What is digital IUPAC? Tools for an increasingly digital research culture

Ms. Leah McEwen
Cornell University

Dr. Fatima Mustafa
Texas A&M San Antonio

Dr. Ian Bruno
Cambridge Crystallographic Data Centre

ChemVoices

JOIN US ON: May 31 @ 12:00 pm (EST)
REGISTER TODAY: https://bit.ly/WorldFAIR

ChemVoices

- A partnership between IUPAC and IYCN and was created to showcase the talents and impact of early-career scientists worldwide.
- It is a platform to discuss issues that are relevant and of immediate concern to early-career scientists.
- Today's Webinar: **Q&A** to introduce **Digital IUPAC** for younger chemists

Leah McEwen
Cornell University
IUPAC CPCDS
USA

- Leah holds an **M.S. in Nutritional Biochemistry** from Cornell University and an **M.L.S. in Library and Information Science** from Emporia State University.
- She has been the **Chemistry Librarian at Cornell since** 1999, supporting information discovery and data management needs for the research community.
- She is an active contributor to national and international data initiatives, **founding the Chemistry Research Data Interest Group (DIG Chemistry)** of the Research Data Alliance (RDA) and organizing thematic programs on chemical data standards.
- She is currently **chair of the Committee on Publications and Cheminformatics Data Standards of the International Union of Pure and Applied Chemistry (IUPAC)**, responsible for the design and implementation of digital standards and **lead on the WorldFAIR Chemistry** project to advance FAIR data practices in Chemistry.

Ian Bruno
CCDC
IUPAC CPCDS
UK

- Ian has a BSc in Chemistry from Durham University and a PhD in **Information Science** from the University of Sheffield. He has worked at the Cambridge Crystallographic Data Centre (**CCDC**) since 1993 in a variety of roles covering CCDC's activities in software development, scientific research and data management.

- In his current role as **Director of Data Initiatives**, he is responsible for advancing strategies that will make CCDC's data more discoverable and reusable by the scientific community. A specific focus is on the adoption and development of community principles, standards and sustainability models to support these aims.

- Ian is an active participant in a range of global initiatives including the **Research Data Alliance, the International Union of Crystallography, the InChI Trust and the International Union of Pure and Applied Chemistry**.

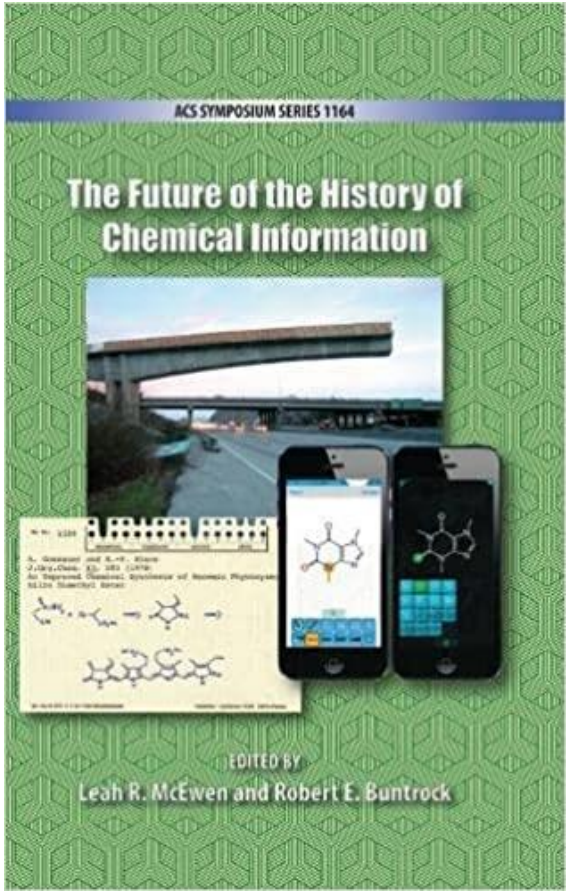
- He is the current **lead of the WorldFAIR Chemistry sub-committee** developing "**Reporting Guidance**".

Poll #1

How do you best describe yourself?

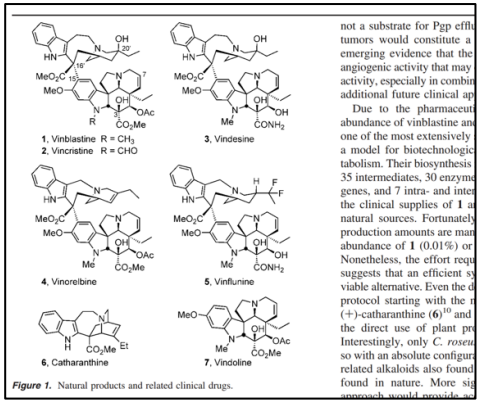
- A younger chemist (below 35 yrs or within 6 years of your last degree or training)
- A chemist
- Chemical data user but not a chemist

Archives to Machine-readable representation?



https://doi.org/10.1021/bk-2014-1164

Traditional reporting practice for spectral data



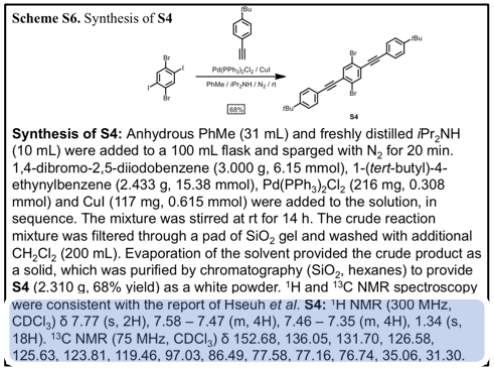
1, Vinblastine R = CH_3
2, Vincristine R = CHO
3, Vindesine
4, Vinorelbine
5, Vinflunine
6, Catharanthine
7, Vindoline

Figure 1. Natural products and related clinical drugs.

not a substrate for Pgp efflu
tumors would constitute a
emerging evidence that the
angiogenic activity that may
activity, especially in combin
additional future clinical ap
Due to the pharmaceut
abundance of vinblastine an
one of the most extensively
a model for biotechnologic
tabolism. Their biosynthesis
35 intermediates, 30 enzyme
genes, and 7 intra- and inter
the clinical supplies of **1** a
natural sources. Fortunately
production amounts are mar
abundance of **1** (0.01%) or
Nonetheless, the effort requ
suggests that an efficient sy
viable alternative. Even the d
protocol starting with the r
(+)-catharanthine (**6**)[10] and
the direct use of plant pr
Interestingly, only *C. roseu*
so with an absolute configur
related alkaloids also found
found in nature. More sig
approach would provide ac

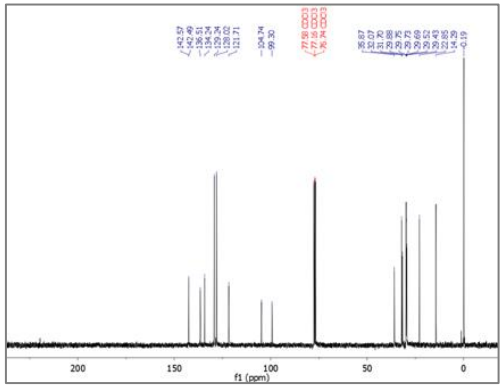
Data most often static and embedded in PDF articles and SI files.

Chemical structures & reactions also.



