. It simplifies exchange of metadata, it is extensible through the use of style sheets (manufacturer dependent) and it can be human readable. On the other hand, there are no agreed data dictionaries, manufacturer dependent style sheets are not future safe, some developers of XML solutions have started to add binary fields, and there are multiple diverging developments such as CML, SpectroML, GAML, CIDX, UnitsML, MathML, DXML, and ATOMML.

There have been four meetings so far concerned with the IUPAC XML in Chemistry project. The first was the discussion group meeting held in Brisbane, Australia on July 4, 2001. The second was a conference in Cambridge, UK, on. January 24-25, 2002. That was followed by the CAS/IUPAC Conference on Chemical Identifiers and XML for Chemistry held in Columbus, Ohio USA on July 1, 2002. The fourth meeting is reported in these proceedings. The ASTM Analytical Markup Project (AnIML) is being run by the ASTM E13.15 Subcommittee.

The proposed "layered" structural model is as follows:



Davies showed tables of the AnIML techniques proposal:

Version 1:

Title	Abbreviation	Responsible Standards Organization
Infrared spectroscopy	IR	IUPAC CPEP subcommittee on electronic data standards
Nuclear magnetic resonance spectrometry	NMR	IUPAC CPEP subcommittee on electronic data standards
Mass spectrometry	MS	IUPAC CPEP subcommittee on electronic data standards (including labels from ASTM Standard E2077)
Chromatography	CHROME	ASTM E13.15 with IUPAC CPEP subcommittee on electronic data standards (taken from ASTM Standard E1947)
Ultraviolet and visible spectroscopy	UV/VIS	NIST (taken from SpectroML)
Ion mobility spectrometry	IMS	IUPAC CPEP subcommittee on electronic data standards and International Society for Ion Mobility Spectrometry (ISIMS)

Version 2

Title	Abbreviation	Responsible Standards Organization
Electron paramagnetic resonance/electron spin resonance spectrometry	EPR/ESR	IUPAC CPEP subcommittee on electronic data standards, EPR/ESR Task Group (Robert Lancashire)
Near infrared spectrometry	NIR	IUPAC CPEP subcommittee on electronic data standards, NIR and Chemometrics Task Group (Gerrard Downey, President International Council for NIR Spectroscopy)
Crystallization	CRYST	International Union of Crystallography

Version 3

Title	Abbreviation	Responsible Standards Organization
Chemometrics data	CHEMOMETRICS	IUPAC CPEP subcommittee on electronic data standards, NIR and Chemometrics Task Group (Gerrard Downey, President International Council for NIR Spectroscopy)

Davies gave an example of a good implementation: IMS IUPAC/JCAMP-DX support. Third party software such as that from ACD/Labs, and Galactic, and an Internet Explorer MIME plug-in, and a Netscape MIME plug-in can all read the data from each other's applications even if the user knows nothing about ion mobility spectra. ACD/Labs, for example, could pull out all the metadata however much or little they knew about ion mobility spectra.

This standardization exercise will reap rewards. IUPAC will show leadership in chemical data standards. The quality of data from internal consistency among glossaries will enable computer validation. Information will be more accessible: linkable by more than a primary key. A single worldwide "root"' for basic chemical terms will be established. A namespace for IUPAC XML projects will be established at the URL http://www.iupac.org/namespaces/…

Discussion

Davies' talk was followed by some discussion about context.

Jeremy Frey: context matters.

Alan McNaught: context is part of the definition.

Fabienne Meyers: the context is not in the Gold Book.

Alan McNaught: it *is* included if ambiguity is possible.

IUPAC and the Gold Book
Alan McNaught, Royal Society of Chemistry (RSC)
adm@rsc.org

The IUPAC "Gold Book" is a compendium of everything that is not nomenclature of chemical compounds (although it does have class names). It is one of the series of IUPAC "Color Books" on chemical nomenclature, terminology, symbols and units and collects together terminology definitions from IUPAC recommendations already published in *Pure and Applied Chemistry* and in the other Color Books. Terminology definitions published by IUPAC are drafted by international committees of experts in the appropriate chemistry sub-disciplines, and ratified by IUPAC's Interdivisional Committee on Terminology, Nomenclature and Symbols (ICTNS). The Gold Book is a collection of nearly 7000 terms, with authoritative definitions, spanning the whole range of chemistry. Chemists went through all the IUPAC recommendations and pulled out definitions. Conflicts were found and were resolved before the Gold Book was first printed in 1985. A substantial improvement to the book appeared in both print and electronic versions 10 years later. The electronic version was unfortunately in PDF format only; Aubrey Jenkins is now leading a project to update the electronic version.

Discussion

David Martinsen: The Gold Book is 10 years out of date. If someone wants to find a definition when he is surfing the Web, he may not find it.

Alan McNaught: It would be desirable for the electronic Gold Book to be updated faster. The function of the Gold Book might be changing: specialized definitions should perhaps be in there. All this needs to be discussed this week.

Jeremy Frey: As the tools to link to dictionaries improve it may be possible to find a term and alert readers to the fact that it is not yet approved. Highly trusted definitions are slower to establish.

Tony Davies: Separate the tools and the presentation.

Aubrey Jenkins: Until recently, nothing post-1995 was in the Gold Book but post-1995 terms *are* now being added. I have been working on it for less than a year and a lot of work is involved. It is too early to say how soon the Gold Book will be up to date to 2004.

Gary Kramer: What is the availability of the colored books?

Alan McNaught: Two (identical) electronic sources of the Gold Book are on the Web, on the RSC site and the IUPAC site. The PDF is searchable. The Green Book is not yet electronic and is out of print. It will be available soon from RSC in hard copy and the PDF version will be released six months later. The Gold Book does not contain data dictionaries from other IUPAC projects or sources.

Units Markup Language

Bob Dragoset, NIST
dragoset@nist.gov

Units Markup Language (UnitsML) is a proposed method of representing scientific units of measure in XML. The original motivation for this project came from initial efforts by Frank Olken and John McCarthy of the Lawrence Berkeley National Laboratory (LBNL). Current collaborators are:

Barry Taylor (emeritus; NIST, units)
Michael McLay (NIST, XML)
Karen Olsen (NIST, programming)
Frank Olken (LBNL)
Peter Murray-Rust (CML - Chemical Markup Language)
Allen Razdow (Mathsoft Inc./Mathcad)

The project is funded in part by NIST's Systems Integration for Manufacturing Applications (SIMA) program (http://www.nist.gov/sima).

The vision is to make XML schema(s) available for incorporating UnitsML into XML documents for encoding scientific units of measure. The collaborators also aim to build an extensive repository of units containing XML schemas and information on units, quantities, and prefixes, and to design schemas and a repository to facilitate unit information processing, i.e., validating documents for self-consistent usage of units, quantities, and prefixes, and unit conversion and manipulation. They will also develop an ontology for scientific units.

There are seven SI base units and quantities:

Unit Name	Unit Symbol	Base Quantity	Quantity Symbol	Dimension Symbol
meter	m	length	l	**L**
kilogram	kg	mass	m	**M**
second	s	time	t	**T**
ampere	A	electric current	I	**I**
kelvin	K	thermodynamic temperature	T	Θ
mole	mol	amount of substance	n	**N**
candela	cd	luminous intensity	l_v	**J**

Dragoset gave two simple examples of instance documents, i.e., ones containing base SI units:

1.
```
<physicalQuantity
        name="tableLength">1.75 m</physicalQuantity>
```

2.
```
<physicalQuantity name="tableLength">
    <value>1.75</value>
    <units>meter</units>
</physicalQuantity>
```

A not so simple example contains derived units:

```
3.      <physicalQuantity name="specialAcceleration">
            <value>100</value>
            <units>mm·µs^-2</units>
        </physicalQuantity>
```

He also gave an example of usage within a document:

```
<physicalQuantity       name="specialAcceleration"
        value="100"
        unitIDRef="acc01" />
```

and definition within a document:

```
<derivedUnit    unitID="acc01">
        <unitsML:units>
            <unit name="meter" prefix="milli" power="1" system="SI" />
            <unit name="second" prefix="micro" power="-2" system="SI" />
        </unitsML:units>
</derivedUnit>
```

Alternatively, one could use unit and prefix symbols rather than unit and prefix names.

The units repository contains attributes, including symbol, UID (unique identifier), and language. The database also includes quantities, prefixes, and symbols, and the ability to "tag" units for a specific industry. UnitsML database search is under development. Users will be able to download units information by all or specific units, all or specific quantities, unit system, and unit type, and download units information in a specific language (e.g., English US) and in various formats (XML, HTML, and ASCII). The team plans to tailor the interface to output user-selected elements and attributes. As an example, Dragoset showed the database output for "joule"

```
<units>
  <unit unitID="UnitsML000013:01:00" asciiSymbol="J">
    <system name="SI" type="special derived"/>
    <name lang="en-US">joule</name>
    <quantityRef name="energy"/>
    <quantityRef name="work"/>
    <quantityRef name="quantity of heat"/>
    <representation id="00">
      <RepUnit name="kilogram" power="1" />
      <RepUnit name="meter" power="2" />
      <RepUnit name="second" power="-2" />
    </representation>
  </unit>
</units>
```

Computer-to-computer communication is enabled. Programs exist which are command line tools for transferring files with URL syntax, e.g., the open source, free software "cURL". The following query example returns the XML for the "joule" shown above:

http://unitsml-i.nist.gov/cgi-bin/UnitsML/search.php?
UID=
&UnitName=joule
&Quantity=all
&SI=on
&Type=all
&Language=en-US
&Format=0
&Reps=

An advantage of UnitsML, and the unit repository, is that the units repository is matched to markup language, allowing design of schemas and a repository to facilitate unit information processing. Industry-specific, customized data dictionaries can be made. However UnitsML is not yet a standard but is under development. In future, the developers of UnitsML aim to develop an ontology for scientific units, complete the units repository, version 1.0, and complete simple schemas for incorporating UnitsML into other XML documents.

Discussion

Alan McNaught asked where this fits in the world view and what the home is for something like this. CODATA? ICSU? David Martinsen suggested NISO. In the discussion it was stated that a good thing about XML is that it could be output in another format. This could be useful for an IUPAC UnitsML. ICSU has no history of this sort of standardization effort except through CODATA. This work affects multiple disciplines: physics, chemistry, biology etc. Should a bid be put in to ICSU?

Chemistry and the Crystallographic Information File (CIF)
David Brown, McMaster University, Hamilton, Ontario, Canada
idbrown@mcmaster.ca

CIF is used for the transfer of crystallographic information between laboratories, journals and databases and for the archiving of crystallographic information in journals (*Acta Cryst.*) and databases (e.g., the Protein Data Bank). Although it was originally designed to report information on crystal structures, it is being expanded to include derived chemical information. There will be some overlap with CML.

CIF uses the STAR (Self-Defining Text Archive and Retrieval) syntax (http://journals.iucr.org/iucr-top/cif/standard/cifstd1.html) and crystallographic dictionaries. Each item of information is given using two tags, the dataname, followed by its value, e.g.,

_cell_length_a 16.235(3)

Only characters of the ASCII set are used thus insuring maximum transferability and stability, and all units are given in SI except Angstroms. Loops are used to give tabular material. The datanames are defined in CIF dictionaries, seven of which have currently been approved. They are also written as STAR files, so the same software can read the CIFs and their dictionaries with the datanames used in dictionaries being defined in a Dictionary Definition Language (DDL). Dictionary entries list the datanames and their properties (i.e., enumeration, relationships between items) allowing CIFs to be computer-validated against their dictionaries. Future dictionaries will include expressions which will allow derived values to be machine-calculated from more primitive values.

CIF has a simpler structure than XML and is less verbose. Thus the processing time of large files is reduced and an ASCII listing of the CIF is easy to read. The dictionaries are owned by the International Union of Crystallography (IUCr) and managed by COMCIFS, the COmmittee for the Maintenance of the CIF Standard. A single dialect is accepted across the discipline by using

concepts that are rigorously defined in the extensive suite of dictionaries, an arrangement that works well for a mature field like crystallography. Unfortunately, software support is weak because of the difficulty in persuading programmers to use the high level CIF features.

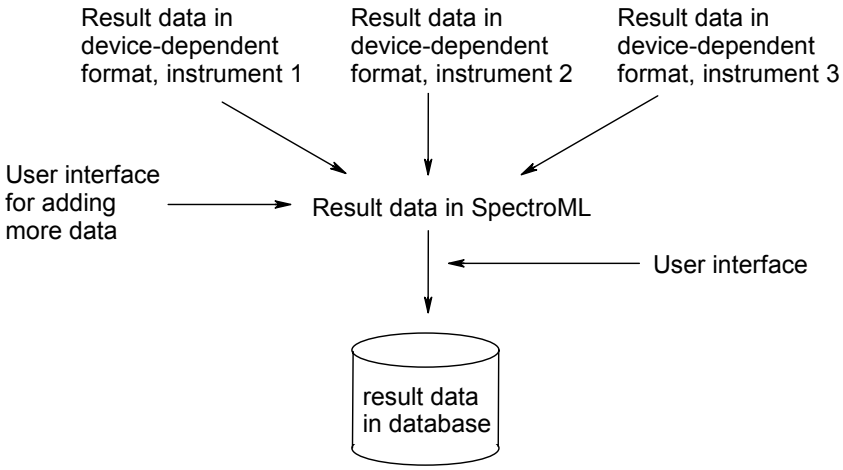
The items currently defined in CIF are crystallographic (unit cell, space group, atomic coordinates, interatomic distances) but CIF is being expanded to include chemical information (identification of molecules or complexes, bond lengths, formal charge, chirality, molecular structure (2D and 3D) and molecular symmetry). Brown attended the meeting hoping to get an understanding of the handling of chemical concepts in XML, to ensure a seamless interface between XML and CIF.

Discussion

Alan McNaught asked whether there was any chemistry already in CIF. David Brown replied that a chemical name can be given, but is not always included.

SpectroML and AnIML for molecular spectroscopy and chromatography data
Gary W. Kramer, NIST
gary.kramer@nist.gov

SpectroML is a markup language for spectroscopy data based on XML, JCAMP-DX (IUPAC), ANDI/Net CDF (ASTM), ThermoGalactic's GRAMS and SPC file formats, data definitions from instrument manufacturers, and ASTM definitions. SpectroML was originally defined for ultraviolet and visible spectroscopy data. Kramer gave a diagram of the application of SpectroML at NIST.



Subcommittee E13.15 on Analytical Data Management has been created under ASTM Committee E13 on Molecular Spectroscopy and Chromatography to develop a markup language for analytical chemistry result data. Known as AnIML, for Analytical Information Markup Language, this project will provide a uniform standard for the interchange and storage of all analytical chemistry result data and its associated metadata. ASTM E13.15 also has jurisdiction over the Andi (Analytical Data Interchange) standards for chromatography and mass spectrometry, LECIS (Laboratory Equipment Control Interface Specification), and the LIMS Guide and validation standards.

SpectroML and GAML (Generalized Analytical Markup Language from Thermo) served as starting points for AnIML. Instrument manufacturers, software companies, and others have been asked to join the initiative. This does not mean starting over. Once AnIML is in place, translators can be written to bridge current data sets to the new standard, which has been developed in a way that makes it extensible to multiple techniques yet avoids duplication of effort.

The conceptual model was shown by Davies earlier: layered by core, technique, vendor, enterprise and user. There is a need for software that allows a user simply to look at various traces and spectra without having to have the proprietary instrument software resident on his machine. The core structure of AnIML makes it possible to view data in a Web browser (e.g., Internet Explorer or Netscape) with simple add-on applets. The outer layers are built on the core and provide places to put technique-specific metadata, vendor-specific terms, business rules, etc.

Kramer reported on the progress of AnIML. There has been good, general participation in E13.15, and collaboration with the CPEP subcommittee on electronic data standards. IUPAC data dictionaries are being reused. The committee has worked on the structure and representation of data elements and on nomenclature of items within AnIML e.g., "experiment". Burkhard Schaefer was responsible for the core schema implementation. Dominik Poetz has implemented one example of a technique layer implementation, for liquid chromatography with diode-array detection. Creon LabControl are doing the technique layer implementations using both the JCAMP-DX and Andi data dictionaries. Application Programming Interface (API) developments, applications, and applets are not yet done. Kramer cited the following URLs where further information can be obtained:

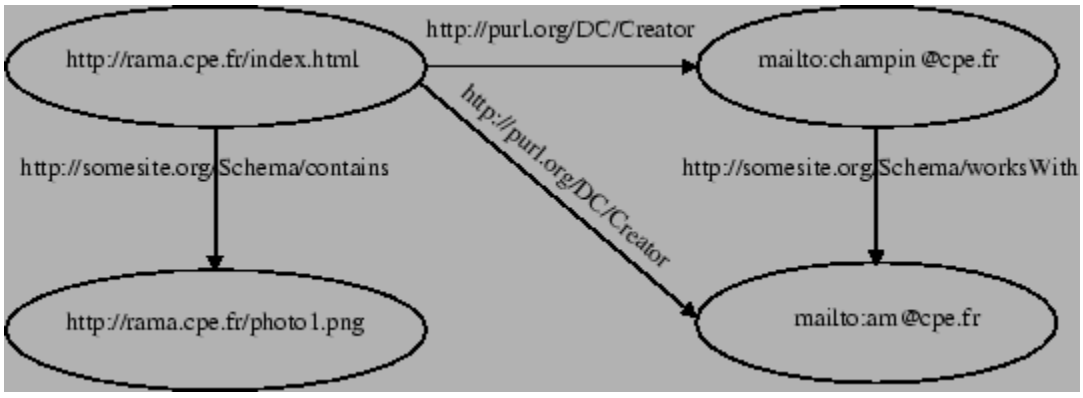
AnIML: http://animl.sourceforge.net and http://sourceforge.net/projects/animl/
Spectra ML and GAML: http://www.xml.org (click on xml registry, then click on chemistry)
GAML: http://www.gaml.org
LECIS: http://www.lecis.org

Capturing chemical information
Miloslav Nič and Jiří Jirát, Institute of Chemical Technology, Prague, Czech Republic
nicmila@systinet.com or miloslav.nic@vscht.cz and jiri.jirat@systinet.com or jiri.jirat@vscht.cz

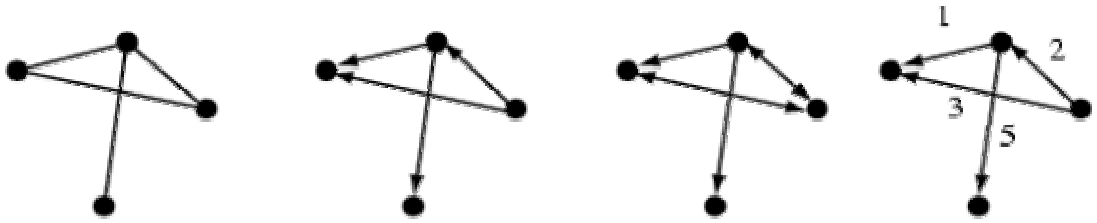
Nič spoke first. Contestants for the most frequently asked questions are "Where is it?", "Where have I seen it?" and "Where should I store it?". Chemical variants of these frequently asked questions are "I remember I read about it somewhere, but where?", "Where have I seen the structure (spectrum, data, ...)?" and "How should I record my data?" Answers might be found in the Semantic Web. According to Tim Berners-Lee, James Hendler, and Ora Lassila, the Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

The Resource Description Framework (RDF) is a language for representing information about resources in the World Wide Web. In the RDF model, "resources" are all things being described by RDF expressions, a "property" is a specific aspect, characteristic, attribute, or relation used to describe a resource, and an "RDF statement" is a specific resource together with a named property, plus the value of that property for that resource.

