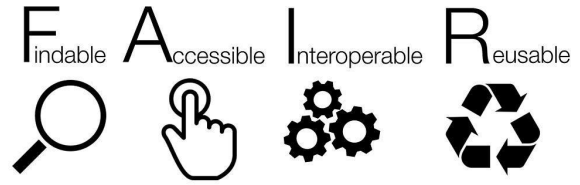
Guidance on FAIR Chemical Data Reporting

Leah McEwen, Cornell University
Ian Bruno, Cambridge Crystallographic Data Centre

WorldFAIR Chemistry
Committee on Publications & Cheminformatics Data Standards

ACS National Meeting, 27 March 2023

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

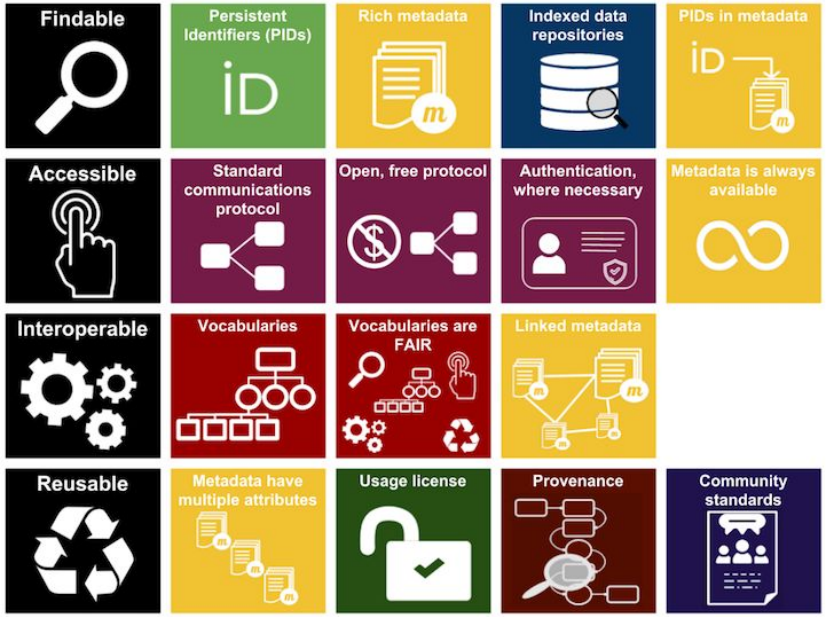


Image credit: Australian National Data Service, CC-BY 4.0
https://www.ands.org.au/working-with-data/fairdata/training

IUPAC World FAIR Chemistry

*Making IUPAC Assets **FAIR***

WorldFAIR

Reporting Guidance

Develop Guidance Protocols for Handling Chemical Data FAIRly

Training Cookbook

Develop online materials on how to manage digital data files and content

Protocol Services

Develop web-based services to confirm chemical identity & machine-readability of chemical data

