CODATA 22
CAPE TOWN

SCIENTIFIC DATA AND SUSTAINABLE DEVELOPMENT
22ND INTERNATIONAL CODATA CONFERENCE
24-27 OCTOBER, 2010
STELLENBOSCH, CAPE TOWN, SOUTH AFRICA

Data Standards within IUPAC and the role of CPEP.

(Co-ordinating Data Standards: The Perspective of Scientific Unions)

Prof. Robert J. Lancashire
Department of Chemistry,
The University of the West Indies,
Mona Campus, Kingston 7, Jamaica

e-mail:robert.lancashire@uwimona.edu.jm

Summary

- **Introduction to IUPAC and CPEP**
- **Introduction to JCAMP-DX and SEDS**
- **XML Initiatives**
  - **AnIML, ThermoML**
  - **The “Colour Books”**
  - **The InChI project**
- **Future Prospects**
- **Acknowledgements**

IUPAC- International Union of Pure and Applied Chemistry http://www.iupac.org

The work of IUPAC is done almost entirely by volunteer scientists (approximately 1400 from many countries) who serve on its task groups, subcommittees and committees.

Its scientific work is largely conducted under a formal project system, in which proposals submitted by chemists worldwide are peer-reviewed and, if meritorious, are approved and supported.

The most recent (2008-2009) biennial report lists IUPAC's six long range goals and provides illustrations of actions taken towards meeting those goals.

Goal 1

IUPAC will provide leadership as a worldwide scientific organization that objectively addresses global issues involving the chemical sciences.

Goal 2

IUPAC will facilitate the advancement of research in the chemical sciences through the tools that it provides for international standardization and scientific discussion.

Goal 3

IUPAC will assist chemistry-related industry in its contribution to sustainable development, wealth creation, and improvement in the quality of life.

Goal 4

IUPAC will foster communication among individual chemists and scientific organizations, with special emphasis on the needs of chemists in developing countries.

Goal 5

IUPAC will utilize its global perspective and network to contribute to the enhancement of chemistry education, the career development of young chemical scientists, and the public appreciation of chemistry.

Goal 6

IUPAC will broaden its national membership base and will seek the maximum feasible diversity in membership of IUPAC bodies in terms of geography, gender, and age.

CPEP – IUPAC Committee on Printed and Electronic Publications

The terms of reference for **CPEP** include offering advice to the President, Executive Committee, other Standing Committees, Divisions, and Commissions on all aspects of the design, implementation, production and dissemination of printed and electronic publications, including computerized databases of all sorts, and to promote the compatibility of electronic transmission and storage of information.

CPEP

In 1996, CPEP evolved from the Committee on Publications that had been in existence since at least 1969.

2010-2011 Membership

Chair: Martinsen, David

Secretary: Lancashire, Robert J.

Titular members:

- Bachrach, Steven M.
- Batchelor, Colin
- Deplanque, René
- Lawlor, Bonnie
- Nic, Miloslav

Ex Officio: Bull, James R.

JCAMP-DX

Joint Committee on Atomic and Molecular Physical Data - Data Exchange

JCAMP-DX originally started as a Task Force on Spectral Data Portability under the direction of Paul A. Wilks, Jr., at the Pittsburgh Conference (Pittcon) of 1983.

By the late 1980s it was sponsored by the American Chemical Society, the American Physical Society, the American Society for Mass Spectrometry, the American Society for Testing and Materials, the Coblentz Society, the Optical Society of America, the Society for Applied Spectroscopy, and the Spectroscopy Society of Canada. The committee developed several spectroscopic data exchange protocols.

In 1995, IUPAC took over the responsibility for the JCAMP-DX range of data standards and initially this was handled by a working party until 2003 when it evolved into **SEDS** (the subcommittee of electronic data standards of CPEP).

SEDS - IUPAC Subcommittee on Spectroscopic Data Standards

2010-2011 Membership

Chair: Davies, Antony N.

Secretary: Lampen, Peter

Members:

- Lancashire, Robert J.
- McIntyre, Peter
- Rutledge, Douglas N.

JCAMP-DX protocols

Key to Success

The key to the successful development of internationally recognized data standards is the open collaboration between industry and user groups. This has been the first, and one of the toughest, hurdles always placed in front of teams wanting to work on the IUPAC JCAMP-DX data standards.

Too much effort has been wasted in the past by diverse bodies developing various data "standards" without first getting agreement of the relevant industries to actually implement the results of their efforts.

All JCAMP-DX reports are made available for comment as “drafts” before being finally published.

