WorldFAIR Chemistry: FAIR-enabling resources

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

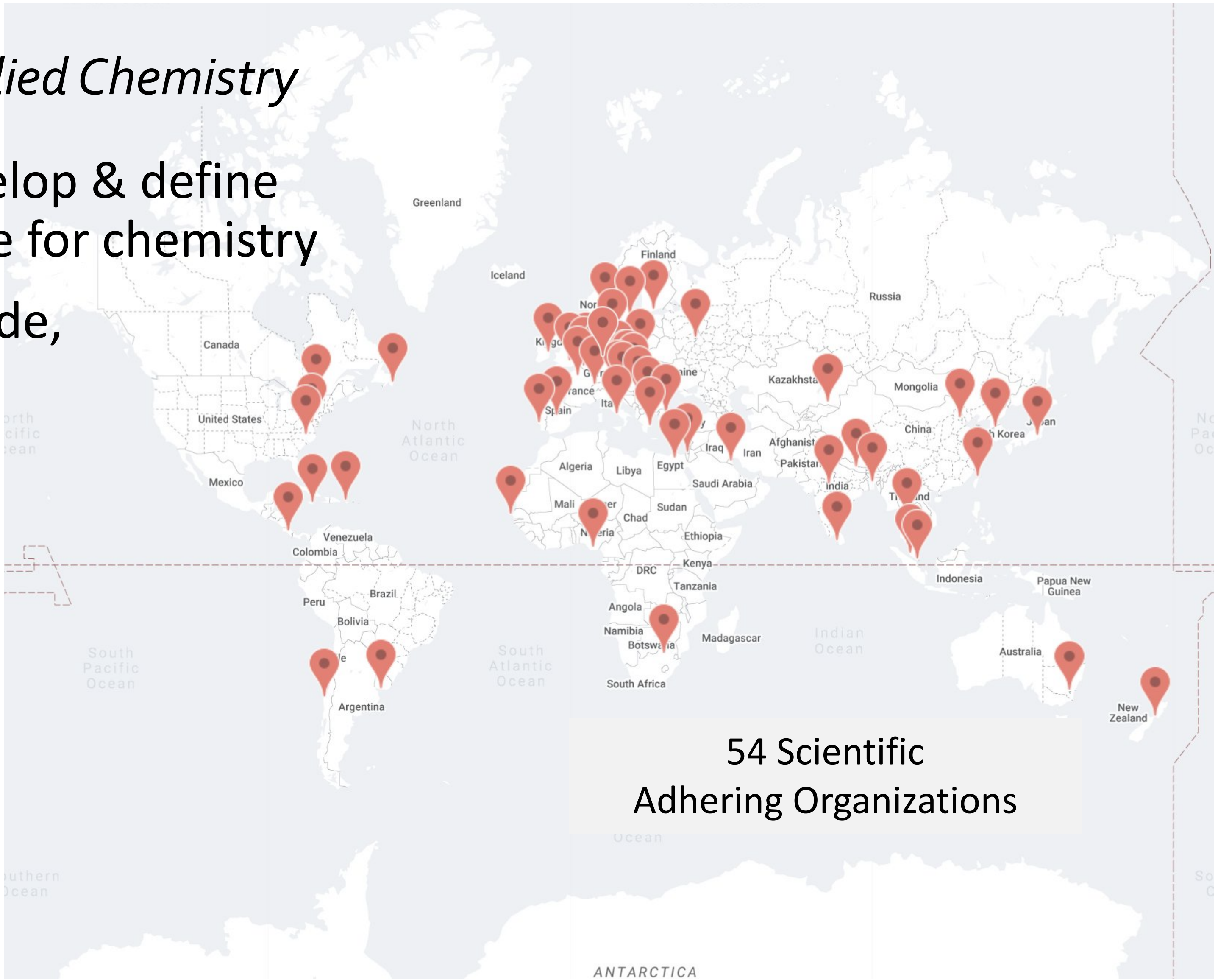
PARC FAIR data webinar series
2023.04.20

What is IUPAC?

