WorldFAIR

Project Title	Global cooperation on FAIR data policy and practice
Project Acronym	WorldFAIR
Grant Agreement No	101058393
Instrument	HORIZON-WIDERA-2021-ERA-01
Topic, type of action	HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions
Start Date of Project	2022-06-01
Duration of Project	24 months
Project Website	http://worldfair-project.eu

D3.1 Digital guidance for Chemistry FAIR data policy and practice

Work Package	WP03 - Chemistry
Lead Author (Org)	Ian Bruno (Cambridge Crystallographic Data Centre), Leah McEwen (Cornell University)
Contributing Author(s) (Org)	Evan Bolton (U.S. National Center for Biotechnology Information), Stuart Chalk (University of North Florida), Richard Hartshorn (University of Canterbury), Fatima Mustafa (IUPAC).

Due Date	31.05.2023
Date	29.05.2023
Version	1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.7887283

Dissemination Level

☒ PU: Public

□ PP: Restricted to other programme participants (including the Commission)

□ RE: Restricted to a group specified by the consortium (including the Commission)

□ CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	09.05.2023	Ian Bruno, Leah McEwen	Draft for internal review
0.2	24.05.2023	Iseult Lynch, Alexander Prent	Internal review of draft
0.3	29.05.2023	Ian Bruno, Leah McEwen	Finalised draft
1	30.05.2023	Laura Molloy	Copy edited and formatted for deposit

Disclaimer

WorldFAIR has received funding from the European Commission's WIDERA coordination and support programme under the Grant Agreement no. 101058393. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Abbreviations and Acronyms

ACS	American Chemical Society
AI/ML	Artificial Intelligence/Machine Learning
API	Application Programming Interface
BIPM	The International Bureau of Weights and Measures
CDIF	Cross-disciplinary interoperability framework
CEINT	Center for the Environmental Implications of NanoTechnology
CHMO	Chemical Methods Ontology
ChEBI	Chemical Entities of Biological Interest
CIF	Crystallographic Information File
CSD	Cambridge Structural Database
DOI	Digital Object Identifier
EBI	European Bioinformatics Institute
ELN	Electronic Laboratory Notebook
EMMO	Elementary Multiperspective Material Ontology
FAIR	Findable, Accessible, Interoperable, Reusable
FERs	FAIR Enabling Resources
FIP	FAIR Implementation Profile
GUM	Guide for estimation of Uncertainty in Measurement
ICSD	Inorganic Crystal Structure Database
InChI	International Chemical Identifier
IUCr	International Union of Crystallography
IUPAC	International Union of Pure and Applied Chemistry
JSON-LD	Javascript Object Notation – Linked Data
LIMS	Laboratory Information Management Systems
mmCIF	Macromolecular Crystallographic Information File
MOF	Metal-organic Framework

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

NanoCommons	Nano-Knowledge Community
NanoFASE	Nanomaterial Fate and Speciation in the Environment
NFDI4Chem	Nationale Forschungsdaten Infrastruktur, Chemistry Consortium
NIST	National Institute of Standards and Technology
PEARs	Publishers, Editors, Authors, Reviewers
PHP	Hypertext Preprocessor scripting language *(recursive acronym)*
PID	Persistence Identifier
PSDI	The UK Physical Data Sciences Infrastructure Project
R&D	Research & Development
RDA	Research Data Alliance
RDaF	Research Data Framework
RDM	Research Data Management
RIPE	Reliability, Interbitibility, Processability, Exchanagibilty
RXNO	Reaction Ontology
SI	The International System of Units
SMILES	Simplified Molecular-Input Line-Entry System
TDE	ThermoData Engine
UVCBs	Substances of unknown or variable composition, complex reaction products, or biological materials
VAMDC	Virtual Atomic and Molecular Data Centre Consortium
WFC	WorldFAIR Chemistry

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Executive Summary

The overarching goal of the WorldFAIR Chemistry Work Package (WP03) is to support the use of chemical data standards in research workflows, between and across disciplines. This will enable downstream data reuse through provision of practical direction and resources. The aim of this deliverable is to establish a framework that can be used by policymakers and developers of services and tools to support FAIR (Findable, Accessible, Interoperable and Reusable) reporting of chemical data. Specific objectives are to highlight applicability of existing standards at a practical level and to identify gaps that need to be addressed to achieve wider data re-use goals.

This report reviews some of the critical and persistent issues around documentation of chemical information. These were identified through a series of webinar panels on the theme: “What is a chemical?”, and through other conferences, workshops, and ongoing collaborative projects run as part of the WorldFAIR project and by the International Union of Pure and Applied Chemistry (IUPAC, the lead organisation of WP03).

Chemicals are everywhere and every tangible object has a chemical nature that impacts its use and behaviour in the environment. As chemical data and chemical principles are increasingly applied broadly across disciplines, the range of representations and contexts for chemical substances and data become more diverse and less easy to precisely define. Molecular entities are fundamental to our understanding of material properties and underlie the configuration of many chemical data models and resources, but it is also critical to look beyond the molecule to particles, surfaces, and states, and their behaviour under different conditions. Few chemistry-related disciplines have mature standards, and better practices in data reporting and interoperability are needed across the board, in both industry and academia. This will allow sharing and reuse requirements to be met in relation to international chemicals management policies and sustainable development goals.