Types of Guidance

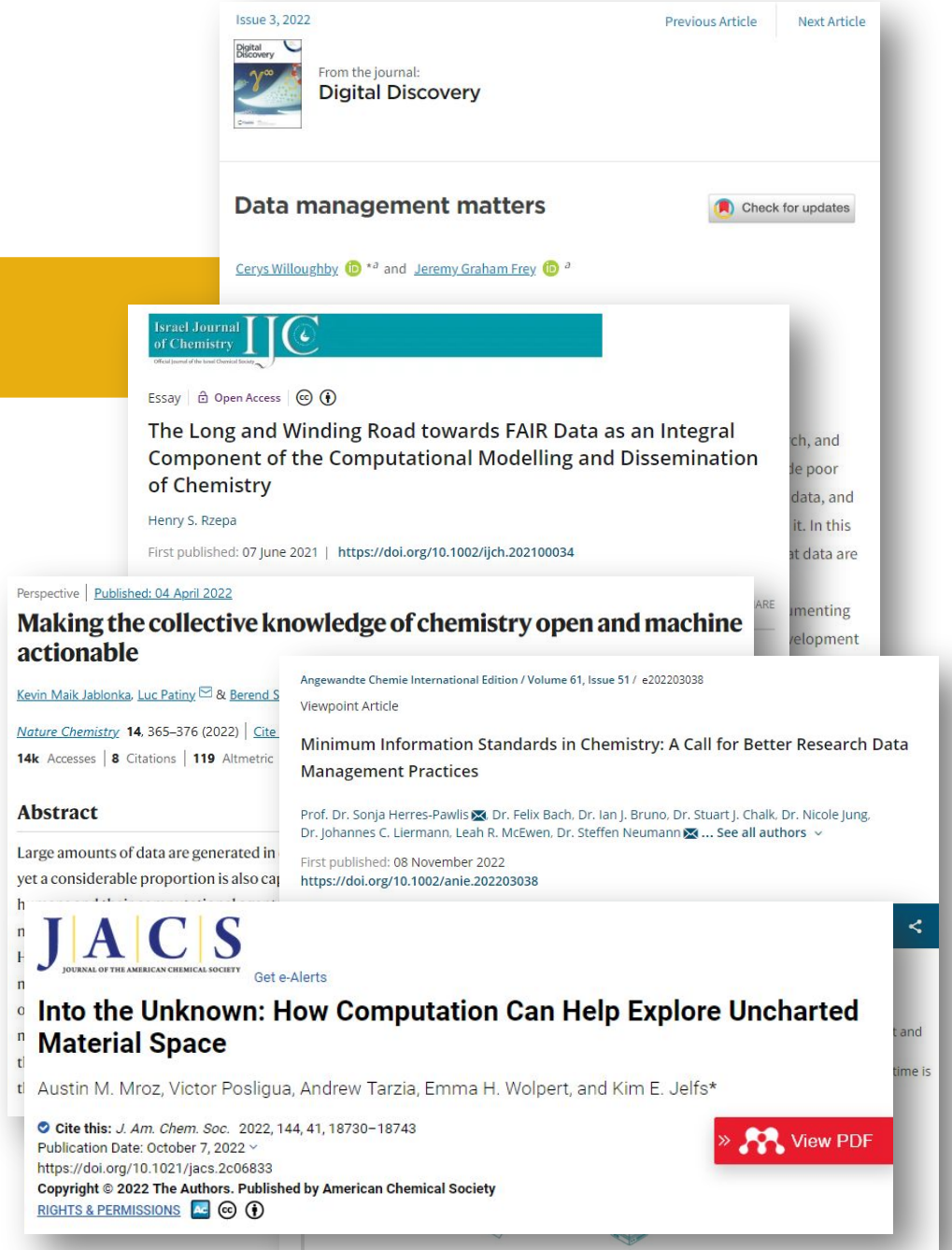
Framing: Over-arching goals

- Addressing global sustainability challenges
- Providing global economic benefits
- Increased transparency and reproducibility
- Enabling inter-disciplinary collaboration
- Universal equitable access to knowledge

Types of Guidance

Perspectives: Desired end state

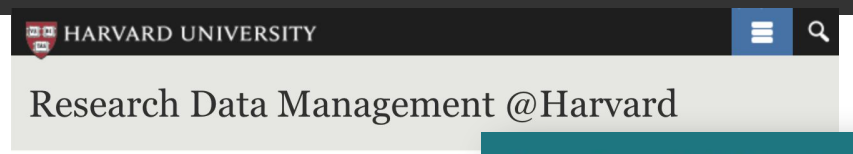
- Community standards for data and information
- Education in best practices
- Tools and infrastructure
- Mandatory data deposition
- Publication of failed experiments
- ELNs as enablers
- High-throughput computational platforms
- Consistency in policies and practice
- Liberate data – make it FAIR

