WorldFAIR Chemistry: Aligning IUPAC Standards with FAIR Data Practices

Leah McEwen, Cornell University Library
IUPAC Committee on Publications and Cheminformatics Data Standards

WorldFAIR Webinar
2023.09.13

Chemistry is everywhere: Chemistry & UN Sustainability Goals

- Enhance access of unprivileged people to services
- Increase agricultural production & food safety
- Prevent & treat disease
- Increase knowledge and job opportunities
- Supporting social justice, in education, workplaces, and product development
- Prevent & treat disease

1 NO POVERTY

2 ZERO HUNGER

3 GOOD HEALTH AND WELL-BEING

4 QUALITY EDUCATION

5 GENDER EQUALITY

6 CLEAN WATER AND SANITATION

- Increase energy efficiency
- Growth of research, industry and economy
- Develop durable products and services
- Creating technologies for all
- Waste management and raw materials consumption
- Innovation in design and increase life span

7 AFFORDABLE AND CLEAN ENERGY

8 DECENT WORK AND ECONOMIC GROWTH

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

10 REDUCED INEQUALITIES

11 SUSTAINABLE CITIES AND COMMUNITIES

12 RESPONSIBLE CONSUMPTION AND PRODUCTION

- Reduce green house gases, and pollutants
- Reduce water eutrophication and pollution
- Natural resources conversation
- Improve chemicals and waste management
- Partnering for sustainable future

13 CLIMATE ACTION

14 LIFE BELOW WATER

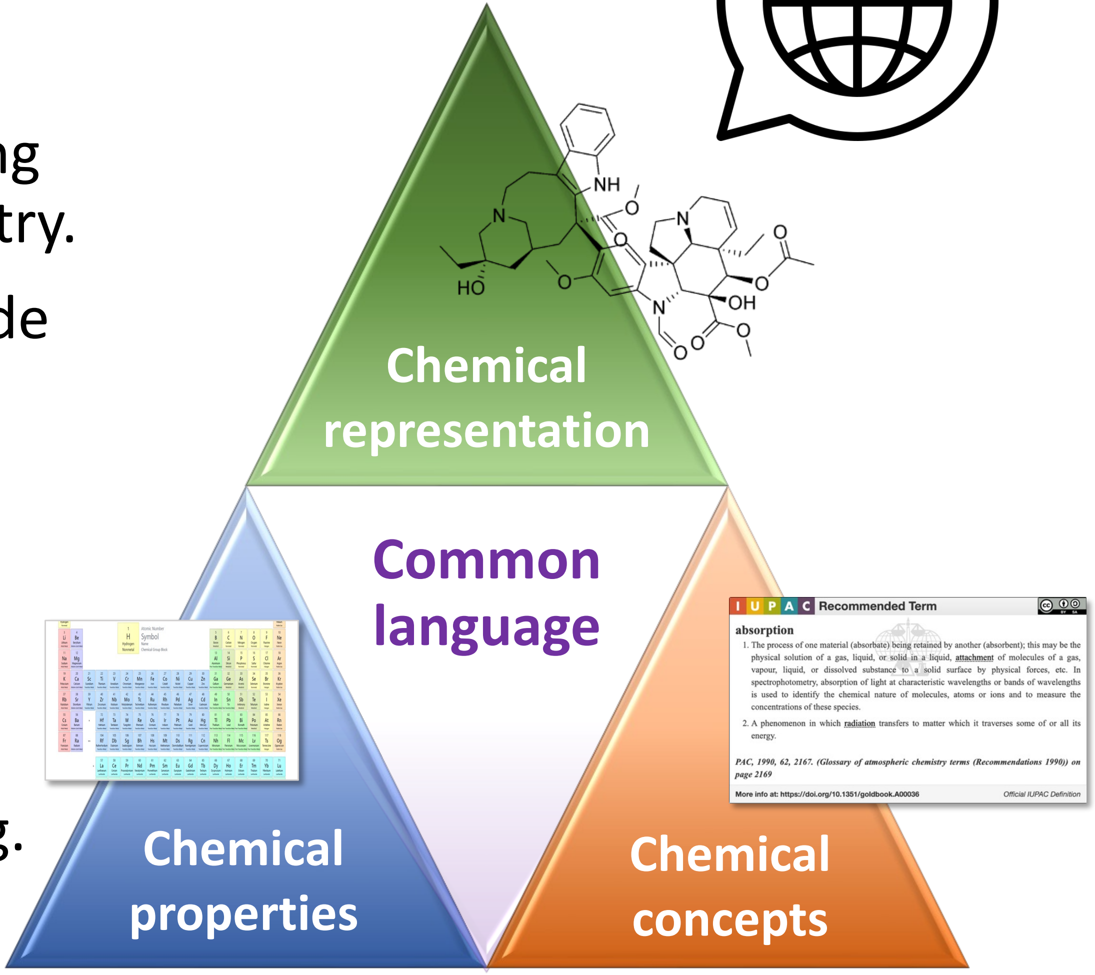
15 LIFE ON LAND

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

17 PARTNERSHIPS FOR THE GOALS

WorldFAIR Chemistry project

- The International Union of Pure and Applied Chemistry (IUPAC) is a standards organization with over 100 yrs of global consensus in defining a common and systematic language for chemistry.
- IUPAC stewards dozens of standards that provide authoritative definitions and parameters for consistent expression of chemical data and information.
- The community needs machine processable representations of these expert defined standards for digital applications and guidance for incorporating these into FAIR data reporting.



