IUPAC

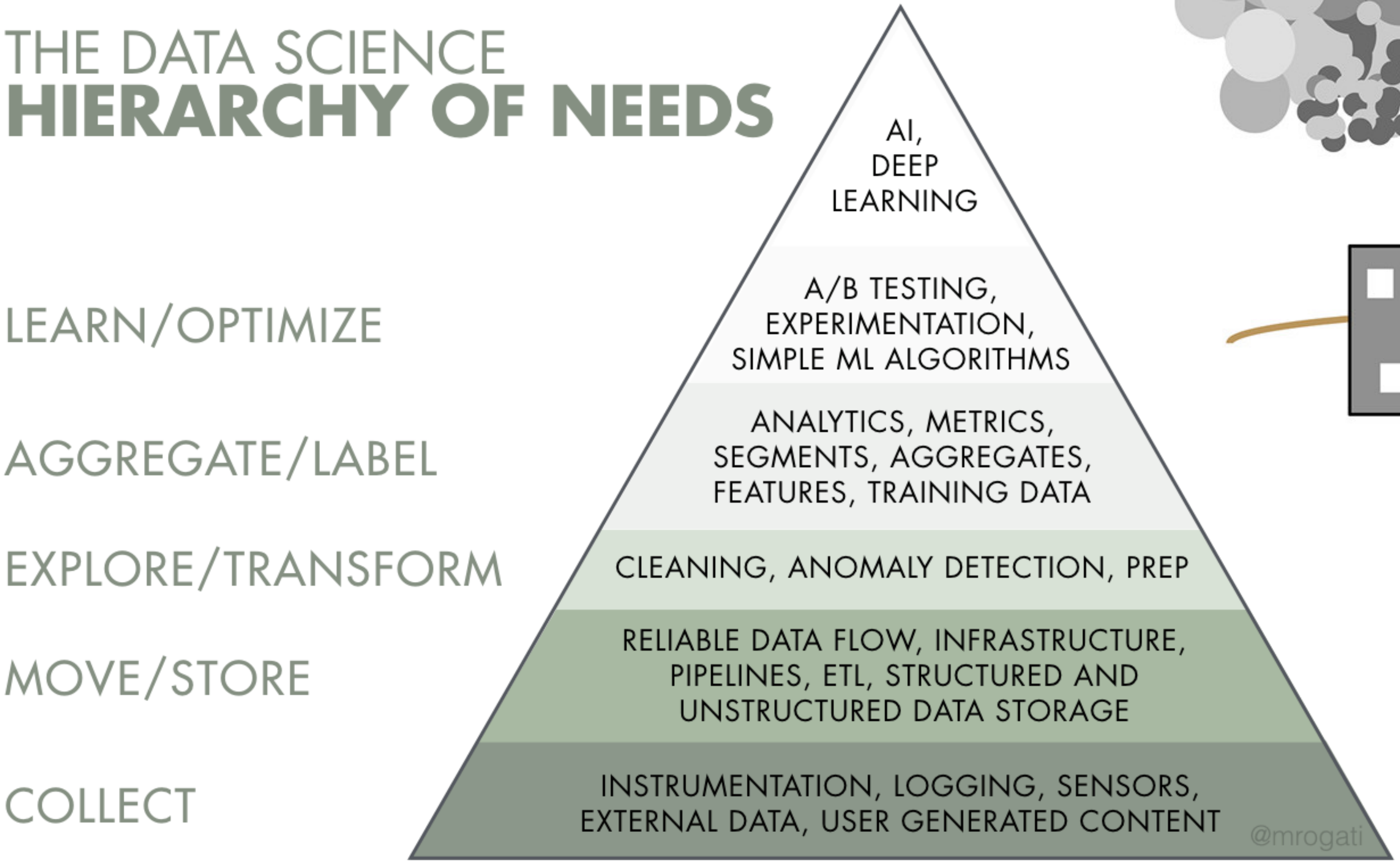
WorldFAIR

WANTED: standard notation for reusable chemical data

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

ChemSpider Webinar
2023.10.17

Chemical data is useful



THE DATA SCIENCE **HIERARCHY OF NEEDS**

LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

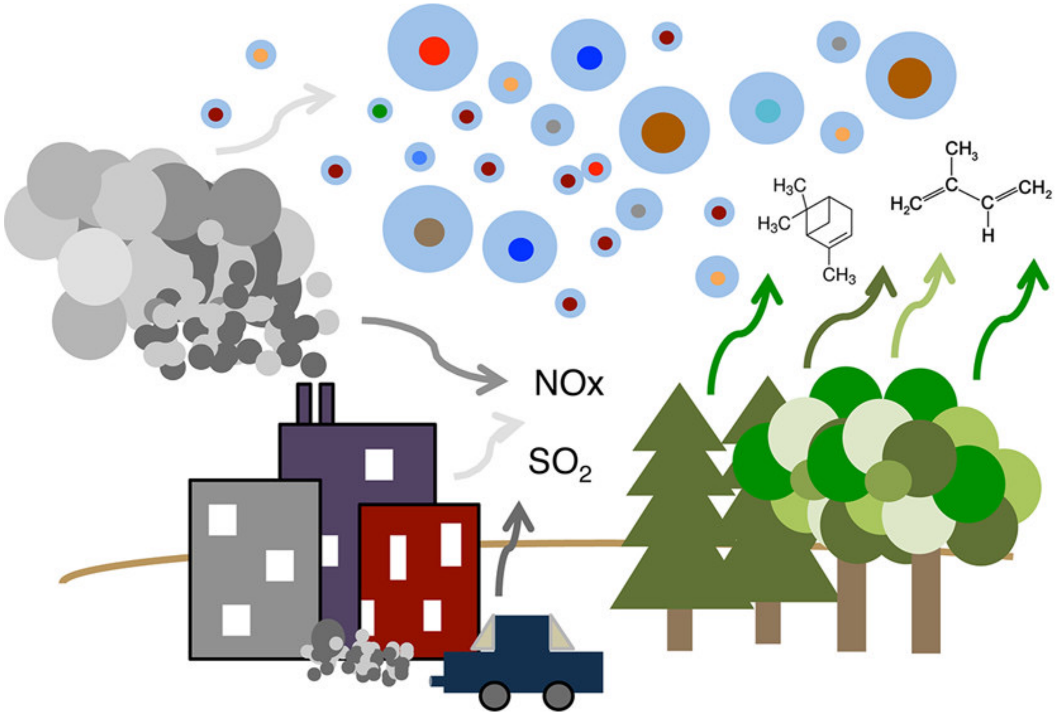
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT

Morgati (2017) hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007



The more data you have, the richer your model

- Breadth (diverse coverage of chemicals)
- Depth (diverse coverage of properties)

Nguyen et al. (2016) doi.org/10.1021/acs.estlett.6b0016

FAIR: “Fully AI-Ready”

Fully AI Ready data are more than accessible digital dataset objects …

- The better quality and precision you have, the stronger the model
- Integrating more data enables more robust discernment of effect from noise

Requirements

- Metadata completeness & consistency
- Data model & domain level description

“Chemically intelligent” notations can help

- Encapsulate formal theories and methods

Standard notations can help even more

DATA SHOULD BE

Findable **Accessible**

Interoperable **Reusable**

BY HUMANS AND MACHINES

Standard Identifiers

Standard Access Protocols

Standard Vocabularies

Standard Metadata Schemas

Indexed Repositories

Key enablers of FAIR

Standards allow us to …

Ascertain fitness-for-purpose

- Dimensions & scope
- Quality & precision

Compile & integrate

Analyze & visualize

Compute & model

> *“A quantitative representation of your subject, however simplified, even in its errors and omissions, is precise. You can think about it rigorously. You can manipulate it and experiment with it.”*
> *~A. W. Crosby, The Measure of Reality*

… more reliably, with less lossy-ness and tedious data cleanup

- *some curation still needed (to apply and validate implementations)*

Are these data **RIPE** for reuse?

Reliable

- can the data be unambiguously positioned relative to the scientific context with the available information provided?

Interpretable

- are data and metadata expressed in a way that is scientifically interpretable and agnostic across local systems (and/or can be converted)?

Processable

- are data and metadata in forms that are processable by common protocols, architectures and infrastructure utilized in the cloud?