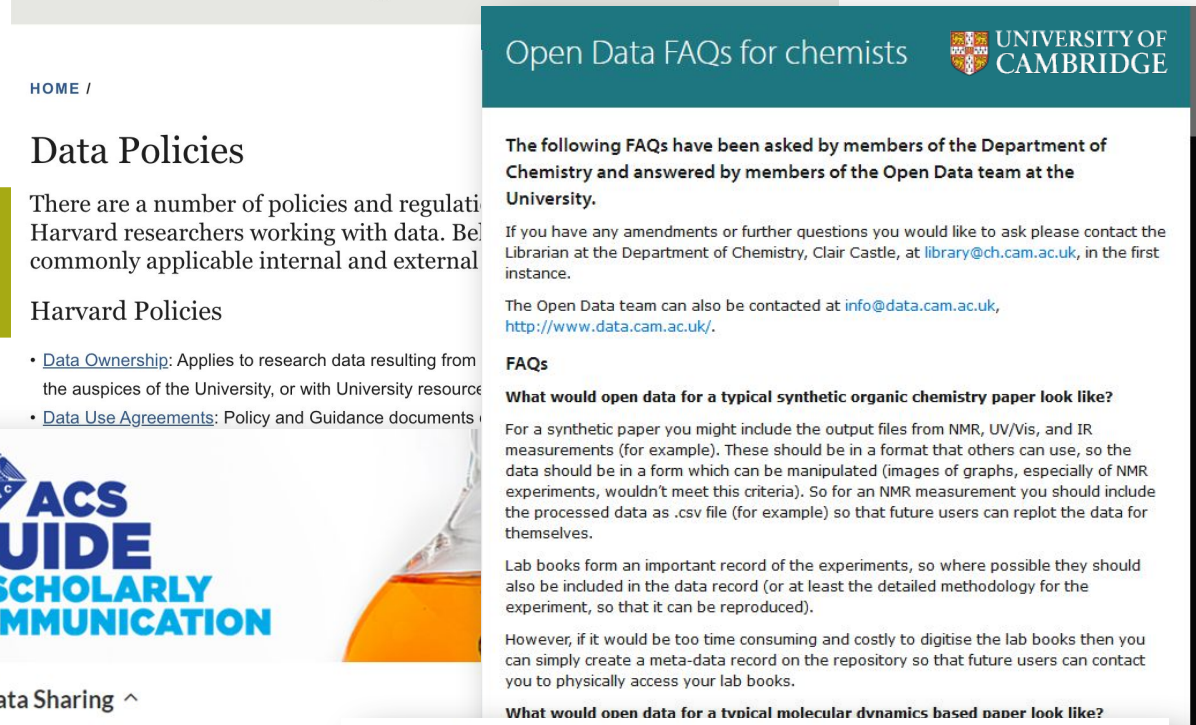


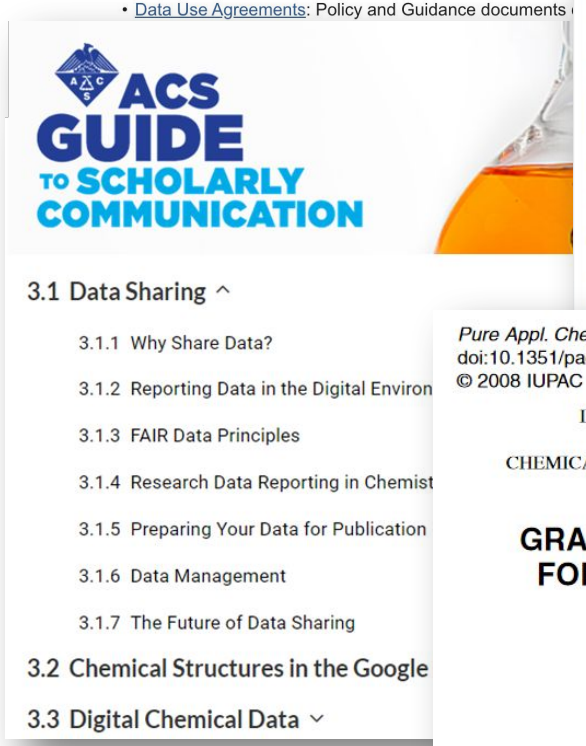
Types of Guidance

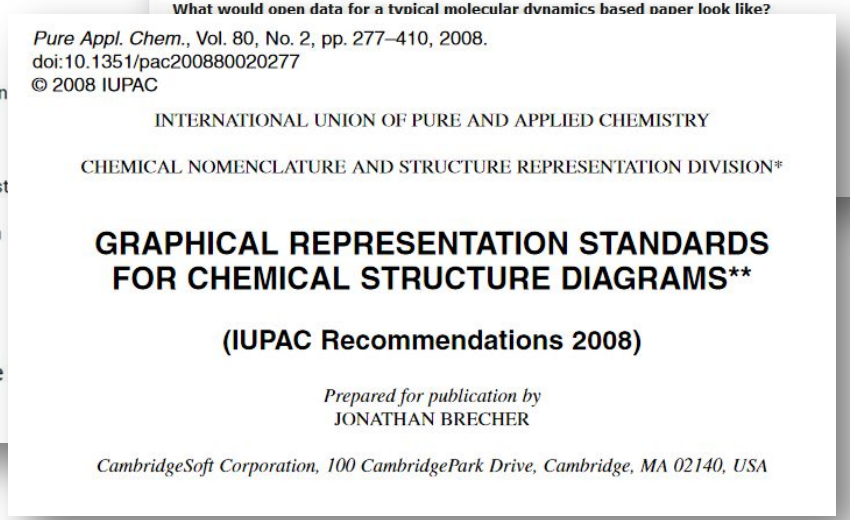
Practical: Specific and targeted

- Professional Societies
- Scientific Unions
- Specific Communities
- Individual Researchers
- Institutional Policies









Types of Guidance

Policies: Enforcing community norms

- Regulatory Guidance
- Government Policies
- Funder Policies
- Journal Policies

Data sharing

Why is data sharing important?

