Current Chemistry Data Activities

9th RDA Plenary Meeting, Barcelona, Spain
5- 7 April 2017

RDA Chemistry Research Data Interest Group (CRDIG)

Enhancing Interoperability across Chemistry, Materials Science, and Photon/Neutron domains through Metadata and Vocabulary
Joint Meeting of IG RDA/CODATA Materials Data, Infrastructure & Interoperability, WG International Materials Resource Registries, IG Chemistry Research Data, IG Research data needs of the Photon and Neutron Science community, IG Metadata

Current Chemistry Data Activity

- **Standard Chemical Identifiers**
  - IUPAC International Chemical Identifier (InChI)
  - Hierarchical Editing Language for Macromolecules (HELM – Pistoia Alliance)
- **Standard File Formats**
  - Development of formats available for spectra
  - Open specifications of de facto standard structure formats (SMILES/SMARTS, MOL/SDfile)
- **Chemistry Terminologies**
  - IUPAC Gold Book project
  - Modelling Terminology for Chemical Safety
- **Chemistry Data Publishing Policies**
  - Comprehensive spreadsheet of journal requirements for chemistry data

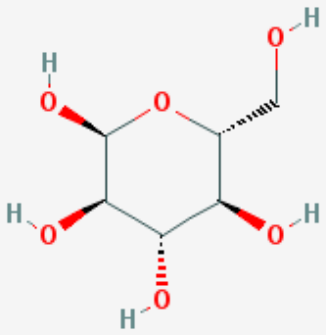
Current Chemistry Data Activity

- **Standard Chemical Identifiers**
  - **IUPAC International Chemical Identifier (InChI)**
  - Hierarchical Editing Language for Macromolecules (HELM – Pistoia Alliance)
- **Standard File Formats**
  - **Development of formats available for spectra**
  - Open specifications of de facto standard structure formats (SMILES/SMARTS, MOL/SDfile)
- **Chemistry Terminologies**
  - **IUPAC Gold Book project**
  - **Modelling Terminology for Chemical Safety**
- **Chemistry Data Publishing Policies**
  - Comprehensive spreadsheet of journal requirements for chemistry data

InChI – What is it?

http://www.iupac.org/inchi

- Identifier
  - International Chemical Identifier
- Standard
  - Initially created by NIST
  - Under auspices of IUPAC
  - Open source, non-proprietary
- Algorithm
  - Normalizes chemical representation
  - Includes ‘hashed’ form called InChIKey



alpha-D-Glucose

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1

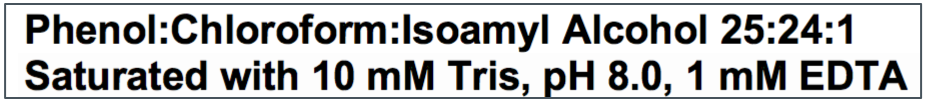
InChIKey=WQZGKKKJIJFFOK-DVKNGEFBSA-N

Layered line notation capturing Version/Type, Chemical formula, Connectivity, Charge/Protonation state, Stereochemistry, Other (e.g., Isotopic)

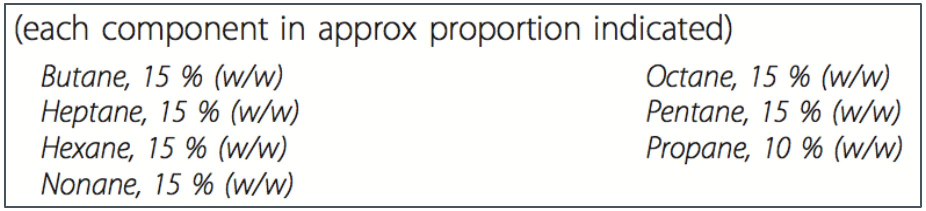
InChI – Current Limitations

- InChI works well for discrete organic molecules
- Current IUPAC/InChI Working groups focussed on improvements for:
  - Organometallics / Inorganics
  - Mixtures
  - Tautomers
  - Reactions (RInChI recently released)
  - Large biomolecules (cf HELM / Pistoia Alliance)
- These and other topics tackled at recent EMBL-EBI Industry Programme Workshop