International Union of Pure and Applied Chemistry

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

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



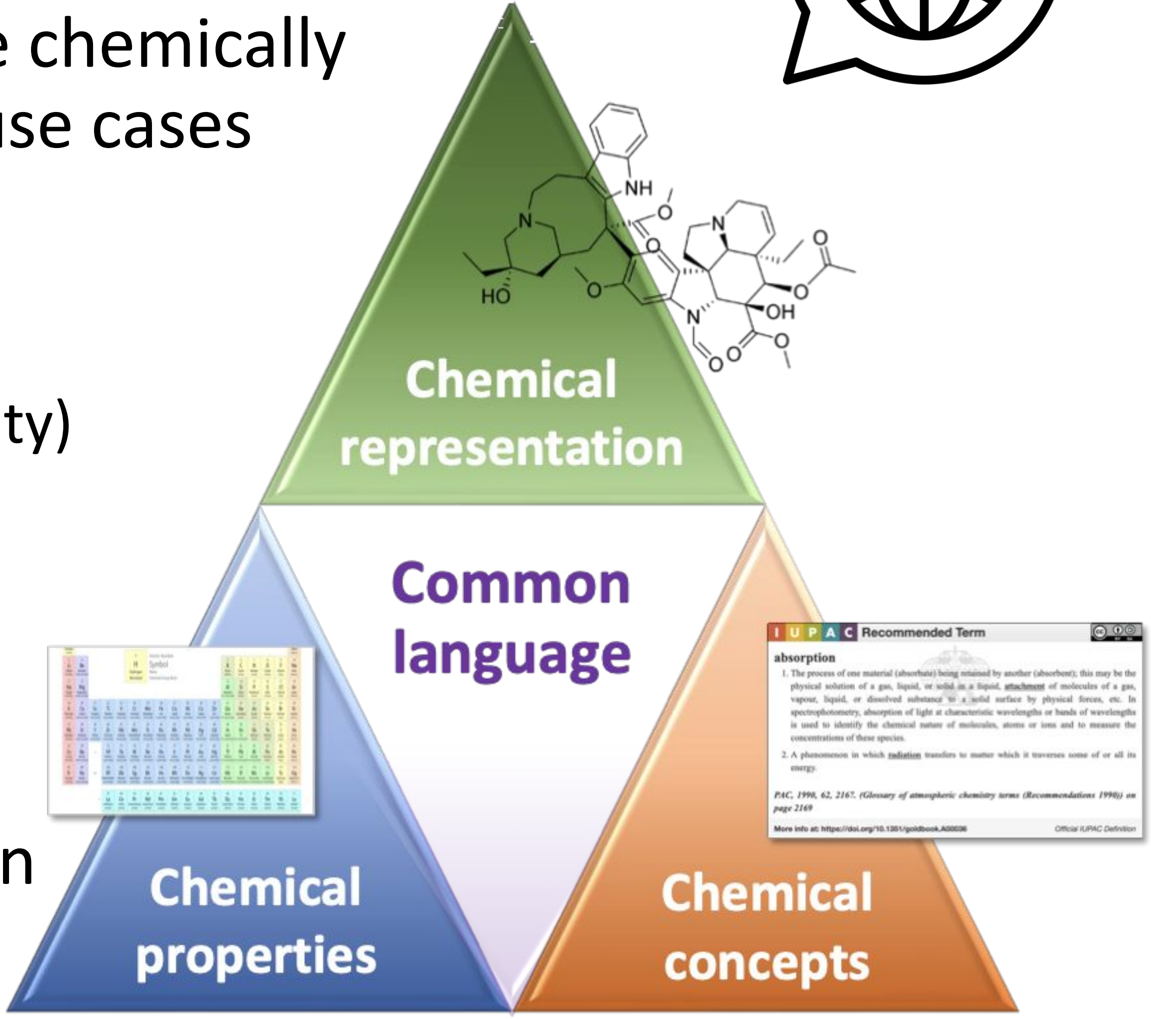
54 Scientific Adhering Organizations

Interpretable chemical data

Outputs from measurements need to be chemically interpretable between systems, across use cases and over time, including:

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

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



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

IUPAC standard definitions and properties

Chemical representation

- Nomenclature
  - Blue Book (organic)
  - Red Book (inorganic)
  - Purple Book (polymer)
- Graphical representation (structures, stereo, reactions)