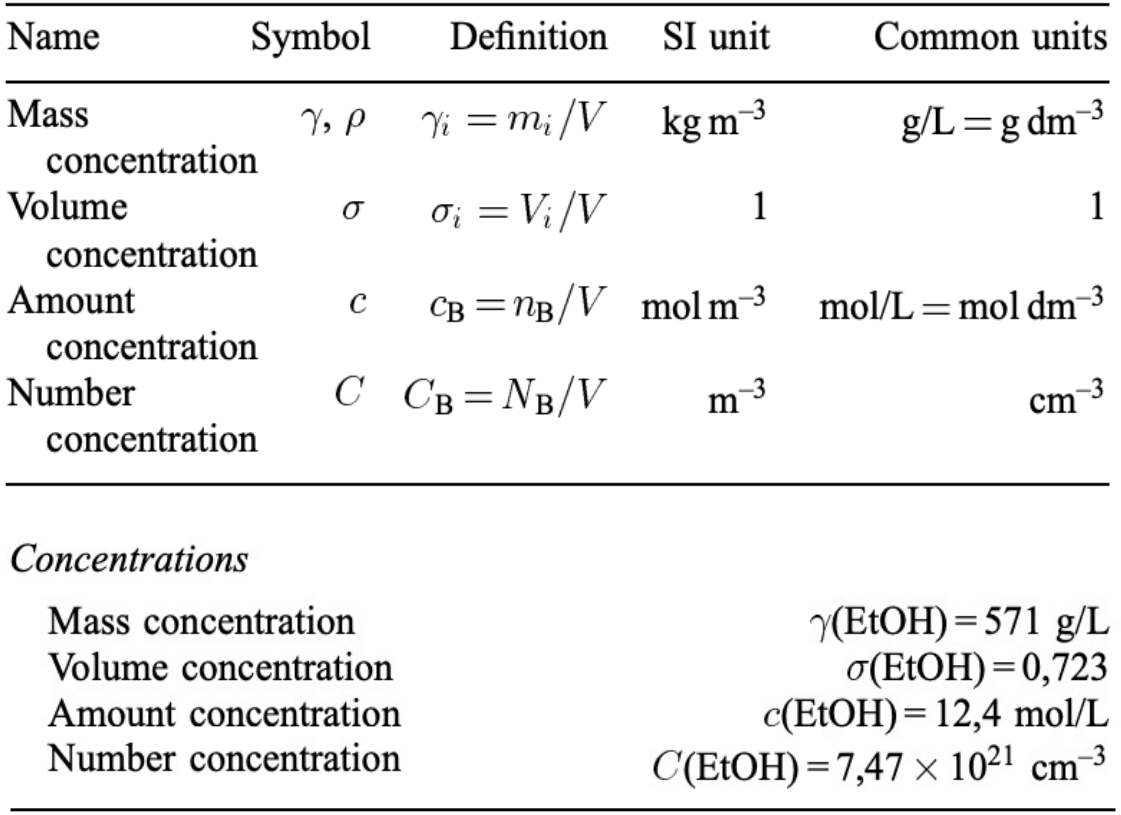
Exchangeable

- are the metadata necessary for finding, accessing, retrieving and processing exposed to APIs via registries, repositories and other information systems?

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

Reliable & Interpretable

- ✓Samples
  - ✓ Chemical composition
  - ✓ Physical state
- ✓Quantities
  - ✓ Equation, symbol, units
  - ✓ Rules (variables, constraints, dependencies)
- ✓Measurements
  - ✓ Principle, method, procedure
  - ✓ Conditions
- ✓Uncertainty
- ✓Provenance