This report additionally considers documentation requirements to achieve FAIR sharing of chemistry data in ways that are Reliable, Interpretable, Processable, and Exchangeable (RIPE), and with minimal loss of quality. Increasing the level of consumable FAIR data depends upon documenting data upstream of sharing to ensure that meaning and quality can be assessed and reassessed appropriately. It is not enough for data objects to be accessible; data need to be accompanied by metadata which provide the contextual information required to enable interoperability and reuse. Fully articulating the scope, structure, and exposure of metadata is critical to enable broader technical mechanisms for programmatic data exchange. The RIPE framework can help research ecosystems across sectors to focus on information requirements, resources and practices required to facilitate provision of data that are mature for sharing, and fully AI-ready across a broad range of use cases. Consistent and comprehensive communication of existing and emerging standards and resources is an important priority to effectively address the challenges confronting meaningful and effective reuse of chemistry data.

Collectively, the chemistry community has over a century of experience in developing and refining standards for communicating high quality chemical information. Explorations undertaken within and alongside WP03 are helping to clarify where these fall short of FAIR ideals, and how we can advance in addressing more complex needs across chemistry and other disciplines. While we have many of the components needed, further refinement of current processes and tools are necessary to enable establishment and use of workflows for sharing quality chemical data, particularly in interdisciplinary contexts. The present focus on the FAIR data principles provides a framework to enable previously well-established chemistry standards to become accessible and applicable for automated programmatic reuse. We envisage this report as a living document evolving over the course of the project, as we further assess IUPAC digital standards to support FAIR chemical data sharing. Future sections are planned that will provide a Roadmap and a Sustainability Blueprint for standards development and adoption as part of our collective recommendations for supporting chemical data reporting policy and practice.

The primary target audience for this report is the range of professionals involved in building and managing systems and services that support process engineers, scientists and other researchers working with data. We will also reach those involved in information management and communication, including professionals in publishing houses, libraries, standards organisations and at other information resources. Additional audiences include chemists, data scientists and other researchers who are actively working with informatics and programmatic applications, and those who are in positions to influence policy that impacts chemical data reporting and exchange.

Other deliverables under development in WorldFAIR Chemistry (WP03) will further demonstrate and facilitate the use of chemistry data standards, including a digital cookbook of interactive recipes demonstrating how to handle chemical data (D3.2 Training package), and protocol specifications for exchanging chemical representations and other metadata via Application Programming Interface (API) services (D3.3 Utility services).

WP03 activities are coordinated through the International Union of Pure and Applied Chemistry (IUPAC), the world authority on chemical nomenclature and terminology that constitute a common global language for communicating chemistry. In the context of the formal IUPAC process for reviewing and adopting consensus standards in chemistry, this work should be regarded as provisional guidance. Complete review and adoption of standards through IUPAC to reach the status of “Recommendation,” which has a specific meaning in the IUPAC lexicon, will occur after WP03 is complete.

Table of contents

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

1. Introduction

The goal of the WorldFAIR Chemistry Work Package (WP03) is to facilitate the use of chemical data standards to enable chemical data reuse. The focus of this deliverable is to provide broad strokes guidance on approaches to descriptive metadata, file formats, unit representation, terminology and other digital motifs for describing chemical data in repositories and other systems and workflows.

Increasing numbers of chemistry-related datasets are being shared in response to research funding mandates and reporting requirements. Many different stakeholders can be involved in processes and workflows to capture, prepare, publish and compile datasets, including researchers, publishers, repositories, software developers, instrument facilities and libraries, among others. To ensure data are discoverable and re-usable will necessitate community-wide practices for describing data that meet the FAIR criteria for machine-readability and utilise domain standards to facilitate long-term interoperability across collections and resources. This report describes a guidance framework for handling and reporting FAIR-enabled chemistry data for different stakeholder roles that are involved in developing policies, practices, products, and services. The chemistry domain is supported by a well-established chemical information professional network and facilitating implementation of standards by tool builders, database developers and other service providers will further empower the research community to share and reuse data.

Guidance pertaining to data reporting and sharing arises in many contexts for different purposes, including from regulatory bodies, national level funding initiatives, journals, repositories, and within communities of practice. The aim here is not to mandate but to provide context for how to address reporting of FAIR chemical data to maximise reusability, building on the International Union of Pure and Applied Chemistry (IUPAC) standards and expertise in standard chemical description. IUPAC has a long history of defining systematic chemical nomenclature and terminology, and evaluating the uncertainty of chemical properties as cornerstones for the application and exchange of modern chemical knowledge. IUPAC continues this effort in the digital space, updating and enhancing existing digital assets, digitalising other critical IUPAC standards, providing guidance on adoption and use, and providing resources and mechanisms for validation of chemical data.