Wonderful world of mixtures

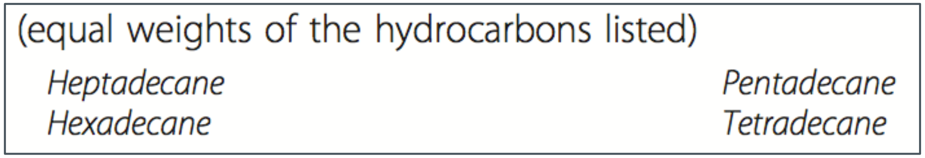


Phenol:Chloroform:Isoamyl Alcohol 25:24:1
Saturated with 10 mM Tris, pH 8.0, 1 mM EDTA



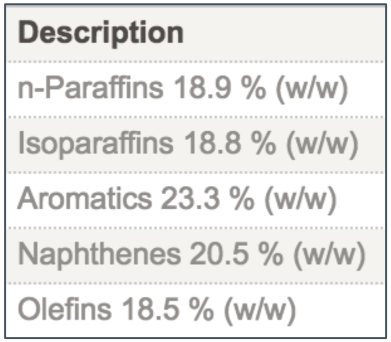
(each component in approx proportion indicated)

Butane, 15 % (w/w)
Heptane, 15 % (w/w)
Hexane, 15 % (w/w)
Nonane, 15 % (w/w)
Octane, 15 % (w/w)
Pentane, 15 % (w/w)
Propane, 10 % (w/w)

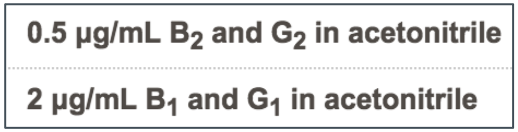


(equal weights of the hydrocarbons listed)

Heptadecane
Hexadecane
Pentadecane
Tetradecane

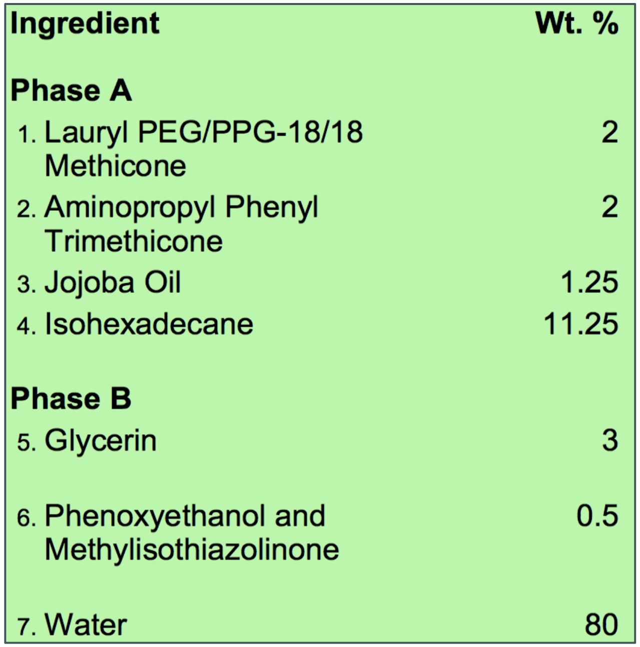


Description
n-Paraffins 18.9 % (w/w)
Isoparaffins 18.8 % (w/w)
Aromatics 23.3 % (w/w)
Naphthenes 20.5 % (w/w)
Olefins 18.5 % (w/w)



0.5 µg/mL B_2 and G_2 in acetonitrile

2 µg/mL B_1 and G_1 in acetonitrile



Ingredient	Wt. %
Phase A	
1. Lauryl PEG/PPG-18/18 Methicone	2
2. Aminopropyl Phenyl Trimethicone	2
3. Jojoba Oil	1.25
4. Isohexadecane	11.25
Phase B	
5. Glycerin	3
6. Phenoxyethanol and Methylisothiazolinone	0.5
7. Water	80

Prototype “MInChI” for “37% wt. Formaldehyde in Water with 10-15% Methanol”:

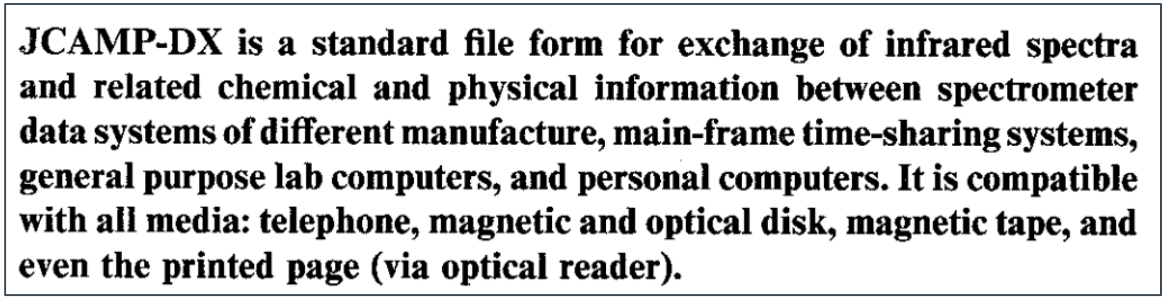
MInChI=0.00.0S/CH2O/c1-2/h1H2&CH4O/c1-2/h2H,1H3&H2O/h1H2/n{1&2&3}/g{37wf-2&10-15vf-2&}

Spectra File Formats

IUPAC Sub-committee on Cheminformatics Data Standards

Charged with reviewing use of chemistry data standards currently available and identifying where there are opportunities to improve these or fill in gaps.