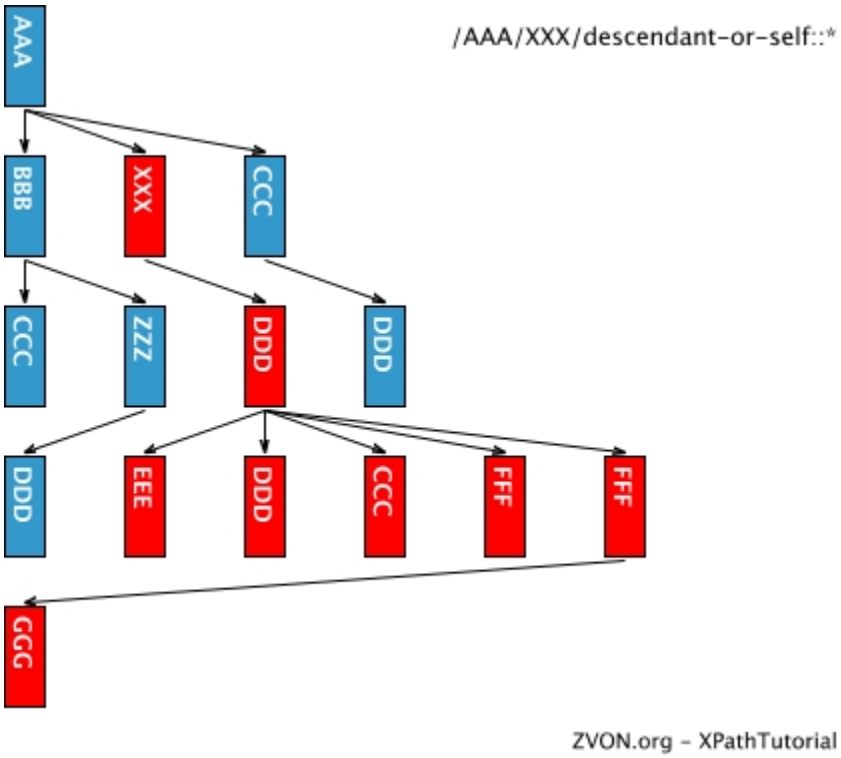
Nič gave an RDF graph example from P.A. Champin's RDF tutorial:



RDF statements can be expressed as a graph. MathWorld defines graph theory as the mathematical study of the properties of the formal mathematical structures called graphs. Graphs are mathematical objects composed of points known as graph vertices or nodes and lines connecting some (possibly empty) subset of them, known as graph edges. Nič gave some graph examples from MathWorld:



He related graphs and XML using the ZVON.org XPath tutorial:



XML is a nice format to express graph relations. Chemists use graphs all the time. An MDL Molfile, for example, illustrates a connection matrix for chemists. For acetone, $CH3.C=O.CH_3$, it is as follows.

Acetone Molfile

```
  -ISIS-  11040309352D

  4  3  0  0  0  0  0  0  0  0999 V2000
   -2.4500   -1.4250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9855   -0.5750    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9855    1.1208    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.4841   -1.4212    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  3  2  0  0  0  0
  1  2  1  0  0  0  0
  2  4  1  0  0  0  0
M  END
```

The PDB (Protein Data Bank) file for acetone also has a connection table:

Acetone pdbfile

```
HEADER    PROTEIN
COMPND

AUTHOR    GENERATED BY OPEN BABEL 1.100.0
ATOM      1 C    UNK     1      -2.450  -1.425   0.000  1.00  0.00
ATOM      2 C    UNK     1      -0.985  -0.575   0.000  1.00  0.00
ATOM      3 O    UNK     1      -0.985   1.121   0.000  1.00  0.00
ATOM      4 C    UNK     1       0.484  -1.421   0.000  1.00  0.00
CONECT    1    2
CONECT    2    3    1    4
CONECT    3    2
CONECT    4    2
MASTER        0    0    0    0    0    0    0    0    4    0    4    0
END
```

This format is very dependent on white spaces. If you forget one “nothing”, your program will be broken. CML (chemical markup language) is nice to use of you are a programmer.

Acetone - CML

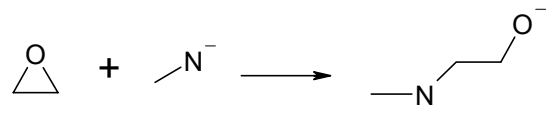
```
<molecule title="" id="m1">
<string title="comment"></comment>
  <atom id="a1">
    <string builtin="elementType">C</string>
      <float builtin="x3">-2.45</float>
      <float builtin="y3">-1.425</float>
      <float builtin="z3">0</float>
  </atom>
  <atom id="a2"><string builtin="elementType">C</string><float builtin="x3">-0.9855</float><float builtin="y3">-
0.575</float><float builtin="z3">0</float></atom>
  <atom id="a3"><string builtin="elementType">O</string><float builtin="x3">-0.9855</float><float
builtin="y3">1.1208</float><float builtin="z3">0</float></atom>
    <atom id="a4"><string builtin="elementType">C</string><float builtin="x3">0.4841</float><float builtin="y3">-
1.4212</float><float builtin="z3">0</float></atom>

  <bondArray>
    <bond>
      <string builtin="atomRef">a1</string>
      <string builtin="atomRef">a2</string>
      <string builtin="order">1</string>
    </bond>
    <bond><string builtin="atomRef">a2</string><string builtin="atomRef">a3</string><string
builtin="order">2</string></bond>
    <bond><string builtin="atomRef">a2</string><string builtin="atomRef">a4</string><string
builtin="order">1</string></bond></bondArray>
</molecule>
```

GTML, graph theory markup language, based on graph theory, addresses some issues that came to light when Nič and Jirát were preparing the Gold Book. The GTML for acetone is somewhat similar to CML.

Acetone - GTML

```
<graph xmlns='http://www.nicmila.org/GTML' type='molecule' id='e'>

<vertex id='a'>
  <symbol>C</symbol>
  <coordinates>
    <x>0.374</x>
    <y>0.217</y>
    <z>0.0</z>
  </coordinates>
</vertex>

<vertex id='b'><symbol>C</symbol><coordinates><x>1.615</x><y>0.938</y><z>0.0</z></coordinates></vertex>

<vertex id='c'><symbol>O</symbol><coordinates><x>1.615</x><y>2.375</y><z>0.0</z></coordinates></vertex>

<vertex id='d'><symbol>C</symbol><coordinates><x>2.86</x><y>0.221</y><z>0.0</z></coordinates></vertex>

<edge id='b-c'>
  <end idref='b'/>
  <end idref='c'/>
  <bond>double</bond>
</edge>

<edge id='a-b'><end idref='a'/><end idref='b'/><bond>single</bond></edge>

<edge id='b-d'><end idref='b'/><end idref='d'/><bond>single</bond></edge>
</graph>
```

Next Nič compared the MDL RXN file and GTML for the reaction:



$RXN

```
     ISIS     110420031425

  2  1
$MOL

  -ISIS-  11040314252D

  3  3  0  0  0  0  0  0  0  0999 V2000
   -2.2542   -1.2250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.5583   -1.2250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4042    0.2454    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  2  1  0  0  0  0
  1  3  1  0  0  0  0
M  END
$MOL
```

```
  -ISIS-  11040314252D

  2  1  0  0  0  0  0  0  0  0999 V2000
    2.4417   -1.3542    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.6375   -0.1542    0.0000 N   0  5  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
M  CHG  1   2  -1
M  END
$MOL

  -ISIS-  11040314252D

  5  4  0  0  0  0  0  0  0  0999 V2000
   11.1000   -0.5292    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   12.7958   -0.5292    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.6417    0.9417    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
   10.2521   -1.9978    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    8.5562   -1.9978    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  1  4  1  0  0  0  0
  2  3  1  0  0  0  0
  4  5  1  0  0  0  0
M  CHG  1   3  -1
M  END
```

GTML reaction

```
<graph xmlns='http://www.nicmila.org/GTML' type='reaction' id='a'>
  <graph xmlns='http://www.nicmila.org/GTML' type='molecule' id='e'>
    <vertex id='b'><symbol>C</symbol><coordinates><x>0.3442</x><y>0.9598</y><z>0.0</z></coordinates></vertex>
    <vertex id='c'><symbol>C</symbol><coordinates><x>1.7812</x><y>0.9598</y><z>0.0</z></coordinates></vertex>
    <vertex id='d'><symbol>O</symbol><coordinates><x>1.0642</x><y>2.2058</y><z>0.0</z></coordinates></vertex>
    <edge id='c-b'><end idref='c'/><end idref='b'/><bond>single</bond></edge>
    <edge id='d-c'><end idref='d'/><end idref='c'/><bond>single</bond></edge>
    <edge id='b-d'><end idref='b'/><end idref='d'/><bond>single</bond></edge>
  </graph>

  <graph xmlns='http://www.nicmila.org/GTML' type='molecule' id='h'>
    <vertex id='f'><symbol>C</symbol><coordinates><x>4.3232</x><y>0.8498</y><z>0.0</z></coordinates></vertex>
    <vertex id='g'><symbol>N</symbol><coordinates><x>5.3372</x><y>1.8668</y><z>0.0</z></coordinates><charge>-
1</charge></vertex>
    <edge id='f-g'><end idref='f'/><end idref='g'/><bond>single</bond></edge>
  </graph>

  <graph xmlns='http://www.nicmila.org/GTML' type='molecule' id='n'>
    <vertex id='i'><symbol>C</symbol><coordinates><x>11.6612</x><y>1.5498</y><z>0.0</z></coordinates></vertex>
    <vertex id='j'><symbol>C</symbol><coordinates><x>13.0982</x><y>1.5498</y><z>0.0</z></coordinates></vertex>
    <vertex id='k'><symbol>O</symbol><coordinates><x>13.8152</x><y>2.7958</y><z>0.0</z></coordinates><charge>-
1</charge></vertex>
    <vertex id='l'><symbol>N</symbol><coordinates><x>10.9422</x><y>0.3048</y><z>0.0</z></coordinates></vertex>
    <vertex id='m'><symbol>C</symbol><coordinates><x>9.5052</x><y>0.3048</y><z>0.0</z></coordinates></vertex>

    <edge id='j-i'><end idref='j'/><end idref='i'/><bond>single</bond></edge>
    <edge id='i-l'><end idref='i'/><end idref='l'/><bond>single</bond></edge>
    <edge id='j-k'><end idref='j'/><end idref='k'/><bond>single</bond></edge>
    <edge id='l-m'><end idref='l'/><end idref='m'/><bond>single</bond></edge>
  </graph>

 <edge id='e-n'><end idref='e' type='initial'/><end idref='n' type='terminal'/></edge>
 <edge id='h-n'><end idref='h' type='initial'/><end idref='n' type='terminal'/></edge>
</graph>
```

A number of chemical structure drawing programs are available free (ISIS Draw, Marvin sketch, ChemSketch) or open (BKChem, JChemPaint). One advantage of ISIS/Draw is the facility to attach data to an object, e.g., atom-number to an atom. The resultant Molfile, however, is cumbersome, e.g., for cyclopropane:

```
  -ISIS-  11080313392D

  3  3  0  0  0  0  0  0  0  0999 V2000
   -4.7583    0.6333    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.2875    0.6333    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -4.0208    1.9120    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  3  1  0  0  0  0
  2  1  1  0  0  0  0
  3  2  1  0  0  0  0
M  STY  3   1 DAT   2 DAT   3 DAT
M  SLB  3   1   1   2   2   3   3
M  SAL   1  1   3
M  SDT   1 atom-number
M  SDD   1    -4.2600    2.1800    DA    ALL  1      5
M  SED   1 1
M  SAL   2  1   2
M  SDT   2 atom-number                F
M  SDD   2    -3.1300    0.2000    DA    ALL  1      5
M  SED   2 2
M  SAL   3  1   1
M  SDT   3 atom-number                F
M  SDD   3    -5.2900    0.2000    DA    ALL  1      5
M  SED   3 3
M  END
```

The cyclopropane format in GTML looks strange but is more compact:

```
<graph xmlns='http://www.nicmila.org/GTML' type='general' id='e'>
  <graph xmlns='http://www.nicmila.org/GTML' type='molecule' id='d'>
    <vertex id='a'><symbol>C</symbol><coordinates><x>0.7263</x><y>0.537</y><z>0.0</z></coordinates></vertex>
    <vertex id='b'><symbol>C</symbol><coordinates><x>1.9723</x><y>0.537</y><z>0.0</z></coordinates></vertex>
    <vertex id='c'><symbol>C</symbol><coordinates><x>1.3513</x><y>1.62</y><z>0.0</z></coordinates></vertex>
    <edge id='a-c'><end idref='a'/><end idref='c'/><bond>single</bond></edge>
    <edge id='b-a'><end idref='b'/><end idref='a'/><bond>single</bond></edge>
    <edge id='c-b'><end idref='c'/><end idref='b'/><bond>single</bond></edge>
  </graph>

  <vertex id='h'>
    <type>atom-number</type>
    <value>3</value>
    <coordinates><x>0.2753</x><y>0.169</y><z>0.0</z></coordinates>
  </vertex>

  <edge id='a-h'>
    <end idref='a' type='initial'/>
    <end idref='h' type='terminal'/>
  </edge>

  <vertex id='g'><type>atom-
number</type><value>2</value><coordinates><x>2.1053</x><y>0.169</y><z>0.0</z></coordinates></vertex>
  <edge id='b-g'><end idref='b' type='initial'/><end idref='g' type='terminal'/></edge>

  <vertex id='f'><type>atom-
number</type><value>1</value><coordinates><x>1.1483</x><y>1.847</y><z>0.0</z></coordinates></vertex>
  <edge id='c-f'><end idref='c' type='initial'/><end idref='f' type='terminal'/></edge>
</graph>
```

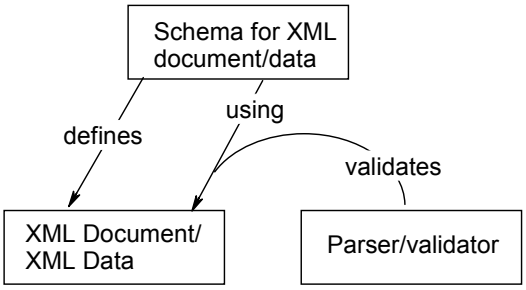
Transformations from Molfile to GTML and *vice versa* can be carried out by mdl2gtml and gtml2mdl. GTML can also be converted to Scalable Vector Graphics (SVG http://www.w3.org/Graphics/SVG) format (by gtml2svg). CML has some disadvantages. It is rich in elements and attributes; separation of layers, procedural processing, and extensibility are problems. The basic layer of CML is chemical: it introduces concepts as molecule or atom and the language is based on these concepts. GTML, on the other hand, is based on an abstract model. Graph theory can represent anything. Chemistry notions are introduced later in higher layers and this fact makes GTML easier to modularize and tailor to different circumstances.

XML is parsable and so on but it will not solve all data problems. XML can be transformed easily to whatever proves to be better five years from now. The semantics of biological data is very rich and requires a very expressive data model. (For a text on XML, bioinformatics and data integration, see *Bioinformatics*, **2001**, *17*, 115.) The expressiveness of the XML data model would probably not be sufficient for molecular biology. Modularity is necessary plus a way of expressing different mark-ups, different concepts, and different opinions.

Graph-XML languages include Graph Mark-up Language (GraphML http://graphml.graphdrawing.org/), Graph eXchange Language (GXL), eXtensible Graph Markup and Modeling Language (XGMML), and Structured Graph Format (SGF). Why use a special language for chemistry? Chemistry needs a combination of directed and undirected edges. The current proposals have some weaknesses for Nič and Jirát's purposes and so some changes would be necessary. There is a also the question of personal taste: it is better to have a new language than a tweaked old one. GTML has the advantages of being based on universal theoretical concepts, recursive, friendly for functional languages, and suitable for expressing vague concepts.

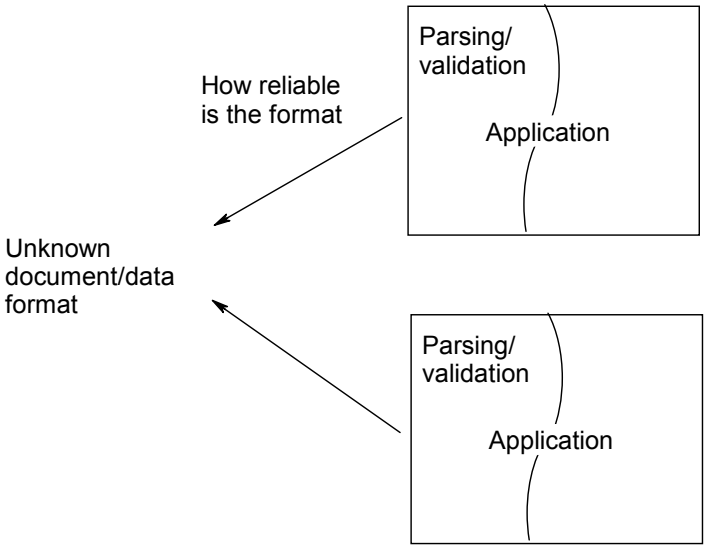
Current research topics in the encoding arena are the Gold Book, other color books, and tutorials and other educational materials. In the decoding field, Nič and Jirát are working on GTML2SVG, including usability studies, GTML2ANY4PVD (GTML to anything suitable for people with visual disabilities), and GTML2ANY4ANY (GTML as a universal transport format both for classical chemistry programs and Artificial Intelligence).

At this point, Jirát took over, covering XML formats and their validation, and the internal structure and output displays of the Gold Book. There are both human- and computer-readable definitions of the XML format. The schema for the XML format has a formalized specification and clearly defined borders for data exchange, i.e., the distinction between what must be validated by the system and what the user must validate himself. Jirát gave a diagram of the roles of the schema.

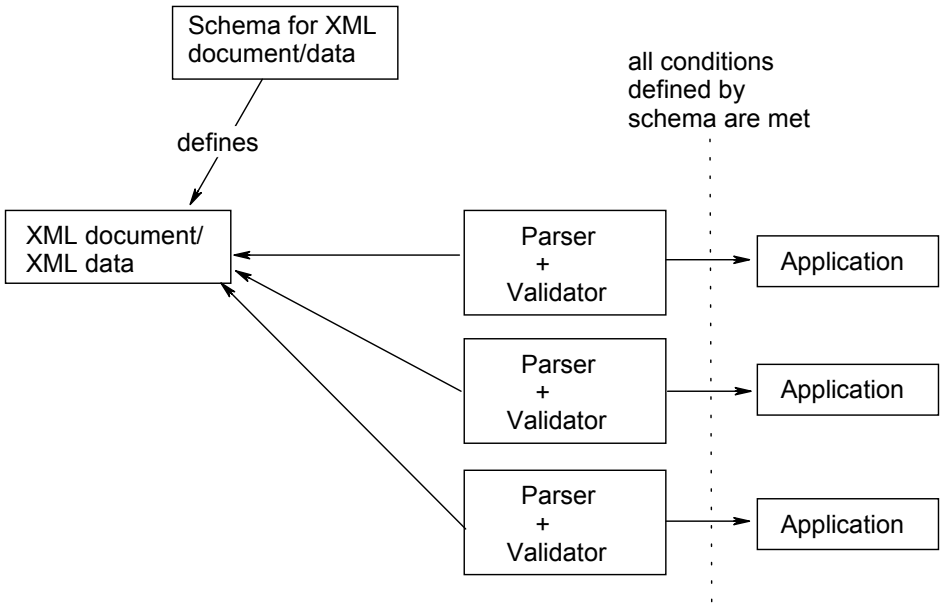


Non-standard and standard XML formats are parsed and validated as follows.