Chemical representation

Common language

Chemical properties

Chemical concepts

IUPAC Recommended Term

absorption

1. The process of one material (absorbate) being retained by another (absorbent); this may be the physical solution of a gas, liquid, or solid in a liquid, attachment of molecules of a gas, vapour, liquid, or dissolved substance to a solid surface by physical forces, etc. In spectrophotometry, absorption of light at characteristic wavelengths or bands of wavelengths is used to identify the chemical nature of molecules, atoms or ions and to measure the concentrations of these species.
2. A phenomenon in which radiation transfers to matter which it traverses some of or all its energy.

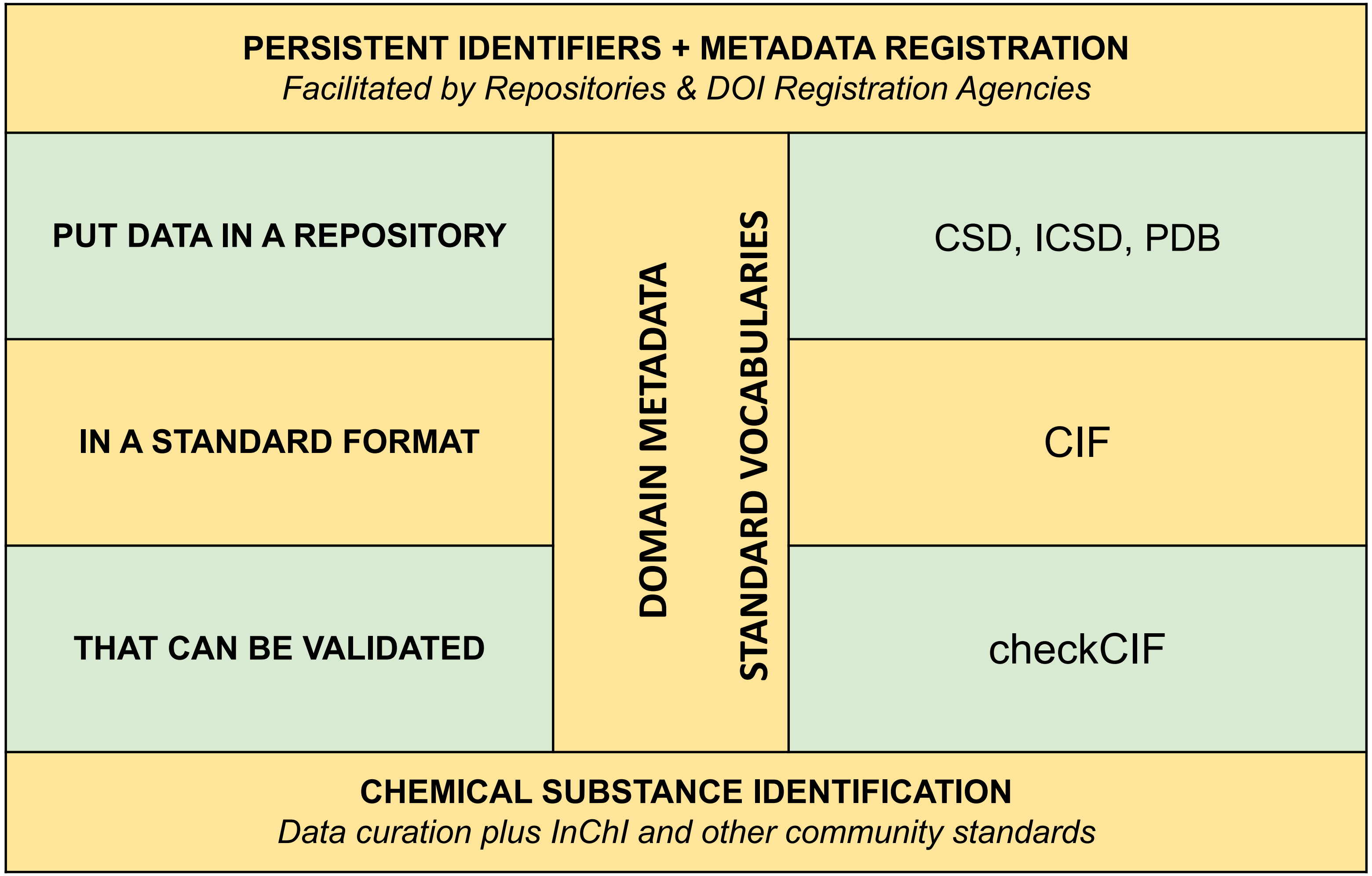
PAC, 1990, 62, 2167. (Glossary of atmospheric chemistry terms (Recommendations 1990)) on page 2169

More info at: https://doi.org/10.1351/goldbook.A00036 *Official IUPAC Definition*

