An XML Namespace for IUPAC

Jeremy Frey, Robert Lancashire, Alan McNaught, Peter Murray-Rust, Henry Rzepa, Steve Stein
(and a dozen more)

CAS/IUPAC Conference on Chemical Identifiers and XML for Chemistry
Columbus, OH
July 1, 2002
<table>
<thead>
<tr>
<th>IUPAC</th>
<th>XML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standards for Chemical Communication</td>
<td>Standards for Representation</td>
</tr>
<tr>
<td>Terminology</td>
<td>Ontology</td>
</tr>
<tr>
<td>Data Identification</td>
<td>Tags</td>
</tr>
<tr>
<td>Standard Data</td>
<td>Instances</td>
</tr>
</tbody>
</table>
XML – Tags and Tools

Data-Type Tree

tags

<A><B><C>data</C></B></A>

XML Serialization

tools

Chemist, Experiment, Database
IUPAC ‘Namespace’

• Nomenclature
• Glossaries
  – Diverse, broad to specialized
• Data Representation
• Chemical Data
  – Periodic Table (relative molar mass)
  – Reference Data

Many in cooperation with other international standards bodies
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
3.1 The international system of units (SI) 69
3.2 Definitions of the SI base units 70
3.3 Names and symbols for the SI base units 71
3.4 SI derived units with special names and symbols 72
3.5 SI derived units for other quantities 73
3.6 SI prefixes 74
3.7 Units in use together with the SI 75
3.8 Atomic units 76
3.9 Dimensionless quantities 77

4 Recommended mathematical symbols 81
   4.1 Printing of numbers and mathematical symbols 83
   4.2 Symbols, operators and functions 84

5 Fundamental physical constants 87

6 Properties of particles, elements and nuclides 91
   6.1 Properties of some particles 93
   6.2 Standard atomic weights of the elements 1991 94
   6.3 Properties of nuclides 98

7 Conversion of units 105
   7.1 The use of quantity calculus 107
   7.2 Conversion tables for units 110
      (Pressure conversion factors 166; Energy conversion factors inside back cover)
   7.3 The esu, emu, Gaussian and atomic unit systems 117
   7.4 Transformation of equations of electromagnetic theory between the SI,
      the four-quantity irrational form, and the Gaussian form 122

8 Abbreviations and acronyms 125
### 7.2 CONVERSION TABLES FOR UNITS

The table below gives conversion factors from a variety of units to the corresponding SI unit. Examples of the use of this table have already been given in the preceding section. For each physical quantity the name is given, followed by the recommended symbol(s). Then the SI unit is given, followed by the esu, emu, Gaussian unit (Gau), atomic unit (au), and other units in common use, with their conversion factors to SI. The constant $\alpha$ which occurs in some of the electromagnetic conversion factors is the (exact) pure number $2.99792458 \times 10^8 = c_0/(\text{cm s}^{-1})$.

The inclusion of non-SI units in this table should not be taken to imply that their use is to be encouraged. With some exceptions, SI units are always to be preferred to non-SI units. However, since many of the units below are to be found in the scientific literature, it is convenient to tabulate their relation to the SI.

For convenience units in the esu and Gaussian systems are quoted in terms of the four dimensions length, mass, time, and electric charge, by including the franklin (Fr) as an abbreviation for the electrostatic unit of charge and $4\pi\varepsilon_0$ as a constant with dimensions (charge$^2$/energy $\times$ length). This gives each physical quantity the same dimensions in all systems, so that all conversion factors are pure numbers. The factors $4\pi\varepsilon_0 = 4\pi\varepsilon_0 = 1$ Fr r esu of charge $= \text{erg cm}^{-2} = \text{cm}^{-2} \text{g}^{-1} \text{s}^{-2} \text{ Fr}^{-1}$, and 4 Fr esu to recover esu expressions in terms of these base units (see section 7.3 below). The symbol Fr should be regarded as a compact representation of esu of charge.