JCAMP-DX References

Technique specific Limited Term Task Groups develop new JCAMP-DX standards in their scientific disciplines.

JCAMP-DX vs 4.24 for **IR** : *Applied Spectroscopy*, 1988, 48(1), 151-162.
JCAMP-CS vs 3.7 : *Applied Spectroscopy*, 1991, 45, 4.
JCAMP-DX vs 4.24 for IR: *Pure & Applied Chemistry*, 1991, 63, 1781-92.
for **NMR** vs 5.0 : *Applied Spectroscopy*, 1993, 47, 1093-1099.
for **MS** vs 5.0 : *Applied Spectroscopy*, 1994, 48, 1545-1552.
extension 5.01 for **NMR** : *Pure & Applied Chemistry*, Vol. 71, No. 8, pp. 1549-1556, 1999.
for **IMS** vs 5.01 : *Pure & Applied Chemistry*, Vol 73, No. 11, pp 1765-1782, 2001
for **NMR** pulse sequences : *Pure & Applied Chemistry*, Vol 73, No. 11, pp 1749-1764, 2001
for **EMR** vs 5.01 : *Pure and Applied Chemistry*,78(03), 613-631, 2006

Most recent task group is for the project:
"Standardization of Data and Meta-data formats for Circular Dichroism and Synchrotron Radiation Circular Dichroism Spectroscopy, and interface with the Protein Circular Dichroism Data Bank". Approved starting date 1st September 2010.

XML initiatives

- AnIML (Analytical Information Markup Language)

This developing XML standard for analytical chemistry data is suggested as a long term replacement for JCAMP-DX.

Collaboration exists with the ASTM E13.15 AnIML project.

- ***CML** (Chemical Markup Language)*

Non IUPAC project, but with dedicated following and support.

- ThermoML

- The Colour Books

- InChI

ThermoML - an XML based IUPAC standard for storage and exchange of experimental thermophysical and thermochemical property data. It was fully described in (*Pure Appl. Chem.*, 2006, 78, 541-612).

ThermoML covers essentially all experimentally determined thermodynamic and transport property data (more than 120 properties) for pure compounds, multicomponent mixtures, and chemical reactions (including change-of-state and equilibrium). Although the focus of ThermoML is properties determined by direct experimental measurement, ThermoML does cover key derived property data such as azeotropic properties, Henry's Law constants, virial coefficients (for pure compounds and mixtures), activities and activity coefficients, fugacities and fugacity coefficients, and standard properties derived from high-precision adiabatic heat-capacity calorimetry.

ThermoML continued

The ThermoML structure represents a balanced combination of hierarchical and relational elements. The ThermoML schema structure explicitly incorporates structural elements related to basic principles of phenomenological thermodynamics: thermochemical and thermophysical (equilibrium and transport) properties, state variables, system constraints, phases, and units. Meta- and numerical-data records are grouped into 'nested blocks' of information corresponding to data sets. The metadata records precede numerical data information, providing a robust foundation for generating 'header' records for any relational database where ThermoML-formatted files could be incorporated. The structural features of the ThermoML metadata records ensure unambiguous interpretation of numerical data as well as data-quality control based on the Gibbs Phase Rule.

The "Colour" Book series

Information on the latest printed IUPAC colour books:
http://www.iupac.org/publications/books/seriestitles/nomenclature.html

- **(Gold book)** The IUPAC Compendium of Chemical Terminology, 1997
-online XML version at http://goldbook.iupac.org/index.html

- **(Green book)** Quantities, Units and Symbols in Physical Chemistry, 2007

- **(Red book)** Nomenclature of Inorganic Chemistry, 2005

- **(Blue book)** Nomenclature of Organic Chemistry, 1993

- **(Purple book)** Macromolecular Nomenclature, 1991

- **(Orange book)** Compendium of Analytical Nomenclature, 1997,
- online PDF version 2002