Scheme S6. Synthesis of **S4**

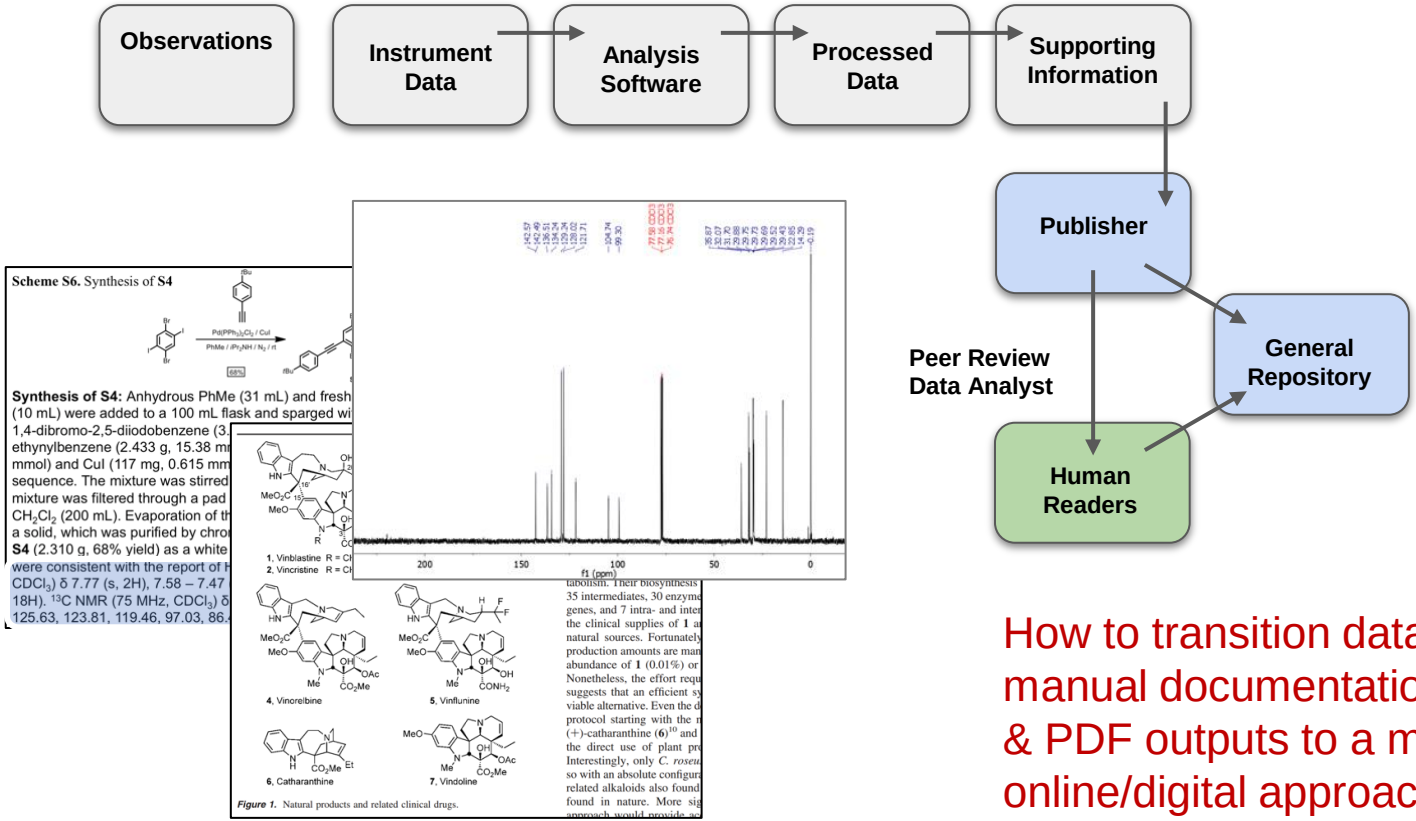
Synthesis of S4: Anhydrous PhMe (31 mL) and freshly distilled iPr$_2$NH (10 mL) were added to a 100 mL flask and sparged with N_2 for 20 min. 1,4-dibromo-2,5-diiodobenzene (3.000 g, 6.15 mmol), 1-(*tert*-butyl)-4-ethynylbenzene (2.433 g, 15.38 mmol), $Pd(PPh_3)_2Cl_2$ (216 mg, 0.308 mmol) and CuI (117 mg, 0.615 mmol) were added to the solution, in sequence. The mixture was stirred at rt for 14 h. The crude reaction mixture was filtered through a pad of SiO_2 gel and washed with additional CH_2Cl_2 (200 mL). Evaporation of the solvent provided the crude product as a solid, which was purified by chromatography (SiO_2, hexanes) to provide **S4** (2.310 g, 68% yield) as a white powder. ^{1}H and ^{13}C NMR spectroscopy were consistent with the report of Hseuh *et al.* **S4:** ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.77 (s, 2H), 7.58 – 7.47 (m, 4H), 7.46 – 7.35 (m, 4H), 1.34 (s, 18H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 152.68, 136.05, 131.70, 126.58, 125.63, 123.81, 119.46, 97.03, 86.49, 77.58, 77.16, 76.74, 35.06, 31.30.



Many challenges with this approach:

- Does not allow readers and indexers to easily reproduce structures
- Difficult to search & discover
- Difficult/impossible for software to interpret
- Limited ways to link chemical structures to data and external sources for further analysis

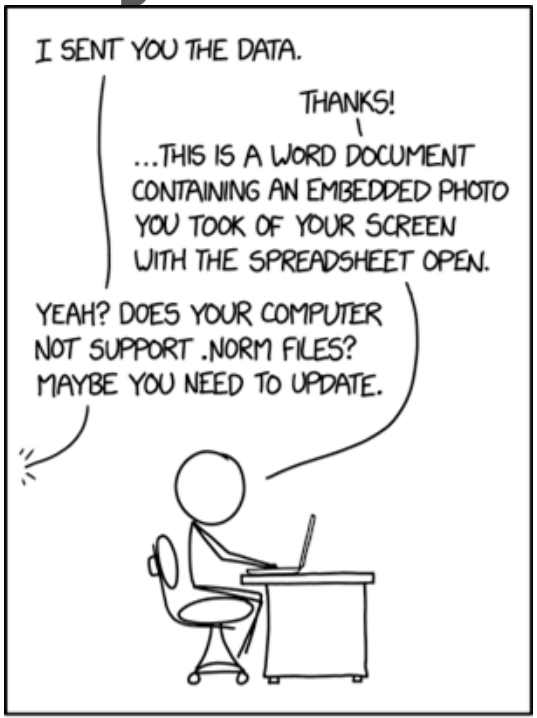
Current chemistry data preparation workflow



How to transition data from manual documentation & PDF outputs to a more online/digital approach?

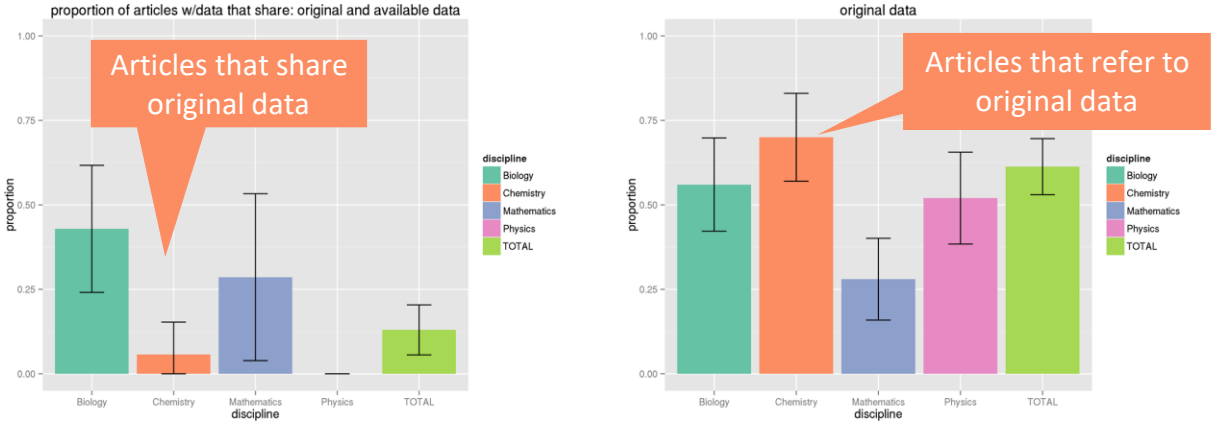
Data publication today

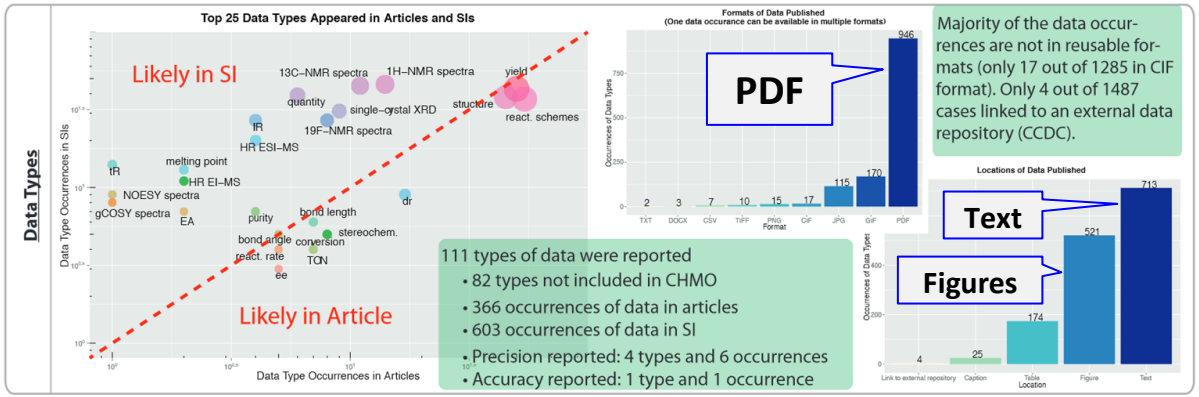
- Challenges
- Opportunities
- Needs



https://xkcd.com/2116

Chemists commonly “share” data…





Womack, Ryan P. (2015). Research Data in Core Journals in Biology Chemistry, Mathematics, and Physics. *PLOS One.* https://doi.org/10.1371/journal.pone.0143460

… but primarily in Supporting Information as a PDF or as a static image in the manuscript…