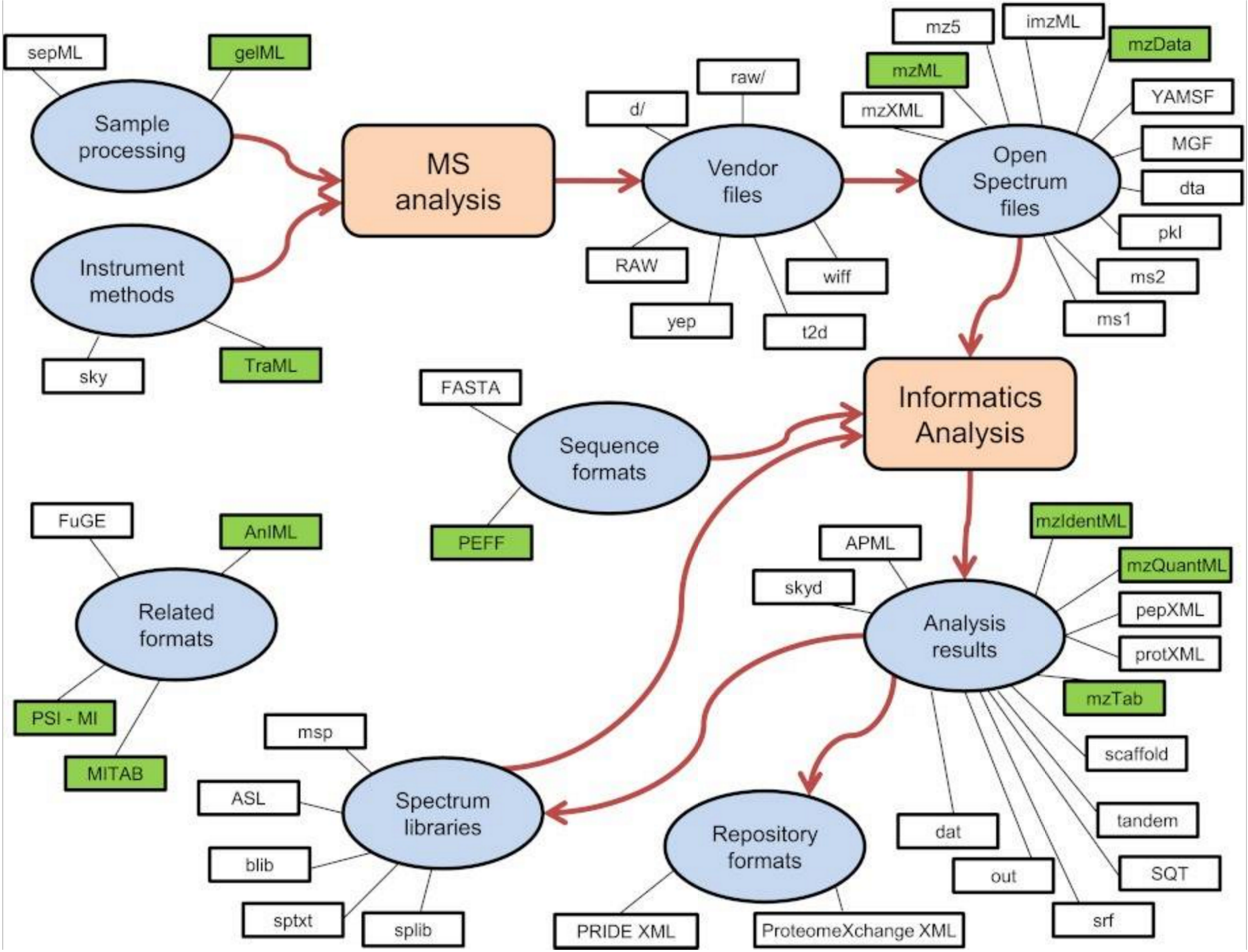
Review of Spectra File Formats



JCAMP-DX is a standard file form for exchange of infrared spectra and related chemical and physical information between spectrometer data systems of different manufacture, main-frame time-sharing systems, general purpose lab computers, and personal computers. It is compatible with all media: telephone, magnetic and optical disk, magnetic tape, and even the printed page (via optical reader).

JCAMP-DX: A Standard Form for Exchange of Infrared Spectra in Computer Readable Form
Applied Spectroscopy
(1988)

- JCAMP-DX: in need of maintenance and updating
- Various vendor-specific versions exist to accommodate additional needs
- Other formats since developed: AniML, mzXML, etc., etc.



*Overview graph of the **mass spectrometry proteomics formats** discussed in Deutsch (2012) Mol Cell Proteomics. 2012 Dec; 11(12): 1612–1621.*

IUPAC Chemical Terminology

- **Blue Book**: Nomenclature of Organic Chemistry
- **Red Book**: Nomenclature of Inorganic Chemistry
- **White Book**: Biochemical Nomenclature
- **Orange Book**: Analytical Terminology
- **Purple Book**: Compendium of Polymer Terminology and Nomenclature
- **Silver Book**: Compendium of Terminology and Nomenclature of Properties Clinical Laboratory Sciences
- **Green Book**: Quantities, Units and Symbols in Physical Chemistry