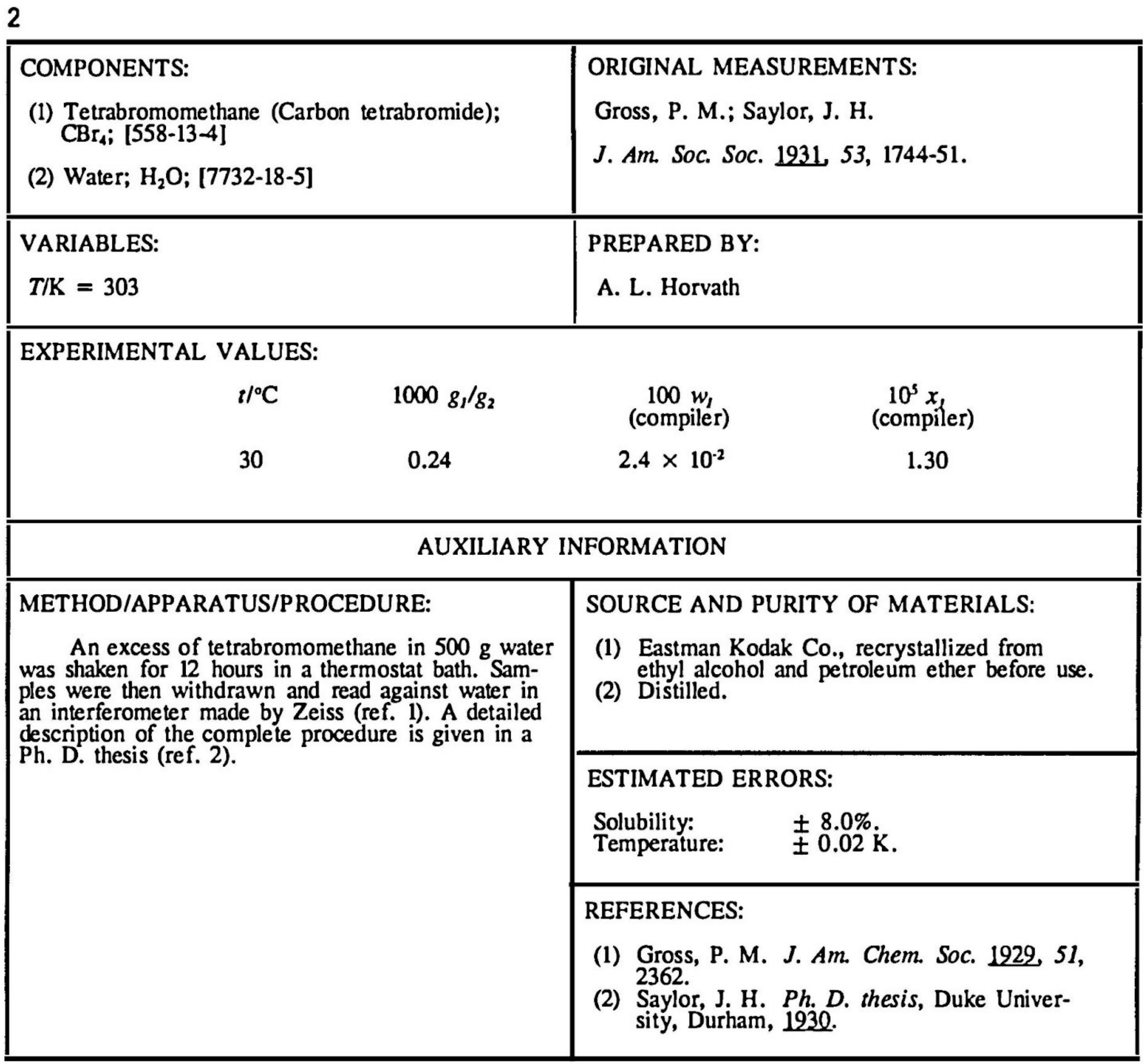
Name	Symbol	Definition	SI unit	Common units
Mass concentration	γ, ρ	$\gamma_i = m_i/V$	$\mathrm{kg\,m^{-3}}$	$\mathrm{g/L = g\,dm^{-3}}$
Volume concentration	σ	$\sigma_i = V_i/V$	1	1
Amount concentration	c	$c_\mathrm{B} = n_\mathrm{B}/V$	$\mathrm{mol\,m^{-3}}$	$\mathrm{mol/L = mol\,dm^{-3}}$
Number concentration	C	$C_\mathrm{B} = N_\mathrm{B}/V$	$\mathrm{m^{-3}}$	$\mathrm{cm^{-3}}$

Concentrations

Mass concentration	$\gamma(\mathrm{EtOH}) = 571\ \mathrm{g/L}$
Volume concentration	$\sigma(\mathrm{EtOH}) = 0{,}723$
Amount concentration	$c(\mathrm{EtOH}) = 12{,}4\ \mathrm{mol/L}$
Number concentration	$C(\mathrm{EtOH}) = 7{,}47 \times 10^{21}\ \mathrm{cm^{-3}}$

Cvitaˇs (1996) doi.org/10.1088/0026-1394/33/1/5

60_3



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.

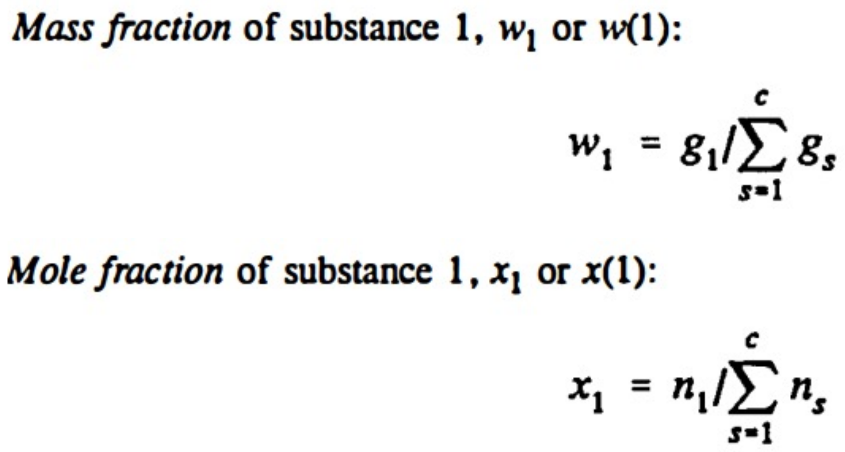
ESTIMATED ERRORS:

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.