In approaching this brief, we are also cognisant of extensive histories of digital representation in the field of cheminformatics and other chemistry sub-disciplines, including existing community best practices and solutions for representation of chemical structure, systems and data that can provide the foundations for any definitive recommendations. Chemistry intersects with many domains and we are aware of some data related issues in other disciplines but better understanding of use cases is needed and we are beginning to further elaborate these through interactions with other WorldFAIR work packages. We are also inspired by, and collaborating with, a variety of exemplar community projects and emerging initiatives that will provide insights into more specific needs for implementation. Insights from these collaborations and the broader community are discussed in the landscape overview in Section 2.

Parallel to this are WorldFAIR cross-package activities to assess FAIR implementation, which we are applying to ascertain the suitability and usability of IUPAC digital standards for enabling FAIR attributes of chemical data in other systems. The emerging developments on a cross-disciplinary interoperability framework of common motifs and infrastructure will also be informative, and IUPAC and the broader community can tap into these to facilitate digital data exchange. Applicability of IUPAC standards and other community resources are discussed under implementation frameworks in Section 3. As these collective analyses progress, we will formulate a Roadmap and a Sustainability Blueprint for further digital standards development in IUPAC.

Consistent communication of the range of resources that exist and how to utilise these in different contexts for reporting FAIR chemical data was found to be most needed in supporting the formulation of policies and guidelines. Therefore, we offer here a framework, summarised in Section 4, as a starting point for the community to develop context-specific guidance based on principles and standards for reporting quality chemical data developed by IUPAC and other authoritative bodies. The emphasis is on highlighting where specific standards exist to enable data sharing policies at a practical level, and where there are gaps that need to be addressed. We anticipate that this framework will continue to evolve as emerging standards mature and broader implementation informs best practices.

2. Landscape overview: Contextualising guidance

There is a long history of reporting and compiling chemical-related data, spanning over 200 years of published sources[1,2]. For practical reasons, data have been collected in many separate resources developed in different fields for various purposes. Challenges curating data from the literature into these bespoke data systems have prompted calls, at various times, for better reporting practices. The data standards and practices developed over time by these groups provide a robust foundation for chemistry to facilitate quality data reporting and sharing in a networked environment.

This section attempts to contextualise the landscape for reporting chemical data and lays out some of the critical and persistent issues around documentation from a high level across a number of perspectives. Sources of discussion in this section derive from a number of conferences, workshops, and collaborative projects, including through IUPAC and more generally in the chemical information and data science communities (see Appendices 5.1 - 5.3 for summaries of recent activities). Interested readers are encouraged to follow through on the many resources and publications referenced throughout the report.

[1] Beilstein, F. R. *Handbuch der Organischen Chemie*, 1st ed.; Leopold Voss, 1881; *(data from 1771).*

[2] Gmelin, L. *Handbuch der anorganischen Chemie*, 1st ed.; Franz Varrentrapp, 1817; *(data from 1772).*

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

2.1. What is a chemical?

What is a chemical? While chemicals are everywhere and every tangible object has a chemical nature, what and how this is described can vary widely, depending on the field, the context, and the scope of study. This fundamental question is relevant for FAIR data sharing and reuse because it impacts description of the objects of study in research, the variables under consideration and the interpretation and application of data to produce knowledge.

Data come from measurements made on physical samples, but interpretation and organisation of data in a chemical context is generally processed around representation of molecules. For chemical data to retain its utility, information describing the sample and representing the chemical(s) involved needs to be tied to the data derived from the sample. Chemistry and chemical samples are often represented and described at both macroscopic and molecular levels. Different views of the same chemical may include different information and may not be fully reducible to a single expression without some loss of information.

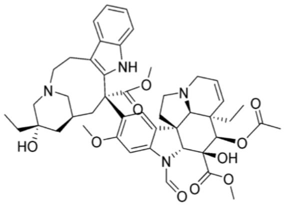
2.1.1. Chemical representation

Chemical representation systems are systematically defined languages used by chemists to encapsulate conceptual models of molecules. The most powerful systems utilise two- and three-dimensional spatial representations. These are often expressed visually for chemists as diagrams but are also expressed numerically in a variety of modalities for machine processing. Various alpha-numeric linear notations, both human readable and machine-processable, have also been developed for convenience in text-base scenarios (e.g. tables, searching, indexes, metadata). See Figure 1 for several examples of chemical representation in common use, including the IUPAC standard nomenclature and InChI string notations as well as other community conventions.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Chemical Structure Representation (various depictions)

Vincristine
(trivial name)