GOAL: align chemical data standards with FAIR

FAIR attributes	Functionality	Chemical notations (examples)
Findable metadata schema	Indexing, matching	InChI, nomenclature
	Searching	Chemical notations (e.g., SMILES), terms (e.g., properties, methods)
Accessible retrieval protocols	Searching, retrieving (APIs) *(consistent across systems)*	Chemical structure resolver *(general spec underway in WFC)*
Interoperable knowledge representations, vocabularies, metadata references	File formats for chemical entities and experimental measurements	SDF, CIF, ThermoML, JCAMP-DX, mzML
	Referrable terms and definitions	Gold Book, VIM, MeSH
	Classification, modeling	CHMO, RXNO, ChEBI, *FAIRSpec*
Reusable validation services	Completeness, consistency	checkCIF

FAIR data implementation in Crystallography



PERSISTENT IDENTIFIERS + METADATA REGISTRATION
Facilitated by Repositories & DOI Registration Agencies

PUT DATA IN A REPOSITORY	**DOMAIN METADATA** **STANDARD VOCABULARIES**	CSD, ICSD, PDB
IN A STANDARD FORMAT		CIF
THAT CAN BE VALIDATED		checkCIF

CHEMICAL SUBSTANCE IDENTIFICATION
Data curation plus InChI and other community standards

Stakeholders

Researchers
Institutions
Funders

Publishers
Editors
Reviewers

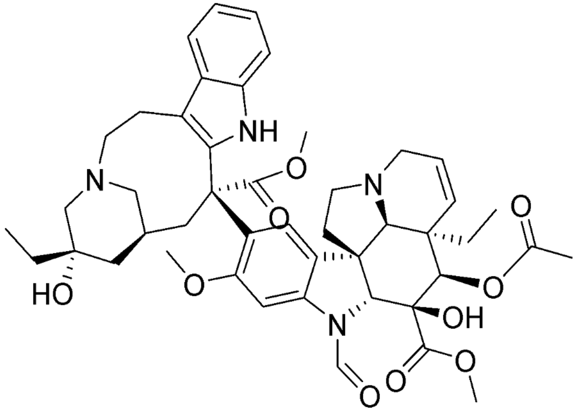
Repositories
Tool Providers
Instrument Manufacturers

Scientific Unions
Professional Societies

Chemical Structure Representation

Vincristine

(trivial name)



IUPAC name - standardized nomenclature

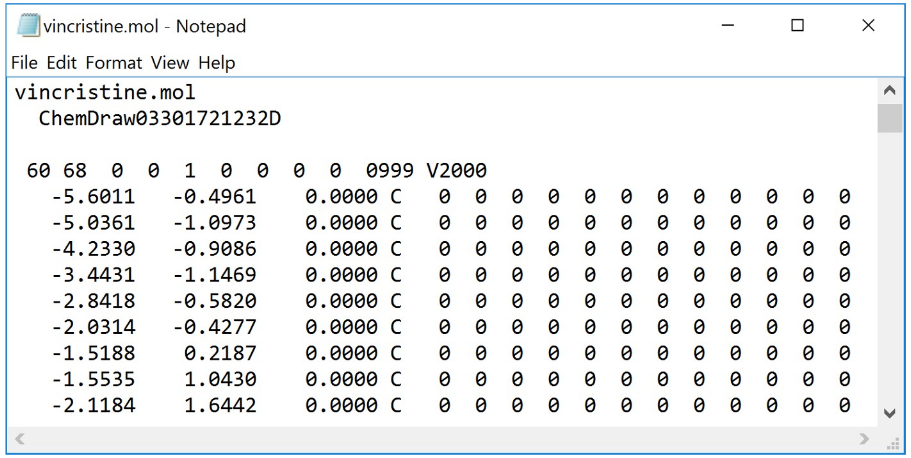
(3a*R*,3a1*R*,4*R*,5*S*,5a*R*,10b*R*)-Methyl 4-acetoxy-3a-ethyl-9-((5*S*,7*S*,9*S*)-5-ethyl-5-hydroxy-9-(methoxycarbonyl)-2,4,5,6,7,8,9,10-octahydro-1*H*-3,7-methano[1]azacycloundecino[5,4-*b*]indol-9-yl)-6-formyl-5-hydroxy-8-methoxy-3a,3a1,4,5,5a,6,11,12-octahydro-1*H*-indolizino[8,1-cd]carbazole-5-carboxylate

SMILES – linear notation for searching, substrucutures *defacto* use, efforts underway to standardize

CC[C@@]1(C[C@@H]2C[C@@](c3c(c4ccccc4[nH]3)CC[N@@](C2)C1)(c5cc6c(cc5OC)N([C@@H]7[C@]68CCN9[C@H]8[C@@](C=CC9)([C@H]([C@@]7(C(=O)OC)O)OC(=O)C)CC)C=O)C(=O)OC)O

Adapted from Scalfani & McEwen, 2019, https://osf.io/psq7k

Molfile – connection table for data exchange, *defacto* use, not yet standardized



```
vincristine.mol
  ChemDraw03301721232D

 60 68  0  0  1  0  0  0  0  0999 V2000
   -5.6011   -0.4961    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -5.0361   -1.0973    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -4.2330   -0.9086    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4431   -1.1469    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.8418   -0.5820    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0314   -0.4277    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.5188    0.2187    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.5535    1.0430    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.1184    1.6442    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
```