Chemical terminology

- Orange Book (analytical)
- Silver Book (clinical)
- White Book (biochemical)
- Green Book (physical)

Chemical properties

- Periodic Table (CIAAW tables)
- Solubility Data Series
- Atmospheric kinetics datasheets
- Polymerization kinetics dataset
- Stability constants dataset

Machine-processable *(to some degree)*

- InChI notations
  - InChIKey
  - RInChI
  - *MInChI*
  - *NInChI*
- *SMILES+ notation*
- HELM notation
  - *Glycans notation*

- Gold Book (compendium)
- NPU terminology for clinical chemistry
- *Green book digital quantities & symbols*
- *DRUM digital units*

- JCAMP-DX spectra format
- ThermoML format
- *AIF adsorption format*
- *FAIRSpec metadata principles*
- *MAPT metadata schema*
- *Solubility metadata schema*
- *Dissociation constants dataset*
- *Atmospheric kinetics dataset*
- *Polymerization kinetics database*

IUPAC: FAIR-enabling resources for Chemistry

- Premise: IUPAC standards provide the scientific backbone for accurate chemical data exchange

- Target: digitalize IUPAC standards and map to FAIR attributes
  - How do these digital standards enable FAIR re-use ?
  - Are IUPAC standards FAIR and accessible for programmatic re-use ?

- Goal: enable application of chemistry description at multiple levels
  - **Concise information** for registration and general discovery
  - **Consistent representation** for exchange and integration
  - **Rich metadata** for analysis & prediction

What makes a FAIR-enabling resource (FER)?

FER implementation	Chemistry considerations	IUPAC notes
I1: **knowledge representation language** Metadata: rules for machine-processable expression Data: formats for machine-processable encapsulation of critical data components	Graphical representation guidelines, nomenclature rules, InChI algorithm, OpenSMILES specification, HELM specification; JCAMP-DX, ThermoML, CIF, AIF, molfile/SDF family, NMReData	Guidance and test sets for implementation of nomenclature rules, graphical representation standards, other formats and metadata schema in development
I2: **structured vocabulary** Metadata: referrable concepts for annotation Data: regularized labels for fields/components	Gold Book, VIM (metrology generally), MeSH, CIF dictionaries (crystallography), STRENDA	Align Gold Book metadata with Dublin Core and SKOS practices
I3: **semantic model** Metadata: ontologies, classification Data: meaningful relationships of data components	CHMO, RXNO, MOP, ClassyFire, ChEBI Models of meaning: CTAB (connectivity), PToE (periodicity), FAIRSpec (spectra-structure), QSAR (structure-activity)	Project to digitalize Periodic Table underway; do we need to a standardize CTAB/CT file specification? QSAR?

What makes a FAIR-enabling resource (FER)?

FER implementation	Chemistry considerations	IUPAC notes
F1: globally unique, persistent, resolvable **identifier service**	Most use general DOI services for data files; various chemical database, record identifiers (e.g. CAS RNs, PubChem CIDs) but not universal	Register more IUPAC datasets, descriptors, etc.
F2: **metadata schemas** for findability	InChI, SMILES, HELM, chemical names, polymer names, named reactions, properties, methodologies, etc.	Core metadata components for discovery (e.g, MMI); *micro-metadata schema?*
F3: **metadata-data linking specification**	Provided in registered metadata associated with DOI services; often managed locally	Register more IUPAC datasets, descriptors, etc.
F4: **registry service** for publishing datasets and metadata records	Pubchem, ChemSpider, DSSTox, ChEMBL, VAMDC Also use general DOI services for citation	Register more IUPAC datasets, descriptors, etc.

What makes a FAIR-enabling resource (FER)?

FER implementation	Chemistry considerations	IUPAC notes
A1: **standardized communications protocol**	Most use http/https, ftp, Shibboleth, etc.; CoreTrustSeal best practices for repos	IUPAC needs hosting repositories (some partner agencies)
		Do we need chemistry criteria for API protocols (e.g., use of chemical linear notations)?
A2: **metadata preservation policy**	Use cases: CBI (confidential business information), retraction	Develop trusted repository criteria?