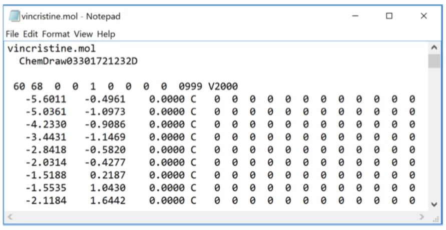
IUPAC name - standardized nomenclature

(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



```
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

Figure 1. Various chemical representation depictions in common use for vincristine, an active ingredient in chemotherapy medication. (Source: modified from Scalfani & McEwen, 2019, https://osf.io/psq7k, CC-BY 4.0 International.)

Chemical information systems will often store 2D chemical structure diagrams as ‘connection tables’ of atoms and bonds. This representation lends itself to graph theory based on analysing relationships between ‘nodes’ and ‘edges’ (e.g. atoms and bonds). The field of cheminformatics has exploited this conceptual model to detect various degrees of isomorphism, or sameness, between molecular entities, and to enable structure-based searching[3]. These methods open up opportunities for different approaches to notation that emphasise different relationships; for example, functional groups associated with desired reactivities or properties. Vast numbers of comparisons can be managed computationally, but as different preferred notations are stored and exchanged across published datasets and in code, dialects of this basic diagrammatical chemical language proliferate[4]. This is a fundamental challenge in achieving interoperability across chemical datasets.

Such digital representation systems for chemical structure have been refined over time to work well for discrete organic molecules where the behaviour of covalent bonds between atoms are constrained by well defined models derived from much-observed chemical precedent. However, limitations exist in capturing more complex topology and different types of intramolecular

[3] Willet, P. Chemoinformatics: a history. *WIREs Comp. Mol. Sci.* 2011, 1(1), 46-56. https://doi.org/10.1002/wcms.1
[4] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* 2011, 3, 33. https://doi.org/10.1186/1758-2946-3-33

relationships and these are magnified as the domain of application extends towards inorganic chemical entities. There are no agreed conventions for representing bonding to metals that reflect delocalised, dative and multi-centre pi bonding concepts commonly used in the expression of organometallic and coordination compounds. Conventions for representing stereochemistry around metal centres exist but are not universally adopted and other aspects of stereochemistry such as atropisomerism[5] present a broader challenge. Community consensus on how to handle the many different tautomeric transformations[6] that can be observed has yet to be established, particularly in the context of generation of standard identifiers[7].

Representation of macromolecules such as biomolecules and polymers present challenges notating relationships between different or repeating subunits. This is further complicated in multiple dimensions, as is commonly the case in metal-organic frameworks (MOFs) for example. Other relationships involving multiple molecular components where there is also scope for improvement in machine-actionable representations include mechanically interlocked molecules and systems where components and composition may be unknown or variable (e.g. Chemical Substances of Unknown or Variable Composition, Complex Reaction Products and Biological Materials: UVCBs)[8].

Chemists use different modes of chemical representation to emphasise different chemical features. Many of these representations can be resolvable more or less through translation with adequate contextual description (i.e. rich metadata). However, some approaches convey different information and conversion between them may be vulnerable to loss of information and lead to ambiguity about structure and identity. All modes of chemical representation embody the interpretation of models of chemical behaviour of interest to those who have assigned them. These different views of the same chemical lead to the consequence that representing a chemical is unlikely to ever be completely reducible to a single expression, or even a 'family' of expressions that could constitute one formal digital representation for one chemical.

2.1.2. Chemical systems