What is research data?

Our data sharing policy

Recommended repositories

Choosing a repository

t specific repositories

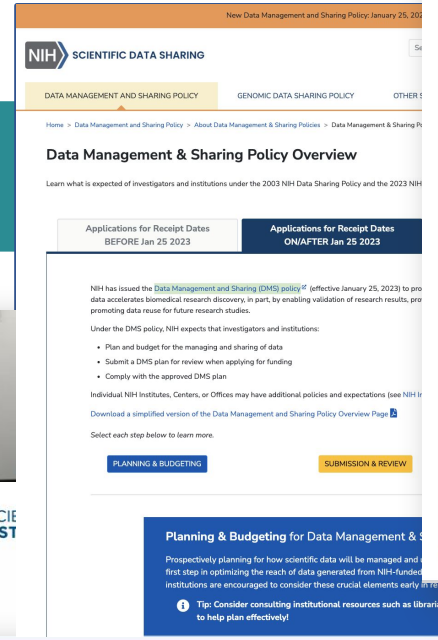
al repositories

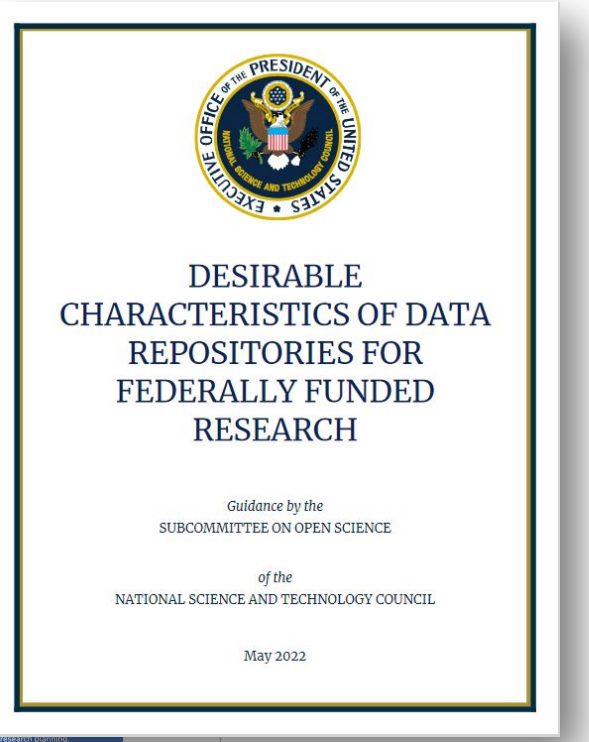
ility Statements

ftware citation

data

software and code





Wiley's Data Sharing Policies

	Data availability statement is published[1]	Data has been shared[2]	Data has been peer reviewed[3]	Example Wiley journals
Encourages Data Sharing	Optional	Optional	Optional	
Expects Data Sharing	Required	Optional	Optional	British Journal of Social Psychology
Mandates Data Sharing	Required	Required	Optional	Ecology and Evolution
Mandates Data Sharing and Peer Reviews Data	Required	Required	Required	Geoscience Data Journal American Journal of

In Horizon Europe, *beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the **FAIR principles***, and should at least do the following:

- Prepare a Data Management Plan (DMP) and keep it updated throughout the course of the project
- Deposit data in a trusted repository and provide open access to it ('as open as possible, as closed as necessary')
- Provide information (via the same repository) about any research output or any other tools and instruments needed to re-use or validate the data

Types of Guidance

Framing: Over-arching goals

Perspectives: Desired end state

Practical: Specific and targeted

Policies: Enforcing community norms

IG

Data policy standardisation and implementation IG

Taxonomy:

RDA RESEARCH DATA ALLIANCE

Developing a Research Data Policy Framework for All Journals and Publishers

Authors: Iain Hrynaszkiewicz, Natasha Simons, Azhar Hussain, Rebecca Grant, Simon Goudie

Framework recognises that different journals and different domains may be at different levels of maturity

Different requirements (or features), varying different degrees of mandating (○ ●)