InChI - formal descriptor standard for identifying, canonical matching and linking of structures

InChI=1S/C46H56N4O10/c1-7-42(55)22-28-23-45(40(53)58-5,36-30(14-18-48(24-28)25-42)29-12-9-10-13-33(29)47-36)32-20-31-34(21-35(32)57-4)50(26-51)38-44(31)16-19-49-17-11-15-43(8-2,37(44)49)39(60-27(3)52)46(38,56)41(54)59-6/h9-13,15,20-21,26,28,37-39,47,55-56H,7-8,14,16-19,22-25H2,1-6H3/t28-,37+,38-,39-,42+,43-,44-,45+,46+/m1/s1

InChIKey: OGWKCGZFUXNPDA-XQKSVPLYSA-N

Expression of chemical data

Mole fraction of substance 1, x_1 or $x(1)$:

$$x_1 = n_1 / \sum_{s=1}^{c} n_s$$

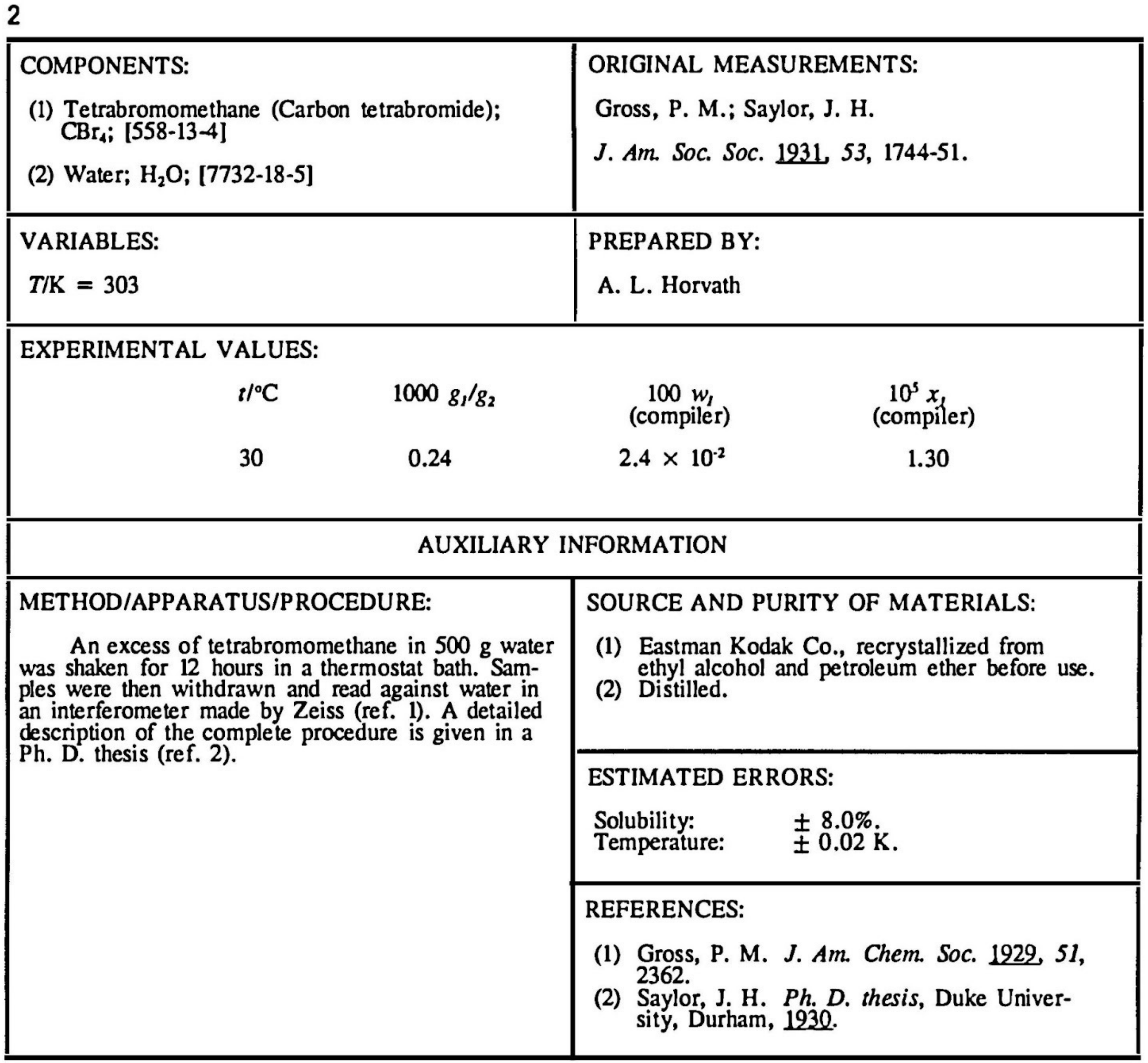
Mass fraction of substance 1, w_1 or $w(1)$:

$$w_1 = g_1 / \sum_{s=1}^{c} g_s$$

Molality of solute 1 in a solvent 2, m_1:

$$m_1 = n_1 / n_2 \, M_2$$

60_3



2

COMPONENTS:

(1) Tetrabromomethane (Carbon tetrabromide); CBr_4; [558-13-4]

(2) Water; H_2O; [7732-18-5]

ORIGINAL MEASUREMENTS:

Gross, P. M.; Saylor, J. H.