The meaning of chemistry revolves around hypothetical concepts, including the notion of chemicals as molecular entities with immutable properties. Being able to chemically interpret information about a real-world sample depends on multiple pieces of information that enable the user to tie the sample and observed measurement to the definitions of these concepts. The relationships of these

[5] Astropisomerism: stereoisomerism arising because of hindered rotation about a single bond.

[6] Tautomeric transformations: interconversion between structural isomers of a chemical compound that differ only in the position of protons and electrons.

[7] Dhaked, D. K.; Ihlenfeldt, W.; Patel, H.; Delannée, V.; Nicklaus, M. C. 'Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2'. *J. Chem. Info. Model.* 2020, 60(3), 1253–75. https://doi.org/10.1021/acs.jcim.9b01080

[8] Lai, A.; Clark, A. M.; Escher, B. I.; Fernandez, M.; McEwen, L. R.; Tian, Z.; Wang, Z.; Schymanski, E. L. *Environ. Sci. Technol.* 2022, 56(12) 7448–7466. https://doi.org/10.1021/acs.est.2c00321

concepts are framed by the chemical system. The chemical system is defined by our understanding of the composition and state of the sample under the conditions of the experiment.

The three dimensions of composition, state and conditions are necessary to describe any chemical system. Each of these can be further characterised and the collective scenario will determine the macro-properties of the sample to hand. The interactions among these dimensions can be studied and represented in various ways. For example, phase is used to describe the easily recognizable qualities of gas, liquid or solid that indicate uniform chemical composition and physical state within that phase. The nature of the phase can be altered by change in conditions, such as temperature and/or pressure (see Figure 2). Factors that impact composition, state and conditions are critical to capture for any chemical measurement.

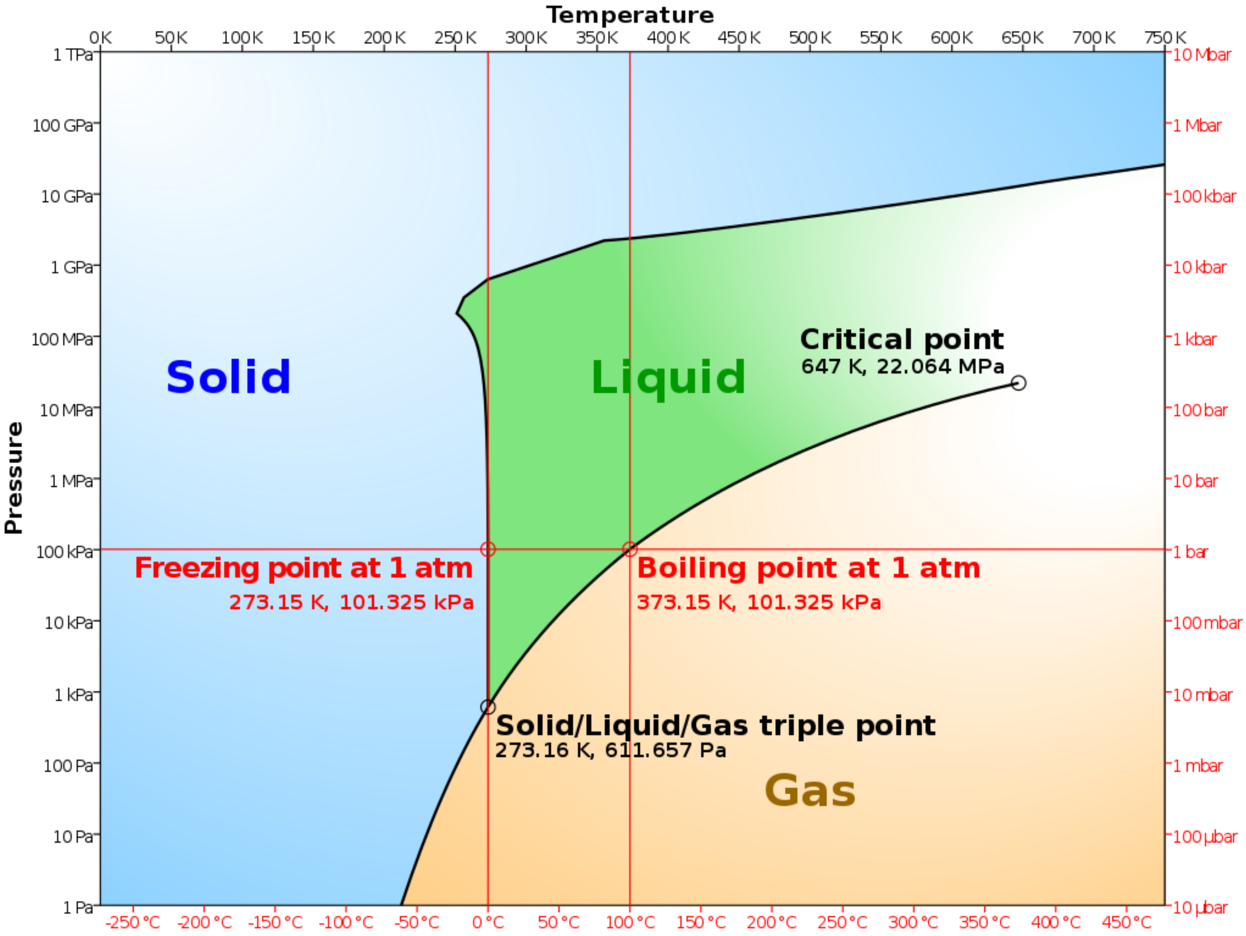


Figure 2. Simplified temperature/pressure phase change diagram for water (note multiple units provided for both temperature and pressure). (Source: Wikipedia, Cmglee, CC-BY-SA 3.0. https://commons.wikimedia.org/wiki/File:Phase_diagram_of_water_simplified.svg, accessed 20230429.)

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Measurement samples constitute chemical systems of variable complexity. The sample origin may be the product mix from a synthesis reaction, or the output of a batch process, or collected from a natural environment. Even so-called ‘pure’ single compounds are idealised concepts. In the real world, all samples are mixtures of different components, which can be known or characterised to some extent. More complex materials may involve multiple phases, or mixtures of sub-mixtures that exhibit different properties and have an overall impact on the response of the material to different conditions. Experimental data are generally obtained as the result of a particular experiment on a specific (batch of) sample that has been characterised relative to its origin and composition.

IUPAC defines a number of concepts to describe chemical entities and their various relationships in a system, at both the level of molecular entities and of distinct chemical species present in a sample[9]. These distinctions can be useful for associating measured activity data with specific molecular features, or modelling behaviour of different chemical species at a macro-scale. However, there is very little consistency in reporting detailed information about chemical systems. Even relatively simple mixtures such as solutions are too often described only in terms of the component of most interest - generally the solute - or the active compound in a more complex drug formulation.

2.1.3. Chemical data

There are a large number of measurement types that pertain to chemical systems. Among these are measurements assessing physical state, chemical composition, molecular characterisation, and various biological activities. Numerous physical quantities are used in chemistry to define these parameters, and their expressions are summarised in a concise guide provided by IUPAC[10]. This wide range of properties enables the study of the material world in both physical and chemical terms. Units provide a frame of reference for how a given quantity is expressed (e.g. starting point, scale). The international system of units or SI base unit for the mole, expressing the amount of substance in terms of numbers of molecules, is of particular relevance for bridging sample and molecular levels[11].