Non-standard formats:



Standard XML formats:

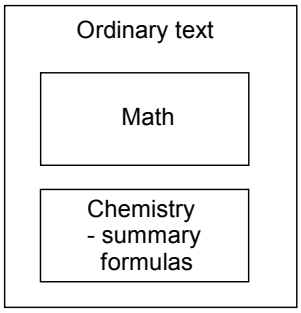


If there were one XML format for everything, there would be thousands of elements and attributes, and too many opportunities for mistakes, omissions, redundancies, discrepancies, gaps, and overlaps. Just one format would be impossible to implement and use. In order to combine chemistry, mathematics, physics etc., a modular approach is preferable, using small and compact languages distinguished by “namespaces”.

To introduce the concept of namespace, Jirát gave an example.

Syntax:

```
<text:aaa xmlns:text = “http://text”>
        <text:bbb/>
                <text:ccc/>
        </text:bbb>
        <<math:BBB xmlns:math = “http://math”>
                <math:CCC/>
        </math:BBB>
        <summary:x111 xmlns:summary = “http://summary”>
                <summary:x222/>
        </summary:x111>
</text:aaa>
```



Each XML language has its own namespace and schema (http://xml.coverpages.org/schemas.html). Validation is selective and focuses on a particular namespace; unknown namespaces are ignored or rejected. Applications process only elements from a selected namespace, e.g., an application might collect only chemical formulas, and omit everything else.

The different namespaces used in the Gold Book are ordinary text (unassigned yet); W3C MathML (for equations and symbols); W3C SVG (for pictures and graphs); summary formulas; and graph markup language (for structural formulas and reactions). For the Gold Book, it was felt that the more rigid the validation, the safer and more reliable the processing would be. One type of schema was insufficient; a combination of various approaches was the most powerful methodology. RELAX NG (REgular LAnguage description for XML, Next Generation, pronounced “relaxing”) plays a key role in overall structure definition and data types. A schematron plays a supplementary role, applying constraints across the XML tree. The XML DTD (Document Type Definition) is used only for MathML entities.

Jirát summarized the history of the validation process. SGML (Standard Generalized Markup Language, the predecessor of XML) had a DTD. The XML DTD is part of the XML specification. It has certain limitations: problems with namespaces, weak data types, and missing advanced constructs. The XML schema is designed mostly for database validations whereas Relax NG is designed for validation of trees (XML documents). The XML schema is a W3C standard. It is a huge specification, designed for software-software data exchange, and has severe limitations for document definitions. RELAX NG is specified by OASIS (Organization for the Advancement of Structured Information Standards, http://www.oasis-open.org). It is simple, yet powerful, and can cover both documents and data definitions. ISO/IEC FDIS 19757-2 “Document Schema Definition Language (DSDL) - Part 2: Regular grammar based validation” covers RELAX NG.

RELAX NG is simple, and easy to learn. It can partner with a separate datatyping language, such as W3C schema data types, which allows the possibility of introducing new data types. XML schema is very hard to learn and understand. James Clark, Chairman of the OASIS RELAX NG technical committee has commented: “The approach to handling data types in W3C XML Schema is totally lacking in modularity. W3C XML Schema is tied to the single collection of data types defined in Part 2 of W3C XML Schema.”

Data types are not just numbers. Identifiers such as CAS Registry Numbers, Digital Object Identifiers (DOI), ISBN, ISSN with check digits, control sum, parentheses, or redundant information cannot be checked by regular expressions. RELAX NG and XML Schema examples for atoms are:

RELAX NG
atom.xml
atom.rng
XML Schema
atom-sch.xml
atom-2.xsd

The coding for an H^{2+} atom is:

```
<summary xmlns='…'>
<atom>
        <symbol>H</symbol>
        <charge>2</charge>
</atom>
</summary>
```

Is this acceptable? The computer would accept it, a human would not understand and a chemist might query it. The Schematron rule for an H^{2+} atom is:

```
<sch:pattern name="Charge of hydrogen" id="charge-H">
  <sch:rule context="chem:atom[chem:symbol = 'H']">
    <sch:assert test="not(chem:charge) or chem:charge &lt; 2">
      <sch:p>
      Error: element "H" has charge: <sch:value-of
select="chem:charge"/>
      </sch:p>
    </sch:assert>
  </sch:rule>
</sch:pattern>
```

The code above represents a hydrogen atom with charge 2+ (which is nonsense). The computer would accept any number, but Nič and Jirát's methodology can provide a Schematron rule, which checks the following:

if the atom is H (<sch:rule context="chem:atom[chem:symbol = 'H']">),
then the charge (if present) must be less than 2 (<sch:assert test="not(chem:charge) or chem:charge < 2">).

Thus the computer can perform more advanced tests. More Schematron examples are summary.sch and graph.sch.

Jirát described the generation of the Gold Book in XHTML. The data were stored as more than 6000 separate documents. Concatenating all the items was necessary for building indexes, and for complete cross- and back-references. An XHTML document was created for each item, so more than 6000 documents were created. Structural formulas and reactions were generated as SVG and/or bitmap images. SVGs can be rendered to bitmap images, for use by those who do not have an SVG plug-in, or for use in printing. Indexes were generated for chemistry, mathematical symbols, equations, shortcuts, and sections.

The Gold Book (http://www.iupac.org/publications/compendium/index.html) is searchable at http://www.chemsoc.org/cgi-shell/empower.exe?DB=goldbook. Towards the end of the lecture, Jirát showed some screen displays. First he showed the definition of "wavefunction", a Gold Book item as an XHTML document. Features of the system include automated checking of references, "cited by" (creation of back-references), various indexes and multiple output formats.

The Gold Book introductory page lists the indexes. The mathematical one indexes symbols and equations. The physics index covers units and values of quantities. The chemistry index is for looking up summary formulas, structures and reactions. There are also several general indexes. The first is a list of all items in alphabetical order. “Sections” are of the type “in” (e.g., “in analytical chemistry”, or “in chromatography”) or “of” (e.g., “of a polymer” or “of luminescence”). Shortcuts (HPLC, LAMES etc.) are also listed, and obsolete terms (e.g., “natural lifetime”) and quoted texts are also indexed. Certain items in the text are italicized but are not explained. Jirát suggested that these could provide a list of candidates for new entries in the Gold Book. He asked for feedback on any other indexes that might be useful.

He showed a screen where values of quantities were listed, e.g.,

750K mesogenic pitch
752K first-orderphase transition
760K delayed coking process
etc.

The words were hyperlinked to other pages. Jirát also showed the page in the index of summary formulas with entries for “acetals, “acid rain in atmospheric chemistry”, and “activated carbon”.

The various Gold Book output formats are XHTML plus MathML plus SVG; XHTML plus MathML plus (SVG as bitmap); XHTML plus (MathML as bitmap) plus SVG; XHTML plus (MathML as bitmap)+(SVG as bitmap); PDF…; and HTML…Jirát suggested that XML formats for data exchange might also be considered.

There remain some problems to be solved. One is semantics of objects such as units, and symbols of quantities, and constants in MathML content markup. Numbers with precision in content markup, e.g. 1.023($\pm$ 2), and shortcuts, e.g., ranges (a = 100-200 nm), and enumerations (x_1, …, x_n), are also problems, and how does one express that a symbol has a constant value? Jirát concluded with two items for discussion. One concerned semantics of MathML objects: might a dictionary of units and quantities in the OpenMath approach be based on the Green Book? The second was similar chemical dictionaries: definition files in XML format on the IUPAC.org Web page.

Discussion

Steve Stein and Tony Davies raised migration issues. Solutions that work only with specialized software will not be used or survive.

Steve Stein: I am disappointed with XMLSpy for CAS Registry Numbers etc.

Tony Davies: Extra data types had to be written for NetCDF for Alpha chips. It is no good when you have to put something into Oracle as a blob (binary large object).

Michael Frenkel: Either you have to use existing technology or you develop your own. There are only two choices.

Alan McNaught: Concentrate on the content not the technology.

Miloslav Nič: It was five months work converting PDF to HTML and doing the corrections.

Peter Linstrom: You can easily go from a highly defined format to weak semantics but not *vice versa*. I recommend to IUPAC that we capture the information knowing more rather than less.

Jeremy Frey: I have not been able to download Jumbo etc. from Peter Murray-Rust’s site and use it but Peter is now working to ensure that CML works for everyone who downloads it.

ThermoML and IUPAC Activities in Standardizing Thermodynamic Data Communications

Michael Frenkel, Thermodynamics Research Center (TRC), National Institute of Standards and Technology (NIST), Boulder, Colorado
frenkel@boulder.nist.gov

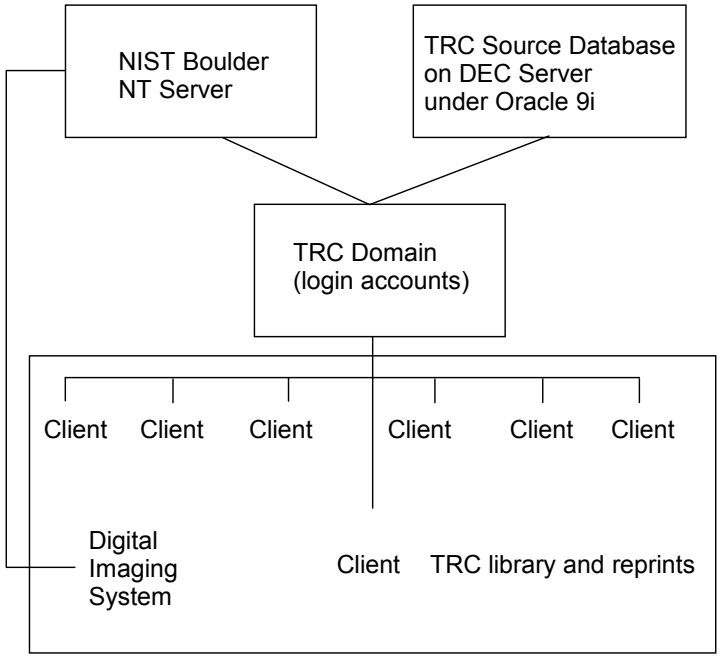
TRC is developing DDE (Dynamic Data Evaluation) with a large user base. How should they return the data to the public domain? XML, which is platform-independent, seemed the best technology to use. ThermoML is an XML-based approach to storing and exchanging thermophysical and thermochemical data. It was developed in close cooperation with project 991 of the Design Institute for Physical Properties (DIPPR) of the American Institute of Chemical Engineers (AIChE). The scope is properties of pure compounds, mixtures, and chemical reactions. Meta and numerical data records are grouped into “nested blocks”. Elements of the Gibbs Phase Rule are at the core of the schema. IUPAC terminology is used for meta and numerical data tagging. Very limited use is made of abbreviations. There are various methods of numerical data presentation. The types of data to be covered are experimental, critically evaluated, and predicted. Extensive validation is being carried out with the TRC Source database which has more than 5,000 data sets from more than 3,000 publications.

[A discussion began at this point.
Peter Linstrom: What do you mean by validation?
Michael Frenkel: Going from Source to ThermoML and back to Source.
Bill Milne: Has IUPAC defined “validation”?
Steve Stein: Validation of the schema and validation of the data are different.]

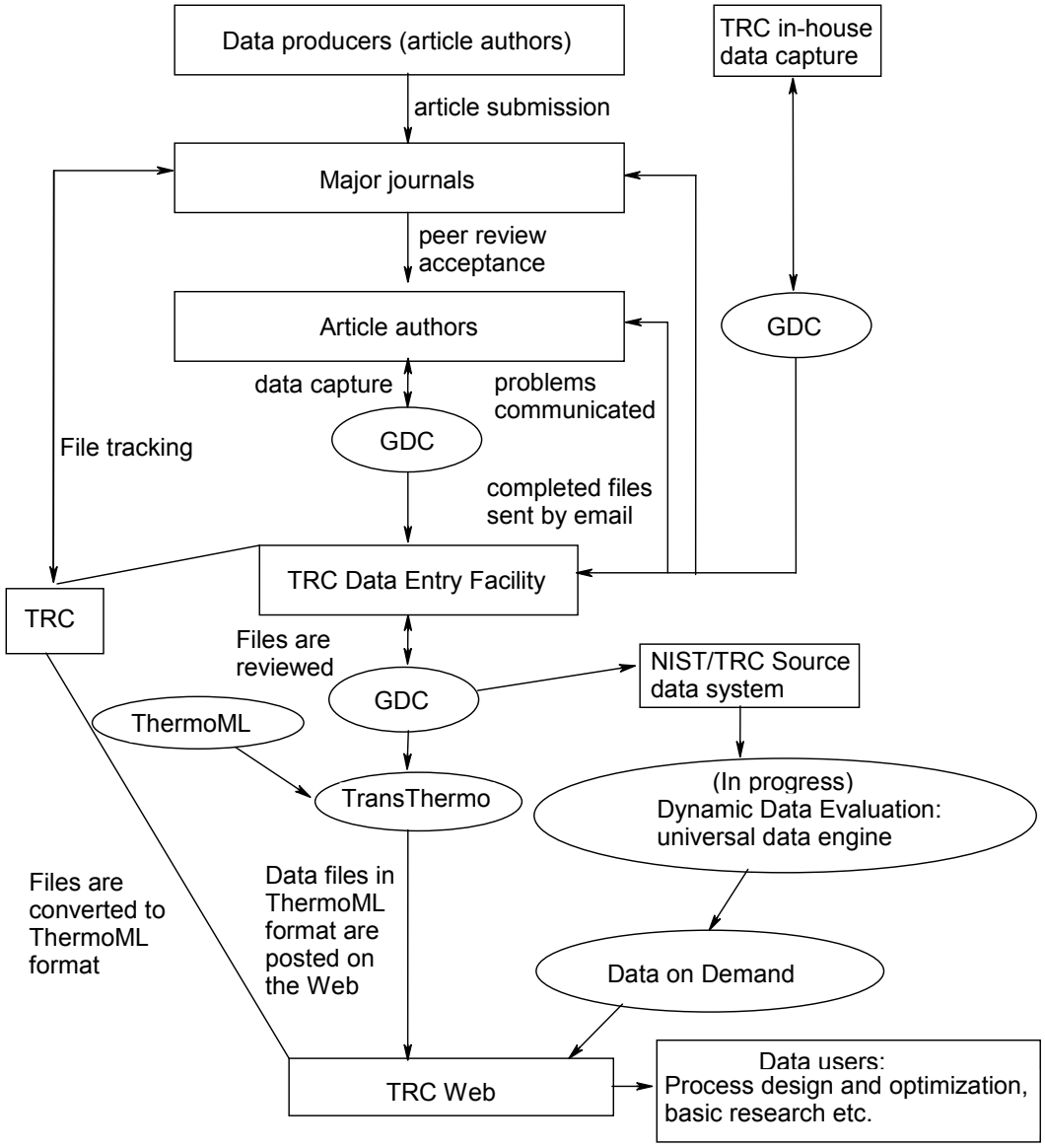
The ThermoML framework for experimental data was published in *Journal of Chemical and Engineering Data*, **2003**, *48*, 2-13. A ThermoML extension to cover various measures of uncertainty conforms to the Guide to the Expression of Uncertainty in Measurement, ISO, International Organization for Standardization, October 1993. A paper about this has been submitted to the *Journal of Chemical and Engineering Data* and accepted for publication. The last major extension is planned to be completed at the end of 2003, to cover critically evaluated and predicted data. A combination of GDC (Guided Data Capture) and ThermoML is used to generate ThermoML files for the data submitted by the authors, posted on the TRC Web site. Cooperation with other journals is planned to be taking place by the end of 2003.

A review of ThermoML for experimental data has been published by Frenkel *et al.* in *J. Chem. Eng. Data* **2003**, *48*, 2-13. This gives schema for the general structure of ThermoML, compound and sample descriptions, the data set structure, property descriptions, property types, and numerical values of property. Representation of uncertainties (specification of uncertainties, and uncertainty values) is described by Chirico *et al.* in *J. Chem. Eng. Data* **2003**, *48*, 1344-1359.

Frenkel gave a diagram of the operational structure of the TRC data entry facility:

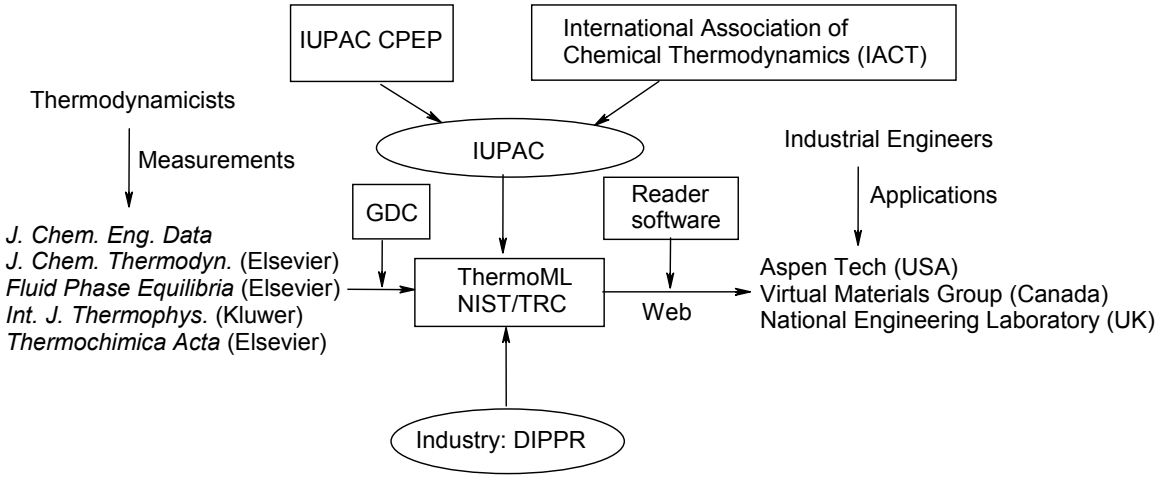


An Editorial by Kenneth Marsh at *J. Chem. Eng. Data* **2003**, *48*, 1 describes the new process for data submission and dissemination. Authors of manuscripts with data that can be captured with GDC will be expected to submit their data to NIST using the GDC software. Marsh stated: “The implementation of these new data archival and electronic dissemination mechanisms will provide enormous benefits to the industrial and academic communities using thermophysical property data and will provide an example for the development of XML formats for use with other data types”. The following procedure is operational with the *Journal of Chemical and Engineering Data*.