Expression of chemical data

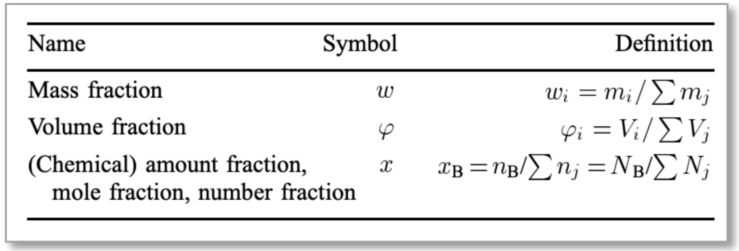


Mass fraction of substance 1, w_1 or $w(1)$:

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

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

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



Name	Symbol	Definition
Mass fraction	w	$w_i = m_i/\sum m_j$
Volume fraction	φ	$\varphi_i = V_i/\sum V_j$
(Chemical) amount fraction, mole fraction, number fraction	x	$x_B = n_B/\sum n_j = N_B/\sum N_j$

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

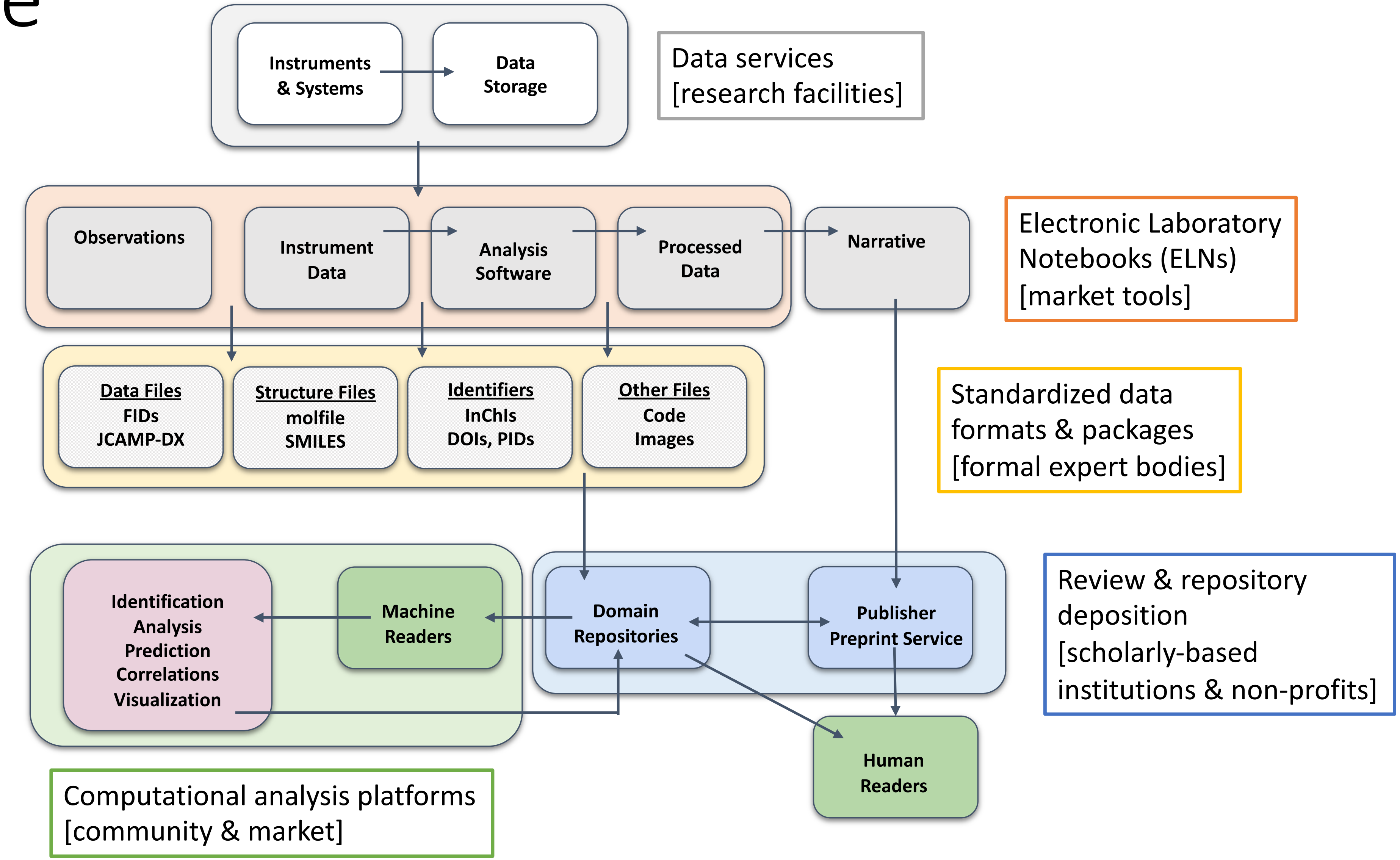
Processable