J. Am. Soc. Soc. 1931, *53*, 1744-51.

VARIABLES:

T/K = 303

PREPARED BY:

A. L. Horvath

EXPERIMENTAL VALUES:

t/°C	1000 g_1/g_2	100 w_1 (compiler)	10^5 x_1 (compiler)
30	0.24	2.4×10^{-2}	1.30

AUXILIARY INFORMATION

METHOD/APPARATUS/PROCEDURE:

An excess of tetrabromomethane in 500 g water was shaken for 12 hours in a thermostat bath. Samples were then withdrawn and read against water in an interferometer made by Zeiss (ref. 1). A detailed description of the complete procedure is given in a Ph. D. thesis (ref. 2).

SOURCE AND PURITY OF MATERIALS:

(1) Eastman Kodak Co., recrystallized from ethyl alcohol and petroleum ether before use.
(2) Distilled.

ESTIMATED ERRORS:

Solubility:	± 8.0%.
Temperature:	± 0.02 K.

REFERENCES:

(1) Gross, P. M. *J. Am. Chem. Soc.* 1929, *51*, 2362.
(2) Saylor, J. H. *Ph. D. thesis*, Duke University, Durham, 1930.

Horvath & Getzen (1995) IUPAC Solubility Data Series, Vol. 60

Data reporting guidance

Framing: Over-arching goals

Perspectives: Desired end state

Practical: Specific and targeted

Policies: Enforcing community norms

Science as an open enterprise

June 2012

THE ROYAL SOCIETY

unesco

UNESCO Recommendation on Open Science

The National Academies of SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research

Perspective | Published: 04 April 2022

Making the collective knowledge of chemistry open and machine actionable

Kevin Maik Jablonka, Luc Patiny & Berend Smit

Nature Chemistry 14, 365–376 (2022)

ACS GUIDE TO SCHOLARLY COMMUNICATION

Israel Journal of Chemistry IJC

Essay | Open Access

The Long and Winding Road towards FAIR Data as an Integral Component of the Computational Modelling and Dissemination of Chemistry

Henry S. Rzepa

Open Data FAQs for chemists

UNIVERSITY OF CAMBRIDGE

Data sharing

Guidance for best practice and reproducibility of experimental data

Data sharing

ROYAL SOCIETY OF CHEMISTRY

New Data Management and Sharing Policy: January 25, 2023. Learn More.

NIH SCIENTIFIC DATA SHARING

Search
NIH Staff | FAQ | Contacts & Help
DATA MANAGEMENT AND SHARING POLICY
GENOMIC DATA SHARING POLICY
OTHER SHARING POLICIES
ACCESSING DATA
ABOUT

Home > Data Management and Sharing Policy > About Data Management & Sharing Policies > Data Management & Sharing Policy Overview

Data Management & Sharing Policy Overview

IUPAC does not dictate policy; the aim is to provide how-to guidance on reporting chemical data robustly in support of data exchange.

WorldFAIR Chemistry Deliverable 3.1

Digital guidance for Chemistry FAIR data policy and practice

- Landscape overview
  - What is a Chemical?
  - Chemical data across disciplines
  - Community level strategies
  - Open Science and Data Sharing Guidance
- Implementation frameworks
  - RIPE for sharing: Reliable, Interpretable, Processible, Exchangeable
  - IUPAC Standards as FAIR-Enabling Resources
  - Aligning across interoperability frameworks
  - Ecosystems of implementation

Nanomaterials
Materials Science
Earth Sciences
Astrochemistry & physics
Oceanography
Environmental Sciences
Life Sciences

"We emphasize the critical need for software and infrastructure developers, repositories, publishers and others who are building systems and services, to actively incorporate, use and reference chemical data standards in workflows, policies and guidelines."