What makes a FAIR-enabling resource (FER)?

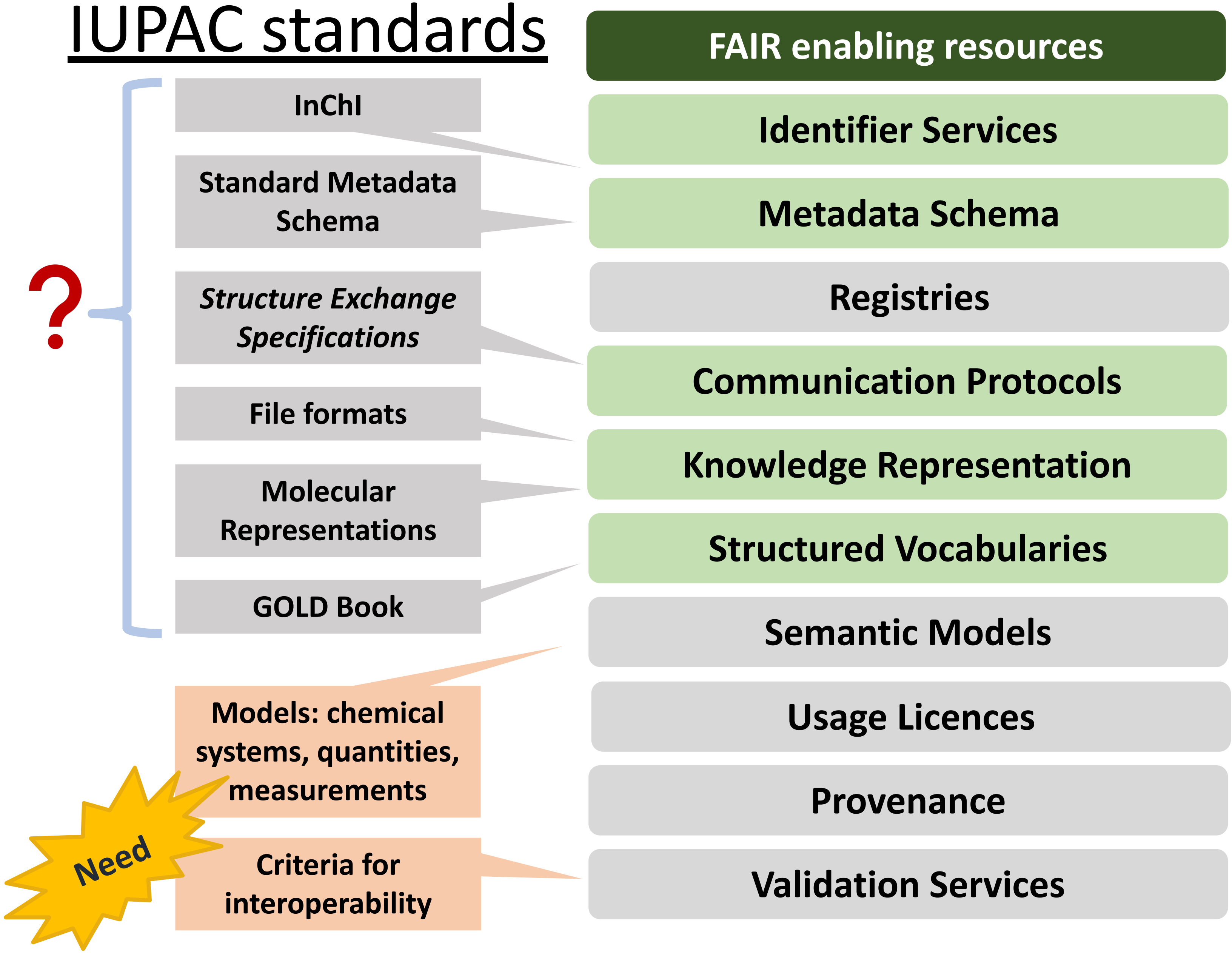
FER implementation	Chemistry considerations	IUPAC notes
R1.1: **usage license**	Emerging publisher guidelines (DAS levels)	Existing policy is for open dissemination of standards w/approval requested for derivatives and commercial use (i.e., CC-BY-NC-ND); relax for metadata*
R1.2: **metadata schema** for describing provenance	Chemistry examples: CAS tracks deleted & superseded CAS RNs; PubChem tracks data sources up to 2 levels; Gold Book tracks sources of definitions, superseded terms	Track: versioning of standards, official copy of record, source of original experimental data, subsequent evaluation processes
R1.3: **validation service** for syntax and semantics	Example: checkCIF (Generic example: OpenRefine)	Validation mechanisms for all IUPAC digital standard formats and rule-sets (e.g., nomenclature, graphical representation)