Chemical entities

Chemical data formats

Semantic terminologies

Data models

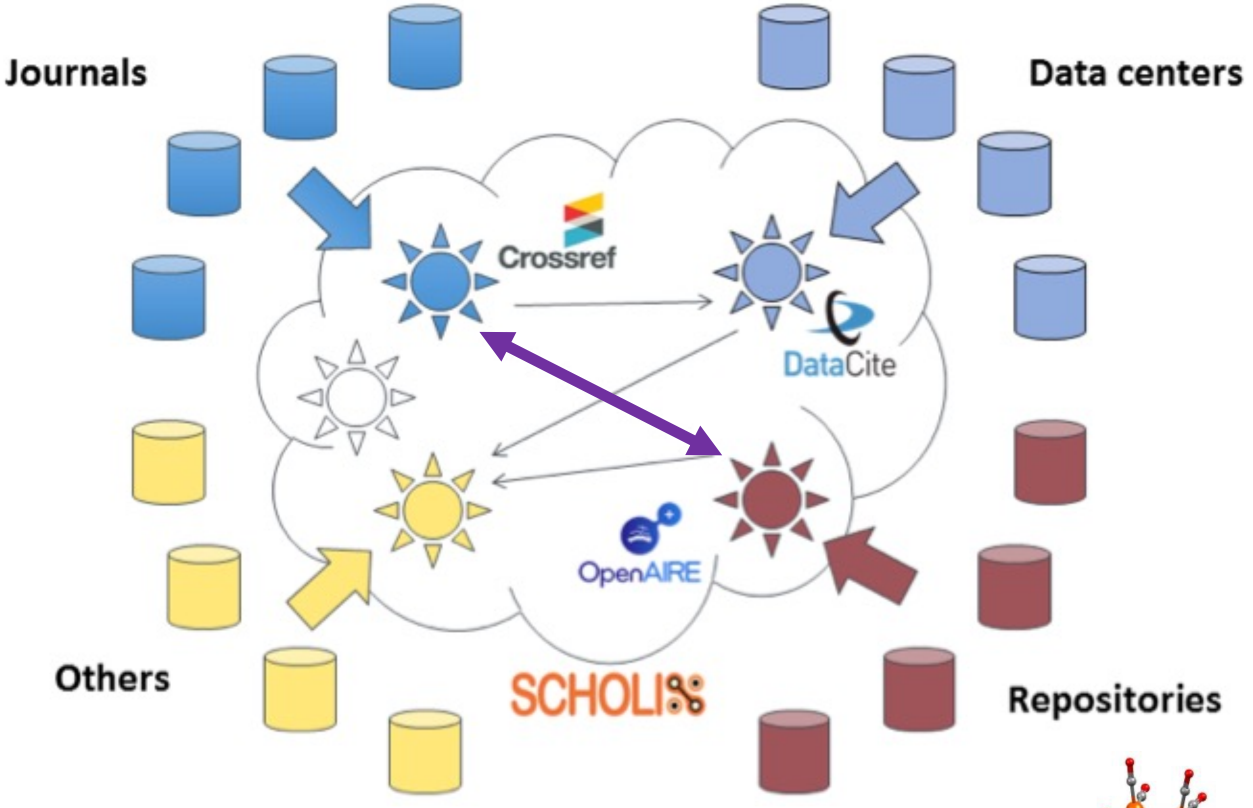


Exchangeable

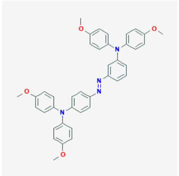
- Domain metadata in APIs and DOIs
  - PIDs and notations
- Chemical entities
  - Searching
  - Integration
- Navigating differences and ambiguity
  - Resolving
  - Mapping

Windhager, et al. (2007)
Eu. J. In. Chem. 4462
DOI: 10.1002/ejic.200700465

Adapted from Burton et al. (2017)
doi.org/10.1045/january2017-burton



https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/inchikey/YFRZGCZAMIBHIS-UHFFFAOYSA-N/png