WorldFAIR Project D3.1, https://doi.org/10.5281/zenodo.7887283

RIPE: well-defined chemical data are broadly reusable

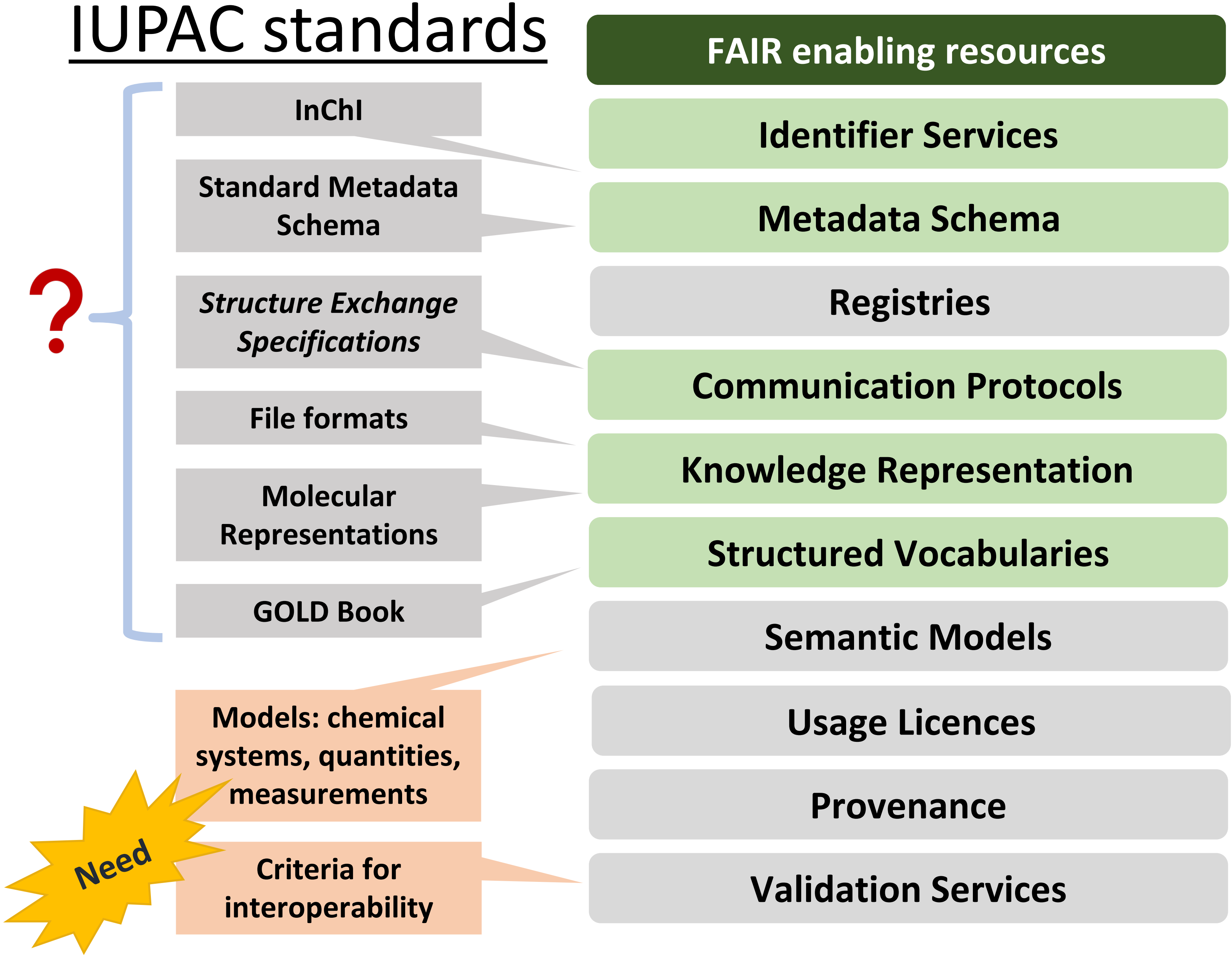
RIPE 4 sharing	Chemical data	Standard definitions (examples)
Reliable information for samples & measurements	Samples: identity of substance(s), sample description (provenance, purity, state)	nomenclature (Blue/Red/Purple books), graphical representation, InChI
	Measurements: techniques, conditions, calibrations, uncertainties	Terminology for analytical chemistry (Orange book), metrology (VIM)
Interpretable scientific expression	Results: quantities, units, calculations, dependencies, processing/derivation	Notations, symbols, terminology for physical chemistry (Green book)
Processable formatted for machines	File formats, validation	SDF, CIF, ThermoML, JCAMP-DX, mzML
	Referrable terms, ontologies	Gold Book, CHMO, RXNO, ChEBI
	Data models, metadata schema	FAIRSpec, *Solubility*, *Periodic Table*
Exchangeable metadata online	Registered metadata for indexing chemicals	InChIs, standard terms/notations
	Standardized exchange APIs for chemicals	*Chemical structure API specification*

(items in italics are in progress)

WorldFAIR Project D3.1 (Table 1)
https://doi.org/10.5281/zenodo.7887283

IUPAC standards

Are these digital standards FAIR for programmatic access and reuse?



?

