An Open Standard for Chemical Structure Representation

The IUPAC Chemical Identifier

S. E. Stein, S. R. Heller, D. V. Tchekhovskoi
Physical & Chemical Properties Division
NIST

“Classification and Categorization”
International Chemical Information Conference
Nimes, France
October 22, 2003
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.
Current Project

Chemical Nomenclature and Structure Representation Division (VIII)

Number: 2000-025-1-800

Title: IUPAC Chemical Identifier (IChI)

Task Group
Chairman: A. McNaught

Members: R. Heller and S. Stein

Completion Date: 2003
Communication of Chemical Identity

- Human
  - Verbal – Common name
  - Text – Systematic/Common name
  - Pictorial – Structure diagram

- Computers
  - Electronic - Precise standards
Digital ‘ Naming ’ of Chemicals

• Chemical structure is the true ‘ identifier ’

• But, structure representations are not unique or convenient for computers.

• So, convert structure to a unique ‘ name ’ by fixed algorithms
  – The Iupac C H emical Identifier (IChI)
Customer Needs

- “Authors”
  - Precise
  - Convention-free
  - Wide coverage

- “Readers”
  - Robust
  - Variable specificity
  - Long life

- “Publishers” (Software)
  - Ready access
Two Problems

• Chemicals
  – Fast isomerization (tautomeration)
  – Ill-defined connectivity

• Chemists
  – Differing conventions
    • Depends on discipline, education and convenience
  – Imprecision/uncertainty
3 Steps to IChI

- **Chemistry**
  - ‘Normalize’ Input Structure
    - Implement chemical rules

- **Math**
  - ‘Canonicalize’ (label the atoms)
    - Equivalent atoms get the same label

- **Format**
  - ‘Serialize’ Labeled Structure
    - Output as character string (‘name’)

Normalize

Simplify

- Divide structure into ‘layers’
  - Each layer ‘refines’ structure

- Ignore ‘Electron Density’
  - Use simple ‘connectivity’ only
  - Ignore bond type and electron location

- Stereochemistry
  - $sp^2$ and $sp^3$ only
  - Free rotation around single bonds
  - No Z/E stereo for small rings (default)
“Layers”

Chemical Substances

- formula
- connectivity
- stereo
- isotope
4 Connectivity ‘Sublayers’

• Disconnect metals and H-atoms
  – Skeleton

• Reconnect fixed H-atoms
  – Tautomerism

• Reconnect mobile H-atoms (optional)
  – All connections fixed

• Reconnect metals (optional)
  – Represent bonds to metals
Tautomer Sublayer

H-migration between 1,3-heteroatoms
Stereochemical Sublayers

- $sp^2$ – double bond
- $sp^3$ – tetrahedral
- {others added later}
- relative, absolute or racemic
Münchnones

Simplify - Ignore Electrons
Simplify - Limit Stereo

Assume Free Rotation Around Single Bonds

No Conformers
Simplify – Double Bond Stereo

Rules

Ignore stereo for small rings
Nitrobenzene

![Nitrobenzene molecule with canonical numbering]

<table>
<thead>
<tr>
<th>Description</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>formula</td>
<td>C_{6}H_{5}NO_{2}</td>
</tr>
<tr>
<td>connectivity</td>
<td>8-7 (9) 6-4-2-1-3-5-6</td>
</tr>
<tr>
<td>H-atoms</td>
<td>1-5H</td>
</tr>
<tr>
<td>charges</td>
<td></td>
</tr>
</tbody>
</table>
MSG
(tautomeric)

<table>
<thead>
<tr>
<th>Description</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>formula</td>
<td>C5H8NO4·Na</td>
</tr>
<tr>
<td>connectivity</td>
<td>6-3 (5 (9) 10) 1-2-4 (7) 8;</td>
</tr>
<tr>
<td>H-atoms</td>
<td>1-2H2, 3H, 6H2 (H-, 7, 8, 9, 10);</td>
</tr>
<tr>
<td>stereo sp³</td>
<td>3-;</td>
</tr>
<tr>
<td>charges</td>
<td>-1; +1</td>
</tr>
</tbody>
</table>
MSG (fixed)

<table>
<thead>
<tr>
<th>Description</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>formula</td>
<td>C5H8NO4.Na</td>
</tr>
<tr>
<td>connectivity</td>
<td>6-3 (5 (9) 10) 1-2-4 (7) 8;</td>
</tr>
<tr>
<td>H-atoms</td>
<td>1-2H2, 3H, 6H2 (H-, 7, 8, 9, 10);</td>
</tr>
<tr>
<td>stereo sp³</td>
<td>3-;</td>
</tr>
<tr>
<td>H-atoms fixed</td>
<td>7H;</td>
</tr>
<tr>
<td>stereo sp³</td>
<td>3-;</td>
</tr>
<tr>
<td>charges</td>
<td>-1;+1</td>
</tr>
</tbody>
</table>
Ferrocene