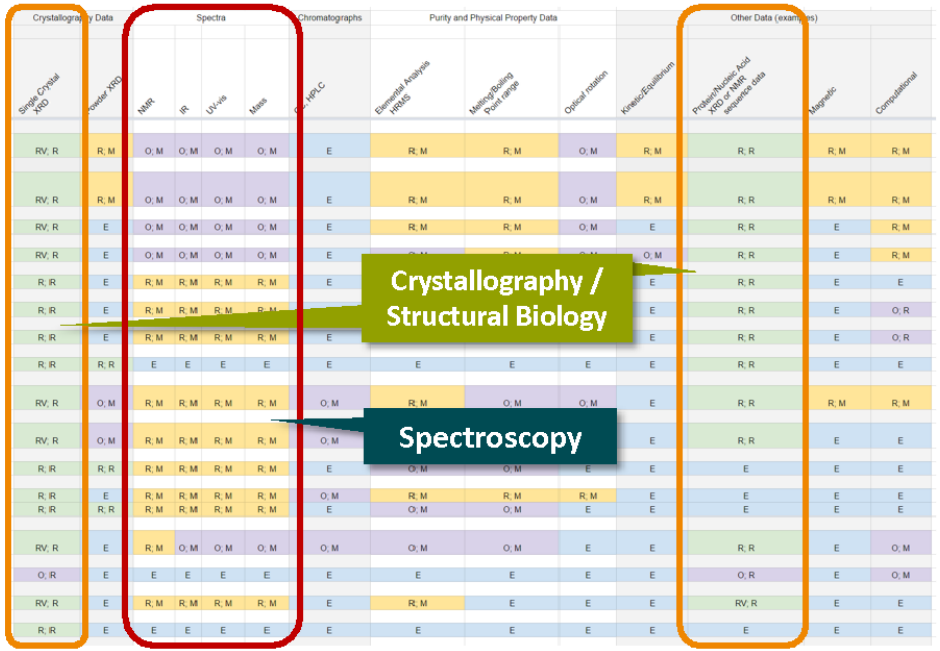
Thielen, Joanna, & Li, Ye. (2015). *Profiling common types of research data and methods published by organic synthesis chemists at the University of Michigan. Paper presented at the SLA 2015 Annual Conference & Info Expo, Boston, MA.* http://hdl.handle.net/2027.42/111832

CHMO = Chemical Methods Ontology

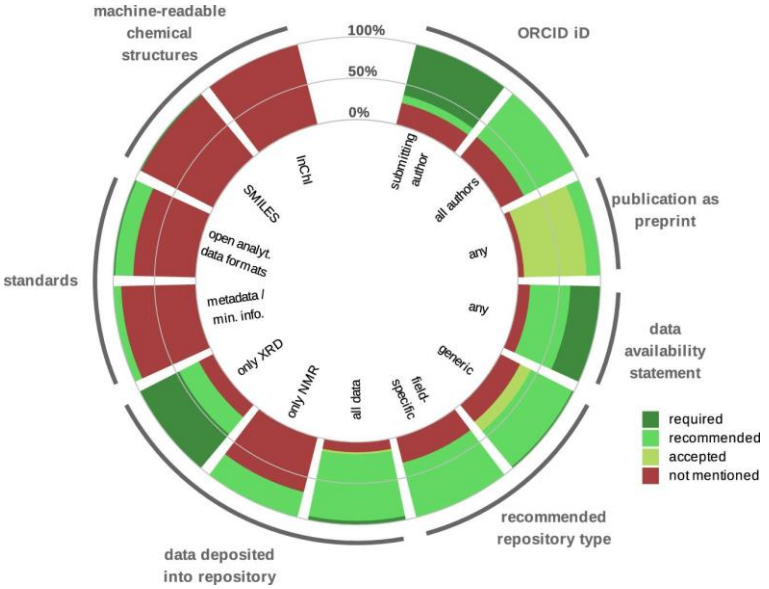
Chemistry journal data policies?

Different Chemistry Data Types

Different Chemistry Journals



- *Vincent Scalfani, RDA CRDIG Open Meeting, ACS Spring Meeting, San Francisco, March 2017*
  https://doi.org/10.6084/m9.figshare.8870144.v1

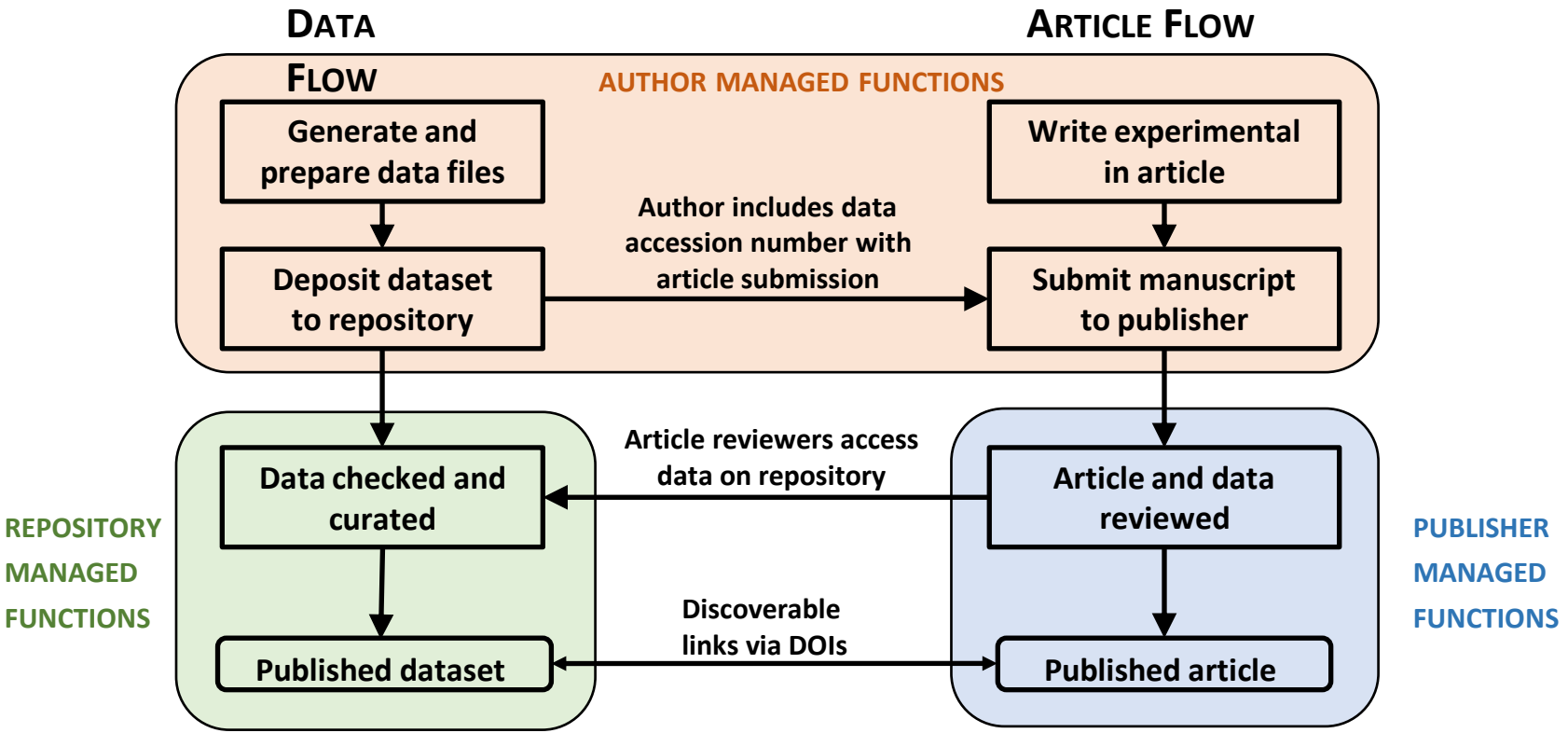


Parks *et al.* (2023) https://doi.org/10.1515/pac-2022-1001

Journal Policy

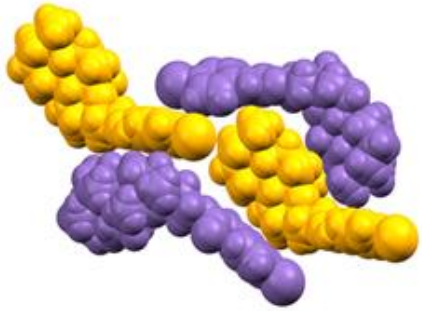
- Required in Repository
- Required in Manuscript
- Optional in Manuscript
- Not required

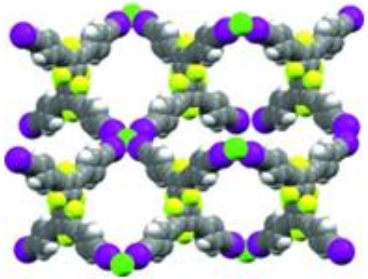
Preparing data files and manuscript in parallel

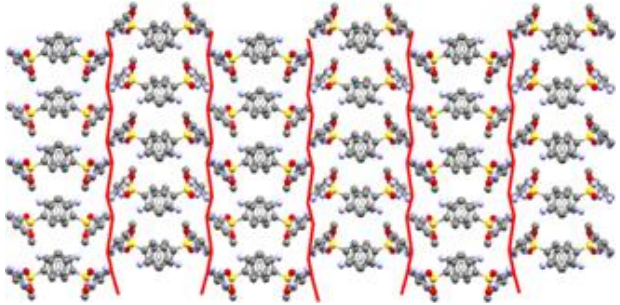


Data publication exemplars:

Learning from crystallography

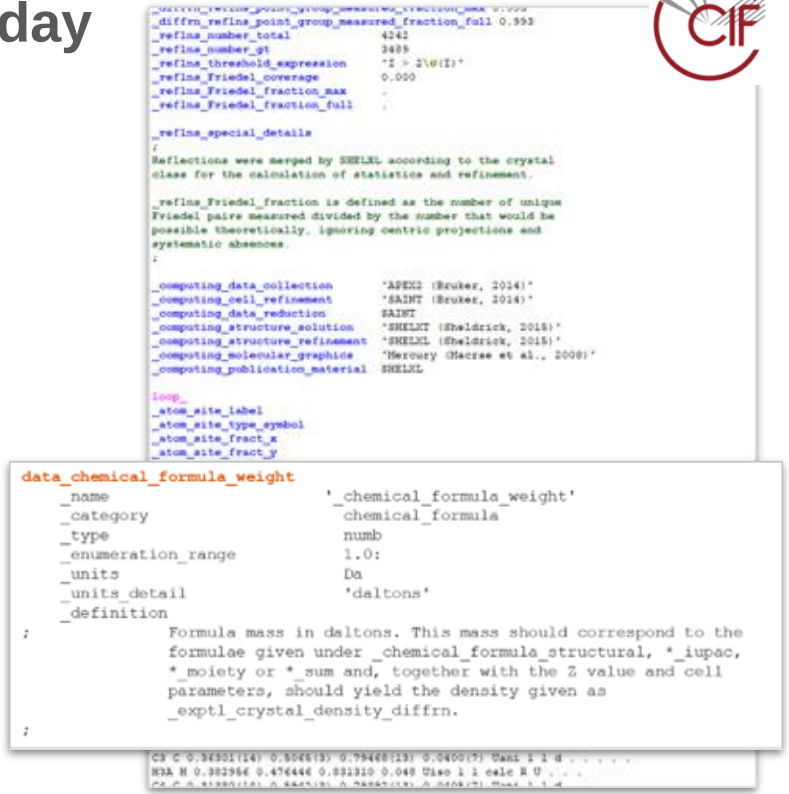






Data publication in crystallography today

- Around ~60,000 small molecule crystal structure datasets published annually
- Majority associated with one of ~16,000 journal articles although some published independently
- All made discoverable and accessible through indexed data resources such as the CSD



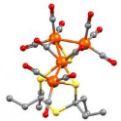
Key-enablers

- Standard (semantic) file format (CIF)
- Adoption of standards by software and tools
- Domain-oriented data repositories
- Joined up data repository / journal workflows
- Standard identifiers for research objects and people

Identifiers, Data Citation and Metadata

DataCite
FIND, ACCESS, AND REUSE DATA

- DOIs uniquely and persistently identify digital objects
- Typically accompanied by provenance and other metadata
- Foundation for formalising data citation and interoperability



doi:10.5517/CCPHZ37

DC1 Data Citation Principles

Data should be considered legitimate, citable products of research…

https://www.force11.org/datacitation

Dataset Publication

CCDC 610092: Experimental Crystal Structure Determination. **A. Crystallographer**, *Cambridge Crystallographic Data Centre* (2007)
http://dx.doi.org/10.5517/ccngvdb

DATA CITATION INDEX

Metadata associated with DOIs for datasets are made openly available for harvesting through DataCite APIs by others.

Scholeexplorer
The Data Literature Interlinking Service
SCHOLIX

ORCID IDs for Researchers

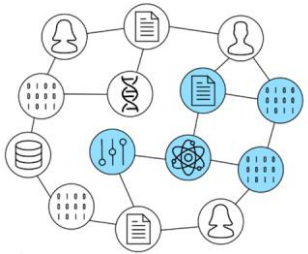
At least 50% of Crystallographers currently provide their ORCID ID

Andrew Bond
ORCID ID
https://orcid.org/0000-0002-1744-0489

Other uses of standard identifiers

- Organisations (ROR)
- Instruments (PIDINST)
- Software (DOIs)
- Samples/specimens (IGSN)
- Antibodies/organisms (RRIDs)
- Chemical structures (InChI)

Capturing **related Identifiers** in metadata is a key enabler of interoperability.



PID graph with digital objects connected by PIDs

https://www.project-freya.eu/en/blogs/blogs/the-pid-graph

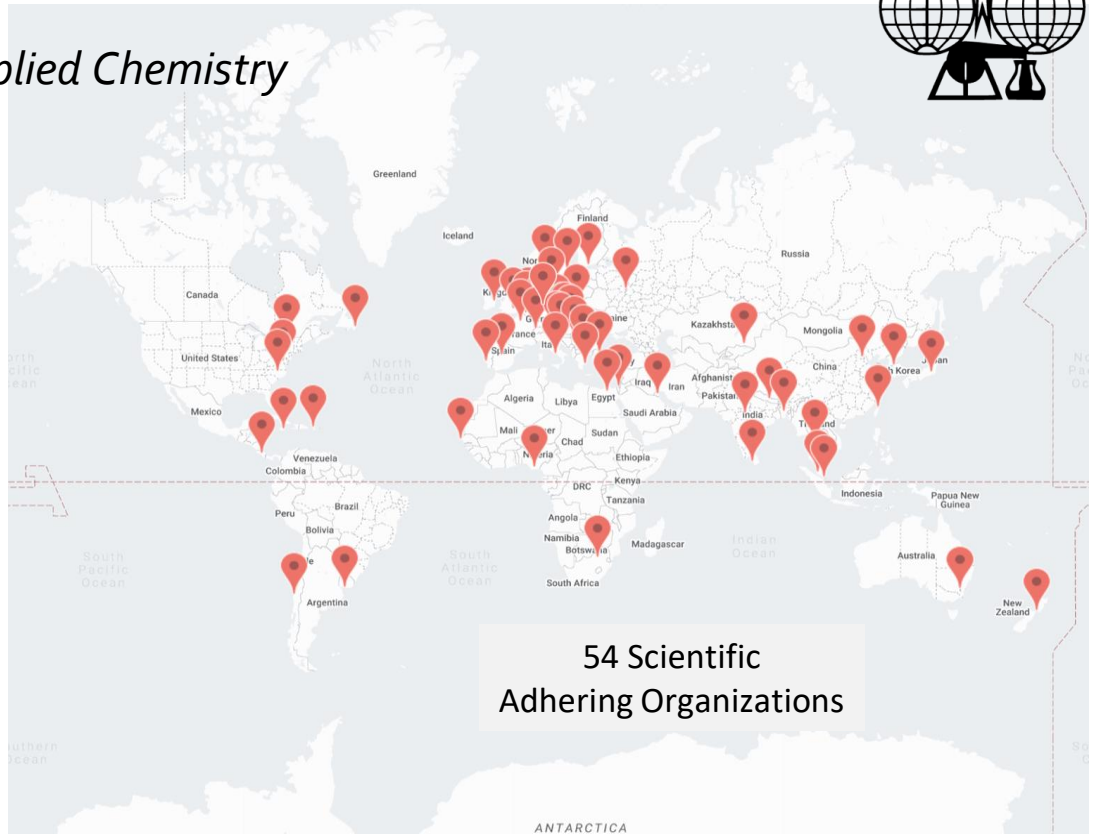
Tell us about IUPAC and why its mission is relevant

IUPAC Mission

International Union of Pure and Applied Chemistry

The global organization that provides **objective scientific expertise** and develops the **essential tools for the application and communication of chemical knowledge** for the benefit of humankind and the world.

- 2000 scientific experts worldwide, drawn from scientific societies
- Pure and applied – research, industry, policy, education
- Core values – neutral, open, transparent provenance, sustainable process, consensus

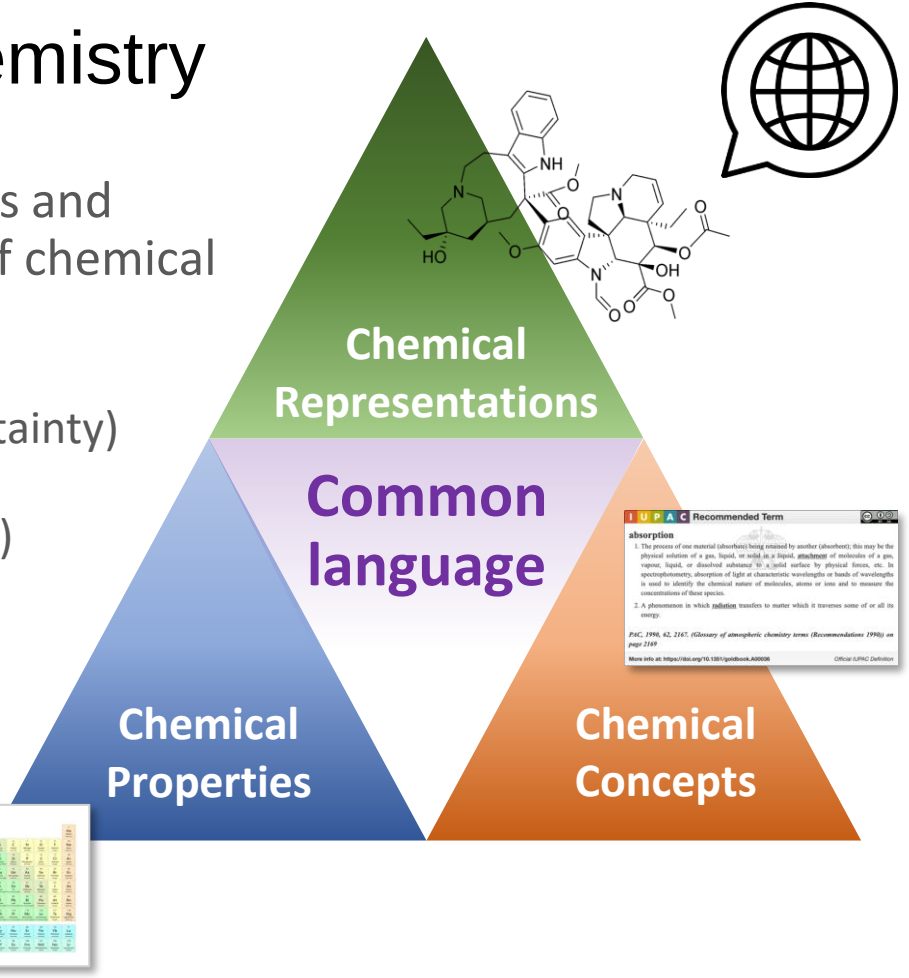


Common language for chemistry

IUPAC provides authoritative definitions and parameters for consistent expression of chemical data and information