Progression from requiring information to be provided to checking and enforcing

14 journal research data policy features arranged as 6 policy types (tiers)

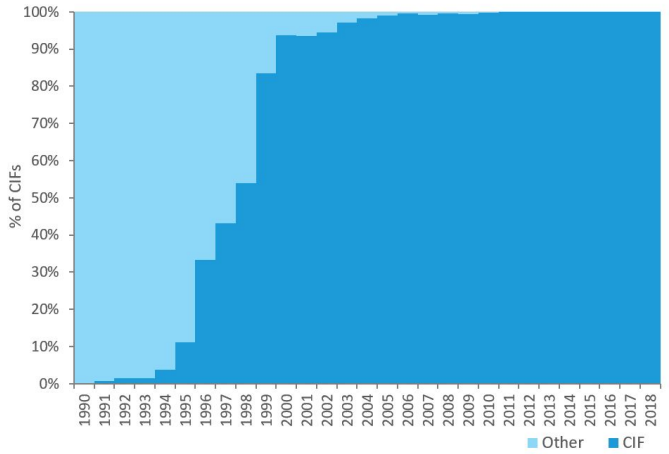
	Policy 01	Policy 02	Policy 03	Policy 04	Policy 05	Policy 06
Definition of the research data	○	○	○	○	○	○
Exceptions to policy	○	○	○	○	●	●
Embargoes	○	○	○	●	●	●
Supplementary materials	○	○	○	●	●	●
Data repositories	○	○	○	●	●	●
Data citation	○	○	○	○	●	●
Data licensing	○	○	○	○	○	○
Researcher/ author support	●	●	●	●	●	●
Data availability statements		○	●	●	●	●
Data formats and standards				○	○	●
Mandatory data sharing (specific data types)				●	●	●
Mandatory data sharing (all papers)				○	●	●
Peer review of data				○	○	●
Data Management Plans (DMPs)				○	○	○

Crystallography Journey

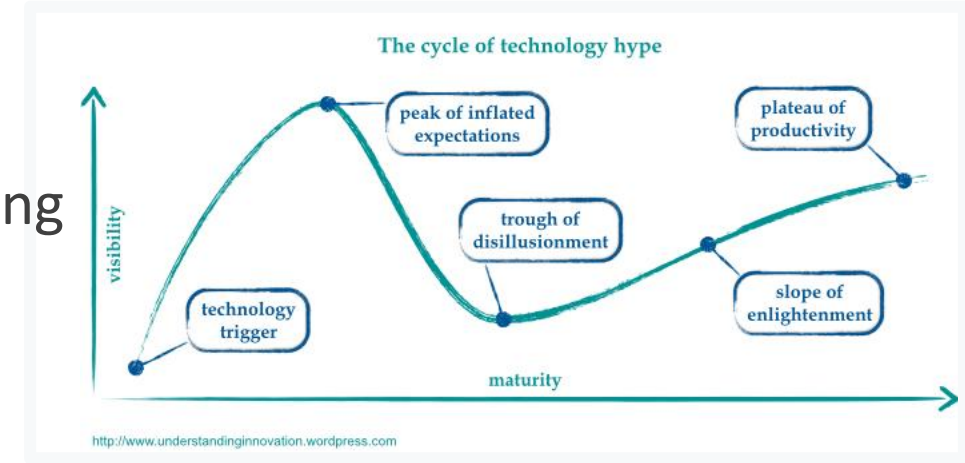
1940s	Data published in tables in journal articles
1960s	Data abstracted into electronic databases
1990s	Data standards (CIF) established
2000s	Cultivation of data validation and publication workflows
2010s	Deposition in data repository pre-publication the default