Windhager, et al. (2008)
CCDC: 640959
DOI: 10.5517/ccphz37

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

Integrating data across domains

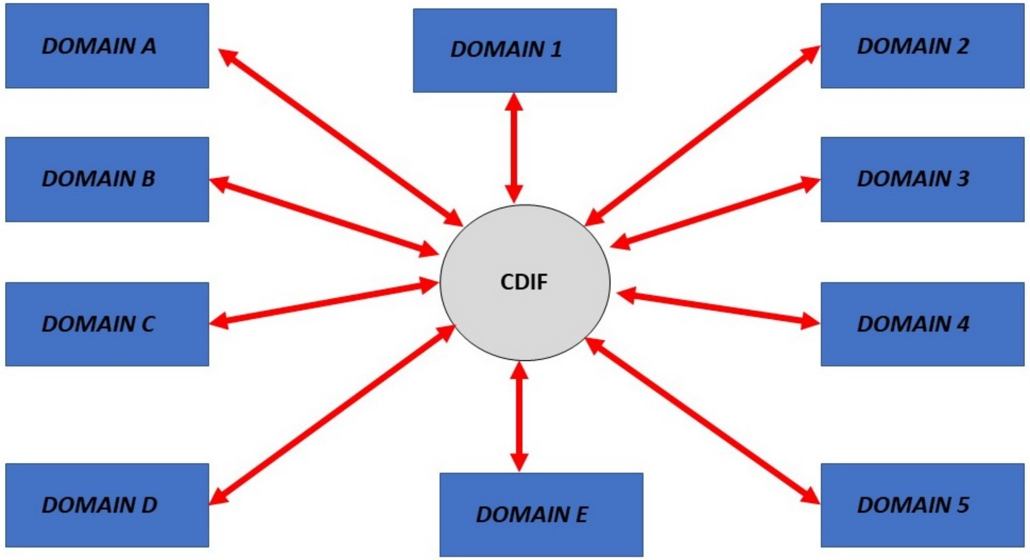
Chemical substance: integration by chemical identification
➥ *standard chemical identifier*

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

WorldFAIR: data standards for digital reuse

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

RDA RESEARCH DATA ALLIANCE

Digital motifs of scientific standards

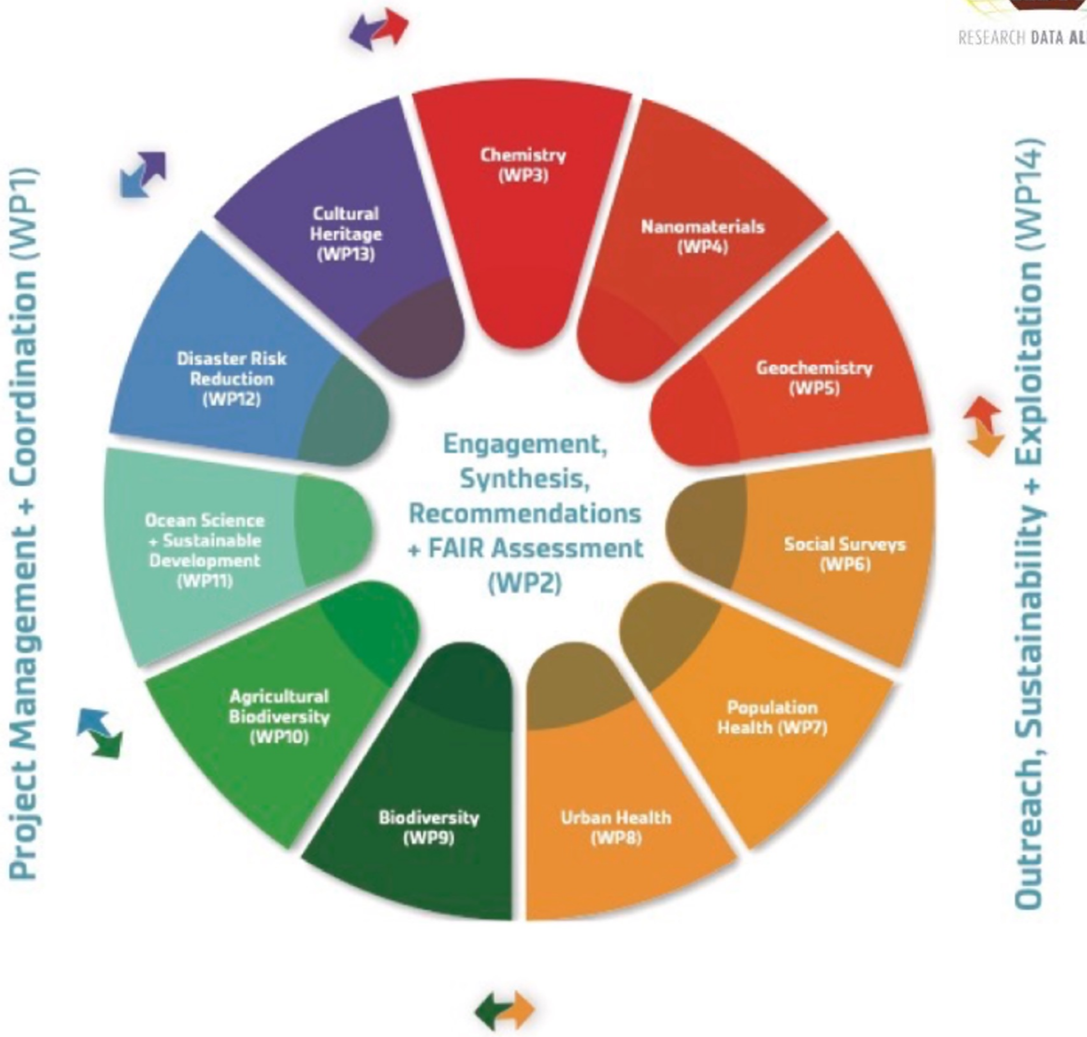
- Aligned with FAIR principles
- Aligned with common high-level protocols and architectures in the cloud