InChI

Standard Metadata Schema

Structure Exchange Specifications

File formats

Molecular Representations

GOLD Book

Models: chemical systems, quantities, measurements

Criteria for interoperability

Need

FAIR enabling resources

Identifier Services

Metadata Schema

Registries

Communication Protocols

Knowledge Representation

Structured Vocabularies

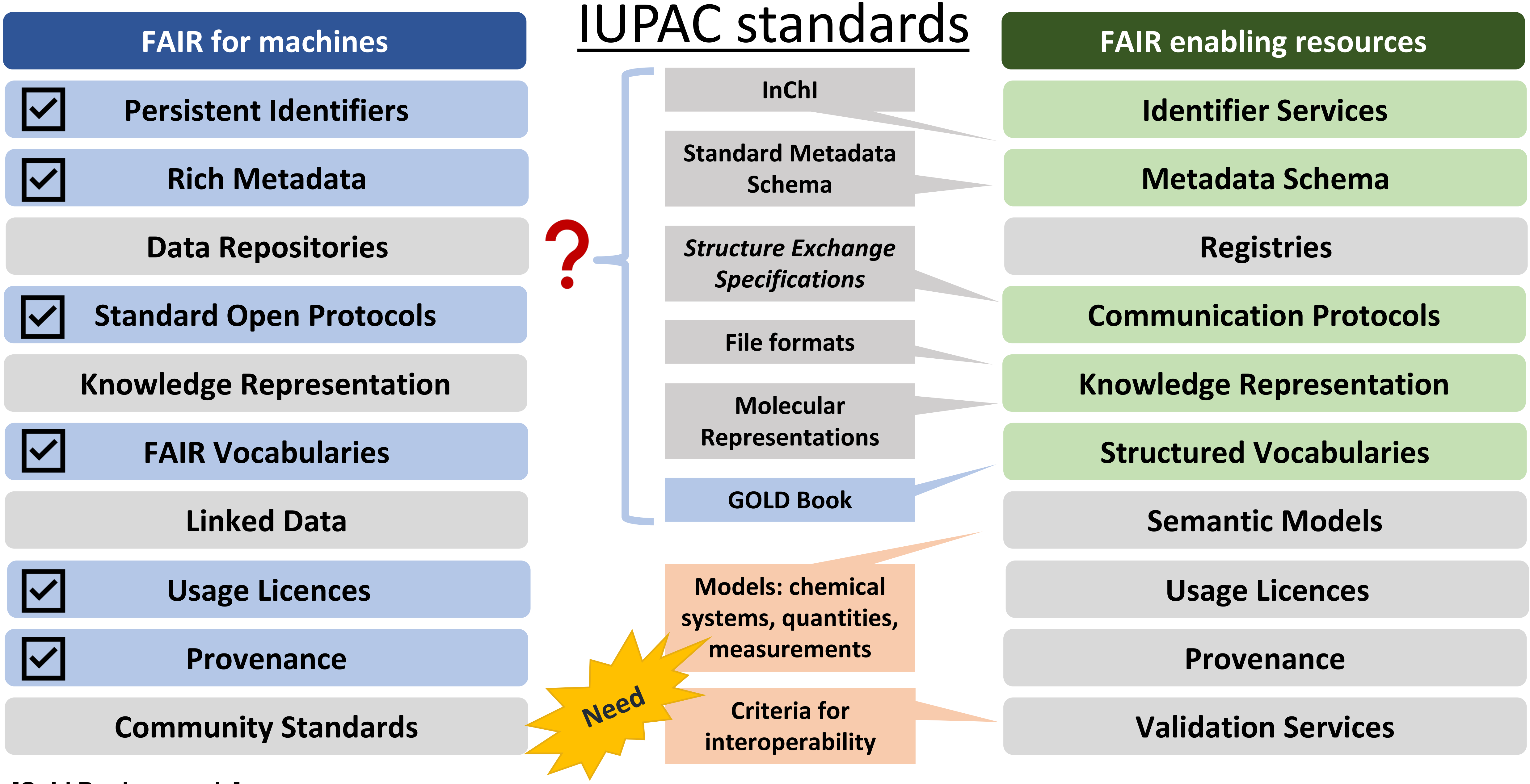
Semantic Models

Usage Licences

Provenance

Validation Services

IUPAC standards



FAIR for machines

- ☑ Persistent Identifiers
- ☑ Rich Metadata
- Data Repositories
- ☑ Standard Open Protocols
- Knowledge Representation
- ☑ FAIR Vocabularies
- Linked Data
- ☑ Usage Licences
- ☑ Provenance
- Community Standards

?

- InChI
- Standard Metadata Schema
- *Structure Exchange Specifications*
- File formats
- Molecular Representations
- GOLD Book

Need

- Models: chemical systems, quantities, measurements
- Criteria for interoperability

FAIR enabling resources

- Identifier Services
- Metadata Schema
- Registries
- Communication Protocols
- Knowledge Representation
- Structured Vocabularies
- Semantic Models
- Usage Licences
- Provenance
- Validation Services

[Gold Book example]

Desired points of cross-domain data integration

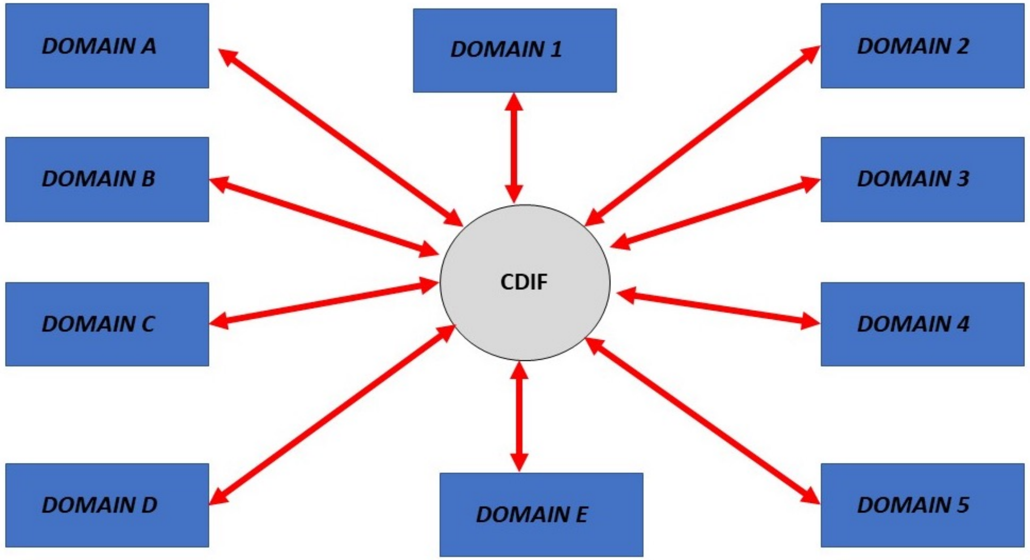
Chemical substance: integration by chemical identification
➥ *standard chemical identifier*