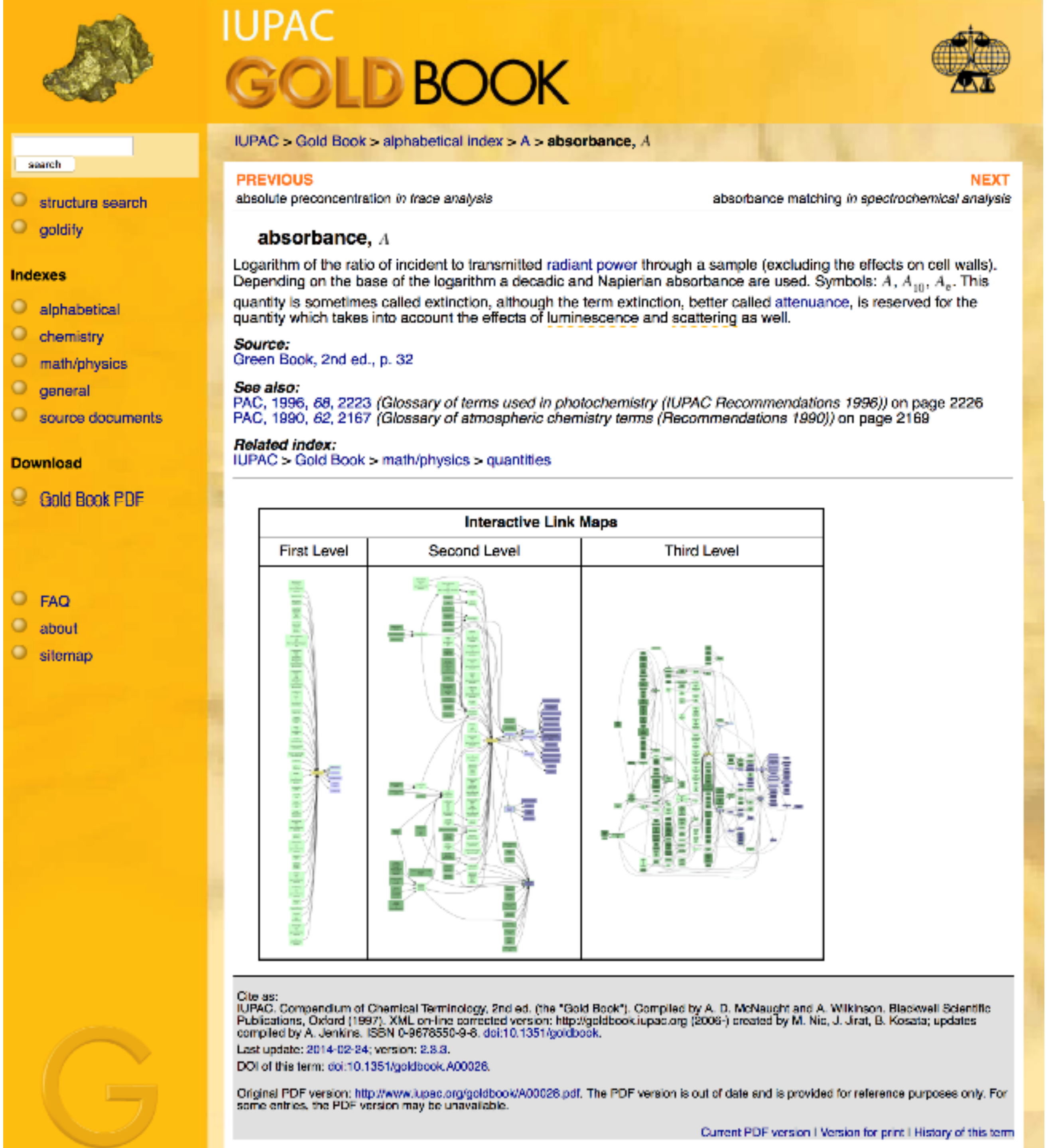
http://iupac.org/what-we-do/books/color-books/

“Digital” Chemical Terminology

goldbook.iupac.org

> 7000 terms with authoritative definitions, spanning the whole range of chemistry

Source documents include *IUPAC Color Books* and recommendations published in *Pure and Applied Chemistry*



IUPAC > Gold Book > alphabetical index > A > **absorbance, A**

PREVIOUS
absolute preconcentration *in trace analysis*

NEXT
absorbance matching *in spectrochemical analysis*

absorbance, A

Logarithm of the ratio of incident to transmitted radiant power through a sample (excluding the effects on cell walls). Depending on the base of the logarithm a decadic and Napierian absorbance are used. Symbols: A, A_{10}, A_e. This quantity is sometimes called extinction, although the term extinction, better called attenuance, is reserved for the quantity which takes into account the effects of luminescence and scattering as well.

Source:
Green Book, 2nd ed., p. 32

See also:
PAC, 1996, *68*, 2223 *(Glossary of terms used in photochemistry (IUPAC Recommendations 1996))* on page 2226
PAC, 1990, *62*, 2167 *(Glossary of atmospheric chemistry terms (Recommendations 1990))* on page 2169

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

Cite as:
IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. doi:10.1351/goldbook.

Last update: 2014-02-24; version: 2.3.3.