<table>
<thead>
<tr>
<th>Description</th>
<th>Layers</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>formula</td>
<td>2C5H5.Fe</td>
<td>C10H10Fe</td>
</tr>
<tr>
<td>connectivity</td>
<td>2*1-2-4-5-3-1;</td>
<td>1-2-4-5-3 (1)11 (1,2,4,5)6-7 (11)9 (11)10 (11)8 (6)11</td>
</tr>
<tr>
<td>H-atoms</td>
<td>2*1-5H;</td>
<td>1-10H</td>
</tr>
<tr>
<td>charges</td>
<td>2*-1;+2</td>
<td></td>
</tr>
</tbody>
</table>
Byproducts:
Stereogenic Centers and Equivalent Atoms

- Aids structure validation
Auxiliary Output

• Warnings/Errors
  – Unusual valences
  – Unrecognized input

• ‘Reversibility’
  – Coordinates
  – Bond/Charge Location
Beta Testing
Performance:
Most Challenging NCI-NIH Structure

50 ms – 2 GHz PC
IChI FAQs

• How can you represent chemistry without electrons?
  – Chemistry is not represented, just identity
  – Whole molecule properties are easily added.

• Do big molecules have big IChIs?
  – Yes, just like systematic names

• How to handle other tautomers, substructures,..?
  – Other software

• Is IChI reversible?
  – Partly - contains only data needed for ‘naming’
  – Auxiliary fields can carry other information

• Is IChI extensible?
  – New layers can add refinement
ICH Capabilities

- Identify compounds at the known level of detail
- Convention-free (mostly)
- Generate quickly from structure
- Contains all essential connectivity information
- Simple ASCII representation
Current Project

Committee on Printed and Electronic Publications

Started

Oct. 2002

Number: 2002-022-1-024

Title: Standard XML data dictionaries for chemistry

Task Group
Chairman: Steve Stein

Members: Kirill Degtyarenko, Jeremy Frey, Francois Gilardoni, Jiri Jirat, Robert Lancashire, Alan McNaught, Peter Murray-Rust, Miloslav Nic, and Henry Rzepa

Completion Date: 2005
Utility of Digital ‘Dictionary’

- Traceability
  - Clarity (especially for computers)
- Indexing
  - Effective ‘keywording’
- Accuracy
  - Error checking
- Automated Processing
Goal Color Books as Source of Basic Chemical Terms in XML

- Why IUPAC?
  - International Acceptance
  - Comprehensive
  - Open Process
  - Long-standing
  - Part of its mission
nuclear fusion reaction
A reaction between two light nuclei resulting in the production of a nuclear species heavier than either initial nucleus.
1982, 54, 1543

nuclear graphite
A polycrystalline graphite material for use in nuclear reactor cores consisting of graphitic carbon of very high chemical purity. High purity is needed to avoid absorption of low-energy neutrons and the production of undesirable radioactive species.
Notes:
Apart from the absence of neutron-absorbing impurities, modern reactor graphites are also characterized by a high degree of graphitization and no preferred bulk orientation. Such properties increase the dimensional stability of the nuclear graphite at high temperatures and in a high flux of neutrons. The term nuclear graphite is often, but incorrectly, used for any graphitic material in a nuclear reactor, even if it serves only for structural purposes.
1995, 67, 498

nuclear isomers
Nucleides having the same mass number and atomic number, but occupying different nuclear energy states.
1982, 54, 1545

nuclear level
One of the energy values at which a nucleus can exist for an appreciable time (> 10^{-22} s).
1982, 54, 1547

nuclear magneton
Electromagnetic fundamental physical constant \( \mu_N = (m_e m_p \mu_B) = 5.050 \times 10^{-27} \text{ J T}^{-1} \), where \( m_e \) is the electron rest mass, \( m_p \) the proton rest mass and \( \mu_B \) the Bohr magneton.
CODATA Bull., 1986, 63, 1

nuclear transition
For a nucleus a change from one quantized energy state into another or a nuclear transformation.
1982, 54, 1553

nucleating agent
A material either added to or present in a system, which induces either homogeneous or heterogeneous nucleation.
1972, 31, 608

nucleation (in colloid chemistry)
The process by which nuclei are formed in solution. The condensation of a single chemical compound is called homogeneous nucleation. The simultaneous condensation of more than one compound is called simultaneous nucleation. The condensation of a compound on a foreign substance is called heterogeneous nucleation.
O.B. 84; see also 1972, 31, 608

nucleation and growth
A process in a phase transition in which nuclei of a new phase are first formed, followed by the propagation of the new phase at a faster rate.
See continuous precipitation, discontinuous precipitation.
1994, 66, 587

nucleic acids
Macromolecules, the major organic matter of the nuclei of biological cells, made up of nucleotide units, and hydrolysable into certain pyrimidine or purine bases (usually adenine, cytosine, guanine, thymine, uracil), d-ribose or 2-deoxy-d-ribose and phosphoric acid.
\[
\begin{align*}
\text{base} & : \text{H or OH} \\
\text{base} & : \text{H or OH}
\end{align*}
\]
IUPAC Compendium of Chemical Terminology

This online version of the IUPAC Compendium of Chemical Terminology corresponds to the second edition (1997), compiled by Alan D. McNaught and Andrew Wilkinson (Royal Society of Chemistry, Cambridge, UK).