Crystallography Journey

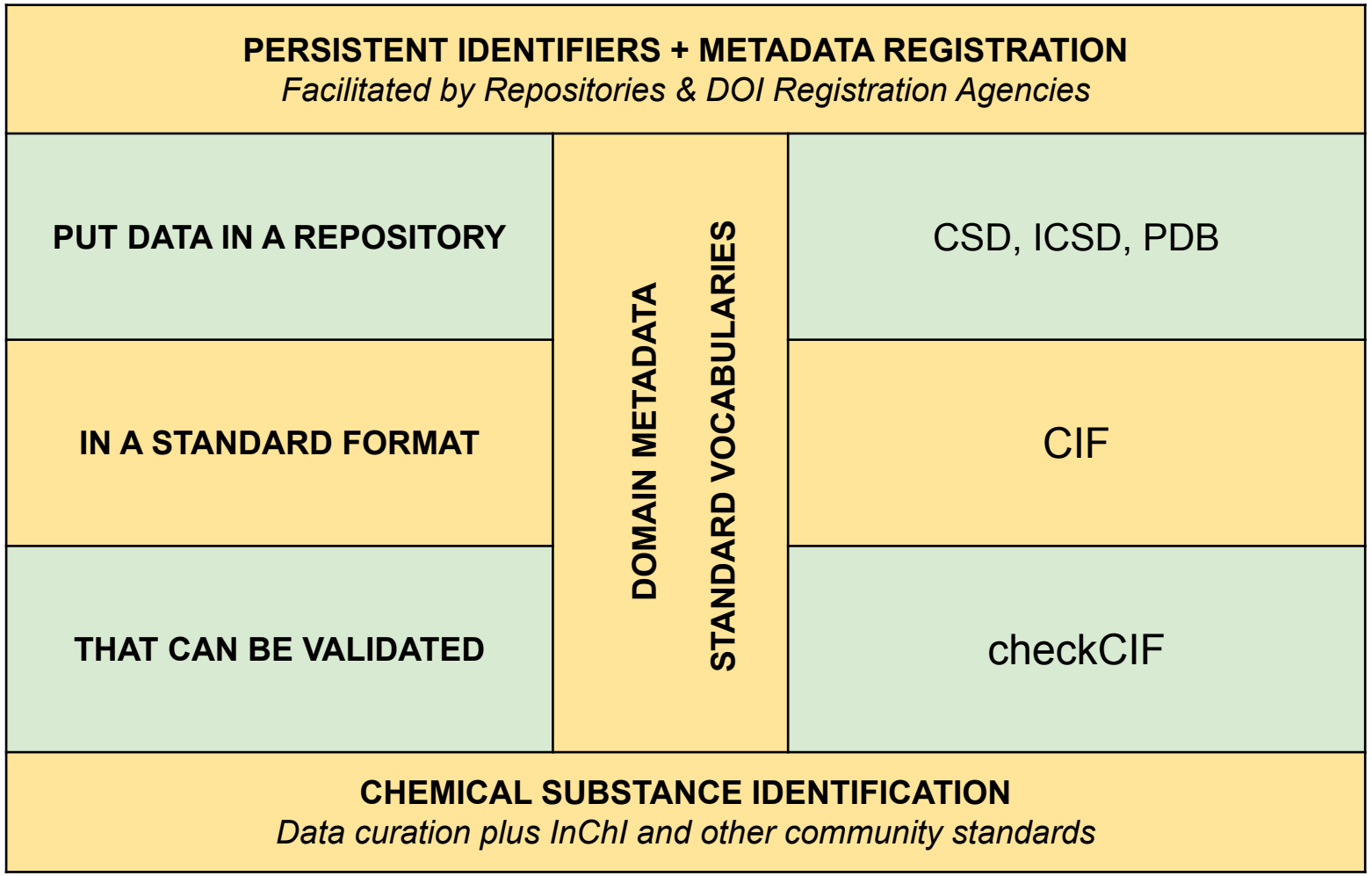
- **Initially CIFs were embedded in PDFs and Word documents**



- **Lasting uptake of CIF and digital deposition catalysed by:**
  - Leadership by IUCr – principles and practice
  - Adoption of CIF by software tools
  - Validation services
  - Joined up repository/publisher workflows
  - Aggregation of data enabling reuse for generating new knowledge



FAIR Data Implementation in Crystallography



Stakeholders

Researchers
Institutions
Funders

Publishers
Editors
Reviewers

Repositories
Tool Providers
Instrument Manufacturers

Scientific Unions
Professional Societies

What has adopting FAIR enabled for crystallography?

- Validation services checking for consistency and completeness
- Extension of deposition services from organic to inorganic small molecules
- Aggregation and linking of data across general information resources

Satisfied by established community practices and workflows

Enabled by adoption of established CIF, DOIs

- Chemistry-based discovery, reuse and linking

Enabled as a result of further enrichment by repository data curators

*A **chemical crystallography** dataset has FAIR **crystallographic** attributes when deposited – it rarely has FAIR **chemical** attributes.*

Describing Diverse Chemistry Datasets Across Distributed Data Resources

RDA Plenary 20, Gothenburg, Sweden, 23 March 2023

Key themes and ideas to emerge:

- Minimal information standards – what is enough for future reuse?
- Data born FAIR (future smart labs) vs made FAIR when published
- Importance of sample – common challenge across domains
- Capturing provenance and process – enabling reproducibility
- Characterising analytical techniques – cataloguing existing standards
- Use cases to drive the need for discovery and interoperability across resources

Points of 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*

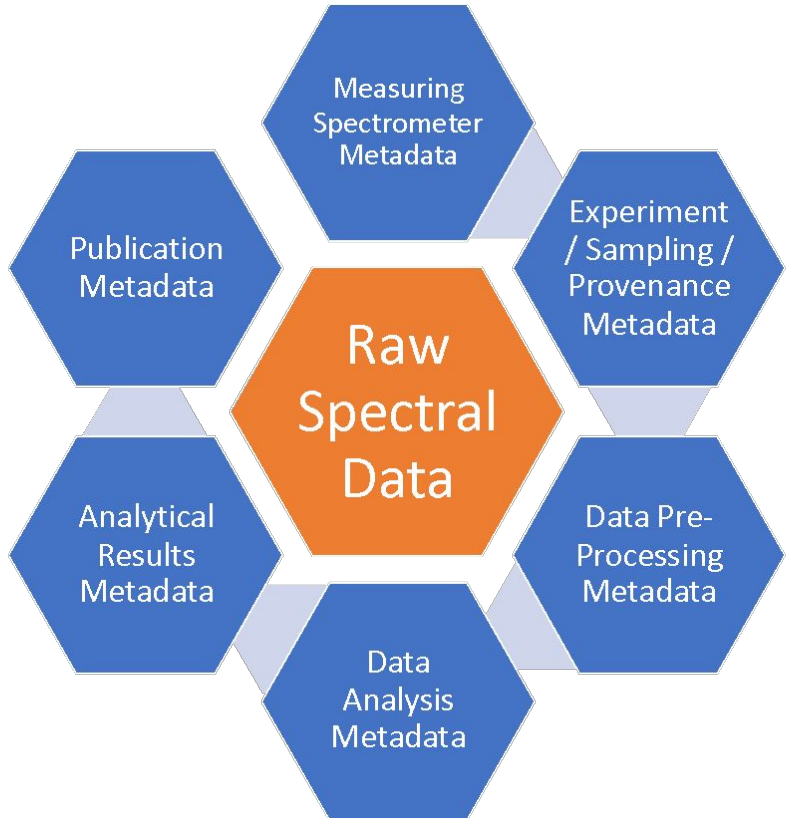
Units: integration of quantities➥ *standard units of measure*