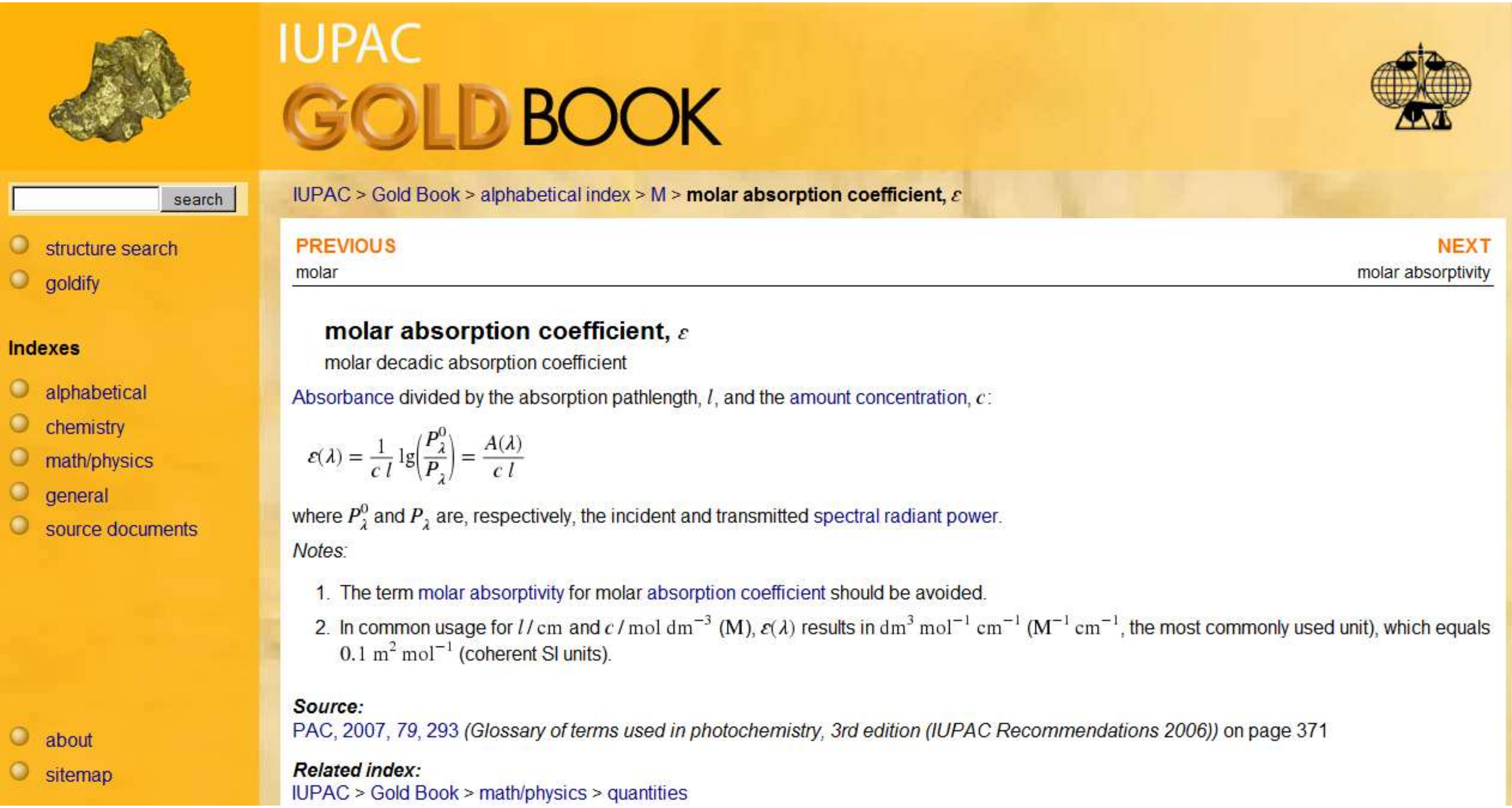
- **(Silver Book)** Compendium of Terminology and Nomenclature of Properties in Clinical Laboratory Sciences, 1995

- **(White book)** Biochemical Nomenclature and Related Documents, 1992

The “Colour” Books

The IUPAC Compendium of Chemical Terminology **(Gold Book)** belongs to the IUPAC “Colour Book” series and comprises over 6000 terminology definitions that were published in *Pure and Applied Chemistry* and other “Colour Books”. The last printed version of the Gold Book was in 1997. In 2002 the project "Standard XML data dictionaries for chemistry" began and 2005 saw the first public preview release of an online XML Gold Book.

http://goldbook.iupac.org/



IUPAC > Gold Book > alphabetical index > M > **molar absorption coefficient, ε**

molar absorption coefficient, ε

molar decadic absorption coefficient

Absorbance divided by the absorption pathlength, l, and the amount concentration, c:

$$\varepsilon(\lambda) = \frac{1}{c\,l}\lg\left(\frac{P_\lambda^0}{P_\lambda}\right) = \frac{A(\lambda)}{c\,l}$$

where P_λ^0 and P_λ are, respectively, the incident and transmitted spectral radiant power.