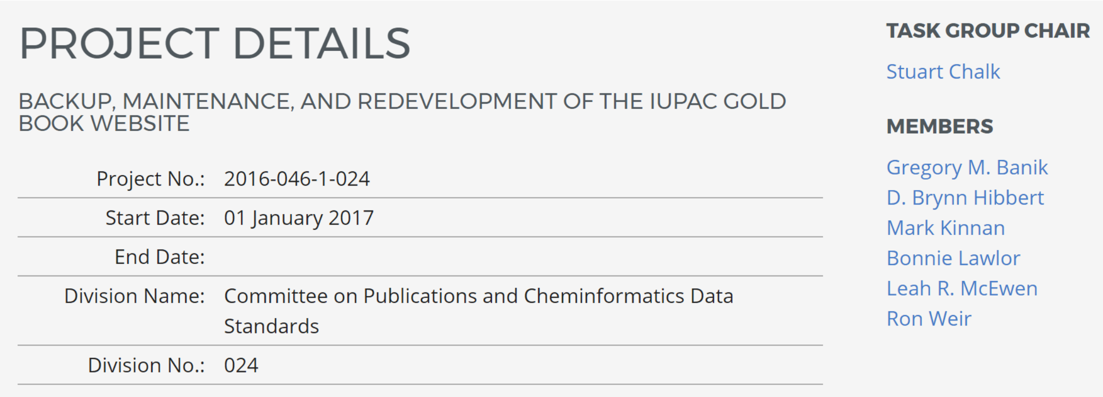
DOI of this term: doi:10.1351/goldbook.A00028.

Original PDF version: http://www.iupac.org/goldbook/A00028.pdf. The PDF version is out of date and is provided for reference purposes only. For some entries, the PDF version may be unavailable.

Current PDF version | Version for print | History of this term

Current IUPAC Gold Book Project

- Goals are to create
  - a stable, modern version of the current Gold Book website
  - a **downloadable vocabulary** of Gold Book terms
  - a simple website to administer updates to Gold Book terms
  - a simple **Application Programming Interface** (API)
- These activities are intended to stabilize and prepare the Gold Book website for future application



PROJECT DETAILS

BACKUP, MAINTENANCE, AND REDEVELOPMENT OF THE IUPAC GOLD BOOK WEBSITE

Project No.:	2016-046-1-024
Start Date:	01 January 2017
End Date:	
Division Name:	Committee on Publications and Cheminformatics Data Standards
Division No.:	024