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 A. Davies, IUPAC FAIRSpec

Adapted from K. Lehnert, OneGeochemistry

FAIR Chemistry Data – General Concepts

FINDABILITY

Generic object identifiers

DOIs for example

Metadata Registries

With high-level chemistry metadata

E.g. InChIs in DataCite

Data Repositories

Ideally chemistry-aware

ACCESSIBILITY

Open Protocols

With chemistry-aware APIs

Authentication and Authorisation where needed

REUSABILITY

Community Licences

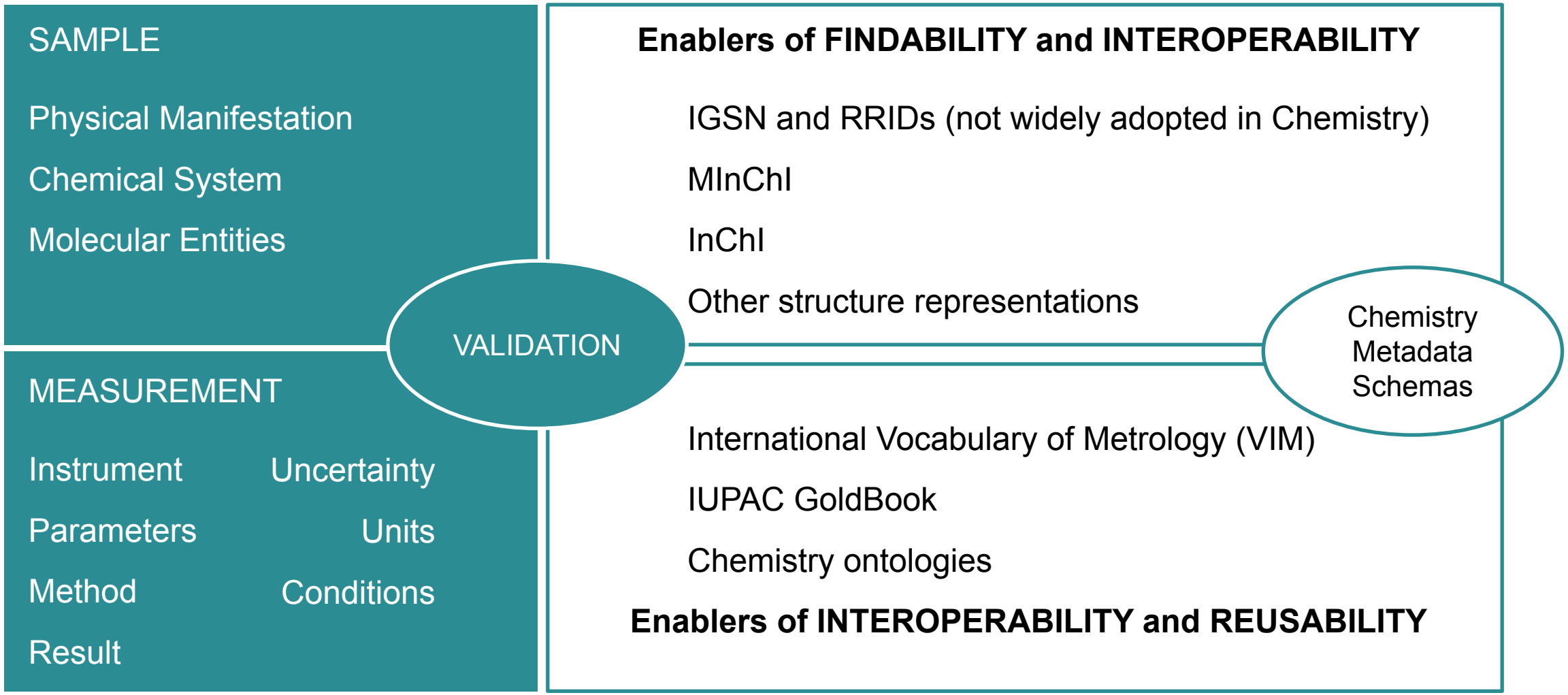
Ideally CC0 or CC-BY

Non-derivative (ND) for reference data

INTEROPERABILITY

Cross-domain Interoperability Framework (CDIF): *A WorldFAIR Focus*

FAIR Chemistry Data – Key Concepts



Well defined chemical data are broadly reusable

RIPE for 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*

Are these digital standards FAIR for programmatic access and reuse?

IUPAC 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

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

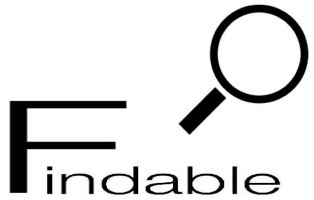
- Models: chemical systems, quantites, measurements
- Criteria for interoperability

FAIR enabling resources

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

[Gold Book example]

WE ARE FAIR ENABLERS



	PIDs & registered metadata	**Domain repositories**	**Open standard formats**	**Verified, licensed**
Repositories	- standard chemistry descriptors - key metadata	- standard chemistry APIs - authentication and authorization	- standard formats, terminology, ontologies - metadata relationships	- standardized validation - transparent licensing
Software (tools)	- generate standard chemistry descriptors	- standard chemistry APIs (e.g., instrument to ELN)	- standard descriptors in native formats - link data/metadata	- metadata extraction
Support services	- cross-linking data and publications	- facilitate deposit - data preparation checklist	- how-to support for using file formats - metadata templates	- data review - process guide
Researchers	- templates to collect metadata	- select repository & upload	- assemble data files - document which formats used	- prepare ReadMe

https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

What Next for Guidance

- **Where do you or your researchers feel all at sea?**
  For managing? For sharing? For validation? For reuse?
- **What new chemistry guidance would help you in your role?**
- **What existing chemistry guidance are you aware of?**
- **Are there specific communities we should be engaging with?**
  Engaged: IUPAC, RDA, NFDI4Chem, PSDI, Geochemists, Nano – who else?
  E.g. Data Curation Network? Library communities? Industry groups?

Guidance on FAIR Chemical Data Reporting

WorldFAIR Chemistry

ACS National Meeting, 27 March 2023

Contact & Follow US

iupac.org

@FAIRChemistry
@iupac

tinyurl.com/iupacchan

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.

WorldFAIR

Guidance on FAIR Chemical Data Reporting

WorldFAIR Chemistry

ACS National Meeting, 27 March 2023

Discussion

Shared notes link:

https://bit.ly/ACSChemNotes

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.