- Quantities (i.e., units, symbols)
- Equations, models (e.g., systems, uncertainty)
- Notations (e.g., chemical substances)
- Terms (e.g., properties, processes, roles)

100 yrs of global consensus to develop & define a common and systematic language for chemistry

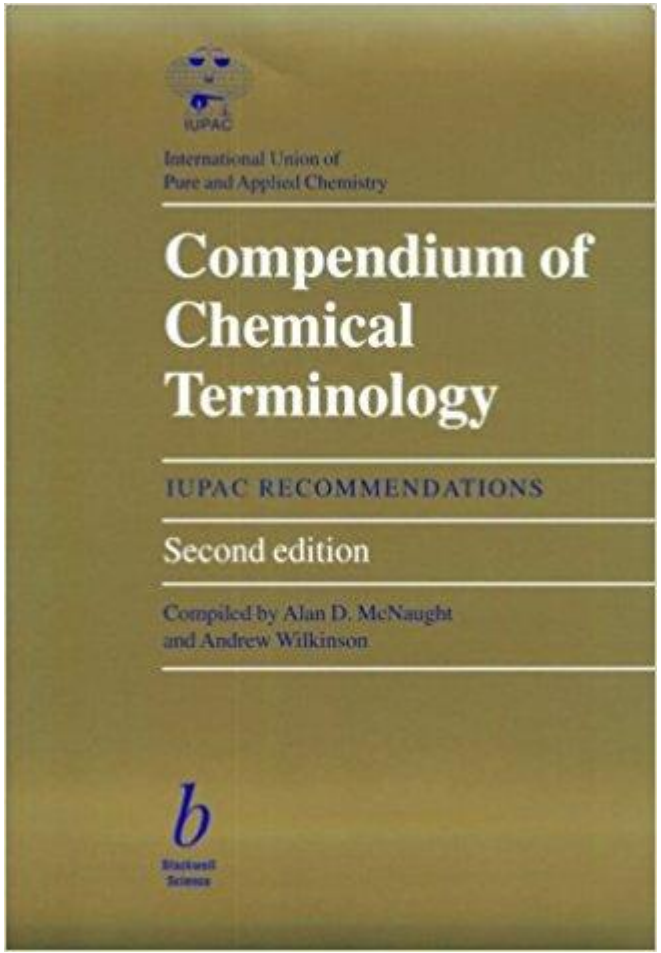


Poll #2

How familiar are you with the efforts of IUPAC to translate standards into the digital domain?

- Not at all familiar
- Not very familiar
- Somewhat familiar
- Familiar
- Very familiar

Tell us about IUPAC and digital standards



Migrating to formats that are actionable by computers?

IUPAC has refined the rules, logic & language of chemistry for trained chemists for over a century.

Computers are not chemists; they need the rules explained in machine logic and languages.

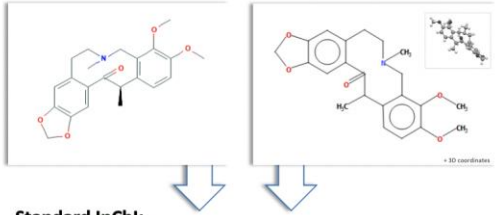
https://xkcd.com/1838

https://xkcd.com/927

InChI: IUPAC International Chemical Identifier

InChI TRUST

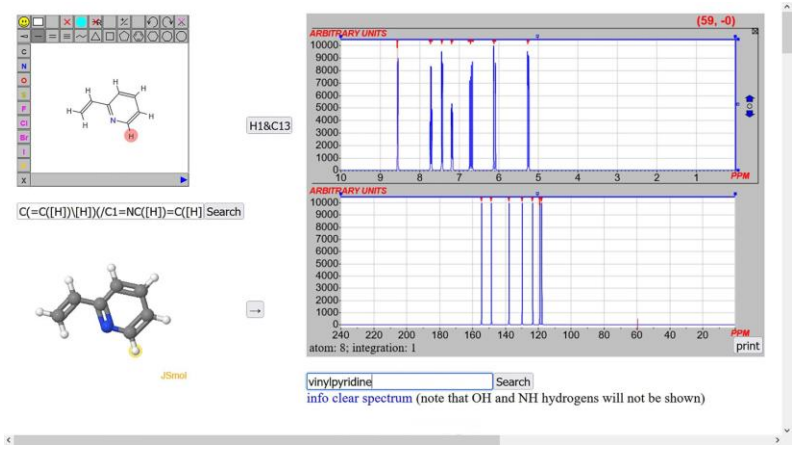
- Small molecule organics
- Reactions
- Mixtures
- *In progress:*
  - *Organometallics*
  - *Large molecules*
  - *Stereoisomers*
  - *Tautomers*
  - *Nanomaterials*



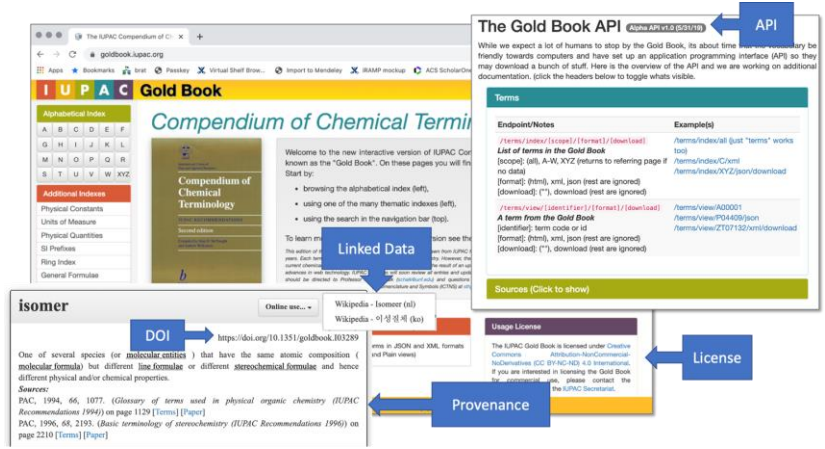
Standard InChI:
`InChI=1S/C22H25NO5/c1-13-15-5-6-18(25-3)22(26-4)17(15)11-23(2)8-7-14-9-19-20(28-12-27-19)10-16(14)21(13)24/h5-6,9-10,13H,7-8,11-12H2,1-4H3/t13-/m0/s1`

Standard InChIKey: `YRAXYGEEARNTF-ZDUSSCGKSA-N`

JCAMP-DX: IUPAC Spectral Data Exchange Format



Gold Book: IUPAC Formal Chemical Terminology



Poll #3

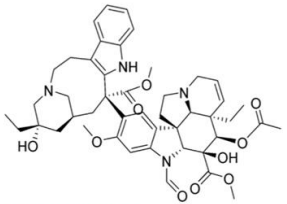
Had you previously heard about digital structure representation formats and identifiers (for example, InChI, SMILES, MOLfile)**?**

1. Yes, and I use one of them at least
2. Yes, but never used any
3. Not at all
4. Other (comment in few words)

Structure identity and representation

- Why is it important?
- What do we need?

Vincristine
(trivial name)



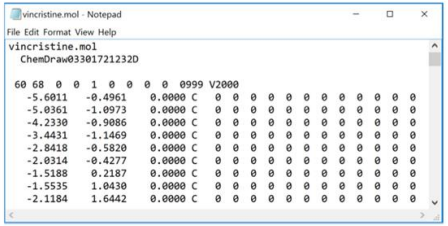
IUPAC name - standardized nomenclature

(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

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



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

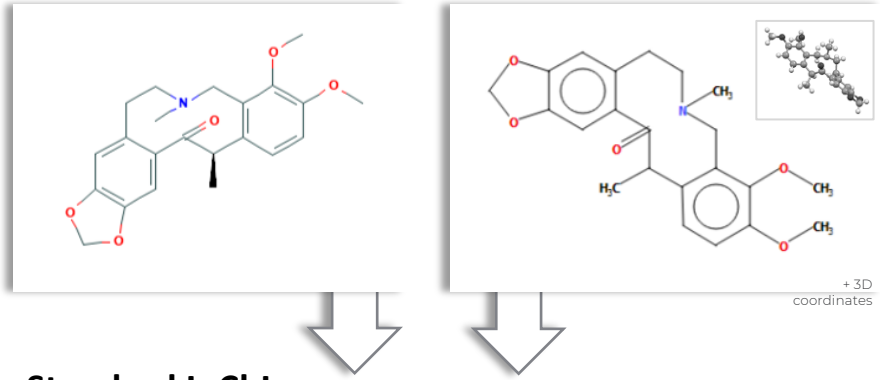
InChIKey: OGWKCGZFUXNPDA-XQKSVPLYSA-N

Generally well-defined for discrete organic molecules - efforts ongoing to make reliable for other chemical entities

InChI: The IUPAC International Chemical Identifier

InChITRUST

https://www.inchi-trust.org/



- A non-proprietary identifier for chemical substances
- Provides a standard way to encode molecular information
- Facilitates the search for such information in databases and on the web
- Enables linking of diverse data compilations.