[9] Perrin, C. L.; Agranat, I.; Bagno, A.; Braslavsky, S. E.; Fernandes, P. A.; Gal, J.; Lloyd-Jones, G. C.; Mayr, H.; Murdoch, J. R.; Nudelman, N. S.; Radom, L.; Rappoport, Z.; Ruasse, M.; Siehl, H.; Takeuchi, Y.; Tidwell, T. T.; Uggerud, E.; Williams, I. H. Glossary of terms used in physical organic chemistry (IUPAC Recommendations 2021). *Pure App. Chem.* 2022, 94(4), 353-534. https://doi.org/10.1515/pac-2018-1010

[10] Stohner, J.; Quack, M. *A Concise Summary of Quantities, Units and Symbols in Physical Chemistry.* 2011, International Union of Pure and Applied Chemistry. https://publications.iupac.org/ci/2011/3304/July11_green-sup-4p.pdf (accessed 20230429).

[11] IUPAC Gold Book. IUPAC – mole (M03980). International Union of Pure and Applied Chemistry; 2019. https://doi.org/10.1351/goldbook.M03980

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

As with chemical representation, many conventions are evident in expressing quantities for chemical and physical properties. Chemical analysis is widespread across many fields, and it is inevitable that property measurements are reported in a variety of ways, including use of different units, symbols and other notations[12]. A considerable level of detail can be encapsulated in the formatting of notations that may not be readily apparent to users in different fields, not to mention being lost to machine algorithms. Distinctions between these conventions may be semantic as well as syntactic and they are not always directly comparable. Interconversion between different expressions may require additional property characterisation that may or may not be known, especially in the case of mixed or impure substances[13].

Tolerance for error in downstream applications is another consideration for re-use of reported data. Chemical measurements can be conducted in different ways using a variety of methods and techniques, and results will inevitably deviate from the 'true' property value of an idealised system. Measurement accuracy can be impacted by both random and systematic sources of error, and various approaches are used to assess and limit possible contributing factors; for example, by calibrating instruments and use of control samples. Estimating the contribution of various sources of error can provide an indication of the uncertainty of the measured value, and its contribution to the overall propagation of error in further functions utilising the measure value. This is an important piece of information for gauging fitness-for-purpose in applications that require a high level of precision or confidence[14].

Compiling and integrating chemical data can provide a more complete picture of the characteristics of an idealised chemical system. Repeated measurements can be compared and critically evaluated based on factors of error. IUPAC and other authoritative bodies provide expert assessment of best available values for various chemical and physical properties that can be very useful for comparing measurements from new samples that are less well characterised. These 'standard' values can also be used as training and validation sets for Artificial Intelligence and Machine Learning (AI/ML) applications and other models.

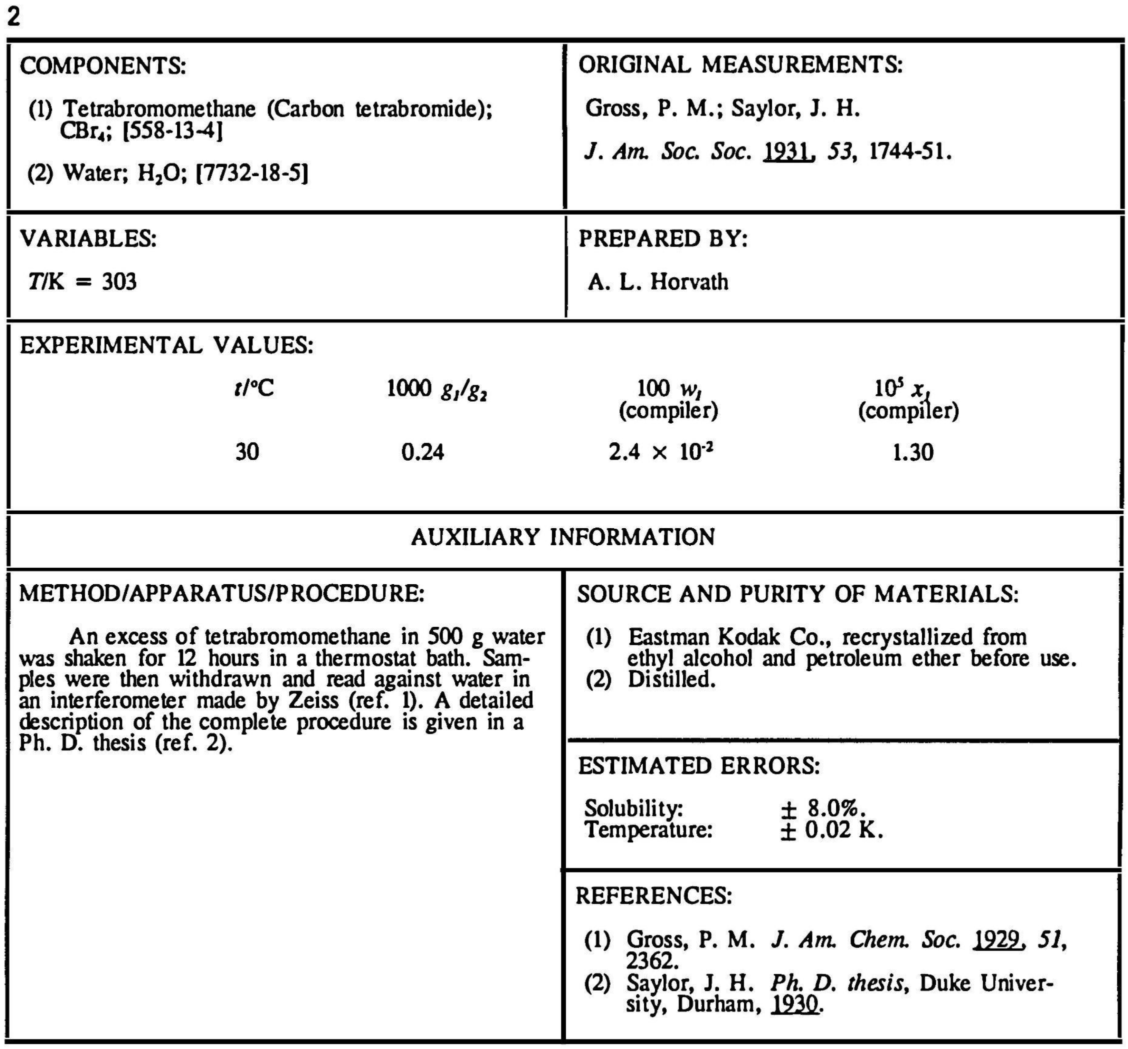
Crucial to both compilation and subsequent re-use of chemical data is unambiguous description of the data and measurement parameters. How measurements are defined and articulated can impact the association of the results with features of the chemical system and what further information can be derived. The range of characteristics of a chemical system that may be of interest and that can

[12] Cvita˘s, T. Quantities Describing Compositions of Mixtures. *Metrologia*. 1996, 33, 35–39. https://doi.org/10.1088/0026-1394/33/1/5
[13] Battino, R.; Clever, H. L.; Fogg, P. G. T.; Young, C. L. Introduction to the Solubility Data Series; Solubility of Gases in Liquids. In, *Carbon Dioxide in Water and Aqueous Electrolyte Solutions*; Scharlin, P., Ed.; Solubility Data Series, Vol. 62; International Union of Pure and Applied Chemistry, Oxford University Press, 1996; pp vi-xiv.