Please note that to browse or search this compendium, you will need Adobe Acrobat Reader.

BROWSE (on IUPAC main site)

SEARCH:

Submit

Sort by

SEARCH - hosted by the Royal Society of Chemistry

Muscat engine includes right-hand wildcard searching.
osmotic coefficient, \( \phi \)

Quantity characterizing the deviation of the solvent from ideal behaviour referenced to Raoult’s law. The osmotic coefficient on a molality basis is defined by:

\[
\phi = \frac{\mu_A^o - \mu_A}{RTM_A \sum m_i}
\]

and on an amount fraction basis by:

\[
\phi = \frac{\mu_A^o - \mu_A}{RT \ln x_A}
\]

where \( \mu_A^o \) and \( \mu_A \) are the chemical potentials of the solvent as a pure substance and in solution, respectively, \( M_A \) is its molar mass, \( x_A \) its amount fraction, \( R \) the gas constant and \( T \) the temperature. The latter osmotic coefficient is sometimes called the rational osmotic coefficient.

G.B. 51, 1994, 66, 546

The Gold Book is ‘Indexed’ on the Web
Gold Book in XML

- Provide uniform chemical terminology for XML documents
- Root for digital ‘tags’ in chemistry
- Model for future IUPAC recommendations
Gold Book – PDF to XML
(implicit to explicit)

• Text
  – ‘Tag’ data and relations

• Chemical Structures
  – To connection tables/CML/SVG

• Equations
  – To MathML

• Figures & Complex Schemes
  – Redraw in SVG
IUPAC Gold book

General indexes

- All items - alphabetical index (start with index on left)
- List of "in" sections
  in analytical chemistry, in chromatography ...
- List of "of" sections
  of a polymer, of luminescence ...

Math indexes

- List of math symbols
- List of units

Chemistry indexes

- List of formulas (linear)
- List of (mostly) structure formulas
- List of structure formulas with double bonds
- List of structure formulas with triple bonds
- List of graphical reactions

Miloslav Nic, Jiri Jirat, Czech Republic
Some structures were convertible

chain transfer

\[
\begin{align*}
\text{chain transfer} & : \\
\text{chain transfer} & \\
\text{chain transfer} & : \\
\end{align*}
\]
1 Physical quantities and units 1
   1.1 Physical quantities and quantity calculus 3
   1.2 Base physical quantities and derived physical quantities 4
   1.3 Symbols for physical quantities and units 5
   1.4 Use of the words 'extensive', 'intensive', 'specific' and 'molar' 7
   1.5 Products and quotients of physical quantities and units 8

2 Tables of physical quantities 9
   2.1 Space and time 11
   2.2 Classical mechanics 12
   2.3 Electricity and magnetism 14
   2.4 Quantum mechanics and quantum chemistry 16
   2.5 Atoms and molecules 20
   2.6 Spectroscopy 23
   2.7 Electromagnetic radiation 30
   2.8 Solid state 36
   2.9 Statistical thermodynamics 39
   2.10 General chemistry 41
   2.11 Chemical thermodynamics 48
   2.12 Chemical kinetics 55
   2.13 Electrochemistry 58
   2.14 Colloid and surface chemistry 63
   2.15 Transport properties 65
Green Book - Promise

Template’ for numeric property validation
– Ensure proper units and representation
– Traceable to IUPAC definition
– Basic Tags for Common Properties
  • Covers 15 ‘fields’ of chemistry
Next

- Nov 12-14 Meeting at NIST

- IChI
  - ‘Final’ Beta Nov. 2003
  - Dissemination
    - Databases, Software
  - Version 2

- XML Data Dictionary
  - Gold Book Conversion
  - Maintenance Method
  - Green Book
Naming follows Recognition

“He called the light Day, and the darkness He called night” (Genesis 1.5)
Green Book - Promise

- Periodic Table and Relative Molar Masses
  - Originating digital source
  - Integrate with relevant IUPAC recommendations

- Provide root of chemical information ‘tree’
  - Spectroscopy, electrochemistry, thermochemistry, catalysis, …
Tautomer Rules

\[
\text{H} \quad \text{L} = \text{Q} = \text{R} \quad \leftrightarrow \quad \text{L} = \text{Q} = \text{R} \\
\text{Q} = \text{C, N, S, P, ...} \\
\text{L, R} = \text{N, O, S, Se, Te}
\]

Salts

\[
\text{–OH} \quad \text{–O}^- \\
\text{Salts}
\]