Standard InChI:

InChI=1S/C22H25NO5/c1-13-15-5-6-18(25-3)22(26-4)17(15)11-23(2)8-7-14-9-19-20(28-12-27-19)10-16(14)21(13)24/h5-6,9-10,13H,7-8,11-12H2,1-4H3/t13-/m0/s1

Standard InChIKey: YRAXYDGEEARNTF-ZDUSSCGKSA-N

InChIKey is a one way hash of the InChI String

InChI layers separately describe formula, connectivity, hydrogen atoms, stereochemistry etc.

InChI

https://www.inchi-trust.org/inchi-videos/

https://www.youtube.com/watch?v=qrCqJ0o4jGs&t=9s

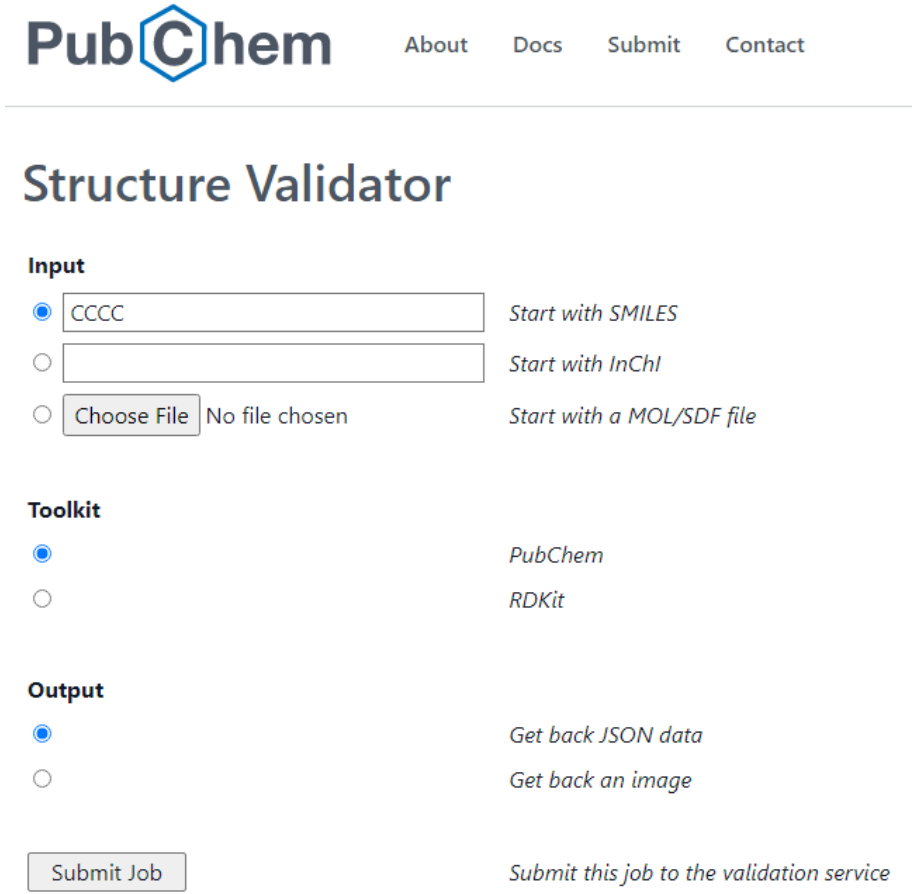
How do I know my structure representation is correct?

Chemistry Protocols

IUPAC WORLDFAIR CHEMISTRY D3.3

D3.3 FAIR Chemistry Protocol Services

bit.ly/ProtServices



Prototype Demonstrator

PubChem About Docs Submit Contact

Prototype Demonstrator

Structure Validator

Input

- (•) CCCC — *Start with SMILES*
- ( ) — *Start with InChI*
- ( ) Choose File No file chosen — *Start with a MOL/SDF file*

Toolkit

- (•) *PubChem*
- ( ) *RDKit*

Output

- (•) *Get back JSON data*
- ( ) *Get back an image*

Submit Job — *Submit this job to the validation service*

Is this structure OK?

PubChem About Docs Submit Contact

Structure Validator

Input

- (•) CCCC — *Start with SMILES*
- ( ) — *Start with InChI*
- ( ) Choose File No file chosen — *Start with a MOL/SDF file*

Toolkit

- (•) *PubChem*
- ( ) *RDKit*

Output

- (•) *Get back JSON data*
- ( ) *Get back an image*

Submit Job — *Submit this job to the validation service*

Is this structure OK?


```
{
 "Result": {
  "Message": "Structure is valid",
  "Statistics": [
   {
    "Type": "DefinedAtomStereo",
    "Value": "0"
   },
   {
    "Type": "UndefinedAtomStereo",
    "Value": "0"
   },
   {
    "Type": "DefinedBondStereo",
    "Value": "0"
   },
   {
    "Type": "UndefinedBondStereo",
    "Value": "0"
   },
   {
    "Type": "HeavyAtoms",
    "Value": "4"
   },
   {
    "Type": "IsotopeAtoms",
    "Value": "0"
   },
   {
    "Type": "CovalentUnits",
    "Value": "1"
   }
  ]
 }
}
```

PubChem About Docs Submit Contact

Prototype Demonstrator

Structure Validator

Input

C([C@@H]1[C@H]([C@@H]([C@H](C(O1 *Start with SMILES*

Start with InChI

Choose File No file chosen *Start with a MOL/SDF file*

Toolkit

PubChem

RDKit

Output

Get back JSON data

Get back an image

Submit Job *Submit this job to the validation service*

Does it have the expected number of (un)defined stereocenters?

PubChem About Docs Submit Contact

Structure Validator

Input

- C([C@@H]1[C@H]([C@@H]([C@H](C(O1 — *Start with SMILES*
- *Start with InChI*
- Choose File No file chosen — *Start with a MOL/SDF file*

Toolkit

- *PubChem*
- *RDKit*

Output

- *Get back JSON data*
- *Get back an image*

Submit Job — *Submit this job to the validation service*

Does it have the expected number of (un)defined stereocenters?


```
{
 "Result": {
  "Message": "Structure is valid",
  "Statistics": [
   {
    "Type": "DefinedAtomStereo",
    "Value": "4"
   },
   {
    "Type": "UndefinedAtomStereo",
    "Value": "1"
   },
   {
    "Type": "DefinedBondStereo",
    "Value": "0"
   },
   {
    "Type": "UndefinedBondStereo",
    "Value": "0"
   },
   {
    "Type": "HeavyAtoms",
    "Value": "12"
   },
   {
    "Type": "IsotopeAtoms",
    "Value": "0"
   },
   {
    "Type": "CovalentUnits",
    "Value": "1"
   }
  ]
 }
}
```


Prototype Demonstrator

PubChem About Docs Submit Contact

Prototype Demonstrator

Structure Validator

Input

- C([C@@H]1[C@H]([C@@H]([C@H](C(O1 — *Start with SMILES*
- *Start with InChI*
- Choose File No file chosen — *Start with a MOL/SDF file*

Toolkit

- *PubChem*
- *RDKit*

Output

- *Get back JSON data*
- *Get back an image*

Submit Job — *Submit this job to the validation service*

Does a drawing look like what the chemist expects?

PubChem About Docs Submit Contact

Prototype Demonstrator

Structure Validator

Input

- C([C@@H]1[C@H]([C@@H]([C@H](C(O1 — *Start with SMILES*
- *Start with InChI*
- Choose File No file chosen — *Start with a MOL/SDF file*

Toolkit

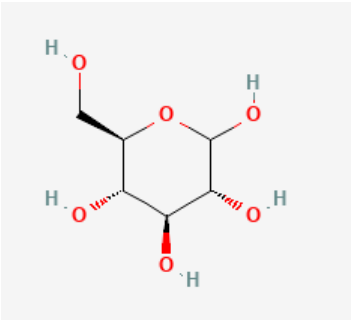
- *PubChem*
- *RDKit*

Output

- *Get back JSON data*
- *Get back an image*

Submit Job — *Submit this job to the validation service*

Does a drawing look like what the chemist expects?



PubChem About Docs Submit Contact

Prototype Demonstrator

Structure Validator

Input

- (•) CC(C)(C)(C)CC — *Start with SMILES*
- ( ) — *Start with InChI*
- ( ) Choose File No file chosen — *Start with a MOL/SDF file*

What if something is wrong?