Notes:

1. The term molar absorptivity for molar absorption coefficient should be avoided.
2. In common usage for l / cm and c / $\mathrm{mol\ dm^{-3}}$ (M), $\varepsilon(\lambda)$ results in $\mathrm{dm^3\ mol^{-1}\ cm^{-1}}$ ($\mathrm{M^{-1}\ cm^{-1}}$, the most commonly used unit), which equals $0.1\ \mathrm{m^2\ mol^{-1}}$ (coherent SI units).

Source:
PAC, 2007, *79*, 293 *(Glossary of terms used in photochemistry, 3rd edition (IUPAC Recommendations 2006))* on page 371

Related index:
IUPAC > Gold Book > math/physics > quantities

InChI- International Chemical Identifier

What is an InChI?

InChIs are character strings that are unique to a chemical structure. They are generated algorithmically by a software program and written in XML comprising different layers and sublayers of information separated by slashes (/).

Each InChI string starts with the InChI version number followed by the main layer. This main layer contains sublayers for chemical formula, atom connections and hydrogen atoms. Depending on the structure of the molecule the main layer may be followed by additional layers e.g. for charge, stereochemical and/or isotopic information.

The InChI project, begun in 2000, has been overseen by an IUPAC Division VIII subcommittee.

The IUPAC InChI quickly gained widespread acceptance and implementation. Unlike other unique identifiers, such as the CAS registry number, the InChI can regenerate the chemical structure with a success rate of over 99%. InChIs are used by major Internet databases (~25 million structures) and are starting to be used by journals. Software developers are providing the identifier in their output. A recent proposed extension is the InChIKey, developed primarily to facilitate use of the InChI by web search engines.

Version 1.03 of the algorithm was released in June 2010

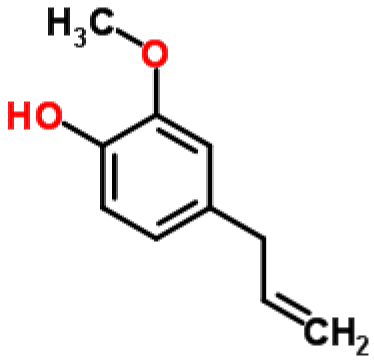
Funding and development of the project is now supported by the InChI Trust.

InChI Trust Website http://www.inchi-trust.org

Mission Statement

- *The InChI Trust develops and supports the non-proprietary IUPAC InChI standard and promotes its uses to the scientific community.*
- *The Trust's goal is to enable the interlinking and combining of chemical, biological and related information, using unique machine-readable chemical structure representations to facilitate and expedite new scientific discoveries.*

Eugenol – from Jamaican pimento leaf



- IUPAC name: 2-methoxy-4-(prop-2-en-1-yl)phenol
- InChI=1S/C10H12O2/c1-3-4-8-5-6-9(11)10(7-8)12-2/h3,5-7,11H,1,4H2,2H3
- InChIKey: RRAFCDWBNXTKKO-UHFFFAOYSA-N

Other IUPAC data

The critically evaluated databases on atmospheric chemistry and water vapour spectroscopy created and maintained by IUPAC are unique and serve as a resource for the regularly updated global climate predictions performed by the Intergovernmental Panel on Climate Change under the auspices of the UN.

Likewise, databases in the area of combustion chemistry and reactive transients such as free radicals are used to understand and model atmospheric pollution.

IUPAC is recognized as the final authority on the naming of elements. The joint IUPAC-IUPAP working party on the discovery of new elements has been reactivated following a considerable number of publications concerning new elements with atomic numbers in the range 112 to 117.

IUPAC Technical Reports

During the period 2008-2009, IUPAC projects led to 27 Recommendations and Technical Reports published in *Pure and Applied Chemistry.* Abstracts and full-text versions of these recommendations and technical reports are available through *PAC* online at no cost. A significant milestone was reached in July 2008 when a full digital archive of all *PAC* articles was completed. This archive, which begins with Volume 1 published in 1960, is easily accessible online, again at no cost. The archive provides a comprehensive published record of the Union’s activities during a decisive period in its history. Anyone can now study 50 years of events, IUPAC projects, and authors with unprecedented ease.

Future Prospects

- The conversion of the other “Colour Books” to online XML documents.
- Continued support for the creation of AnIML protocols for spectroscopic data.

Acknowledgments

Much of the information in these slides has come from the IUPAC web site and the reports available there.

IUPAC - CPEP, SEDS / JCAMP-DX working group.