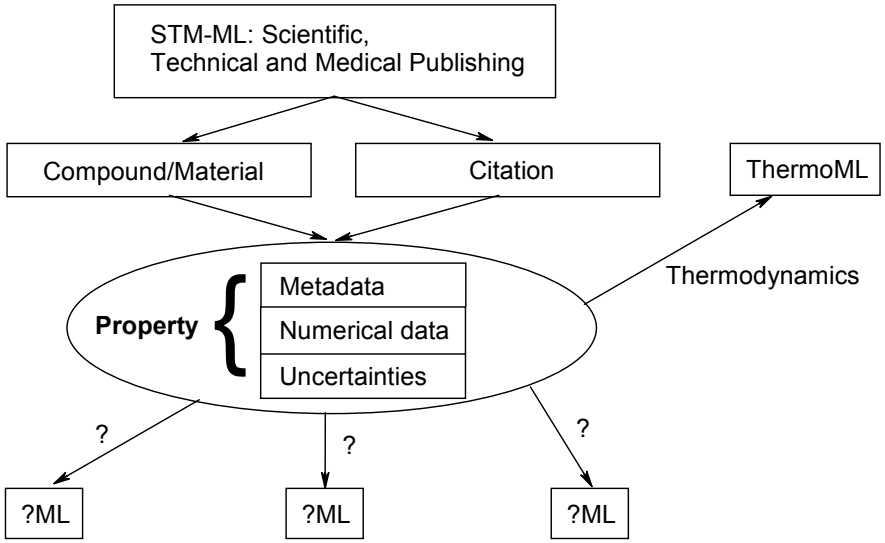


Within an article in the *Journal of Chemical and Engineering Data* there are links to ThermoML files on the Web for that article. In Issue 1 of 2003 there were 3 such articles; in Issue 2, 21 articles, and in Issue 3, 30 articles. Frenkel displayed a ThermoML file available for free download. The current IUPAC project, 2002-055-3-024, “XML-based IUPAC standard for experimental and critically evaluated thermodynamic property data storage and capture”, overseen by CPEP, is due for completion in 2004. The Task Group is chaired by Frenkel.

Frenkel gave a diagram of the global data communication process which currently is designed to involve five journals:



Further journals will be involved in future. Frenkel concluded with a diagram showing modular implementation of XML storage of technical data:



Discussion

David Brown commented on similarities of concept in the TRC system to the system for data capture and validation in crystallography. Michael Frenkel agreed with this statement, mentioning however, that there are 2 major differences: the TRC process is designed for 120 properties and provides XML-based output. Tony Davies saw similarities between the spectroscopic databank and the ThermoML and crystallographic ones. Jeremy Frey was concerned that, since there is only a three-year grant for the spectroscopic databank, people will be reluctant to put data into the system if they think that the system will not survive in the long term.

Protein Data Bank (PDB) and Chemical Compound Data Annotation and Query

T. N. Bhat, NIST
talapady.bhat@nist.gov

The Crystallographic Information File (CIF), a subset of STAR (Self-defining Text Archive and Retrieval format), was described by Brown earlier. The result of the CIF effort was a dictionary of data items sufficient for archiving the small molecule crystallographic experiment and its results.

This dictionary was later expanded by including data items relevant to the macromolecular crystallographic experiment. Version 1.0 of the macromolecular Crystallographic Information File (mmCIF) dictionary was released in June 1997 and updates have taken place since then (http://ndbserver.rutgers.edu/mmcif/). The development of the mmCIF dictionary and the associated Dictionary Definition Language, DDL 2.2.1, are described in The Macromolecular Crystallographic Information File (mmCIF) *Meth. Enzymol.* (1997) 277, 571-590; http://www.sdsc.edu/pb/cif/papers/methenz.html, and STAR/mmCIF: An Extensive Ontology for Macromolecular Structure and Beyond *Bioinformatics* (2000) 16(2), 159-168; http://deposit.pdb.org/mmcif/bioinfo00.pdf.

The mmCIF dictionaries carry three letter codes, and specify atom names, bonds etc. These can be used in searching for similar structures. IUPAC names are included sometimes. There are thousands of compounds in animals or plants, playing a major role in drug discovery, which are not made up only of amino acids or nucleotides. The methods used to name and curate peptides and DNA as fragments cannot be used for chemical compounds and drugs: peptides may be fragmented using amino acids but no such established method is available for chemical compounds. Usually during annotation, chemical compounds are assigned a name based on all the atoms. Such naming conventions are subject to data uniformity problems among fragments. For this reason, annotation techniques (Bhat, T. N.; Bourne, P. *et al.* The PDB Data Uniformity Project *Nucleic Acids Res.* **2001**, *29(1)*: 214-8) and search engines (Berman, H. M.; Westbrook, J. *et al.* The Protein Data Bank *Nucleic Acids Res.* **2000**, *28(1)*, 235-42) are not synchronized in their methods of processing chemical compounds.

Annotation of chemical compounds is often done as a single entity and Web tools query chemical compounds using their fragments which may or may not directly relate to the data dictionaries used for annotation. We need better methods for annotation of chemical compounds uniformly in terms of their fragments. A novel method that addresses some of these issues is being developed (Bhat, T. N.; Prasanna, M. D. Protein Data Bank and Chemical Compound Data Annotation and Query**,** to be published elsewhere).

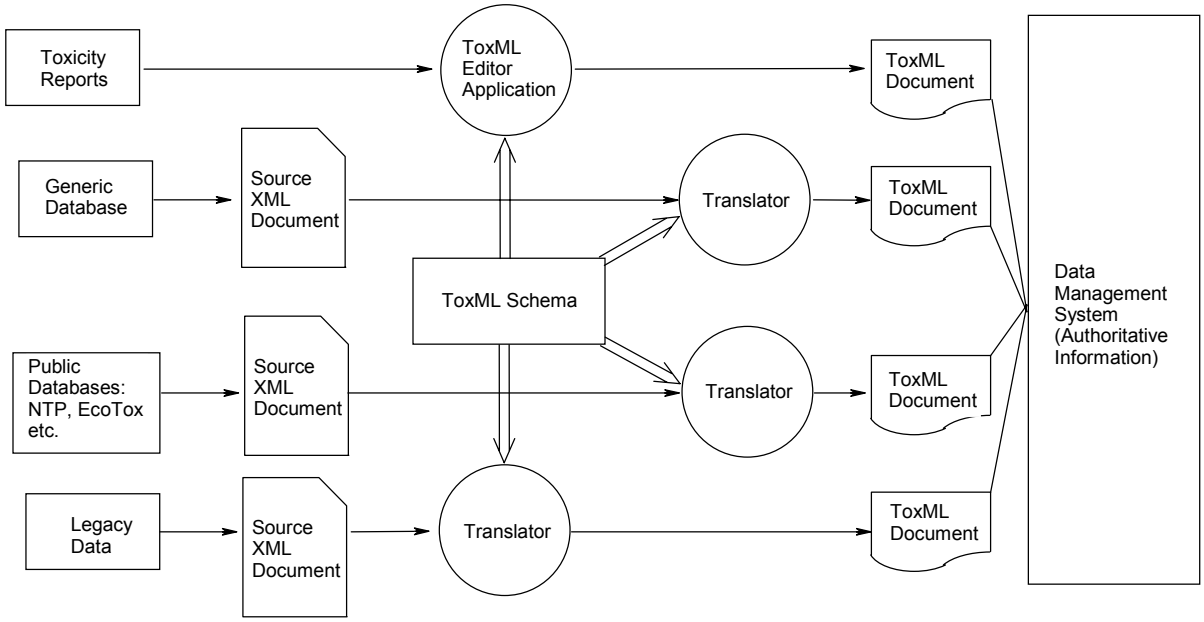
ToxML: Controlled Vocabulary for Toxicity Data Integration and Data Mining
Chihae Yang, LeadScope Inc.
cyang@leadscope.com

LeadScope develops structure-based decision support software (LeadScope Enterprise for data integration, searching, data mining and prediction) and databases (the LeadScope toxicity and LeadScope known drugs content modules).

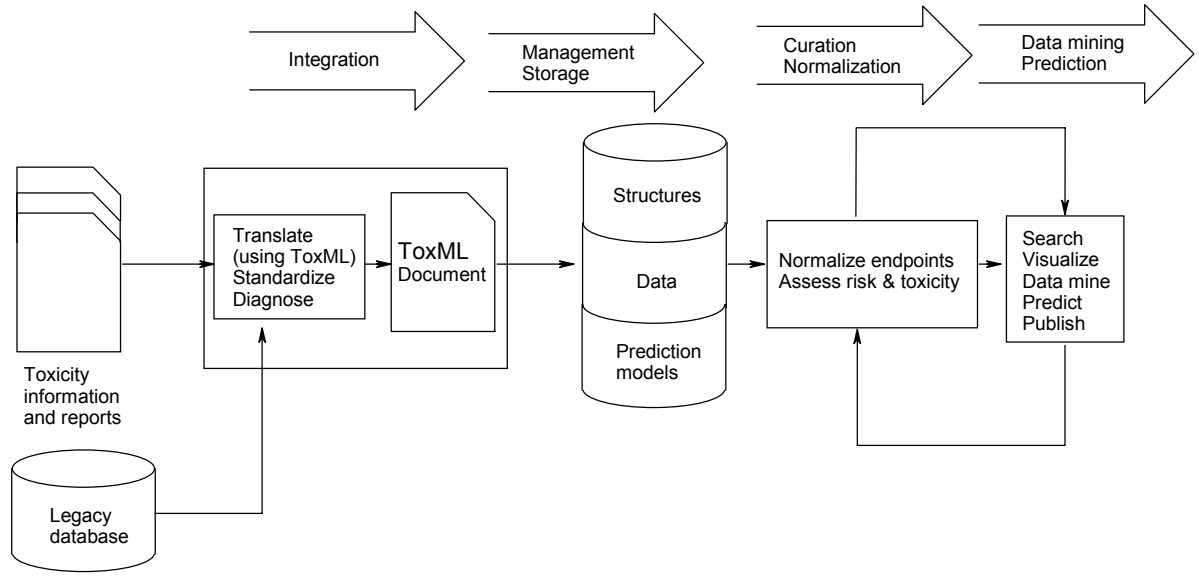
Toxicity data is currently fragmented and disparate, from many diverse sources, characterized by uncontrolled language and data model. It is inaccessible, its consistency in terms of data quality is questionable, and its use for data mining and decision making is limited because of the inadequate quantity, quality, and level of annotation.

The objectives of the project described by Yang were to streamline the processing of toxicity-related information (both regulatory and industrial) by introducing a common language; to enhance searching and retrieval of information by use of a controlled vocabulary and seamless integration; to enable the assessment of data quality; and to enhance decision making, data mining and prediction by giving transparent access to structures and data.

ToxML is an extensible, open standard XML format for toxicity data. It is a representation of detailed toxicity experiments and endpoints, independent of database schema or application. Yang gave a diagram of a standardized toxicity database.



She also discussed the workflow of *in silico* toxicology. A database of structures, data and prediction models is built from the ToxML documents and predictions can then be carried out:



Yang gave detailed examples of the entities involved in carcinogenicity and the Ames test.

A forms-based editor for entering toxicity information is used to convert experimental information to the ToxML standard, check the validity of each section as data is entered in the form, and assess the completeness of the required fields. This editor reads and writes a ToxML document. Yang displayed a typical screen for a document editor for toxicity. On the left hand side was a Window Explorer-like listing and, in the main part of the screen, information about a substance (functional and chemical category for substance type, molecular structure, names and IDs, physical properties, calculated properties, and measured properties) could be entered along with the substance identifier. Yang illustrated an IChI identifier in the section where names and identifiers are entered.

She then selected “Add new study” and was given the option to choose “*in vivo*”, or “*in vitro*”, or “genotoxicity battery” in the main part of the screen. She chose the last and a new screen appeared allowing her to choose a regulatory test (bacterial mutagenesis, or *in vitro* chromosome aberration, or *in vivo* micronucleus, or *in vitro* mammalian mutagenesis) and screening tests (*in vitro* micronucleus, *in vivo* chromosome aberration sister chromatid exchange, HGPRT, UDS, comet, DNA adduct, and germ cell effects). She also displayed the example of a complete ToxML document for Tylenol, occupying more than two pages of coding, including a Molfile, an IChI and toxicity information.

The Green Book “Quantities, Units and Symbols”
Jeremy Frey, University of Southampton
j.g.frey@soton.ac.uk

The first *Manual of Symbols and Terminology for Physicochemical Quantities and Units* (the “Green Book”), was published in 1969, with the objective of securing clarity and precision, and wider agreement in the use of symbols, by chemists in different countries, among physicists, chemists and engineers, and by editors of scientific journals. Subsequent editions of this book (http://www.iupac.org/publications/books/author/mills.html) and of *Quantities, Units and Symbols in Physical Chemistry* have been substantially revised and extended, containing many new resolutions and recommendations from Conférence Générale des Poids et Mésures, the International Union of Pure and Physics, the International Organization for Standardization and IUPAC. The style of the manual has also been changed such that it is no longer simply a book of rules, but more a manual of assistance and advice to meet the everyday needs of the practicing scientist.

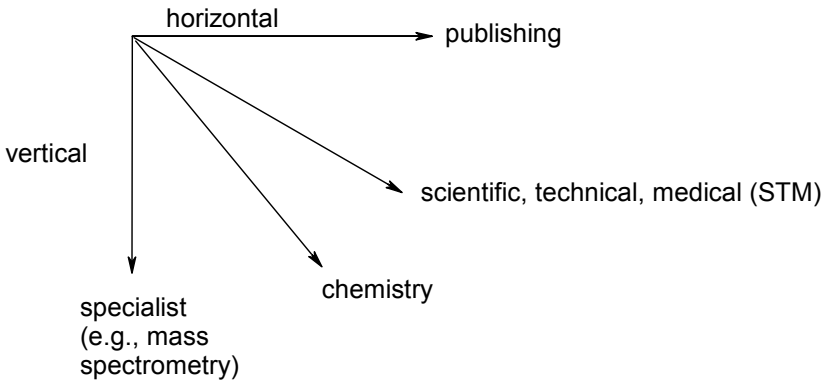
The book’s contents are: Historical Introduction; Physical Quantities and Units; Tables of Physical Quantities; Definitions and Symbols for Units; Recommended Mathematical Symbols; Fundamental Physical Constants; Properties of Particles, Elements and Nuclides; Conversion of Units; Abbreviations and Acronyms; and References.

Unfortunately the Green Book is out of date and until the TeX version is released, work cannot commence on an active implementation such as that done for the Gold Book. There remains a small problem, as even TeX is not completely portable, and when the exact representation of the symbols is crucial, then it is essential to make sure that the correct fonts are also available especially to render special characters such as Greek letters.

In the active version of the Green Book it will be possible to click on an item in an equation and see more information if that item occurs elsewhere in the book. Once the Green Book is out in PDF format Commission I.1 will launch projects to produce a fully online version and to ensure that people are aware of the Green Book. In the past, publicity has not been as good as it should have been.

The Gold and Green Books: Day Two Introduction
Stephen Stein, NIST
steve.stein@nist.gov

Stein gave a diagrammatic view of the horizontal (generalized) and vertical (specialist) applications of terminology:



It is relatively straightforward to develop an agreed-on terminology and XML application in vertical specialist areas. More broadly applicable terminology can be much more difficult to establish because effective terminologies must be acceptable to all of the specialties that use the terms. The Gold Book is a wide container for basic chemistry concepts, devices, procedures, data processing (mathematics), and terms associated with the data. The “container” has descriptors (data annotation), experimental apparatus, sample conditions, and methods. Should the terms be placed in categories? This raises problems concerning terminology “ownership” which can be difficult to solve. Each discipline develops domain-specific terms (metadata) and when multiple disciplines overlap, differences in matching terms must be resolved if these disciplines are to communicate. IUPAC is an authority for generic chemical terms such as cation, chemical formula, elementary reaction, reaction order, elements, and relative molar mass. The Gold and Green Books can be seen as an “upper level” for terminology and properties.

The Gold Book and CML have overlapping data dictionaries (Stein illustrated these as two separate tree structures). “Namespaces” may not solve the problem. Broad, unifying XML applications such as Peter Murray-Rust’s STMML are intended to unite them to make a single document. Stein raised the possibility of conflicts as STMML took input from the Gold Book, CML, UnitsML, ThermoML etc. to make one document. The Green Book promises to be a template for numeric property validation in chemistry, to ensure proper units and representation and basic standards for numeric data tagging. The Green Book is very densely packed. The IUPAC Red Book is the “official” digital source of the periodic table and an IUPAC Commission publishes relative atomic masses. IUPAC Color Books are the root of a chemical information “tree” of spectroscopy, electrochemistry, thermochemistry, catalysis, etc. Properties and units would be traceable to the IUPAC definition, and specialists within IUPAC could reach consensus definitions for such broad-based terminology.

Discussion

Lorrin Garson: DTDs for Standard Generalized Markup Language (SGML) were developed a long time ago. The work came to a halt but it might pay to study the DTDs.

Standards and Conventions for Electronic Interchange of Chemical Data
Peter J. Linstrom, NIST
peter.linstrom@nist.gov

Electronic data interchange standards should support existing standards and conventions. Semantics are more important than syntax. Many problems share common components; so let us standardize the common problems first. Linstrom addressed these issues in turn. Important work in chemical information has already been done, for example compilation of the Gold, Green and Blue Books; there is no need to repeat work which has already been done. Electronic formats which are not semantically compatible with existing standards and conventions will undermine them. An example is the nomenclature for spiro compounds, SP-1.5. Systems which do not

record the information conveyed in physical markup (e.g., superscripts) will not be able to correctly represent such names. If electronic formats cannot support the semantics of existing standards they will undermine them.

Science is incorporated in the semantics, not the syntax, so the syntax should not obscure the science. Meaningful standards are written in human readable language since they will be implemented by humans. The science (semantics) should be clear to all, not just to information science specialists. IT also changes faster than fundamental concepts in science. Semantic precision is important. “Free format” electronic standards are of limited value: compare typesetting with logical markup. The use of meaningful ontologies increases the value of the data for automated processing and meaningful ontologies come from the chemical sciences not from information technology. The XML schema does not have sufficient semantic rigor to capture the semantics of many types of scientific data, e.g., chemical formulas, IUPAC nomenclature, arbitrary precision numbers, and objects (such as reactions) with constraints. Standards should specify the limits of what can be validated with XML schema. An IUPAC name, for example, implies complex constraints.

As for common problems, certain concepts are incorporated in many different chemical informatics problems. Examples are system identification (chemical species, mixtures and reactions, and state and conditions); property identification (ontologies of properties and expressions of numerical values, and uncertainties); and source identification, that is, bibliography. There needs to be only one solution to common problems. There are many advantages to this approach. Some of these problems are more complex than they may initially appear to be; it is better to focus on getting them right the first time. This preserves semantic precision. A single solution also avoids repeated efforts, thus saving time, and produces reusable code, which lowers costs. The IUPAC chemical identifier is an example of the single-solution approach.

“Simple” common problems can actually be complex. Chemical formulas have atom counts indicated by symbols and integers, but what about ordering and grouping, electron counts for ions, and neutron counts for isotopically pure atoms, and how should complexes be handled? The markup of superscripts in IUPAC names is mandated by an existing standard but what about

b-Carotene
#beta;-Carotene
#946;-Carotene
.beta.-Carotene
<<beta>>-Carotene?

All of these notations have been used to represent a valid IUPAC name, β-Carotene. Are any or all of these acceptable?

Common problems in chemical informatics include chemical formulas, names and identifiers; chemical structures (but not the IUPAC identifier); and definitions of properties, states, and instruments and techniques. Many of the common problems represent indexes to chemical information. Construction of interoperable data systems requires that the items in these indexes be standardized. Construction of complex data formats is greatly aided by the availability of suitable high quality building blocks.

In summary, we should support, and build on, the strengths of the existing chemical information standards. Science is independent of the delivery mechanism. [Jeremy Frey corrected “delivery mechanism” to “transport format”.] Electronic data formats are of limited value without semantic precision. Shared problems should be solved first.