Chemical property: integration of property values
➥ *standard property terms*

Measurement: integration by technique, by conditions
➥ *standard definitions*

Units: integration of quantities➥ *standard units of measure*

Material sample: integration by composition, state of matter, space group ➥ *standard classifications/descriptions*



Origin of sample: integration by location, source (e.g., species), named reactions
➥ *standard location metadata, species classification, reaction classification*

Origin of measurement: integration by analyst or lab, by instrument
➥ *PIDs: ORCID, ROR, etc.*

Temporal: integration by date of sample collection, date of measurement
➥ *standard date format*

Adapted from K. Lehnert, OneGeochemistry, RDA P20 (2023)

WorldFAIR

Further work on cross-domain interoperability

WorldFAIR

- **Samples: provenance and identification**
  - Many well-developed identifiers and other semantic descriptions to describe different facets of sample provenance, *can we harmonize?*
  - RDA P21, Birds of a Feather (WP02, WP03, WP04, WP05, WP10, and others) *'Describing Chemical, Physical and Biological samples digitally'*

23-26 OCT 2023 SALZBURG International Data Week A FESTIVAL OF DATA

- **Measurements & quantities**
  - Molecular structures ⇔ Physical systems ⇔ System conditions
  - Daghstuhl workshop, Oct 1-6 (hosted by WP02 and GO FAIR) *'Defining a core metadata framework for cross-domain data sharing and reuse'*

GO FAIR

- **Terminologies => ontologies**
  - Application of authoritative terminologies in semantic frameworks
  - Ontologies4Chem Workshop, Oct 11-12 (hosted by NFDI4Chem)
  - RDA WG proposal: *'Harmonised terminologies and ontologies for FAIR materials data documentation'*

Editors4Chem Workshop, Nov 2

NFDI4Chem

Future Outlook: Challenges & Opportunities

WorldFAIR

- **Sustainability**
  - Parlous shortage of time and resources currently available to develop and maintain standards, policies, guidance and tools needed to enable machine-actionable reporting of chemical research data into the pipeline
  - *Who should be funding the work necessary long term?*
  - IUPAC workshop, Nov 14-15 (hosted by the Pistoia Alliance) *'Sustainable business models for digital standards development'*

Pistoia Alliance

- **Roadmap**
  - WorldFAIR is initiating excellent synergies towards interoperability and assessment but implementation still primarily at the organizational level and bespoke, system by system
  - *What can IUPAC do to continue to enable success in chemistry digital standards, by and for the broader community?*
  - *What does broad adoption and functional chemical FAIR data exchange across domains look like in practice?*
  - *Beyond FAIR: how do we align demand for chemical data and foster more data reuse across sectors to support SDGs?*

WorldFAIR Chemistry Deliverable Prototypes

WorldFAIR

➔ *develop **guidelines, training materials and tools** that facilitate use of standards*

D3.1 FAIR Chemistry Guidance

bit.ly/IUPACDigitalRecommend

D3.2 FAIR Chemistry Training Cookbook

bit.ly/CookFAIR

D3.3 FAIR Chemistry Protocol Services

bit.ly/ProtServices

Sample prototypes

InChI TRUST CCDC NFDI4Chem PSDI PHYSICAL SCIENCES DATA INFRASTRUCTURE PubChem GO FAIR COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL RDA RESEARCH DATA ALLIANCE

WorldFAIR “Global cooperation on FAIR data policy and practice” is funded by the EC HORIZON-WIDERA-2021-ERA-01-41 Coordination and Support Action under Grant Agreement No. 101058393.

Acknowledgements

WorldFAIR

WorldFAIR Chemistry team (iupac.org/project/2022-012-1-024)

IUPAC Secretariat & volunteers

Community collaborators (chemical sciences & beyond)

WorldFAIR project collaborators

WorldFAIR project funders

iupac.org

@FAIRChemistry
@iupac

YouTube https://bit.ly/WhatsAchemical

FAIRChemistry@iupac.org

@iupac.org

zenodo FAIRChemistry Community

WorldFAIR “Global cooperation on FAIR data policy and practice” is funded by the EC HORIZON-WIDERA-2021-ERA-01-41 Coordination and Support Action under Grant Agreement No. 101058393.