Conversion factors are either given exactly (when the ± sign is used), or they are given to the approximation that the corresponding physical constants are known (when the ± sign is used). In the latter case the uncertainty is always less than ±0.5 in the last digit quoted.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Relation to SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>metre (SI unit)</td>
<td>m</td>
<td>m = $10^{-2}$ m</td>
</tr>
<tr>
<td>centimetre (cgs unit)</td>
<td>cm</td>
<td>cm = $10^{-2}$ m</td>
</tr>
<tr>
<td>bohr (au)</td>
<td>a$_0$, b</td>
<td>= $4\pi\varepsilon_0 a_0^2$</td>
</tr>
<tr>
<td>angstrom</td>
<td>Å</td>
<td>= $10^{-10}$ m</td>
</tr>
<tr>
<td>micron</td>
<td>μm</td>
<td>μm = $10^{-6}$ m</td>
</tr>
<tr>
<td>millimicron</td>
<td>μm</td>
<td>μm = $10^{-6}$ m</td>
</tr>
<tr>
<td>x unit</td>
<td>X</td>
<td>≥ 1.002 × 10$^{-1}$ m</td>
</tr>
<tr>
<td>fermi</td>
<td>fm</td>
<td>fm = $10^{-15}$ m</td>
</tr>
<tr>
<td>inch</td>
<td>in</td>
<td>in = 2.54 × 10$^{-2}$ m</td>
</tr>
<tr>
<td>foot</td>
<td>ft</td>
<td>ft = 12 m = 0.3048 m</td>
</tr>
<tr>
<td>yard</td>
<td>yd</td>
<td>yd = 3 ft = 0.9144 m</td>
</tr>
<tr>
<td>mile</td>
<td>mi</td>
<td>mi = 1760 yd = 1609.344 m</td>
</tr>
<tr>
<td>nautical mile</td>
<td>nmi</td>
<td>= 1852 m</td>
</tr>
</tbody>
</table>

- **AA**: atomic absorption
- **AAS**: atomic absorption spectroscopy
- **ac**: alternating current
- **ACM**: adiabatic channel model
- **ACT**: activated complex theory
- **A/D**: analog-to-digital
- **ADC**: analog-to-digital converter
- **AES**: Auger electron spectroscopy
- **AIUPS**: angle-integrated ultraviolet photoelectron spectroscopy
- **AM**: amplitude modulated
- **amu**: atomic mass unit (symbol: u) (see p.75)
- **AO**: atomic orbital
- **APS**: appearance potential spectroscopy
- **ARAES**: angle-resolved Auger electron spectroscopy
- **AS**: Auger spectroscopy
- **ATR**: attenuated total (internal) reflection
- **AU**: astronomical unit (see p.110)
- **au**: atomic unit (see section 7.3, p.120)
- **bcc**: body centred cubic
- **BET**: Brunauer–Emmett–Teller
- **BIS**: brenstrahlung isochromat spectroscopy
- **BM**: Bohr magneton (symbol: μ$_B$, see p.116)
- **bp**: boiling point
- **Btu**: British thermal unit (see p.112)
### 2.5 ATOMS AND MOLECULES

The names and symbols recommended here are in agreement with those recommended by EUPAP [4] and ISO [5]. Additional quantities and symbols used in atomic, nuclear and plasma physics can be found in [4 and 5].