*(human-readable rendering for end users must be consistent)

FAIR chemical datasets/systems/workflows

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

Are these digital standards FAIR for programmatic access and reuse?



IUPAC standards

?

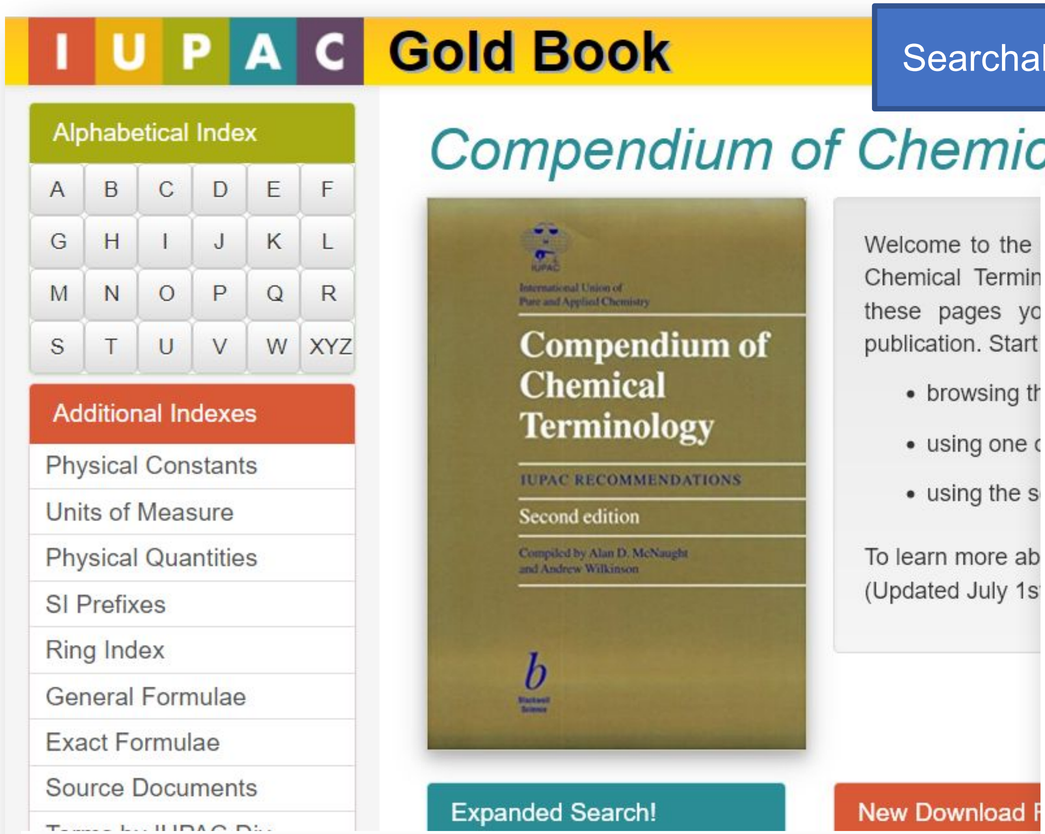
- InChI
- Standard Metadata Schema
- *Structure Exchange Specifications*
- File formats
- Molecular Representations
- GOLD Book

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 Gold Book

Searchable

Search (three chars min

Resources

Alphabetical Index

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	XYZ

Additional Indexes

- Physical Constants
- Units of Measure
- Physical Quantities
- SI Prefixes
- Ring Index
- General Formulae
- Exact Formulae
- Source Documents