The Gold Book: Progress Report

Stephen Stein, NIST
steve.stein@nist.gov

At present the Gold Book on IUPAC.org has each term on a PDF page, looking like a printed page but with links to other definitions. The pages are for display only; they are not easily convertible because of the graphics, symbols and equations they contain. The text is not parsed and there is no metadata. Stein showed a page from the Gold Book. He also showed the IUPAC Web page displayed after a Google search for “osmotic pressure”. This shows how putting the Gold Book on the Web gives it greater importance and use.

Miloslav Nič and Jiřī Jirát have been translating the Gold Book (http://www.nicmila.org/Gold/Output/), as described earlier in this report. The text has been processed automatically, by perceiving and tagging data types and relationships; this work is complete. Equations should also be live objects, after tagging in MathML; most of this tagging work has been done. Simple structures can be translated automatically to connection tables, CML and SVG. So far, 270 structures have been translated; 50-60 very complicated structures remain undone. Figures and complex schemes (bitmap images) need to be redrawn in SVG. Quite a few months work remains. All but about 100 pages of the Gold Book will be ready at the beginning of 2004 but this does not mean that users will be able to point a browser at it and use it in the way they want to use it. Also, Aubrey Jenkins’ updates are not included.

Miloslav Nič and Jiřī Jirát are perfecting the chemical parsing, and working on editing and updating methods, and Web access. Aubrey Jenkins, Alan McNaught and Peter Linstrom (and others at NIST) are redrawing figures and complex schemes, carrying our proofreading, dealing with display issues and data confirmation, and doing updates. Maintenance will involve fixing errors and adding new entries. It is hoped that in future the Gold Book will be used by other MLs, by means of the schema, and by integration of the dictionary *via* STMML. Internal maintenance will include development of authoring procedures. The Gold Book project will provide uniform chemical terminology for XML documents, traceability of terminology, a root for the tagging of chemistry, and a model for future IUPAC recommendations.

Discussion

Alan McNaught: There are many more cross-references in the PDF than in the printed version. Some terms in the Gold Book are used in the Green Book but only as items in equations. The Gold Book gives a definition of them in words. These features of the Green and Gold Books need to be brought together. About 100 or 200 definitions are missing from the translated Gold Book (out of 2000) so the work is 95% complete.

Jeremy Frey: Because of a lack of links, the current version is of little use to the casual user. It should not be released until it is more usable.

Steve Stein: There are now more indexes to the information than there were in the PDF version: there are now 400 pages of indexes. A new dimension is the relationship between terms. This is the direction in which to go in the future but it will not be as simple as might be thought from my earlier comments.

Jeremy Frey: A word may be used in 25 entries but it may not mean the same in all 25 places. We are referring to *internal* links in the glossary.

Tony Davies: Device independence is significant. W3C is working on moving content to WAP phones. At the moment IUPAC is converting a book into all sorts of formats (SVG, CML etc.) A typical application will read all the stored formats and interpret them for the user in the outside world. An application that understands CML, for example, will select the right information for CML and ignore other formats. At this stage we are *capturing* the information, not *displaying* it.

Jiří Jirát and Jeremy Frey discussed kappa, and other font problems.

Steve Bryant: The National Library of Medicine (NLM) has a textbook project in which books have been marked up in XML and PubMed citations have been linked. If IUPAC gave NLM access to the Gold Book, NLM could link it to PubMed.

Michael Frenkel: In the July 2003 issue of *Today's Chemist at Work*, Nancy Maguire reported on the STIX project (http://www.ams.org/STIX)

Lorrin Garson: STIX will make available a font set for more than 8000 characters. Most of them have GIFs.

Jeremy Frey: IUPAC should not say "thou shalt use 'x' now" but rather it should hint "you should note that current practice is 'x' but you might like to start using 'y' soon because in ten years' time it will be more common".

The Use of Graph Theory to Describe Chemical Structures
David Brown, McMaster University, Hamilton, Ontario, Canada
idbrown@mcmaster.ca

All chemical structures that involve localized bonds can be described by bond graphs. These are widely used for describing the structure of molecular compounds and the properties of such graphs are well understood. It is less well appreciated that polar compounds can also be represented by bond graphs providing that both ionic and covalent bonds are included. Such graphs are usually infinite but can be reduced to finite graphs, providing that the infinite graph has translational symmetry as is the case in crystals. Bond graphs of polar compounds are oriented graphs meaning that the bonds carry arrows pointing from the atom with the higher electronegativity (anion) to the atom with the lower electronegativity (cation). In chemistry such oriented graphs are usually bipartite, meaning that every atom can be classified as either an anion or a cation and all the bonds have an anion at one end and a cation at the other.

If we treat each atom as a point charge equal to its formal ionic charge, then all the cations are linked to their neighboring anions by lines of electrostatic field, the total number of such lines linking any two atoms being the electrostatic flux that forms the bond between them. For equilibrium structures this electrostatic bond flux, which is essentially the same as the bond order, can be easily calculated using graph theory, and since the bond flux is found to correlate with bond length, the expected bond lengths can be determined. Graph theory provides a useful formalism in an ontology designed to capture the important features of chemical structure. Brown is willing to write a paper on it for the IUPAC project.

Discussion

Mark Nicklaus asked about a correction for partial charges and quantum mechanics. David Brown said that partial charges are not needed for a description of the chemical bond and quantum mechanical features are introduced empirically through the correlation between bond length and bond flux. Quantum mechanical arguments are, however, necessary to understand complexities such as the Jahn-Teller effect. Steve Stein sees Brown's ideas as useful in the extension of IChI.

The Green Book: Progress Report
Jeremy Frey, University of Southampton
j.g.frey@soton.ac.uk

TeX is not completely portable, and when the exact representation of the symbols is crucial, then it is essential to make sure that the correct fonts are also available especially to render special characters such as Greek letters. You can lead people towards the use of standards but you

cannot push them. Frey showed a table of Green Book items, giving their names, symbols, defining equations, units (with common variants) and notes. When the Green Book is a structured document, the equations will need to be live so that links can be made *via* the symbols. (In the Gold Book, linking is done through words.) Work is being done on an extension to handle uncertainties, giving a range and explanation of each uncertainty and a link to further information. The PDF form of the new version will eventually be published on the Web by RSC but something more live than this version is needed. The TeX version is only available in Zurich at the moment and it is not clear how transportable it will be. There are version control problems. Photochemistry is causing some problems: two different groups of chemists use different terminology. Clinical chemistry will have to be overlooked for the moment.

Discussion

In answer to a question by David Lide, Frey said that a shortened laminated form will be done by the publisher (RSC). There was much discussion on fonts. Tony Davies asked what font was used in *Pure and Applied Chemistry*. Fabienne Meyers said that the Secretariat uses what the author supplies. Details are not known. David Martinsen said that we must separate the content from the presentation of it as much as possible, then define styles based on the content. Steve Stein showed a screen from http://www.nicmila.org/Gold/Output/ and drew attention to the “in” and “of” sections of the indexes (as described in “Capturing chemical information” above). Alan McNaught said that had he and his co-workers known that “in” and “of” indexes would be done, they would have put more information in the original edition.

Discussion Item One. Is This the Right Time to Encourage Data Dictionary Development?

Is this the right time to encourage data dictionary development within IUPAC (or elsewhere)? Or should we let data dictionaries grow naturally? Timing? What about workshops, white papers and projects?

David Brown: Dictionaries are already being developed. Where do people go to answer the question ”What is the standard way to write a chemical formula?"? IUPAC should be the place to find the recommended format for chemical formulas and other terms. It would be good if CIFs conformed to IUPAC recommendations on the formats to be used for chemical formulas and nomenclature in electronic databases.

Steve Stein: In IUPAC no-one takes responsibility for the very broad-based terms. An appeal has been made for recommendations about missing features of the Gold Book but there has been little feedback.

Alan McNaught: IChI will be of some help with regard to formulas.

Tony Davies/Alan McNaught: Once it is feasible to do so, we should have a link in the electronic version of the Gold Book calling for feedback.

Mark Nicklaus: Medicinal chemists use Me and Phe widely.

Peter Linstrom: You cannot support all the different variants that chemists use.

David Lide: You can define how to handle dot disconnects, such as .H_2O, and related issues.

David Brown: We should use the IChI formula to identify compounds and note where it is not unique or cannot be constructed.

David Martinsen: We need to take account of migration. Should IUCr be putting spaces between elements?

Michael Frenkel: Should IUPAC make links to the Gold Book mandatory? The output from a project (data reports, data compilations, and recommendations) in electronic form must have links to the Gold Book and must state whether or not they comply with the Gold Book.

Tony Davies: We must accept that most documents will be in Word.

Gary Mallard: IUPAC needs style guides.

Tony Davies: Volunteers with limited IT knowledge cannot be imposed upon. We must also make sure that Fabienne Meyers does not end up with a lot of work.

David Brown: One should keep in mind the requirements of computer searching.

Tony Davies: People will be using ISIS/Draw to get structures into IChI.

Discussion Item Two. Is Multiplicity of XMLs a Real Problem that has Real Solutions?

Is multiplicity of XMLs a real problem that has real solutions? There are problems for publishers and data providers. What is the role of standards organizations?

Peter Linstrom: Not all of these XMLs fall into IUPAC's purview.

Lorrin Garson: I cannot imagine a universal XML schema for chemistry. ACS will need data in structured formats but not necessarily XML. It is not clear where this is going. In answer to a question from Bill Milne, the technology is not yet mature enough for ACS to list approved XMLs. Thus, Question Two is not *yet* a problem for publishers but it will be.

Alan McNaught: RSC puts all papers into XML and then works only on XML but it is an RSC-specific XML.

David Martinsen: ACS already gets XML from Henry Rzepa and Peter Murray-Rust.

David Brown: This leads us into Question Three.

Discussion Item Three. Will XML Authoring Tools Ever be Available?

Will XML authoring tools ever be available? If so, when, and what should be done in the meantime?

Name unrecorded: The average chemist is not interested in writing XML but there *are* tools that write XML. Reference Manager, for example, does it unknown to the author.

Jeremy Frey: RSC has technology that converts Word format to XML.

Tony Davies: Note Michael Frenkel's diagram of the multiple input formats and the 5 journals, where conversion to ThermoML is carried out. The data flow works. It is the middle bit that is the key. IUPAC has the definitions of what the key terms are for this core/key format.

Michael Frenkel: We should encourage people who have XMLs to produce the tools to author their XMLs but you cannot *require* it.

Steve Stein: Lots of people are relying on Bill Gates. IUPAC project proposals should specify how the XML (or other standard) will be used in the real world. The Gold Book (a tool) has to be maintained.

Alan McNaught: What should we ask Aubrey Jenkins to do?

Miloslav Nič: For the next 1-2 years we will have to do it ourselves because there will not be a user-friendly enough process for anyone else to do it.

Jiří Jirát: One or two people will be able to use XML Spy.

Jeremy Frey: People will *not* submit SVG images.

Steve Stein: Limited targets for Aubrey Jenkins are needed.

Miloslav Nič: Sometimes it is quicker to draw a new picture. I am able to pass some work on to students.

Tony Davies: Take a step back and ask "What is the underlying format of Gold Book entries?".

Steve Stein: Within the next year there is not going to be a tool that anyone can use to submit information to the Gold Book. We are too early. An entry tool is needed for SVG, chemical structures, augmented chemical structures, and text. When will things get better? In one year's time? In two years?

Miloslav Nič: We need chemists to go through our entries.

Steve Stein: We will take care of that.

Marc Nicklaus: I stress the importance of authoring tools. Chemists do not differentiate between JPEG and GIF, let alone understand a Molfile.

Daniel Zaharevitz: Either it must be trivially easy or chemists must be given an incentive. If your compounds get screened for free, it is worth getting a student to figure out what a Molfile is.

Tony Davies: Electronic laboratory notebooks are tools.

Marc Nicklaus: IUPAC must push for tools.

Alan McNaught: Aubrey Jenkins can supply Word text and ChemDraw structures. Is that satisfactory for the moment?

Steve Stein: We will look into electronic laboratory notebooks.

Jesús Salillas Tellaeche: We [Prous Science] take what we are given and convert it into our XML format.

Bill Milne: Anything that puts a burden on authors will fail from the start. You cannot put barriers up for chemists.

Steve Stein: NIST is not a regulatory agency, nor is IUPAC.

Daniel Zaharevitz: There are ways to force scientists to make their data more useful if their work is funded by the taxpayer.

Tony Davies: IUPAC is essentially funded by the chemical industry.

Discussion Item Four. How do we Deliver and Maintain the Namespace?

How do we deliver and maintain the namespace, by schema/DTD, or instance? What about validation tools? What are the applications of most initial concern?

Alan McNaught: Put the namespace on the IUPAC Web site.

Jiří Jirát: Find a way of using a URL to find the appropriate namespace.

David Martinsen: There is a W3C location for registering these things.

Tony Davies: IUPAC should be asserting ownership of the chemical XML namespace to the W3C: this was decided in Brisbane.

Jeremy Frey: But Peter Murray-Rust will not transfer CML to IUPAC.

Discussion Item Five. How do we Raise Awareness and Interactions?

How do we raise awareness and interactions? Suggestions for organizations, publicity and a Web presence are welcome. Peter Murray-Rust arrived at this stage and Steve Stein asked him about publicity.

Peter Murray-Rust: Put it on the IUPAC Web site,

Alan McNaught: You cannot stop people using other standards but you can try to encourage them to use IUPAC ones. Should IUPAC offer application software to help people use XML, e.g., applications that link things to the Gold Book? This is an object for which we should aim.

Tony Davies: JCAMP did not have software in the past but it now intends to supply code through sourceforge (http://www.sourceforge.net).

Alan McNaught: We should not just put something up on the Web and leave it; we must make it usable.

Tony Davies: IUPAC should make a recommendation on the content and technology and report on how to use it.

Discussion Item Six. What about Subdividing the Glossary?

What about subdividing the glossary, with data tags, concepts etc.?

Alan McNaught: ICTNS has put together a standard format for glossaries. We should see what we ought to have.

The IChI Project: Background and Overview
Alan McNaught, RSC
adm@rsc.org

In response to the fact that IUPAC names were becoming ever more complex, Steve Heller came up with a proposal to specify a chemical structure as a computer file. The chemical structure of a compound is its true identifier but structures are not unique or convenient for computers. So the project seeks to convert the structure (in the form of its connection table) to a unique string of characters by algorithms, generating the IUPAC Chemical Identifier (IChI). The first version of the algorithm was released in March 2002 It works only for organic, covalent structures but there have been no problems, or few problems, with it. A further version, which also handles inorganic and organometallic compounds, will be released in 2004. A user of the software draws a structure and the program converts it to a unique string of characters. The string can, in theory, be converted back to a structure.

The IUPAC Chemical Identifier (IChI): Project Objectives
Steve Stein, NIST
steve.stein@nist.gov

The IUPAC Chemical Identifier project, 2000-025-1-800, was set up in 2000 with Alan McNaught as Task Group Chairman and Steve Stein and Steve Heller as members of the Task Group. The project is to be completed by the end of 2003. The IChI algorithm draws on the work of others reported in the literature. An IChI identifies what a compound is. It is not designed to be used for substructure searching, for conversion to a 3D structure, or for any other purpose. In essence, an IChI says “this is what is in the bottle”. The concept has received much publicity and been discussed widely for some years. Together with XML in chemistry, it was the subject of a conference held by IUPAC and Chemical Abstracts Service (CAS) in July 2002. Dmitrii Tchekhovskoi of NIST wrote the IChI algorithm.

The objective of the project as originally stated was to establish a unique label (the IChI) which would be a non-proprietary identifier for chemical substances that could be used in printed and electronic data sources, thus enabling easier linking of diverse data compilations. Various articles have appeared describing it:

David Bradley, "That IChI Feeling", *The Alchemist*, April 24th 2002
http://www.chemweb.com/alchem/articles/1015947904091.html

Michael Freemantle, "Unique Labels for Compounds", *C&EN*, Vol. 80, No 48, 2 Dec 2002
http://pubs.acs.org/isubscribe/journals/cen/80/i48/html/8048sci1.html

"Chemists Synthesize a Single Naming System", *Nature*, **2002**, *417*, 369 (May 23, 2002)
http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v417/n6887/full/417367a_fs.html

Humans communicate chemical identity orally by means of a common or “trivial” name, in text as a systematic or common name, and pictorially as a structure diagram. Computers communicate chemical identity electronically, and precisely. The chemical structure of a compound is its true identifier but structures are not unique or convenient for computers. So the project seeks to convert the structure (in the form of its connection table) to a unique string of characters by fixed algorithms, generating the IChI. Two requirements must be fulfilled. Different compounds must have different identifiers, with all the information needed to distinguish the structures. Any one compound has only one identifier, including only the necessary information for that compound.

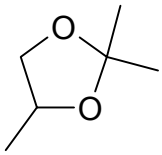
Stein defined the needs of authors, readers, and publishers (i.e., the needs of the customers). Authors need an identification system which is precise, convention-free, and covers a wide range of compounds. Readers need a system which is robust, with variable specificity and a long life. “Publishers”, in the form of software programs, communicate between authors and readers. Ready access is essential.

Initially, Stein foresaw two problems: chemicals react rapidly (i.e., they tautomerize) and may be of ambiguous or uncertain structure; and chemists use differing conventions based on their discipline, education and convenience. The three steps to creating an IChI involve chemistry, mathematics, and conventions. The technical details are described in a later presentation by Dmitrii Tchekhovskoi.

IChI in Action
Peter Murray-Rust, University of Cambridge
pm286@cam.ac.uk

Murray-Rust expressed enthusiasm for IChI and demonstrated its use in searching a Chemical Markup Language (CML) database, using JChemPaint (http://jchempaint.sourceforge.net) an

open source 2D chemical structure editor from Egon Willighagen. (Note the importance of open collaborative software projects in chemistry.) He drew a chemical structure:



He selected “Create IChI” from a pull-down menu, and displayed the ICSTI string, with atom 5’s stereochemistry unknown. He then searched a database for the above structure. The database was an Xindice one from Apache, that holds XML natively and allows fast searches of the XML components and attributes. The primary key for the database entries is IChI but no stereochemistry is included at the moment. Searching 200,000 National Cancer Institute compounds was very fast. Murray-Rust displayed CML, crystal coordinates, Cartesian coordinates, and 2D coordinates. The CML schema was published in Murray-Rust, P.; Rzepa, H.S. Chemical Markup, XML, and the World Wide Web. 4. CML Schema. *J. Chem. Inf. Comput. Sci.* 2003, 43, 757-772. The CML can be read into a 3D display program; in this example, Jmol (http://jmol.sourceforge.net), an open source 3D chemical structure viewer, was used to read the CML. The methodology is scalable to very large databases.

Chemical Databases, Identifiers, and Web Services
Marc C. Nicklaus, Laboratory of Medicinal Chemistry (LMC), National Cancer Institute (NCI), National Institutes of Health (NIH)
mn1@helix.nih.gov

Nicklaus acknowledged the work of Wolf-Dietrich Ihlenfeldt (formerly of Computer Chemistry Center (CCC), Erlangen, Germany), Frank Oellien (formerly of CCC), Bruno Bienfait (formerly of CCC and LMC) and Johannes Voigt (formerly of LMC). He displayed the opening screen from the NCI database enhanced NCI database browser “Cactus” (http://cactus.nci.nih.gov/ncidb2 and http://www2.chemie.uni-erlangen.de/ncidb2) which can be used to search the NCI database by structure or data.