[14] Shaw, D. G.; Bruno, I.; Chalk, S.; Hefter, G.; Hibbert, D. B.; Hutchinson, R. A.; Magalhães, M. C. F.; Magee, J.; McEwen, L. R.; Rumble, J.; Russell, G. T.; Waghorne, E.; Walczyk, T.; Wallington, T. J. Chemical data evaluation: general considerations and approaches for IUPAC projects and the chemistry community (IUPAC Technical Report). *Pure App. Chem.* 2023, *in press*.

impact further interpretation and manipulation of the material under study are dependent on many factors. Thus, documentation practices for chemical data can have a direct impact on the downstream utility of the information, and sufficient granularity of detail as defined by the method of measurement and accepted practice is necessary to retain chemical meaning.

Figure 3 illustrates an example datasheet prepared by a compiler in 1996 summarising information about an original solubility measurement reported in 1931. Note the inclusion of auxiliary information - for example, the estimated error and reference to the source and purity of materials as originally provided by the author. Using this auxiliary information, the compiler was able to convert the units of the original measurement to more updated conventions.

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.

Figure 3. Example summary datasheet of a reported solubility measurement. (Source: Horvath, A. L.; Getzen, F. W.; Eds. Halogenated Methanes with Water. Solubility Data Series, International Union of Pure and Applied Chemistry, Vol. 60, pg 2; Oxford University Press, 1995. Used with permission from IUPAC.)

2.2. Chemical data across disciplines

The chemistry-related needs of activities such as pharmaceutical drug development have driven a lot of community activity relating to machine representation of chemicals and data for half a century[15]. This emphasis has tended to focus on representation of discrete organic molecules and their properties. However, chemistry data are relevant across many other disciplines, including materials science, earth sciences, oceanography, and astrophysics. As chemical data and chemical principles are increasingly applied in other disciplines, the contexts for chemical substances and data and the range of conventions used to represent them becomes more diverse. What is meant by a chemical becomes even less easy to precisely define and manage for interoperability.

The following sections identify some key facets of data practices in disciplines surfaced in the WorldFAIR Chemistry "What is a chemical?" Webinar series, the Research Data Alliance (RDA) P20 Chemistry Research Data session (see Appendices 5.1 and 5.2), outputs from other WorldFAIR work packages and other IUPAC collaborations.

2.2.1. Nanomaterials

Chemical and physical characterisation of nanomaterials is vital to understanding their toxicity and potential impact on the environment. Nanomaterials are typically multi-component and multi-layered, and can be heterogeneous within and between batches. It is important to be able to describe surface properties and how particles interact, as well as core chemical composition along with any impurities that might be present. Further, a nanomaterial may change depending on its surrounding environment and such transformations should ideally be captured.

[15] Willet, P. Chemoinformatics: a history. *WIREs Comp. Mol. Sci.* 2011, 1(1), 46-56. https://doi.org/10.1002/wcms.1

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Standards for the description of nanomaterials were established in 2015[16] and continue to be built upon today through the specification of standard identifiers[17] and work being undertaken through initiatives such as NanoCommons,[18] NanoFASE[19] and CEINT[20].

2.2.2. Materials science

Materials science draws on metallurgy, mineralogy and ceramics. It involves the blending of chemistry, physics and engineering to arrive at a repeatable arrangement of atoms. Small changes in structure can lead to significant changes in properties. Chronology and an understanding of a material's history - the processes applied, the ingredients consumed - is vital. Materials data tend to be sparse, heterogeneous, unstructured, complex and expensive. The Materials Genome Initiative is developing infrastructure to pipeline data into materials discovery and optimisation[21].