Compendium of Chemical Terminology

International Union of Pure and Applied Chemistry

Compendium of Chemical Terminology

IUPAC RECOMMENDATIONS

Second edition

Compiled by Alan D. McNaught and Andrew Wilkinson

Welcome to the Chemical Termin these pages yo publication. Start

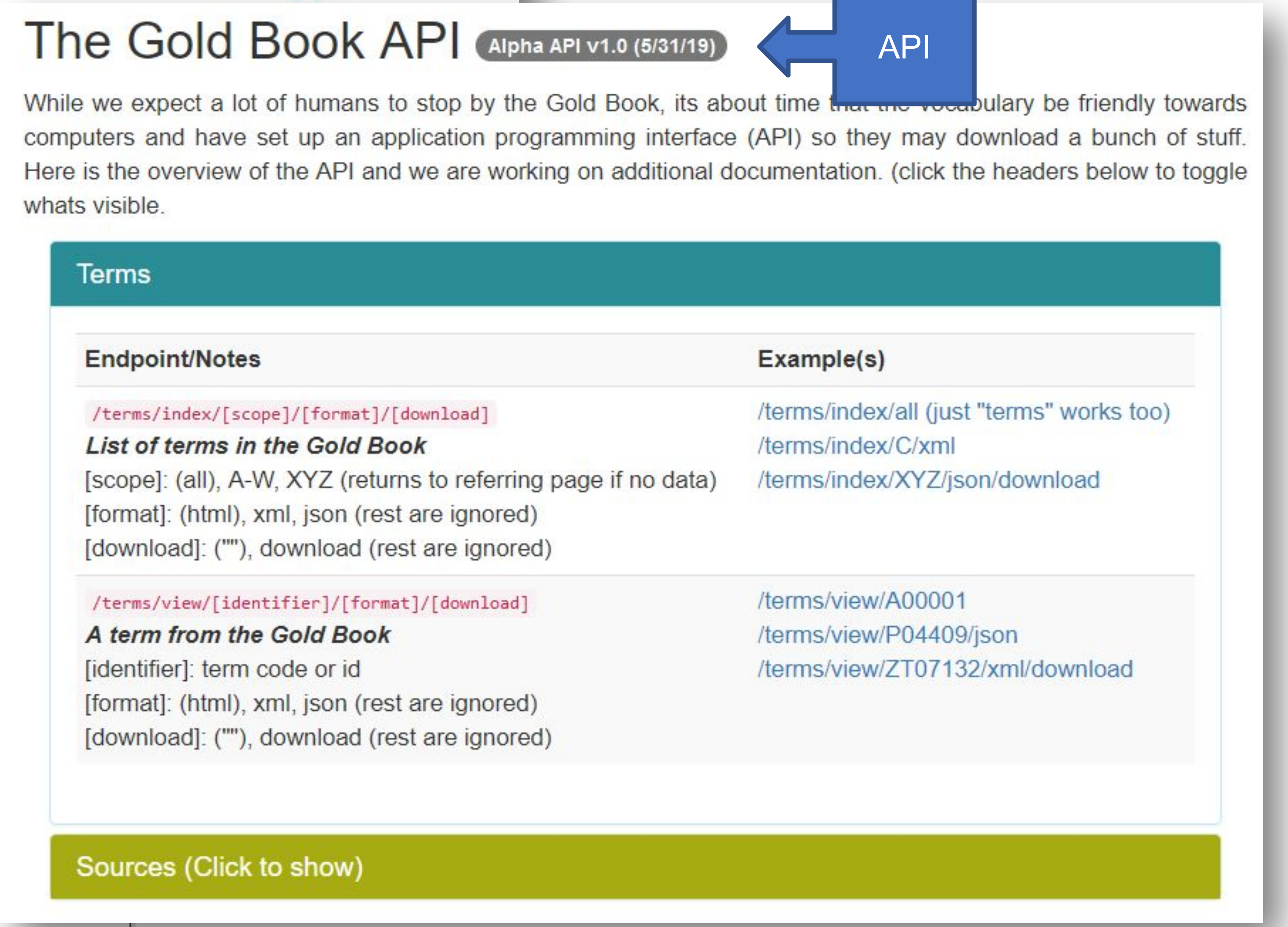
- browsing th
- using one o
- using the s

To learn more ab (Updated July 1s

Expanded Search!