He then showed the form used to enter a query. On the left are pull-down menus for query types: NSC number, CAS Registry Number, formula, molecular weight range, and structure. The query data value is entered in boxes on the left of the screen. In the middle is a “negate” option for each query type. The query fields can be connected by Boolean logic. Nicklaus showed a typical lengthy display (multiple screens) of the structure and data for a compound in the system. AIDS and cancer test results and PASS predictions (Prediction of Activity Spectra for Substances) are given.

The US government produces other databases:

Database	Date	Number of compounds	Availability
NCI Open	September 2003	260,071	Public
NIST MS Library	2002	147,194	
NIST WebBook	April 2003	31,167	Public
NLM ChemIDplus	April 2003	160,590	Public
EPA GCES Database	March 2002	66,347	Public

Nicklaus also listed just a few of the many databases available commercially

Database	Date	Number of compounds
Ambinter	August 2002	487,397
Asinex Gold	September 2003	202,237
Asinex Platinum	September 2003	117,518
Beilstein Natural Products	February 2002	124,701
ChemBridge	February 2002	100,000
ChemDiverse IDC	September 2003	122,684
ChemDiverse CombiLab	September 2003	201,139
ChemNavigator	August 2003	6,954,906

This is a total of 11.6 million compounds (8.35 million unique structures) and the number is growing. There are about 517,500 structures in the combined US Government publicly available databases (including about 2500 in DSSTox not tabulated above) and more will hopefully follow from NIAID, USPTO, FDA and other agencies. Vendor catalogs will be added. Entries will be linked to external services such as PubMed.

In the canonical version of the databases, hydrogens are added, 3D coordinates are calculated, some deficiencies (such as nitro groups) are fixed, and canonical property fields are added. Apart from that, original fields are left untouched for the most part. In future, it is possible that more complete data may be held, for example, CAS Registry Numbers, names etc. may be added. Nicklaus showed a long list of canonical fields including E_ICHI and E_SMILES. (E_ means “ensemble” in CACTVS.) However, IChI is not being used for anything at present.

Determining the overlap of two databases using SDfiles is not easy but overlap of structures can be detected more easily using hash-coded identifiers, including the CACTVS hash codes and new tautomer-invariant hash codes. Eight different variants of CACTVS hash codes are included: all eight combinations are calculated and added to the SD file.

Property name	Tautomer-invariant	Stereo-sensitive	Isotope-sensitive, charge-sensitive	Fragment handling
E_HASHY	no	no	yes	entire ensemble
E_HASHSY	no	yes	yes	entire ensemble
E_TAUTO_HASH	yes	no	yes	entire ensemble
E_STEREO_TAUTO_HASH	yes	yes	yes	entire ensemble
E_MAXFRAG_HASHY	no	no	no	largest fragment only
E_MAXFRAG_HASHSY	no	yes	no	largest fragment only
E_MAXFRAG_HASHTY	yes	no	no	largest fragment only
E_MAXFRAG_HASHSTY	yes	yes	no	largest fragment only

E_STEREO_TAUTO_HASH is the strictest criterion. It distinguishes between different compounds in the way a chemist would, i.e., it does not interpret different tautomers of the same structure as different chemicals. An index file is created for rapid comparison. A “telephone directory” for chemicals has now been created. A SMILES code is appended to each line of the

index file. Very rapid searches, overlap analyses, counts etc. can be carried out with just Unix pipes. For example: uniq –w16 index-file | wc –l determines unique structures among 11.6 million compounds (8.36 million) in 3.5 min on an Athlon XP1900+.

Nicklaus showed the results of an index file test. O, OH_3^+, OH^-, [O], etc. all had different hash codes. Fifteen tautomers of guanine all had the same string. Ring opening and closure tautomers are given *different* hash codes intentionally. The hash code is not dependent on the atom numbering in the SDfile. The hash code and the unique SMILES are the same regardless of which atom is drawn first but the non-canonical SMILES strings differ. But unique SMILES cannot be used to equate different tautomers of the same molecule. In an example of the stereo-sensitive hash code, achiral and racemic versions of one compound had the same hash code but two enantiomers had different hash codes. The hash code is a 64-bit string. No collisions have been found yet; if any *are* found, another 64 bits will be added to the code. Nicklaus' team will make the script they used publicly available. Currently, it is not particularly user-friendly.

Chemical Markup Language (CML)
Peter Murray-Rust, University of Cambridge
pm286@cam.ac.uk

CML is not an IUPAC initiative but it is open source. The CML core supports any micromolecular property; it has built-in molecular, atom, bond, formula, name and atom property features. There is dictionary support for fractional charges and bond properties. CML2 (http://cml.sourceforge.net) is frozen as published in Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the World Wide Web. 4. CML Schema. *J. Chem. Inf. Comput. Sci.* 2003, 43, 757-772. CML2 is more compact than the first version of CML and supports more things. In stereochemistry, tautomers and isotopes it mirrors IChI. All specifications are supported by software: Java classes are generated automatically from the schema.

CML-DOM (CML Document Object Model, see Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the World Wide Web. 2. Information Objects and the CMLDOM. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1113-1123) is instantly compilable from schemas (see http://cml.sourceforge.net). Murray-Rust showed a pdf file with an atom element definition and attributes of the element listed. Every class and element has been generated by machine and can be incorporated in the program. A user could create an AnIML DOM and read in his data. About 4000-5000 Java classes are to be distributed. C++, Python and Fortran are also needed.

Schemas can validate documents and data, can provide documentation and can add functionality, for example, Java classes or an XSLT style sheet. (XSLT is a language for transforming XML documents into other XML documents, http://www.w3.org/TR/xslt. XSLT is designed for use as part of XSL, which is a stylesheet language for XML. In addition to XSLT, XSL includes an XML vocabulary for specifying formatting. XSL specifies the styling of an XML document by using XSLT to describe how the document is transformed into another XML document that uses the formatting vocabulary.)

CML code (in Java, C++ etc.) is automatically compiled from schemas. There are four modules (namespaces): CMLCore, CMLReact, CMLSpect and CMLComp. CMLSpect is meant to interoperate with AnIML. About 100 elements are important and are in XSD (XML Schema Definition) files. A lot of elements are needed for chemistry. Murray-Rust showed attributes for atomArray. The user selects the elements he needs for his application.

Parsers, stylesheets etc. must be supported by running code: Murray-Rust and his co-workers established this principle a long time ago in the XML project. All information is linked to dictionaries: "cml:mpt" is resolved into the CML namespace dictionary and "unit:celsius" is resolved into unit namespace dictionary. There are also dictionaries for MOPAC (a general-purpose semi-empirical quantum mechanics package), SELF (an electronic data format for physicochemical data, http://www.xml-cml.org/self/), chemistry-for-biology, and fragments for

biology. The dictionaries are understandable both by machines and by humans. Murray-Rust showed the MOPAC ISCF(C) (a self-consistent field used in solving the Schrödinger equation). All this code is openly available but is it is in an early version.

Chemical Data Handling in the NIST Chemistry WebBook: a Brief Overview
Peter J. Linstrom, NIST
peter.linstrom@nist.gov

The NIST Chemistry WebBook is a site on the Internet that distributes data from NIST data projects and outside contributors. It is a collection of databases that were not designed with semantic compatibility in mind. It is used by 20,000 people every week: chemists, engineers, educators and students. It contains a wide variety of thermodynamic, ion energetics, spectral, energy level, and Henry's law data; interactive tools including fluid property models and a group additivity model for gas phase thermodynamics; and chemical names, synonyms and structures.

Standards are used where they are available: JCAMP-DX for spectral data, a variant of the REFER format for references, and Molfile for structures. Most data are kept in an extensible format developed internally at NIST. This format is configured by data and unit definition files and is documented with 81 pages of information. (This format was established before XML was widely used.)

The data type definition file defines name, format (symbol, units, data type) class (grouping), method types, and special options. Linstrom showed an example definition: entropy of vaporization at standard temperature and pressure. The unit definition file defines unit types and conversion rules. Converted values have the correct number of digits. Conversion rules can be derived from a combination of rules. The data file indicates source, units, data, metadata and comments. It is organized by compound, reaction, or mixture.

In future there will be updates to existing data sets, and a retention index database and links to data not available in the WebBook will be added. The NIST WebBook will strongly support the IUPAC Chemical Identifier. Identifiers will be available for all species with structures. The WebBook will be searchable *via* the Identifier and will have a tool which generates Identifiers for arbitrary structures submitted *via* a drawing applet or an uploaded Molfile. This tool will allow a user to view XML or download a file.

Discussion

Ture Damhus: What does "extensible" mean?

Peter Linstrom: It is easy to add new data types to the WebBook but it not possible to add certain things from a non-similar system.

Peter Murray-Rust: People are working on extensible ontology languages.

An Identifier for Crystal Phases
I. David Brown, McMaster University, Hamilton, Ontario, Canada
idbrown@mcmaster.ca

The IChI project approached the International Union of Crystallography (IUCr) to see if IUCr had a way of identifying crystallographic phases. The answer was "no" so Brown was asked to establish a committee to provide an identifier. In the current presentation he summarized the results of the committee's discussions to date.

The committee distinguished between *simple* and *compound* identifiers. Simple identifiers are arbitrary numbers assigned by a central organization (e.g. CAS Registry Numbers) whereas a compound identifier is a concatenation of the properties that characterize the material. Compound

identifiers can be constructed by anyone in the field. The IUCr committee has opted for a compound identifier, containing the information that must be known even to assign a simple identifier.

The properties that are sufficient to characterize a crystal phase are the chemical and crystallographic properties, simple identifiers, and conditions, e.g., temperature, pressure, and crystallization process. Chemical identifiers must identify the chemical content, namely the sum formula of the crystal (for inorganic compounds only the relative composition is meaningful); state of matter (gas, liquid, crystal or other solid form, although the project is only interested in crystals); and number of carbon atoms with 0, 1, 2 and 3 attached hydrogen atoms (to distinguish organic isomers other than stereoisomers). Bear in mind that the composition applies to the whole crystal, not just a molecule. It includes solvent of crystallization and other species.

The crystallographic identifier specifies the space group number (independent of the setting chosen), Bravais symbol (redundant if the space group is known, but useful if it is not), Wyckoff sequence, which represents high symmetry positions (useful for inorganic materials but not so useful for organic materials), and reduced cell. The reduced cell is subject to experimental uncertainty and a certain tolerance is needed if a user searches on this.

Simple identifiers are not unique and mostly cover only a subset of compounds. For example the CAS Registry Number is often ambiguous when applied to crystals. Other simple identifiers are the CSD (Cambridge Structural Database) refcode, ICSD (Inorganic Crystal Structure Database) collection number, NIST code, PDB (Protein Data Bank) code, Pauling file code, Pauling file type code, Strukturbericht code, and mineral name (used only for naturally occurring minerals). Recording conditions of temperature and pressure is useful for phases that exist only under non-ambient conditions.

Brown concluded with some general considerations. Some properties may be misassigned and others may not be known. The properties described above, even if known, may not define the phase uniquely. By a suitable choice of search algorithm the identifier of a target phase can be used to restrict the number of phases retrieved. After that, a decision by a scientist is needed. Brown's project was on hold until after the results of the present meeting were available.

Discussion

Steve Heller suggested that the crystal phase might be included in the next stage of IChI. In answer to a question from Peter Murray-Rust, David Brown said that the Wyckoff letter cannot be derived algorithmically.

Proposed IUPAC Project: Graphical Representation Standards for Chemical Structure Diagrams

Bill Town, Kilmorie Consulting
bill.town@kilmorie.com

The objectives of this project are to extend IUPAC's leadership in the development of standard nomenclature and terminology in chemistry into the domain of chemical structure diagrams; to provide a single, comprehensive set of guidelines for creating chemical structure diagrams in printed and in electronic media; and to enable IUPAC to serve as the principal, authoritative source for chemical structure representations.

Existing IUPAC nomenclature recommendations discuss some aspects of chemical structure diagrams tangentially. Existing recommendations on chemical structure diagrams are incomplete, and do not discuss many basic issues, so many organizations have formulated their own guidelines for creating chemical structure diagrams but none of those guidelines is comprehensive. This project would provide a single, comprehensive set of guidelines for creating chemical structure diagrams, which would be a significant benefit to the chemistry community.

The proposed project began informally with a “scoping exercise” led by Jonathan Brecher in which the chemistry community was invited to discuss those aspects of creating chemical structure diagrams which are amenable to standardization through IUPAC recommendations. Some 20 people responded. Some draft recommendations were created during the course of that scoping exercise, and these will serve as a useful starting point. The participants in the exercise also identified several areas that are likely to be contentious, or were incompletely specified; those areas will receive specific attention during the course of the project. Discussions from the scoping exercise are archived at http://groups.yahoo.com/group/iupacstructures/; possible recommendations, for discussion, are at http://www.angelfire.com/sc3/iupacstructures/.

Polymer representations are being are excluded at the moment but, if time permits, polymers will be addressed and the names of one or more of polymer nomenclature experts will be added to the task group. With the ever-increasing importance of electronic publication, this project will consider issues related to the production of chemical structure diagrams both in printed and in electronic media. Where possible, every effort will be made to ensure identical recommendations in all media. The recommendations for the production of chemical structure diagrams will aid in the correct recognition of structural information by the IUPAC Chemical Identifier (IChI) algorithm.

In some cases, the current state of chemistry software may preclude the use of the most preferred styles for chemical structure diagrams. When necessary, this project will provide practical advice for the production of electronic chemical structure diagrams according to the current state of the art of chemistry software. It is likely that this project will also produce a set of recommendations for enhancements that chemistry software developers may undertake to improve that state of the art.

Project team members are:

- Bill Town (Kilmorie Consulting, UK, task group chairman)
- Jonathan Brecher (CambridgeSoft, USA)
- Harry Gottlieb (GSK,USA)
- Richard M. Hartshorn (Canterbury, NZ)
- Gerry Moss (QMUL,UK)
- Peter Murray-Rust (Cambridge, UK)
- József Nyitrai (Budapest, Hungary)
- Warren Powell (retired from CAS, USA)
- Ann Smith (Merck Index, USA)
- Stephen Stein (NIST, USA)
- Keith Taylor (MDL, USA)
- Tony Williams (ACD/Labs, USA)
- Andrey Yerin (ACD Labs, Russia)

Corresponding members are:

- Helen Cooke (GSK, USA)
- Ture Damhus (Novozymes, Denmark)
- Pat Giles (CAS, USA)
- Alan McNaught (RSC, UK)
- Bert Ramsay (Eastern Michigan University, USA)

There is an open invitation for interested persons to join the group as corresponding members. A proposal submitted in October 2003 is awaiting approval by IUPAC. (Postscript: project approved in 2004, with additional team members Sean Conway, Kirill Degtyarenko, and Matt Griffiths.)

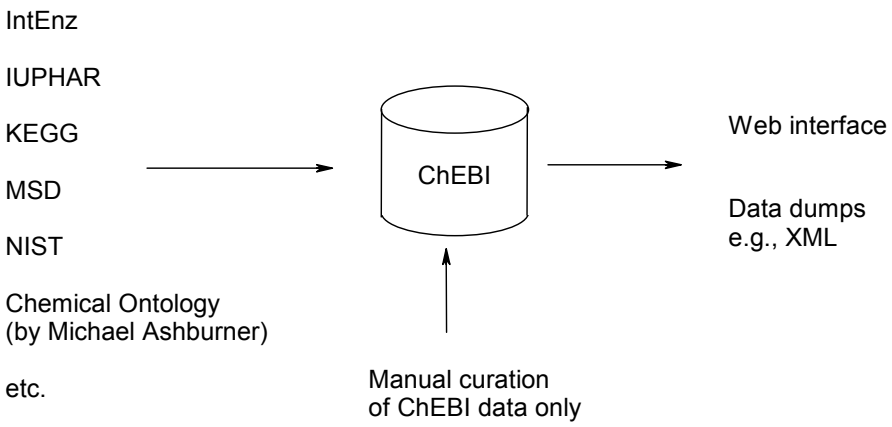
ChEBI. A Reference Database of Biochemical Compounds

Paula de Matos, European Molecular Biology Laboratory-European Bioinformatics Institute, EMBL-EBI
pmatos@ebi.ac.uk

EBI is a center for research and services in bioinformatics. It develops and provides a number of databases (e.g., SwissProt and the EMBL nucleotide database) and other services. It is also hosting an intensive collaboration with GO (http://www.geneontology.org) and Public Library of Science (http://www.publiclibraryofscience.org/). ChEBI (Chemical compounds of Biological Interest, http://www.ebi.ac.uk/chebi) is a database of biochemical compounds aiming to provide a standard of biochemical compounds that will eventually serve other biological databases. It will integrate existing sources and be an instant reference for non-chemists.

It is being developed because EBI has identified a need for a *definitive* set of vocabulary which can be used to serve both EBI's own and external biological databases. The compound dictionaries currently available are either commercial or do not deal with biochemical compounds in a definitive way. ChEBI will provide a free, platform-independent and immediately usable data source.

The principles behind it are akin to those of Michael Ashburner (http://www.geneontology.org). Nothing held in the database must be proprietary or derived from a proprietary source that would limit its free distribution and availability to anyone; and every data item in the database should be fully traceable and explicitly referenced to the original source and version. Although the EBI will provide a Web interface, the entirety of the data should be available to all without constraint as, for example, SQL table dumps, ASCII tables, and XML.



IntEnz is the name for the Integrated relational Enzyme database (http://www.ebi.ac.uk/intenz/index.html) and is the most up-to-date version of the Enzyme Nomenclature. IUPHAR is The International Union of Basic and Clinical Pharmacology (http://www.iuphar.org/). KEGG is the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg/kegg2.html). EBI's Macromolecular Structure Database (MSD, http://www.ebi.ac.uk/msd/) is the European project for the collection, management and distribution of data about macromolecular structures, derived in part from the Protein Data Bank.

The data held in ChEBI include a unique public identifier, names (the ChEBI "approved" name, systematic names such as the preferred IUPAC name, and synonyms), empirical formula (if available), structure (if available) in 2D (e.g., as SMILES, CML, or Molfile) or 3D (e.g. CIF and Molfile), and other identifiers such as IChI, and CAS Registry Numbers. Also stored are cross-references to other public resources (BioCyc, KEGG Ligand, NIST Chemistry WebBook and UM-BBD) and housekeeping data.

BioCyc (http://www.biocyc.org/) is a collection of Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism, with the exception of the MetaCyc database, which is a reference source on metabolic pathways from many organisms. UM-BDD is the University of Minnesota Biocatalysis/Biodegradation Database (http://umbbd.ahc.umn.edu/contact.html), a classification relating entities, namely chemical ontology and the Research Collaboratory for Structural Bioinformatics (RCSB, http://www.rcsb.org/index.html) Molecule Classification, used in ChemPDB (a consistent and enriched library of ligands, small molecules and monomers that are referred to as residues and heterogroups in any PDB entry, http://www.ebi.ac.uk/msd-srv/chempdb/cgi-bin/cgi.pl).

KEGG Ligand and IntEnz (approximately 15 000 entries) have already been loaded into a relational database, and curated manually with nomenclature and merging of redundant entries. ChEBI's first ontology, Chemical Ontology by Michael Ashburner, has been applied. Software tools have been developed for public and curator Web interfaces and the addition of structures.