<table>
<thead>
<tr>
<th><strong>Name</strong></th>
<th><strong>Symbol</strong></th>
<th><strong>Definition</strong></th>
<th><strong>SI unit</strong></th>
<th><strong>Notes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>nucleon number, mass number</td>
<td>$A$</td>
<td>$N = A - Z$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>proton number, atomic number</td>
<td>$Z$</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>neutron number</td>
<td>$N$</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>mass of atom, atomic mass</td>
<td>$m$</td>
<td>$m = m_{12}C_{12}/12$</td>
<td>kg</td>
<td>1, 2</td>
</tr>
<tr>
<td>atomic mass constant</td>
<td>$\alpha$</td>
<td>$\alpha = 4\pi\varepsilon_0\hbar^2/e^2$</td>
<td>m</td>
<td>2</td>
</tr>
<tr>
<td>electron rest mass</td>
<td>$m_e$</td>
<td></td>
<td>kg</td>
<td>2</td>
</tr>
<tr>
<td>elementary charge, proton charge</td>
<td>$e$</td>
<td></td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>Planck constant</td>
<td>$h$</td>
<td>$h = \hbar/2\pi$</td>
<td>Js</td>
<td>2</td>
</tr>
<tr>
<td>Bohr radius</td>
<td>$a_0$</td>
<td>$a_0 = 4\pi\varepsilon_0\hbar^2/m_e e^2$</td>
<td>m</td>
<td>2</td>
</tr>
<tr>
<td>Hartree energy</td>
<td>$E_h$</td>
<td>$E_h = h^2/2m_e$</td>
<td>J</td>
<td>2</td>
</tr>
<tr>
<td>Rydberg constant</td>
<td>$R_n$</td>
<td>$R_n = E_h/2e$</td>
<td>m$^{-1}$</td>
<td></td>
</tr>
<tr>
<td>fine structure constant</td>
<td>$\alpha$</td>
<td>$\alpha = e^2/4\pi\varepsilon_0 m_e e^2$</td>
<td>l</td>
<td>2</td>
</tr>
<tr>
<td>ionization energy</td>
<td>$E_i$</td>
<td></td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>electron affinity</td>
<td>$E_a$</td>
<td></td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>electronegativity</td>
<td>$Z$</td>
<td>$Z = (E_i + E_a)$</td>
<td>J</td>
<td>4</td>
</tr>
<tr>
<td>dissociation energy</td>
<td>$E_d$</td>
<td></td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>from the ground state</td>
<td>$D_g$</td>
<td></td>
<td>J</td>
<td>5</td>
</tr>
<tr>
<td>from the potential minimum</td>
<td>$D_p$</td>
<td></td>
<td>J</td>
<td>5</td>
</tr>
</tbody>
</table>

1. Analogous symbols are used for other particles with subscript $p$ for proton, $n$ for neutron, $s$ for s for atom, N for nucleus, etc.

2. This quantity is also used as an atomic unit; see sections 3.8 and 7.3.

3. $m_e$ is equal to the unified atomic mass unit, with symbol $u$, i.e. $m_e = 1$ u (see section 3.7). In biochemistry the name dalton, with symbol Da, is used for the unified atomic mass unit, although the name and symbol have not been accepted by CGPM.
Logarithm of the Retention Factor

This term is equivalent to the $R_m$ value used in planar chromatography (see $R_m$ value). The symbol $\kappa$ is suggested to express log $k$:

$$\kappa = \log k = \log \left(\frac{1 - R}{R}\right)$$

Retardation Factor ($R$)

The fraction of the sample component in the mobile phase at equilibrium; it is related to the retention factor and other fundamental chromatography terms:

$$R = \frac{1}{k + 1}$$

Relative Retention ($r$)

The ratio of the adjusted or net retention volume (time) or retention factor of a component relative to that of a standard, obtained under identical conditions:

$$r = \frac{V_{R_2}}{V_{R(0)}} = \frac{V_{S_2}}{V_{S(0)}} = \frac{t_{R_2}}{t_{R(0)}} = \frac{k_2}{k_m}$$

Depending on the relative position of the peak corresponding to the standard compound in the chromatogram, the value of $r$ may be smaller, larger or identical to unity.
### TABLE 10.9 Transformation of sample into vapour.
**Terms symbols and units for measurable quantities**

<table>
<thead>
<tr>
<th>Terms</th>
<th>Symbol</th>
<th>Practical Unit</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of liquid consumption</td>
<td>$F_1$</td>
<td>cm$^3$s$^{-1}$</td>
<td>In the usual case of a pneumatic nebulizer $F_1$ is called the rate of liquid aspiration</td>
</tr>
<tr>
<td>Efficiency of nebulization</td>
<td>$\epsilon_n$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Fraction desolvated</td>
<td>$\beta_d$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Fraction volatilized</td>
<td>$\beta_v$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Fraction atomized</td>
<td>$\beta_a$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Efficiency of atomization</td>
<td>$\epsilon_a$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Flame temperature</td>
<td>$T_f$</td>
<td>K</td>
<td>When the temperature varies</td>
</tr>
</tbody>
</table>

#### 12.3.3.2 Tandem Mass Spectrometers

**General**

**Fixed neutral loss (gain) scan**

A scan that determines, in a single instrument, all the parent ion mass/charge ratios which react to the loss or gain of a selected neutral mass.