New Download F

https://goldbook.iupac.org



API

The Gold Book API

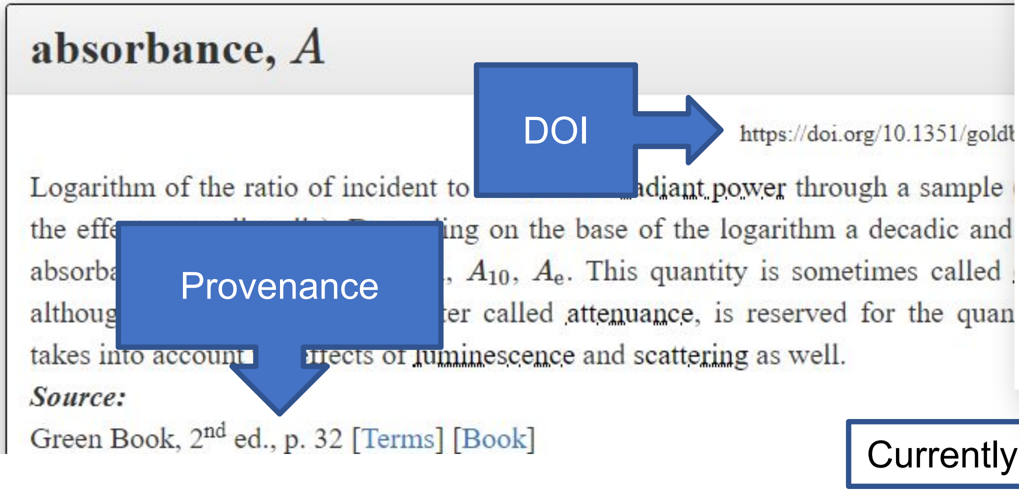
Alpha API v1.0 (5/31/19)

While we expect a lot of humans to stop by the Gold Book, its about time that the vocabulary be friendly towards computers and have set up an application programming interface (API) so they may download a bunch of stuff. Here is the overview of the API and we are working on additional documentation. (click the headers below to toggle whats visible.

Terms

Endpoint/Notes	Example(s)
`/terms/index/[scope]/[format]/[download]` ***List of terms in the Gold Book*** [scope]: (all), A-W, XYZ (returns to referring page if no data) [format]: (html), xml, json (rest are ignored) [download]: (""), download (rest are ignored)	/terms/index/all (just "terms" works too) /terms/index/C/xml /terms/index/XYZ/json/download
`/terms/view/[identifier]/[format]/[download]` ***A term from the Gold Book*** [identifier]: term code or id [format]: (html), xml, json (rest are ignored) [download]: (""), download (rest are ignored)	/terms/view/A00001 /terms/view/P04409/json /terms/view/ZT07132/xml/download

Sources (Click to show)



absorbance, *A*

DOI

https://doi.org/10.1351/goldb

Logarithm of the ratio of incident to radiant power through a sample the effe ing on the base of the logarithm a decadic and absorba , A_{10}, A_e. This quantity is sometimes called although er called attenuance, is reserved for the quan takes into account effects of luminescence and scattering as well.