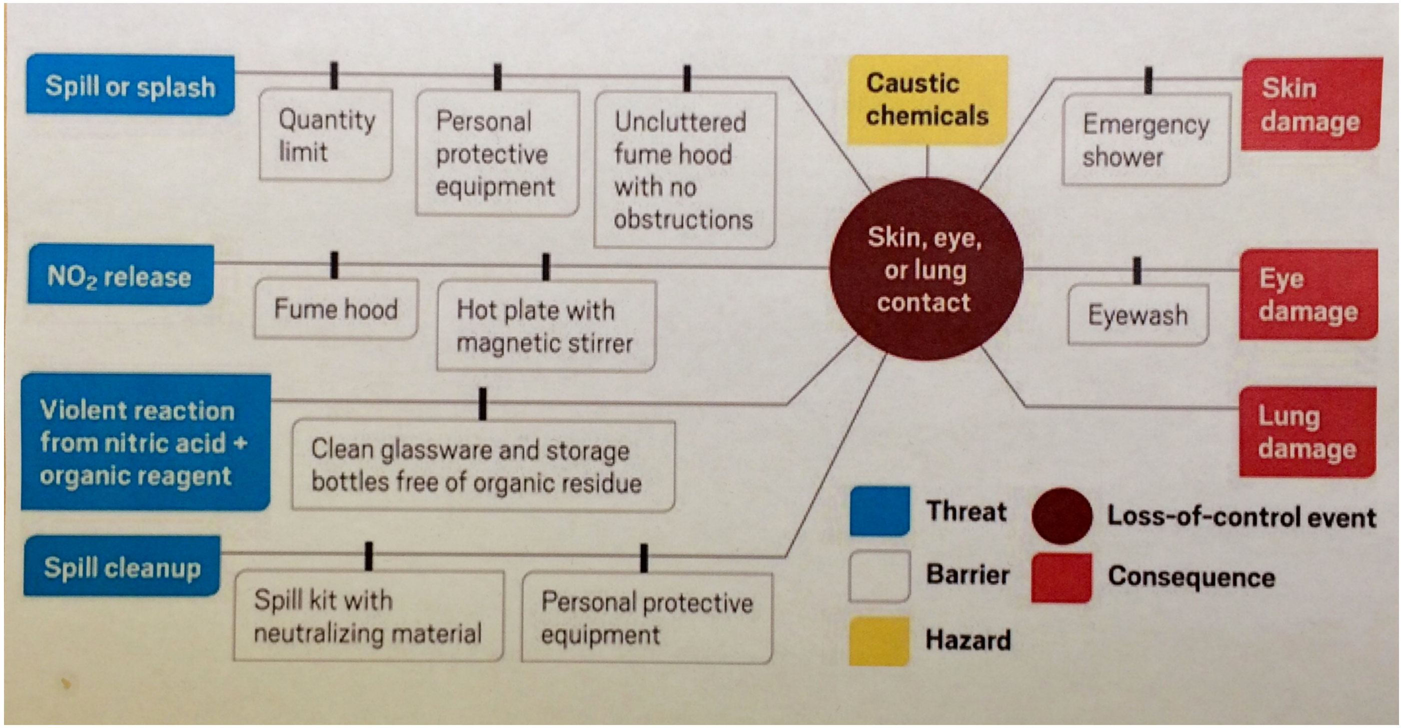
TASK GROUP CHAIR

Stuart Chalk

MEMBERS

Gregory M. Banik
D. Brynn Hibbert
Mark Kinnan
Bonnie Lawlor
Leah R. McEwen
Ron Weir