Best practice

- How to use in data management tools and workflows
- Guidance for policy development

Engagement with broader community

- Case studies in chemistry and neighboring disciplines
- Modeling data integration



Project Management + Coordination (WP1)

Chemistry (WP3)

Nanomaterials (WP4)

Geochemistry (WP5)

Social Surveys (WP6)

Population Health (WP7)

Urban Health (WP8)

Biodiversity (WP9)

Agricultural Biodiversity (WP10)

Ocean Science + Sustainable Development (WP11)

Disaster Risk Reduction (WP12)

Cultural Heritage (WP13)

Engagement, Synthesis, Recommendations + FAIR Assessment (WP2)

Outreach, Sustainability + Exploitation (WP14)

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.

IUPAC standard definitions and properties

<table>
<tr><th>Chemical representation</th><th>Chemical terminology</th><th>Chemical properties</th></tr>
<tr><td>• Nomenclature
  • Blue Book (organic)
  • Red Book (inorganic)
  • Purple Book (polymer)
• Graphical representation (structures, stereo, reactions)</td><td>• Orange Book (analytical)
• Silver Book (clinical)
• White Book (biochemical)
• Green Book (physical)</td><td>• Periodic Table (CIAAW tables)
• Solubility Data Series
• Atmospheric kinetics datasheets
• Polymerization kinetics dataset
• Stability constants dataset</td></tr>
<tr><td colspan="3"><b>Machine-processable <i>(to some degree)</i></b></td></tr>
<tr><td>• InChI notations
  • InChIKey
  • RInChI
  • <i>MInChI</i>
  • <i>NInChI</i>
• <i>SMILES+ notation</i>
• HELM notation
  • <i>Glycans notation</i></td><td>• Gold Book (compendium)
• NPU terminology for clinical chemistry
• <i>Green book digital quantities & symbols</i>
• <i>DRUM digital units</i></td><td>• JCAMP-DX spectra format
• ThermoML format
• <i>AIF adsorption format</i>
• <i>FAIRSpec metadata principles</i>
• <i>MAPT metadata schema</i>
• <i>Solubility metadata schema</i>
• <i>Dissociation constants dataset</i>
• <i>Atmospheric kinetics dataset</i>
• <i>Polymerization kinetics database</i></td></tr>
</table>

IUPAC standards: analog to digital workflows

Users & applications in data context

- Researchers documenting and reporting data
- Repositories aggregating chemical data
- Large databases of chemical substances
- Cheminformatics toolkits
- Chemical drawing and naming programs
- Electronic lab notebooks
- Modelers and data scientists
- Other developers and enablers
- Digital data projects in cognate disciplines

Gaps & challenges

- Gaps: FAIR access to evaluated property datasets, quantity models, semantic classifications
- Challenges: validation, harmonization provenance, licensing, sustainable development, outreach & adoption support

How-to-use support

FAIR Chemistry Guidance

bit.ly/IUPACDigitalRecommend

FAIR Chemistry Training Cookbook

bit.ly/CookFAIR

FAIR Chemistry Protocol Services

bit.ly/ProtServices

Sample prototypes

Community challenges in chemistry

Consumers

+

Curators

Where are our data?

- Are they sustainably hosted and curated?

Can we establish cross-community consensus around data standards?

- Best practices? Adoption? Validation?

What are we willing to pay for? (to sustain data & standards curation)

- Workflow tools? Value add AI & modeling tools?

Can we enable open, crowd funded and supported tools that evolve with the needs of the community?

How are we introducing digital data principles and management to young and early career chemical professionals and other scientists?

Collaborations

WorldFAIR

Digital exchange already predominates scientific communication and is rife for improvement and advancement – lets collaborate!

- 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

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.