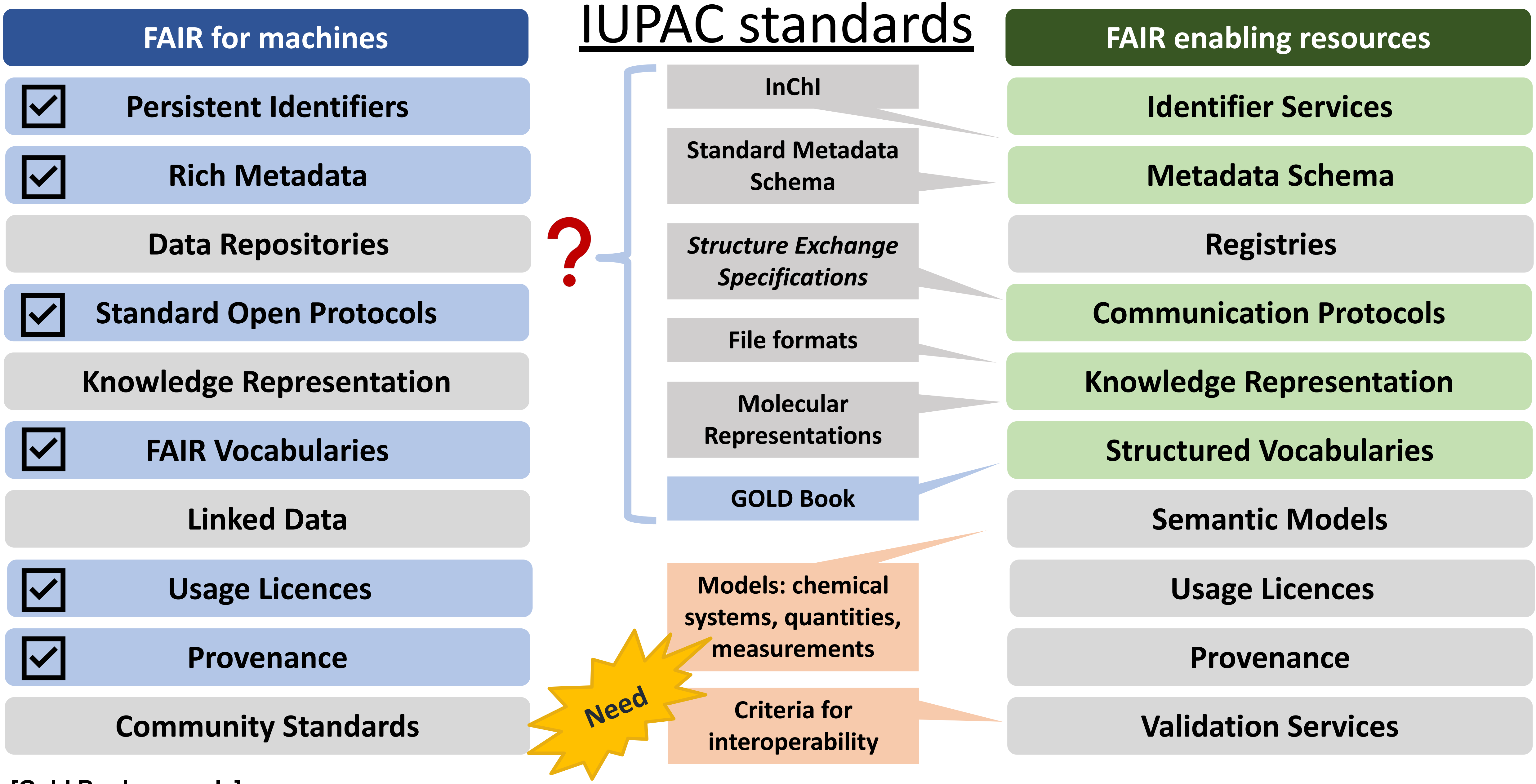
Provenance

Source:

Green Book, 2nd ed., p. 32 [Terms] [Book]

Currently developing the processes and practices needed to ensure definitions are born digital

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]

FIP/FER analysis: friendly suggestions ☺

Domain/broader application

- Full declaration option for R1.3: **validation**
  - Check representation for syntax & semantics; assessment of (meta)data completeness
- Option to reference domain descriptors as metadata components in F2
- Option to reference domain specifications for API protocols (e.g., URI syntax)
- FIP assessment for each FER (FERs also need to be FAIR to reuse, exposes more how-to-use)
- Link to FAIR implementation community profiles from FER profiles as FER curators
- Allow any FER to show up under any attribute in FIP profiles, and push selected types back to the FER profile
- Profile review by responsible organizations and domain technical experts

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*

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 - validation checks
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	- validation check - select license - prepare ReadMe

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

Questions?

iupac.org

FAIRChemistry@iupac.org

@FAIRChemistry
@iupac

@iupac.org

https://bit.ly/WhatsAchemical

FAIRChemistry Community

WorldFAIR

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.