Toolkit

- (•) *PubChem*
- ( ) *RDKit*

Output

- (•) *Get back JSON data*
- ( ) *Get back an image*

Submit Job — *Submit this job to the validation service*

PubChem About Docs Submit Contact

Prototype Demonstrator

Structure Validator

Input

- (●) CC(C)(C)(C)CC — *Start with SMILES*
- ( ) — *Start with InChI*
- ( ) Choose File No file chosen — *Start with a MOL/SDF file*

Toolkit

- (●) *PubChem*
- ( ) *RDKit*

Output

- (●) *Get back JSON data*
- ( ) *Get back an image*

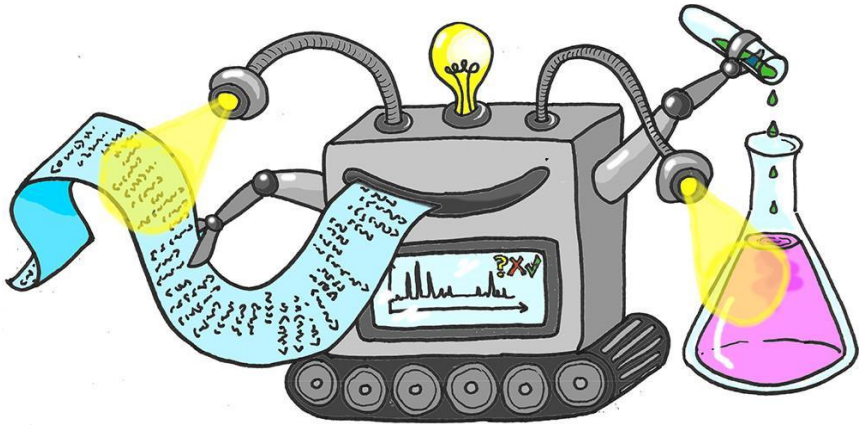
Submit Job — *Submit this job to the validation service*

What if something is wrong?


```
{
 "Fault": {
  "Code": "Invalid",
  "Message": "Structure is not valid",
  "Details": [
   "Record 0: Warning: Detected illegal valence for element \"C\": 5 sigma bonds, 0 pi bonds, 0 charge",
   "Exception: Valence validation failed"
  ]
 }
}
```