Chemistry Safety Terminology



Spill or splash

Quantity limit

Personal protective equipment

Uncluttered fume hood with no obstructions

NO_2 release

Fume hood

Hot plate with magnetic stirrer

Violent reaction from nitric acid + organic reagent

Clean glassware and storage bottles free of organic residue

Spill cleanup

Spill kit with neutralizing material

Personal protective equipment

Caustic chemicals

Skin, eye, or lung contact

Emergency shower

Skin damage

Eyewash

Eye damage

Lung damage

🟦 Threat

□ Barrier

🟨 Hazard

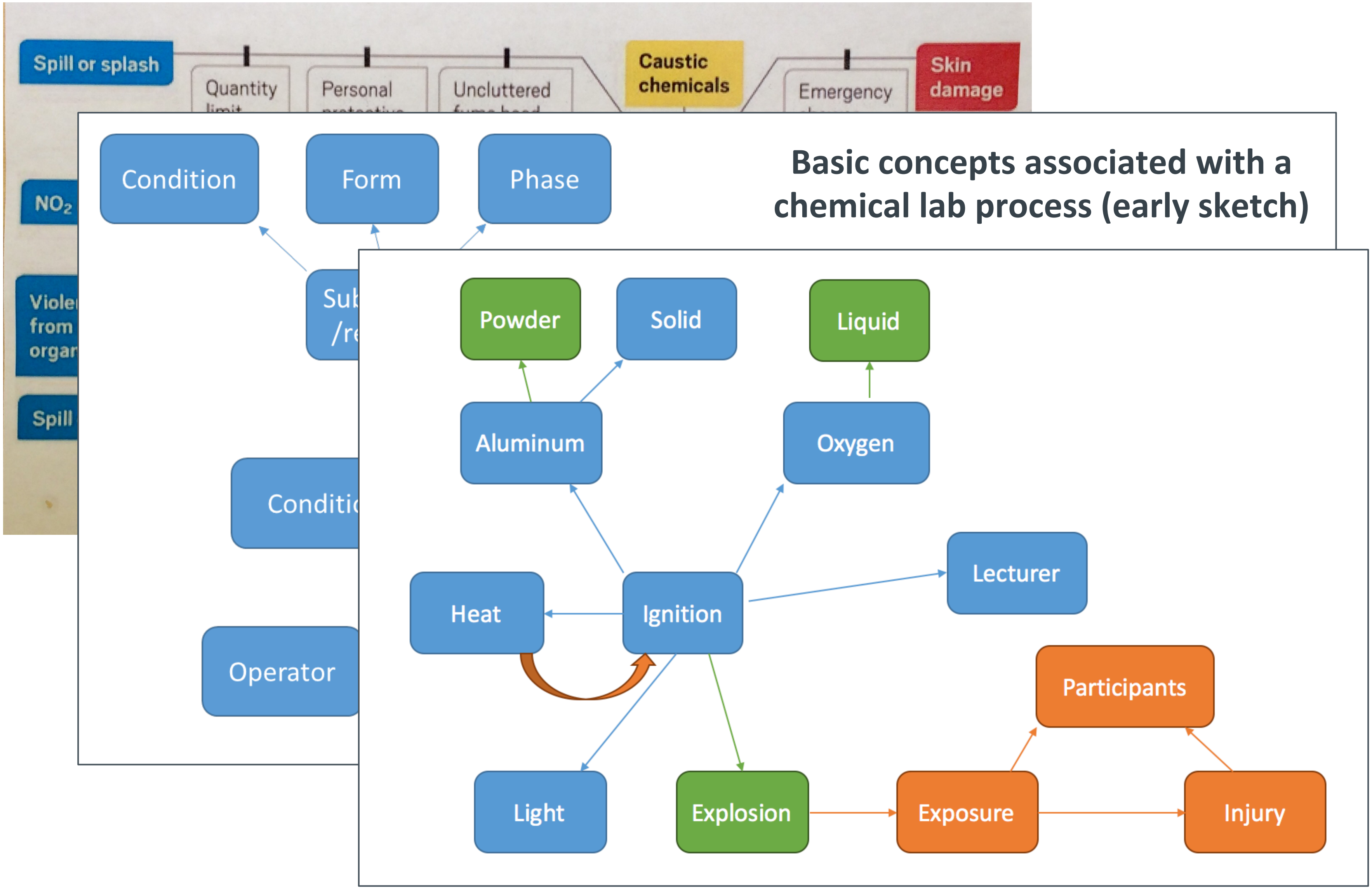
● Loss-of-control event

🟥 Consequence

Hydrogen telluride ignites with cold concentrated nitric acid, sometimes exploding

[Substances / Outcomes / Consequences / Conditions / Operations]

Chemistry Safety Terminology



Spill or splash

Quantity limit

Personal protective

Uncluttered fume hood

Caustic chemicals

Emergency

Skin damage

NO_2

Viole from orgar

Spill

Basic concepts associated with a chemical lab process (early sketch)

Condition

Form

Phase

Sub /re

Conditi

Operator

Powder

Solid

Liquid

Aluminum

Oxygen

Heat

Ignition

Lecturer

Participants

Light

Explosion

Exposure

Injury

Acknowledgements

- **InChI Project and Workshops:** Steve Heller, Evan Bolton, Leah McEwen
- **InChI Mixtures:** Leah McEwen, Gerd Blanke, Alex Clark, John Duffus, Richard Hartshorn, Chris Jakober, Jon LaRue, Andrey Yerin
- **Open Chemical Structure File Formats:** Greg Landrum
- **IUPAC Sub-committee on Cheminformatics Data Standards:** David Martinsen, Leah McEwen, Gregory M. Banik, Ian Bruno, Stuart Chalk, Antony Davies, Jeremy Frey, Robert Lancashire, Ron Weir, Andrey Yerin
- **IUPAC Gold Book Project:** Stuart Chalk
- **Chemistry Safety Terminology:** Leah McEwen, Gary Berg-Cross, Evan Bolton, Ahmed Eliesh, Chris Jakober, Phil Painter, Ralph Stuart
- **Chemistry Data Publishing Policies:** Vincent Scalfani, Tammy Hanna
- **RDA Chemistry Research Data Interest Group:** http://bit.ly/digchem

Forthcoming Chemistry Data Events

RDA 9th Plenary Meeting, Barcelona, April 2017

- *Contributions to sessions on interoperability across disciplines*

Beilstein Symposium – Open Science and the Chemistry Lab of the Future

- *22 – 24 May 2017, Rüdesheim, Germany*

RSC-CICAG meeting on Structure Representation, Liverpool, 22 June 2017

IUPAC World Congress, Sao Paulo, July 2017

- *Special Symposia: Research Data, Big Data, and Chemistry*

InChI/IUPAC Workshop, NIH, Maryland, August 16-18, 2017

- *in conjunction with the InChI Trust*

ACS Fall 2017 Meeting, Washington DC

- *Joint Symposium on “Open Structures”: CSA Trust, ACS (CINF), RDA (CRDIG), IUPAC (CPCDS)*

RDA 10th Plenary Meeting, Montreal, September 2017

To be kept informed join the
RDA Chemistry Research Data Interest Group
http://bit.ly/digchem