EBI wants to use IChI because it can be a unique identifier for the EBI compounds and a standard format for the future, and because IChI is an IUPAC backed project. There are also advantages for IChI if EBI uses it. A unique identifier is more useful if it references public resources. ChEBI, an EBI backed project, could be a platform to launch IChI as the next generation compound identifier for the biological community

All the requirements for initial implementation are fulfilled: batch implementation and software to generate IChI strings. However it would be good of there were the possibility of having other formats (SMILES and CML) as input and if IChI strings could be reverse-engineered to structural images in Molfiles or SMILES. It is hoped that the ChEBI data will be made available *via* SQL table dumps, ASCII and XML, so conversion from the IUPAC XML format would be desirable. The vision is not just that ChEBI will be free, community-backed, and cross-referenced to SwissProt, GO, and IntEnz.ENZYME, but that ChEBI will set the standard.

Discussion

Marc Nicklaus pointed out that the aspartate zwitterion problem is solved by a hash code that can be either stricter or more lenient.

IChI: Day Two Introduction
Stephen Stein, NIST
steve.stein@nist.gov

At the opening of the proceedings, Alan McNaught suggested that the term "IChI" should be changed to "INChI" in recognition of NIST's funding of most of the work to date.

Steve Stein then outlined the concepts behind the identifier. The three steps to creating an IChI involve chemistry, mathematics, and conventions. First the input structure is normalized by implementing chemical rules. The normalized structure is then canonicalized by labeling the atoms; equivalent atoms get the same label. Conventions are then applied: the labeled structure is serialized and output as a character string, or "name".

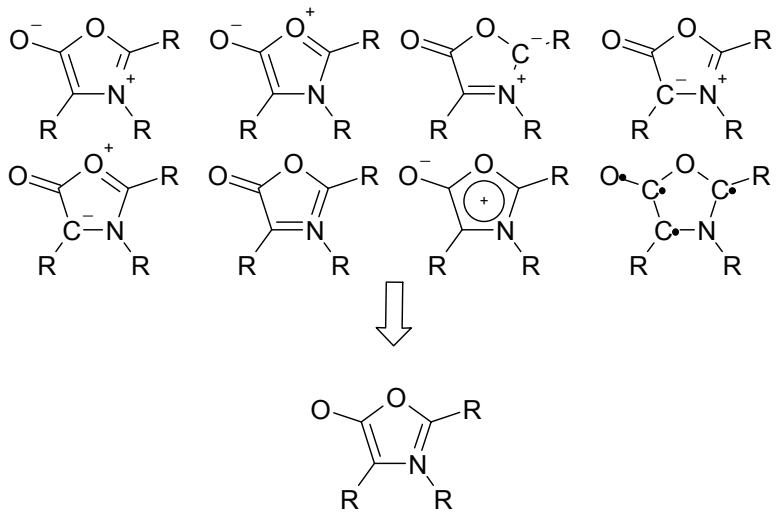
Certain simplifications are applied to normalization. The structure is divided into layers, each layer refining the structure. Electron density is ignored; simple connectivity alone is used. Double, triple and coordination bonds, and odd electrons and charges are ignored. Stereochemical sublayers include sp^2, double bond stereochemistry, and sp^3, tetrahedral stereochemistry. Free rotation around single bonds is assumed and, by default, sp^2 stereochemistry is not assigned in rings containing fewer than 8 atoms.

Structure information is divided into layers. The four basic layers are formula, connectivity, stereochemistry, and isotopic "corrections". Layers are used because they are logical (they

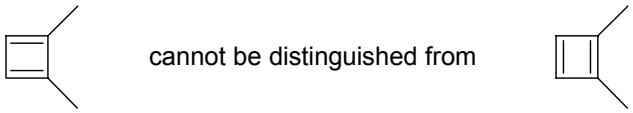
separate the variables) and understandable, with no significant cost; they are flexible for chemists (they represent known levels of information) and more layers (conformation, coordination, etc.) can be added as needed. There is a common connectivity core for all chemicals. For the connectivity sublayers, metals and hydrogen atoms are disconnected; metals are re-connected and hydrogen atoms, non-mobile in non-tautomers, or mobile, distinguishing tautomers, are reconnected. Optionally metals are reconnected, with bonds to metals represented.

Electron density can be ignored because it is not required for compound identification: it represents “excited states”. Pi electrons are responsible for a large portion of interesting chemistry but they are not important for naming. Representations are simplified with regard to delocalization, aromaticity, zwitterions, coordination etc. Stereochemical sublayers include sp^2, double bond stereochemistry, and sp^3, tetrahedral stereochemistry. Other types of stereochemistry will be added later. Relative, absolute and racemic stereoisomers are distinguished. No Z/E stereochemistry is perceived for small rings.

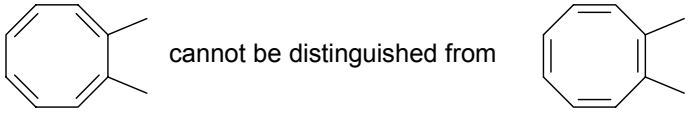
Stein gave the Műnchnones as an example of ignoring electrons:



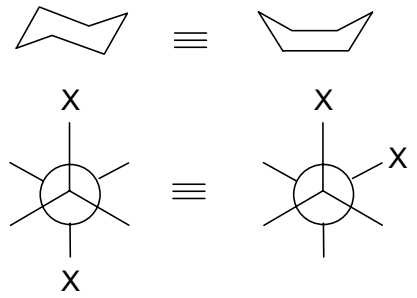
Unfortunately



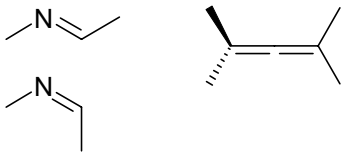
and



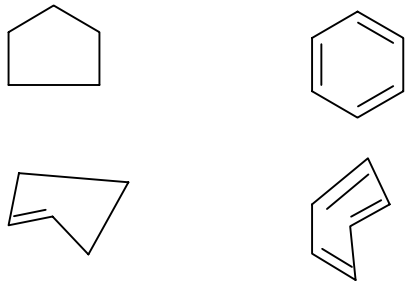
Since free rotation is assumed around single bonds, conformations are ignored:



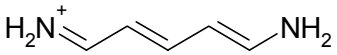
Stereochemistry of sp^2 atoms is simplified by using rules and pi-electrons, as in



and by ignoring Z/E stereochemistry for rings of 7 or fewer atoms:

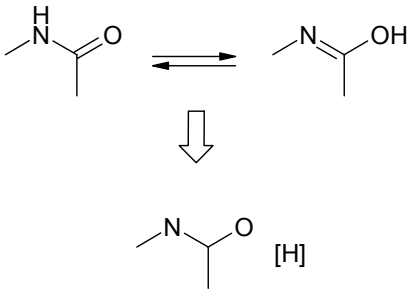


The bond perception will observe that



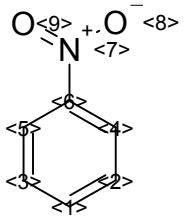
is an incorrect depiction: there is no Z/E stereochemistry in this molecule.

The basic tautomer layer recognizes hydrogen atom migration between 1,3 heteroatoms:



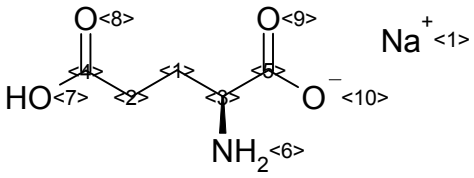
This is similar to the convention used at Chemical Abstracts Service and keto-enol tautomerism is not handled.

Stein gave some examples of canonically numbered structures and their IChI layers.



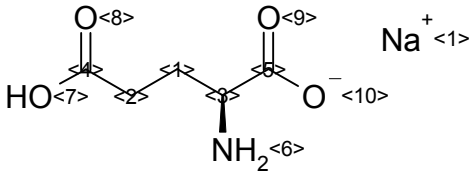
Description	**Layers**
formula	C6H5NO2
connectivity	8-7(9)6-4-2-1-3-5-6
hydrogen atoms	1-5H

charges



tautomeric

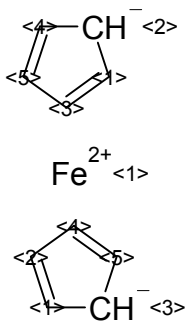
Description	**Layers**
formula	C5H8NO4.Na
connectivity	6-3 (5 (9) 10) 1-2-4(7)8;
hydrogen atoms	1-2H2, 3H, 6H2 (H-,7,8,9,10);
stereo sp^3	3-;
charges	-1;+1



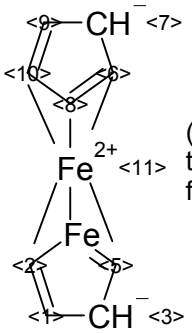
fixed

Description	**Layers**
formula	C5H8NO4.Na
connectivity	6-3 (5 (9) 10) 1-2-4(7)8;
hydrogen atoms	1-2H2, 3H, 6H2 (H-,7,8,9,10);
stereo sp^3	3-;
hydrogen atoms fixed	7H
stereo sp^3	3-;
charges	-1;+1

Disconnected structure



Reconnected structure



(bonds from Fe to *every* C omitted for simplicity)

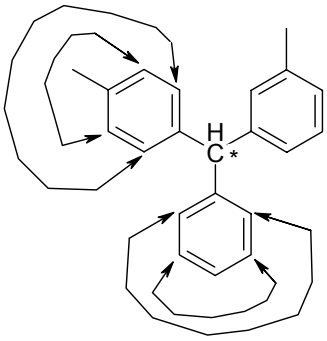
Disconnected structure:

Description	**Layers**
formula	2C5.Fe
connectivity	2*1-2-4-5-3-1;
hydrogen atoms	2*1-5H;
charges	2*-1;+2

Reconnected structure:

Description	**Layers**
formula	C10Fe
connectivity	1-2-4-5-3(1)11(1,2,4,5)6-7(11)9(11)10(11)8(6)11
hydrogen atoms	1-10H

Stereogenic centers and equivalent atoms, e.g.,



are recognized at the canonicalization stage, as a by-product of IChI creation. This assists chemists for structure confirmation. Warnings and error messages are output, for example, for unusual valences or unrecognized input. “Reversibility” would require a knowledge of coordinates and bond/charge locations.

Chemical Structures in European Patents
John Brennan, European Patent Office (EPO)
jbrennan@epo.org

Brennan’s group (pure and applied organic chemistry) handles 15,000 patent applications a year, 5000 of them relating to novel compounds. These patents specifically indicate 100,000-1,000,000

novel chemical structures a year. The EPO has to deal with patent attorneys who often tend to be conservative in their approach to new media.

Electronic filing is obligatory for biosequences. Electronic filing and paying of fees is possible, but is currently done in only a minority of cases. Patents can be inspected online and Esp@cenet offers access to full text and a simple online search facility. The EPO would like to be able to display structures in Esp@cenet. A project concerning electronic structure filing is in progress. A chemical filing tool for patent applicants is being developed by the Documentation/Tools department of the EPO, using XML for text and CML for structures. A legal framework is being proposed and interested parties are being consulted. The EPO is not a regulatory authority; it has to use persuasion.

The electronic structure filing tool accepts structures from a structure data file, a structure editor, or a text file, verifies the structures, and prepares them for filing. The use of XML and CML is in agreement with joint EPO-JPO-USPTO (the European, Japanese and US patent offices) and WIPO (World Intellectual Property Organization) policy to structure patent data in XML. The structure filing tool will convert structural data to the patent office's format.

The new system would have benefits for applicants, the interested public, database producers and the EPO. Initially, the system will be used only for specifically claimed compounds. Later it might be used for claimed reactions, claimed compositions (e.g., sustained release formulations), claimed use of known compounds (second medical use) and maybe even for Markush structures, although the EPO is not interested in electronic submission of Markush structures. The pilot project will progress in 2004 according to the response of applicants.

Discussion

Peter Murray-Rust: The published version of CML could incorporate IChI. There is a "friendly slot" for it.

John Brennan: In the EPO internal search service, a searchable database could be made.

Nomenclature (and IChI?) Issues with Inorganic Compounds
Ture Damhus, Novozymes A/S
tda@novozymes.com

Damhus is a member of the Red Book working group, and ICTNS, and Division VIII from 2004 on. He hopes that he will soon be initiating the inorganic PIN project (preferred names for inorganic compounds).

We need to decide what it is that we want to name (or identify with an IChI) in inorganic chemistry. Consider compounds in a broad sense *versus* molecular graphs and empirical formulas and molecular formulas *versus* structural formulas. A hierarchical approach to naming is necessary, and we should be prepared for any valency, any bonding scheme and any geometry! Damhus presented a large number of challenging examples.

First he cited Cl_4^+, $Cl_2O_2^{\bullet+}$, $Cl_2O_2^{2+}$. These may be easy for IChI but they are hard to name. Steve Stein confirmed that an IChI can be calculated for Cl_4^+ since you only need a formula not a structural formula. Cl_2O_2 could be done with four bonds but you could not show that it was a rectangle. Diamond and graphite would be done as carbon and they would be distinguished by crystal phase, and S_8 and Cl_2 can be done now, but polymers are not handled by IChI at the moment. Tony Davies cautioned against "mission creep".

Damhus gave some more examples:

PCl_3
$AlCl_3$, $[AlCl_3]$; $[Al_2Cl_6]$ (here, IChI may have thing to teach the inorganic group)
CuCl
$[LiAl_4]^-$ or $[Al_4Li]^-$ (pyramid with 4 aluminium atoms in the base; it is not obvious which is the central atom)
$[CAl_3Si]^-$ (with planar tetracoordinate carbon)
TlI_3 ($Tl(I)_3$ or $Tl[I_3]$?)
SbF_5; $[SbF_4]^+[SbF_6]^-$
$(NH_4)_2[Ce(SO_4)_3]$ or $Ce(SO_4)_2 \bullet (NH_4)_2SO_4$
Oxides e.g., K_2O, K_2O_2, KO_2, KO_3 (compare BaO_2, MnO_2, Fe_2O_3 etc.)

Steve Heller said that these examples point up the problems of nomenclature. David Brown said that in organic chemistry you can talk about a molecule but once you get rid of a crystal of inorganic material you do not have the same substance. The key to nomenclature of inorganic compounds is that they only exist in the solid. CuCl, for example, only exists in the solid. Talapady Bhat suggested adding a polymer layer to IChI but Steve Stein and Michael Frenkel said that this would not be easy.

Damhus' next examples were N_4S_4 and As_4S_4. He showed the structures of N_4S_4 and As_4S_4 from the sixth edition of Cotton and Wilkinson's textbook on inorganic chemistry. The S-S distance is longer than a normal S-S single bond distance but short enough to indicate significant direct S…S interaction. The bonds are different in the two compounds. The formula layer of IChI should handle this.

In response to the next example,

$[Cr(en)_3][Ni(CN)_5] \bullet 1.5\ H_2O$ (two geometries of anion),

Steve Stein said that geometries other than sp^2 and sp^3 will be added to IChI later.

Other naming challenges are presented by the following.

Electrides (with magnetic properties)
    $[K(crypt\text{-}222)]^+e^-$
    $[Cs(18\text{-}crown\text{-}6)_2]^+e^-$
Agostic bonds, H bridges etc.
    $[Co_6(CO)_{15}H]^-$ (H in an octahedron of Co atoms)
Clusters
Multiple-layer cages
    $[As@Ni_{12}@As_{20}]^{3-}$ (two interpenetrating polyhedra)

Damhus showed the structure and bonding of the cluster $[Ge_9{=}Ge_9{=}Ge_9{=}Ge_9]^{8-}$, a nanorod that is a linear tetramer of nine-atom germanium clusters, reported in Ugrinov, A.; Sevov, S. C. *Inorg. Chem.* 2003, 42, 5789-5791. He also showed a structure from Wang, Q.-M.; Mak, T. C. W. *Dalton Transactions* **2003**, 25-27, which reports a new triple salt $(PhCH_2NMe_3)_4[Ag_{17}(C_2)_2(CF_3CO_2)_{16}(NO_3)(H_2O)_4]$ that features an asymmetric silver(I) double cage constructed from the edge-sharing of a triangulated dodecahedron and a monocapped square antiprism, and an anionic columnar structure generated with an unprecedented μ_5-ligation mode of the nitrate ion. In this compound, carbon is a kind of bridging ligand, i.e., there are carbide ions in a cage of silver atoms. The IChI numbering might help here, in that it would show which atom is bonded to the metal atom and obviate the use of kappa etc. If you can define the binding, you can do more than just produce an IChI of the formula.

Finally Damhus showed a large number of structures from pages 1580-1582 of the IUPAC 1999 recommendations on nomenclature of organometallic compounds of the transition elements (Salzer, A. *Pure and Applied Chemistry*, **1999**, *71*, 1557-1585,

http://www.iupac.org/reports/1999/7108salzer/index.html). These were organometallic compounds of the transition elements with ligands containing multiple double bonds, where the hapto symbol is used to indicate the number of contiguous ligating atoms of the ligand to the metal. He added that he had listed many problems and he had not even touched on non-stoichiometric compounds. Peter Murray-Rust said that frameworks and symmetry matter in many of these examples and they do not fit easily into the current IChI structure.

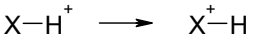
IChI Algorithm: Technical Issues

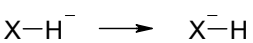
Dmitrii Tchekhovskoi, NIST (co-authors Steve Stein and Steve Heller)
dmitrii.tchekhovskoi@nist.gov

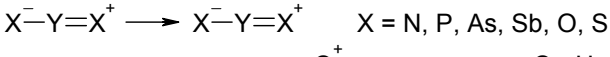
The steps involved in converting a chemical structure to an IChI are

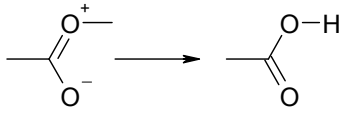
1. Fix the structure drawing
2. Disconnect salts
3. Move protons to neutralize
4. Disconnect metals
5. Eliminate radicals if possible
6. Process charges and tautomers
7. Mark possibly stereogenic atoms and bonds
8. Obtain canonical numbering(s)
9. Output the IChI.

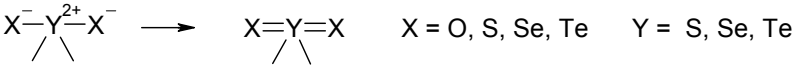
Tchekhovskoi gave some examples of fixing the structure drawing.

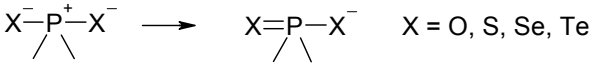




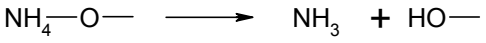


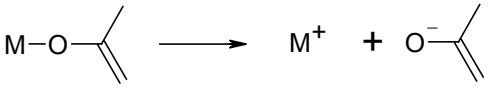
(For example  where X = O and Y = C)





Rules for salt disconnection are as follows.