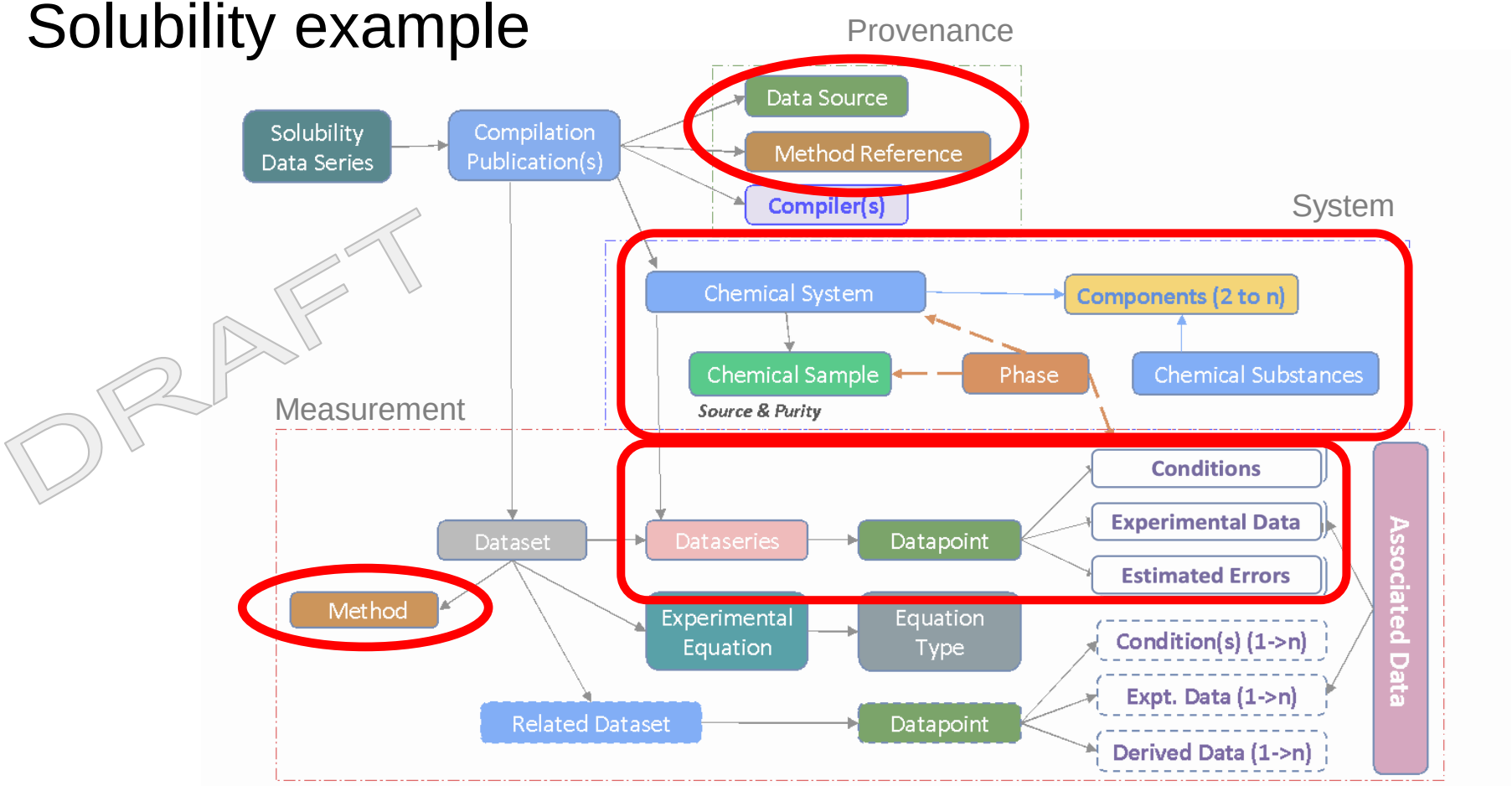
Describing your data

- Samples
- Quantities
- Conditions



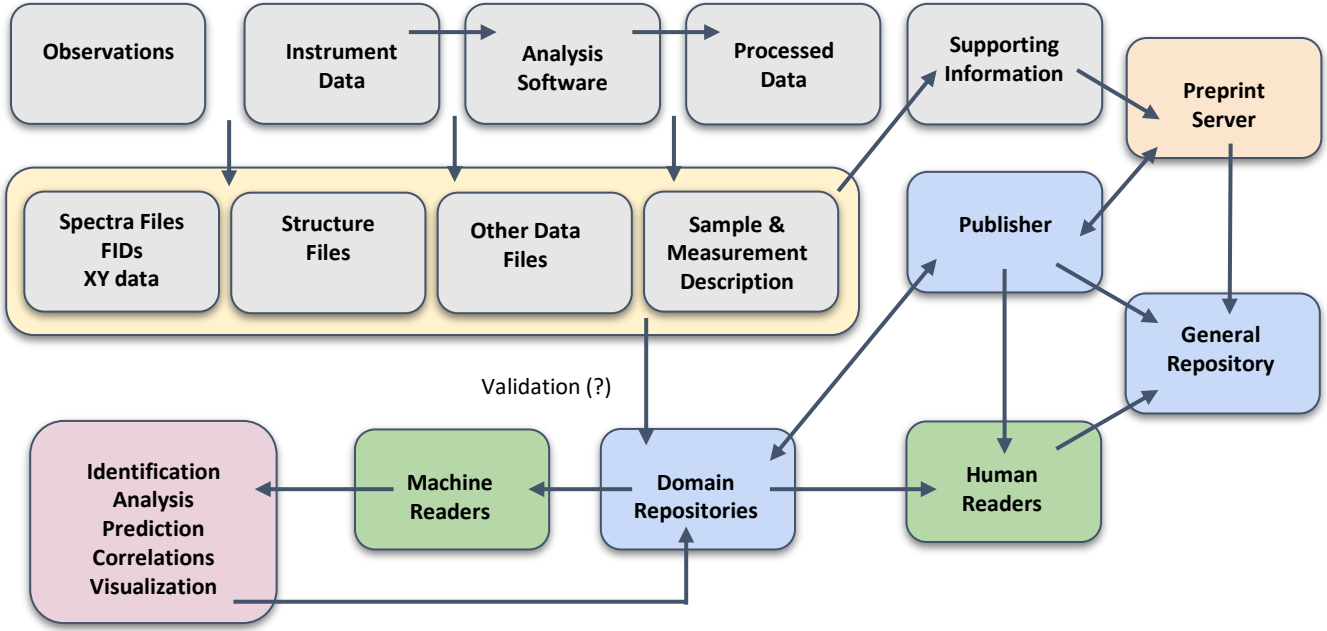
DOI: 10.1021/acscentsci.8b00176

Solubility example

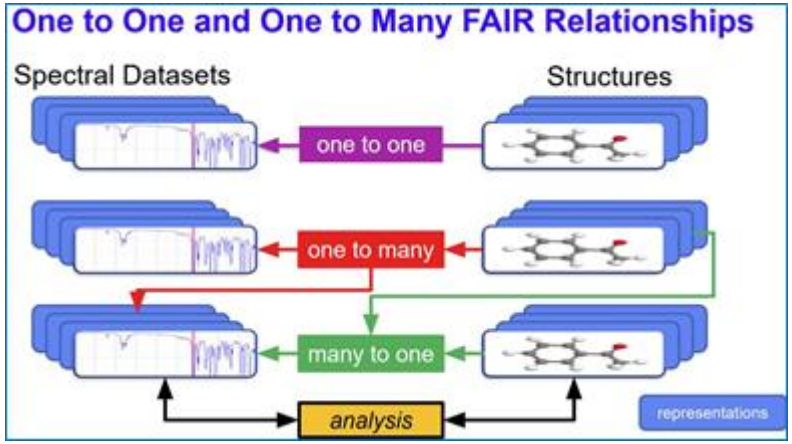


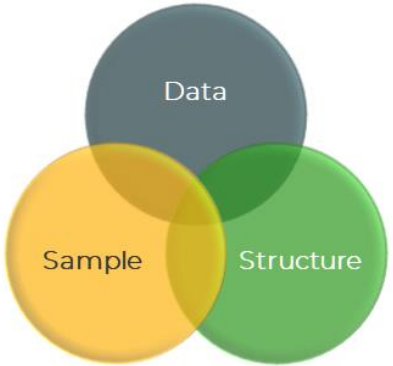
Can we create more streamlined data management and publishing workflows that increases discovery and use?

Electronic Laboratory Notebooks (ELNs)



What are the challenges with describing spectral data, for example?



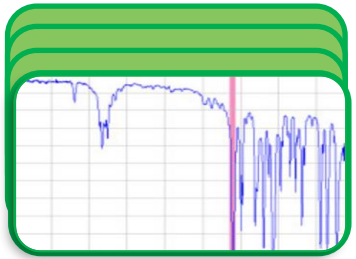


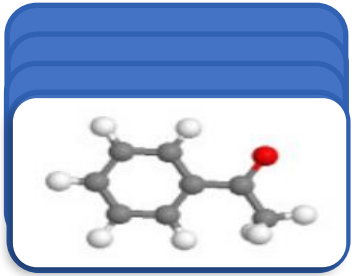
Multitude of proprietary formats

IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles

Robert M. Hanson, Damien Jeannerat, Mark Archibald, Ian J. Bruno, Stuart J. Chalk, Antony N. Davies, Robert J. Lancashire, Jeffrey Lang and Henry S. Rzepa

From the journal Pure and Applied Chemistry
https://doi.org/10.1515/pac-2021-2009
https://doi.org/10.26434/chemrxiv-2022-t783k



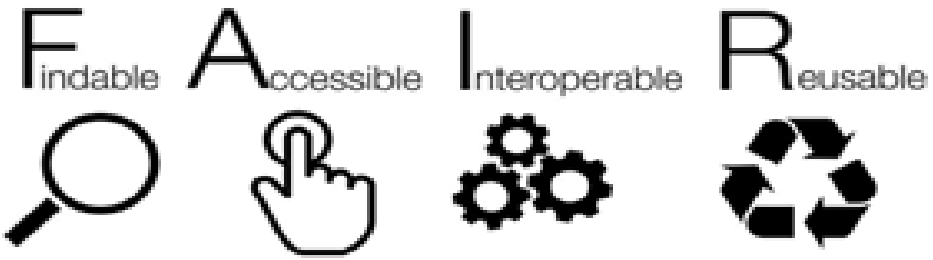


Working towards:

- A set of **guiding principles** defining FAIR in relation to spectroscopic data
- A detailed **data model** for describing the contents of an IUPAC FAIRData Collection and relationships between objects – **sample, data, structures, analysis**
- A specification for a **digital finding aid** - the metadata needed to **locate and reuse** a dataset

Complementary to representation formats such as JCAMP-DX or proprietary alternatives

Encoding best practice: FAIR



The FAIR Data Principles

Comment | OPEN

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).

DATA SHOULD BE

Findable **Accessible**

Interoperable **Reusable**

BY HUMANS AND MACHINES

"FULLY AI READY"

The machine knows what I mean

Standard Identifiers

Standard Access Protocols

Standard Vocabularies

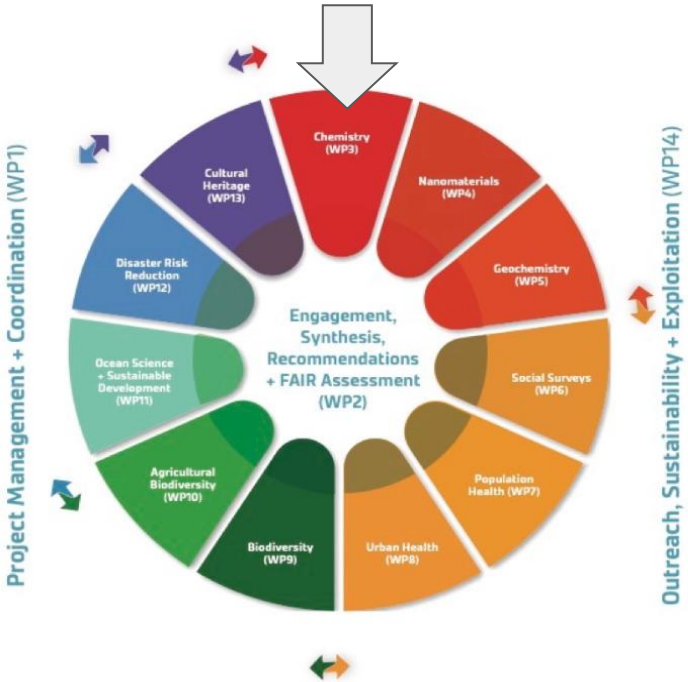
Standard Metadata Schemas

Indexed Repositories

Key enablers of FAIR

WorldFAIR Chemistry

Advancing implementation of the FAIR data principles with standards



D3.1: Digital guidance for Chemistry FAIR data policy & practice
Guidance for implementing standards into FAIR chemical data workflows

D3.2: Chemistry training package
Interactive demos and recipes for handling FAIR chemical data

D3.3: Utility services for Chemistry standards
Protocol specifications for chemical information validation and exchange

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.

Poll #4

How much help do you think chemists need in chemical data management in the digital domain?

1. Very much needed
2. It is an eye opening topic for me, and I think chemists need it
3. It is somewhat needed
4. I am comfortable with my data management workflow

This is all new to me - where can I get help?

Recorded presentation on Cookbook

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

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.

What is a Chemical? Webinar Series

YouTube https://bit.ly/WhatsAchemical

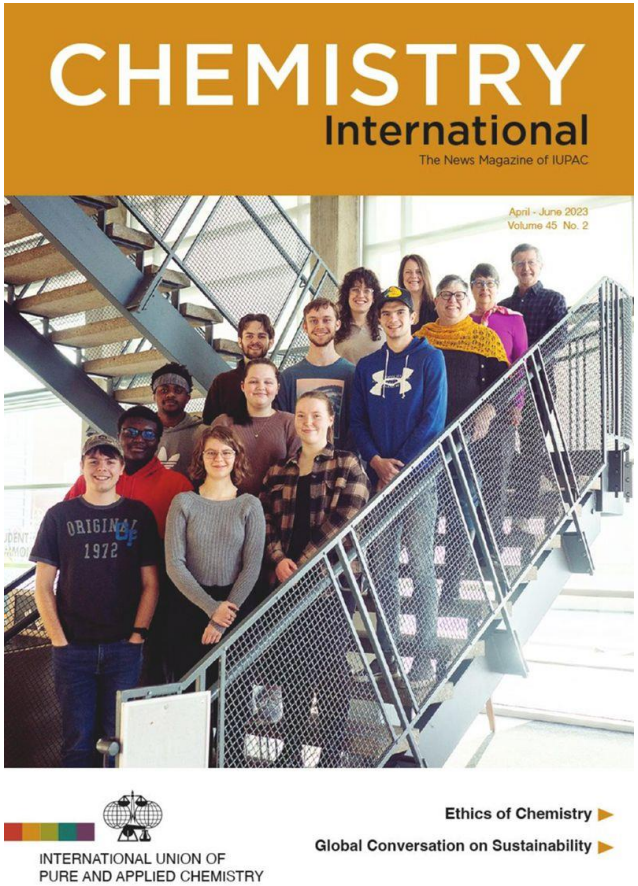
WorldFAIR “Global cooperation on FAIR data policy and practice” is funded by the EC HORIZON-WIDERA-2021-ERA-01-41 Coordination and Support Action under Grant Agreement No. 101058393.

Upcoming Events

1. **IDW Multidisciplinary Session, October 2023**
1. **Fall 2023 workshops**
1. **Many more!**

ChemVoices Past & Upcoming Webinars

http://bit.ly/ChemVoicesP



IUPAC Concentrate Monthly Newsletter

http://bit.ly/IUPACConc

Subscribe

Contact Us

iupac.org

@FAIRChemistry
@iupac

https://bit.ly/WhatsAchemical

FAIRChemistry@iupac.org

@iupac.org

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.