**Fixed neutral loss (gain) spectrum**

A spectrum obtained when data are acquired that determine all the parent ion mass/charge ratios that react by the loss (gain) of a selected neutral mass.
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

cellular graphite
A polygranular 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 graphite material in a nuclear reactor, even if it serves only for structural purposes.
1995, 67, 498

nuclear isomers
Nuclides 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 $\mu_N = (m_e/m_p)\mu_B = 5.050 \times 10^{-27}$ 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.
Varieties of Terms in IUPAC Glossaries

• Data
  – Numeric Quantities (number + dimension)
  – Non-numeric
    • Formulas, reactions, phases, structure

• Concepts
  – Chemicals Properties and Behavior
    • Chemical rules, properties, attributes and behavior
  – Methods
    • Apparatus, procedures, acronyms
  – Reporting
    • Process and Represent Results
  – Non-chemical concepts
    • Use other namespaces one day
What to do?

- Publish glossaries as XML instances (documents)
- Restructure glossaries to serve as ‘namespace’
  - Utilize CML-STM
- Express implicit constraints for data checking
  - Enforce quantity/dimension relations
- (Re)organize data structure
  - Fit neatly in a tree
Publish Glossaries in XML

• Benefits
  – Forces a logical organization on the information
  – Lead by example
  – Supply test documents to developers
  – Information first, uses follow

• Negatives
  – Little present ‘pull’ from users
    • Is it too early?
    • Paucity of end user software tools
  – Leaves too much work
Restructure Glossaries

• Very basic ‘namespace’ required for further growth
• Establish IUPAC as source of XML chemical terminology.

• Is it too early?
XML ‘Instance’

<?xml version="1.0" encoding="ISO-8859-1"?>
<glossary xmlns=http://www.iupac.org/GreenBook
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:schemaLocation="http://www.iupac.org/GreenBook D:\test.xsd">
<quantity name="cartesianCoordinates" pname="Cartesian Space Coordinates" dictRef=""
class="spaceAndTime">
  <i:dimension system="si">
    <i:unit name="m"/>
  </i:dimension>
  <i:symbolSet class="any">
    <i:symbol>x</i:symbol>
    <i:symbol>y</i:symbol>
    <i:symbol>z</i:symbol>
  </i:symbolSet>
</quantity>
</glossary>
Constraints for Data Checking

- Provide XML Schema
- Physical Quantities
  - Ensure correct dimensions
- Chemical Quantities
  - Valid formulas, state symbols,
  - Chemical Constants
    - Periodic Table
<schema targetNamespace="http://www.iupac.org/GreenBook"
xmlns:i="http://www.iupac.org/GreenBook"
xmlns="http://www.w3.org/2001/XMLSchema"
elementFormDefault="unqualified"
attributeFormDefault="unqualified">

<annotation>
  <documentation>
IUPAC test Schema
  </documentation>
</annotation>

<!-- examples of prefix/power templates -->

<!-- examples of prefix/power templates -->

<complexType name="dimensionPrePowType">
  <attribute name="prefix" type="string" use="optional"/>
  <attribute name="pow" type="string" use="required"/>
  <!-- restrict these later to valid values -->
</complexType>

<complexType name="pow1Type">
  <complexContent>
    <restriction base="i:dimensionPrePowType">
      <attribute name="pow" use="required" fixed="1"/>
    </restriction>
  </complexContent>
</complexType>
<physicalProperty xmlns="http://www.iupac.org/GreenBook"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.iupac.org/GreenBook D:\test.xsd">
  <i:energy type="pure">
    <i:value>
      <i:nvalue>12.2</i:nvalue>
      <i:uncert>
        <i:nsignif>3</i:nsignif>
      </i:uncert>
    </i:value>
    <i:units>
      <i:joule pow="1"/>
    </i:units>
  </i:energy>
</physicalProperty>
(Re)organize Data Structure?
Help Wanted

• ‘Domain’ Experts
• Data Distributors
• XML/Software Developers

• ‘International’ UPAC

• steve.stein@nist.gov