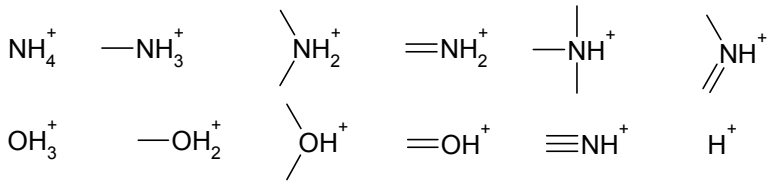


M is a metal

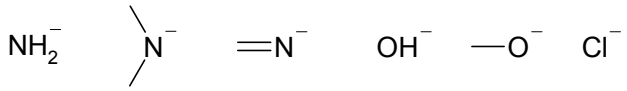
where a metal is defined as any element except:

IIIA	IVA	VA	VIA	VIIA	VIIIA
13	14	15	16	17	18
				H	He
B	C	N	O	F	Ne
	Si	P	S	Cl	Ar
	Ge	As	Se	Br	Kr
			Te	I	Xe
				At	Rn

In the process of moving protons to neutralize the molecule, proton donors are

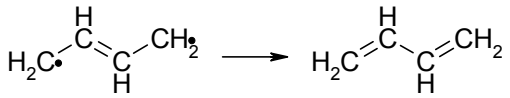


and proton acceptors are

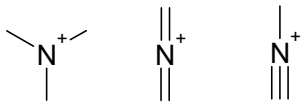


S is treated in the same way as O, and F, Br and I are treated as Cl. Ion pairs (e.g., $=N^+=N^-$ or $=N^+-O^-$) are ignored.

Metals are disconnected according to the definition of metal given above. Charges are adjusted in the case of disconnected F, Cl, Br, I, At, O, S, Se, Te, N, P, As, and B if possible. Radicals are eliminated if possible:



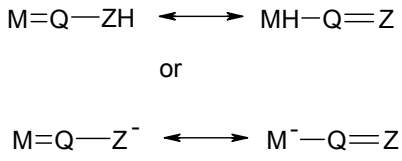
Positive charges which can be moved are perceived. Positive charge donors are



Positive charge acceptors are

P is treated in the same way as N.

According to Mockus, J.; Stobaugh, R. E. The Chemical Abstracts Service Registry System. VII. Tautomerism and Alternating Bonds. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18-22, tautomerism can occur in the following cases:

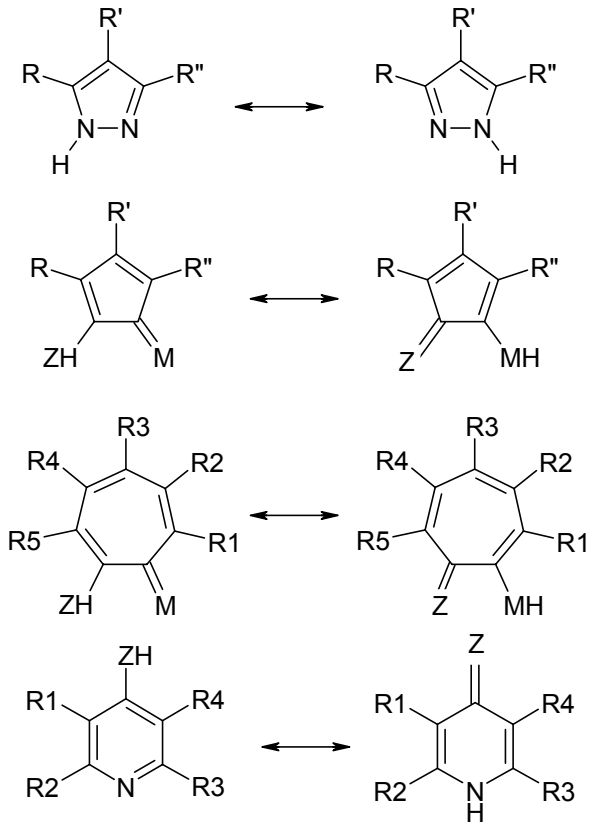


M, Z = N^{III}, O^{II}, S^{II}, Se^{II}, Te^{II} (Roman superscripts designate chemical valence)

Q = C, N, S, P, Sb, As, Se, Te, Br, Cl, I

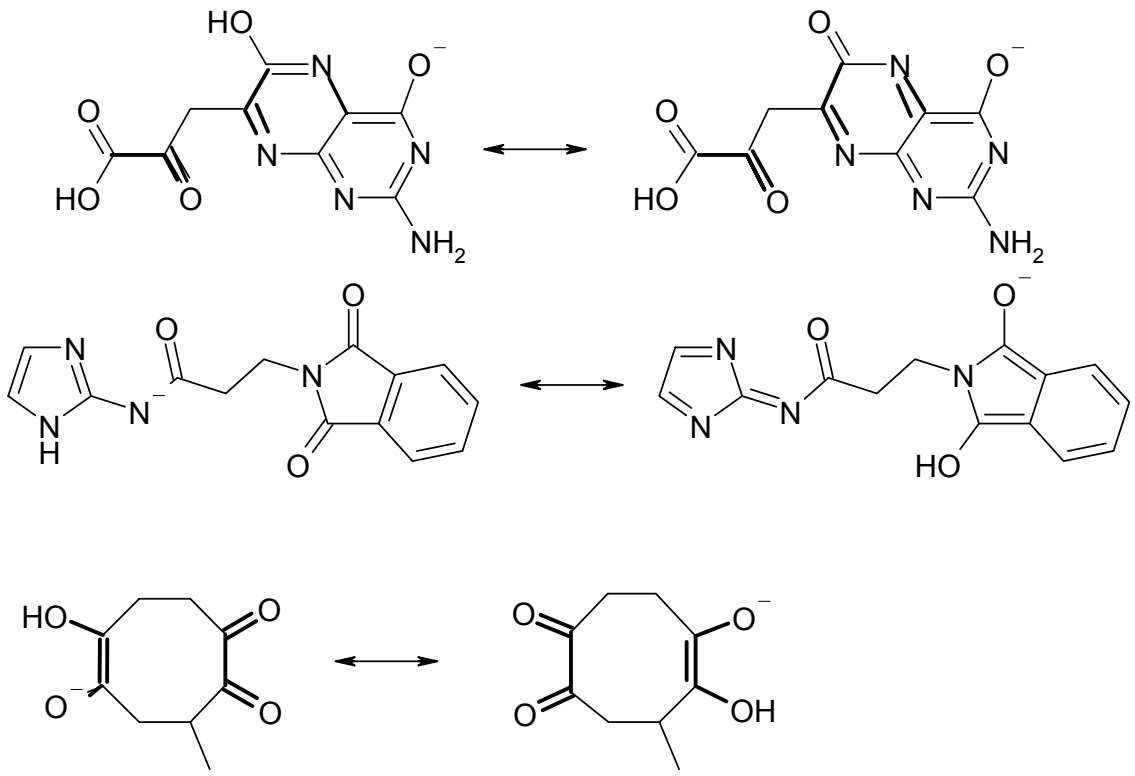
H = protium, deuterium or tritium

(The “double” bond may be a double bond, a bond in a ring with alternating single and double bonds, or a “tautomeric” bond.) This definition has been adopted for IChI but has been extended to



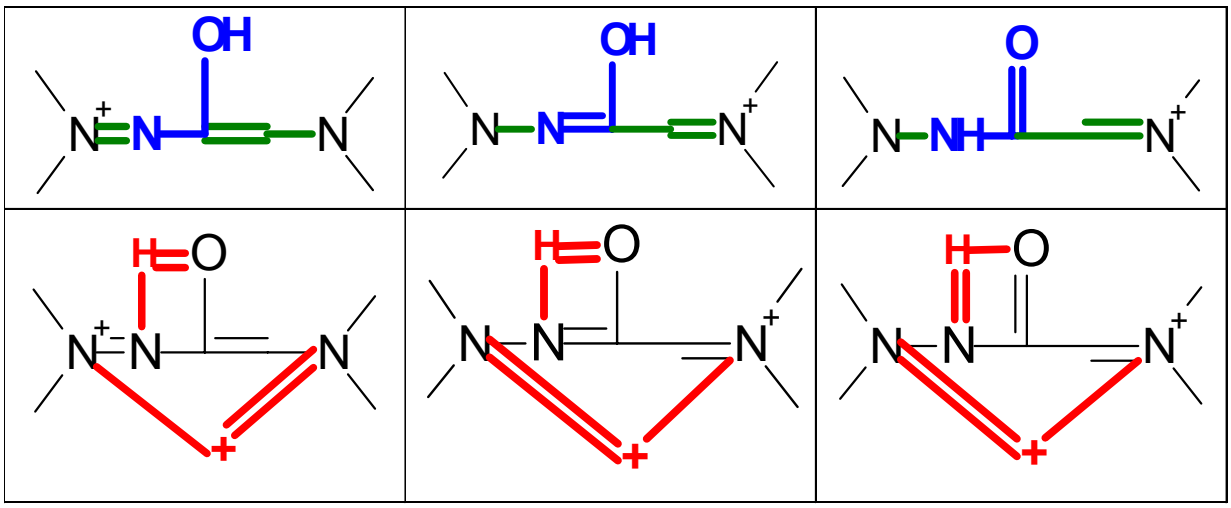
and the related structures with the movable hydrogen atom replaced by a negative charge.

Salt-specific tautomerism is turned on if an “acid” anion is present. Take for example these pairs of equivalent structures:



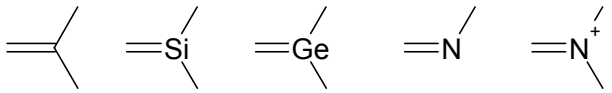
(In the first and last structures above, tautomeric bonds are not emphasized but other changeable bonds are emboldened.)

Charges and tautomerism are represented internally by a mathematical graph theory “trick”. This is an efficient way of handling alternating pathways.



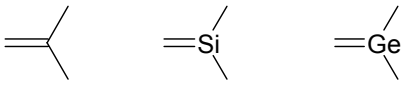
The alternating circuits (made out of single-double bonds) allow testing for changeable bonds. Only those changeable bonds that belong to rings of changeable non-tautomeric bonds are considered as possibly stereogenic. The method of testing for whether a bond is changeable is based on the publication Kocay, W.; Stone, D. An Algorithm for Balanced Flows. *Journal of Combinatorial Mathematics and Combinatorial Computing* **1995**, *19*, 3-31.

Stereochemistry handing is partially based on Blackwood, J. E.; Blower, P. E.; Layten, S. W.; *et al.* Chemical Abstracts Service Chemical Registry System. 13. Enhanced Handling of Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 204-212. The following bonds are treated as possibly stereogenic:

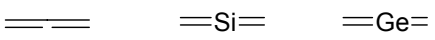


Only one of two atoms connected by a possibly stereogenic double bond is shown. Cumulenes are treated as possibly stereogenic: Stereogenic cumulenes may have up to 3 double bonds.

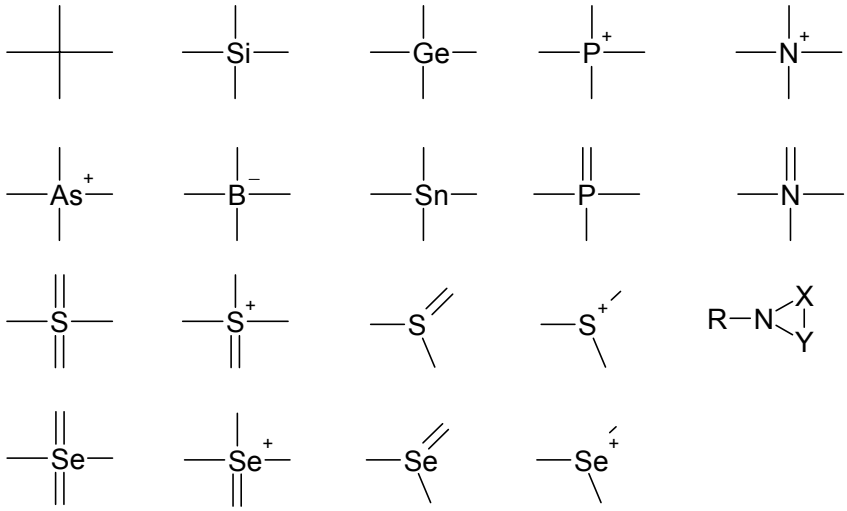
Terminal atoms



Middle atoms



Atoms treated as stereogenic are as follows.



An atom or positive ion of N, P, As, S, or Se is not treated as stereogenic if it has (a) a terminal H atom neighbor or (b) at least two terminal neighbors, $-XH_m$ and $-XH_n$, (n+m>0), where X = O, S, Se, Te, or N.

IChI uses canonical numbers instead of Cahn Ingold Prelog (CIP) priorities: a greater canonical number precedes a smaller one, and terminal hydrogen atoms are preceded by all other atoms. For stereogenic bonds, parity (+) corresponds to "E" and parity (-) corresponds to "Z". For stereogenic (sp^3) atoms, parity (+) corresponds to "R" and parity (-) corresponds to "S". In canonicalization it is assumed that (-) < (+) < (unknown) < (undefined). Note that sp^3 parities do not affect parities of stereogenic bonds and IChI parities are different from those obtained according to CIP rules.

Canonical numbering involves the following steps:

1. Obtain the numbering for a structure without hydrogen atoms
2. Add non-tautomeric H and tautomeric groups
3. Optionally, add tautomeric H in original positions to turn off the tautomerism
4. Optionally, add isotopic features
5. Optionally, add stereo features

At each step one more “string” is minimized while keeping “strings” minimized in previous steps unchanged. Stereo parities are obtained from the canonical numbers used instead of CIP priorities. Steps 1-3 of the canonical numbering, are done using an algorithm modified from McKay, B. D. Practical Graph Isomorphism. *Congressus Numerantium* **1981**, *30*, 45–87.

Discussion

Bill Milne: A paper should be published describing the work done to date.

Discussion

David Lide: I commend IChI. I think it will save people a lot of time.

Steve Stein: We turn now to beta-testing the IChI program. The output could be XML-like or it could be in a version without delimiters, i.e. in plain text form, or it could be compressed. Auxiliary information could be full, minimal or none.

Bill Town: I prefer the fuller version: it is easier to read.

Marc Nicklaus: The version number must be present: it is better at the end.

Steve Stein: You do not want to sort on the version number: put it at the end.

Peter Murray-Rust: How do I know that there is no charge, for example, if a shorter form is used? I do not like implied semantics.

Dmitrii Tchekhovskoi: I suggest that the first four entries be fixed and separated with slashes. A colon separates the label and the content beyond that.

Steve Stein: We do not want to be more verbose than we have to be. Should we have to record “charge null”?

Miloslav Nič: In response to the discussion about advantages of non-XML and XML syntax, I suggest using XML as the basic format and calling a possible plain text variant a compressed version, which can be unambiguously expanded to the long XML.

The majority view was to put the version number *first*. It is still possible to sort.

Steve Bryant: If the IChI algorithm is improved, do not invalidate identifiers made by an older version.

Steve Stein: The first layers are unlikely to change. Version 2 must supply guidance on how to use version 1.

Miloslav Nič: The shorter text string is useful for putting in a publication.

[Who said this?]: People will have legacy version one IChIs and no longer have the program and the structures.

Daniel Zaharevitz: Do not assume that Layer 1 will never change.

Steve Bryant: Attach version tags to each layer.

Peter Murray-Rust: People *will* do things of their own that do not conform. XML can be canonicalized. It can be given an XML digital signature. There is a danger of “null” being read as “unknown” or something else.

Steve Bryant: Add new tags always so that earlier layers are never altered.

Daniel Zaharevitz: But do not perpetuate mistakes.

David Martinsen: There needs to be a reference to the IUPAC specification in the XML schema, or in an IChI file itself, to indicate that the marked-up text was in fact done with IChI, using some version of the IUPAC definition.

Steve Stein: The output gives warnings for connected salt, wedge bonds for non-stereo centers, exceptional valence and Z/E stereo ignored or missing (marking where the double bonds were drawn). The warnings will not be part of the canonical identifier but will be text comments. There will be the option to turn the warnings on or off. Errors are also reported, e.g., meaningless stereo descriptors and unexpected input. Allowable inputs will be Molfile, CML (soon) and, perhaps, SMILES.

[Who?]: Can you draw any structure you like in a Molfile?

Steve Stein: The Molfile format has been published.

Peter Murray-Rust: We will address the problem of representation. There is an identifier for user labeling of a structure in the demonstration version of IChI that should not be included in the real version. Dmitrii Tchekhovskoi needs it for convenience of matching his structures and identifiers. The number is needed for debugging and the like but it is not part of the identifier.

Steve Stein: Confirmatory information that is output is chemical formula, equivalent atoms, stereo centers, and computed formula. If we are to give users the option to draw a pleasing structure from the IChI output, we will need to include x,y, (z) coordinates, positions of double bonds, wedge bonds, and explicit hydrogen atoms.

Alan McNaught: There will be a demand for this.

Steve Stein: CML can do this.

Steve Heller: There are programs to do this.

Miloslav Nič: Do not put these extra output fields into IChI.

Steve Stein: It is auxiliary information. It is relevant to the EPO project and to Bill Town’s structure representation project. The *least* we should be able to offer is the option to draw one meaningful structure.

Peter Murray-Rust: The first objective is to produce an identifier. How much further do we want to go? XML concerns lossless information. CML can be written in such a way as to not lose information. Do we really need to draw pretty structures?

Daniel Zaharevitz: I do not see why anyone wants to redraw the structure. It is an identifier, not a structural identification. Why encourage confusion?

Dmitrii Tchekhovskoi: One IChI can lead to multiple depictions.

Alan McNaught: A naïve person may ask “is this really the IChI for the structure I put in?”.

Steve Stein: All this extra information will be in an optional file, an extra file, not in the identifier. Call it “IChI+”. This is not part of the schema.

Peter Linstrom: To put an IChI into the NIST WebBook we need a MIME type for IChI and a file extension. Is there to be one for IChI and one for text IChI?

Peter Murray-Rust: You can contact Henry Rzepa but, increasingly, software vendors do not recognize MIME types. An IChI will not alert Microsoft to the fact that it is an IChI but .xml fails for the human reader.

Ed: No decision was made. Further investigation is needed.

Ture Damhus: In the past, people have asked for a flow diagram to help them to name compounds. Could IChI have a flow diagram?

Alan McNaught: We cannot be very prescriptive at this stage.

Ture Damhus: How should the user organize his thoughts about exactly what he wants to name?

Steve Stein: There is an overlap here with Bill Town’s project, or maybe the problem is integral to it.

Peter Linstrom: We need not just technical documentation for IChI but an explanation for the bench chemist.

Alan McNaught: Yes, documentation about what the program will and will not do.

Steve Stein: In future Markush structures and other issues might be considered.

Aubrey Jenkins: Lots of polymers, e.g., styrene, are extremely simple structures.

Alan McNaught: We have achieved more than we needed to achieve by December 2003. We need to put a supplementary proposal to IUPAC for future work on IChI.

Steve Stein: Phase information, purity etc. might be included.

Dmitrii Tchekhovskoi: Multiple protonated species in proteins and peptides could be added (an extension in the tautomer arena).

Alan McNaught: The algorithm must be described and an article for *Pure and Applied Chemistry* needs to be written.

Ed. It was decided that Steve Stein and Steve Heller would do this and Alan McNaught would write something for publication in *Chemistry World* in 2004.

Marc Nicklaus: If you want to generate IChI identifiers for a large database, should you do it now, or wait for the next version of the algorithm?

Steve Stein: Do it to raise awareness but the algorithm is not for production use yet.

Peter Murray-Rust: We should think about freezing the algorithm at some stage soon so we can get feedback for the community. Public mailing lists or wikis should be used, in the way that Jonathan Brecher used them. (Editor's note. A wiki is a collaboration tool: a Web site where the pages can be changed and instantly published using only a Web browser. No programming is required. Pages are automatically created and linked to each other. A wiki is a way of creating a document socially. "Wiki wiki" is Hawaiian for "quick".)

Ture Damhus: Think carefully about how you do your documentation for users who are not computer-literate.

Steve Stein: I prefer to have a few people look hard at the algorithm at the moment, not lots of people each doing one or two structures.