2.2.3. Earth sciences

Geochemistry is fundamental to understanding processes in natural systems that impact on our planet including its environment and resources. Data collected in geochemistry are very variable and heterogenous. A sample may be subjected to many analyses and there will be multiple analytes to consider. As well as chemical composition and analytical data, capturing when and where a particular sample was collected and how it has been treated are also vital. This is referred to as sample provenance information.

Cosmochemistry aims to understand the composition of materials collected from or detected and characterised on other planets, including the abundances of elements, isotopes and gases. Methodologies used are new and emerging and the ways they report data are diverse.

[16] The Committee on Data for Science and Technology (CODATA), International Science Council. *Uniform Description System for Materials on the Nanoscale (v.1.0).* Published 9 March 2015. https://codata.org/uniform-description-system-for-materials-on-the-nanoscale-v-1-0-published/ (accessed 20230428).

[17] Lynch, I.; Afantitis, A.; Exner, T.; Himly, M.; Lobaskin, V.; Doganis, P.; Maier, D.; Sanabria, N.; Papadiamantis, A. G.; Rybinska-Fryca, A.; Gromelski M.; Puzyn, T.; Willighagen, E.; Johnston, B. D.; Gulumian, M.; Matzke, M.; Etxabe, A. G.; Bossa, N.; Serra, A.; Liampa, I.; Harper, S.; Tämm, K.; Jensen, A.; Kohonen, P.; Slater, L.; Tsoumanis, A.; Greco, D.; Winkler, D. A.; Sarimveis, H.; Melagraki, G. Can an InChI for Nano Address the Need for a Simplified Representation of Complex Nanomaterials across Experimental and Nanoinformatics Studies? *Nanomat.* 2020, 10(12) 2493. https://doi.org/10.3390/nano10122493

[18] *NanoCommons, Nano-Knowledge Community.* https://www.nanocommons.eu (accessed 20230430).

[19] *Nanomaterial Fate and Speciation in the Environment (NanoFASE).* http://www.nanofase.eu (accessed 20230430).

[20] *Center for the Environmental Implications of NanoTechnology (CEINT).* https://ceint.duke.edu (accessed 20230430).

[21] *Materials Genome Initiative Home Page.* https://www.mgi.gov (accessed 20230430).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Communities within chemistry have been collaborating with Earth Sciences initiatives to explore opportunities for interoperability informed by the FAIR data principles[22] and to help establish new communities that can identify and support the data needs of researchers in Geochemistry[23].

2.2.4. Astrochemistry and physics

In astrochemistry and physics communities, there have been significant efforts over more than ten years to address the interoperability of resources that capture atomic and molecular data observed in outer space[24]. Standard chemical identifiers have been used to enable federated search across resources. The scope of these efforts has extended to address the needs of atmospheric science and physics communities. Representation and terminology standards and machine-actionable dictionaries are being developed to enable more granular representation of states of atoms, molecules and solids[25].

2.2.5. Oceanography

A key element of oceanography is the study of chemical components of the oceans, their reactions and pathways of transformation within the ocean and at its interface with sediment and the atmosphere. Data and information relevant to oceanography can be found in a range of systems and infrastructure to enable interoperability across these has been developed[26]. Priority areas thus far identified by the Oceanography WorldFAIR Work Package (WP11) have highlighted the need for this infrastructure to better interoperate with chemical data hubs that can provide information on chemical entities found in the ocean including marine pollutants and biogeochemical fluxes[27].

2.2.6. Environmental sciences

The characterisation of chemicals is central to the practice of environmental sciences. Similar to the Earth Sciences, linking data to a particular sample collected from a particular place is also important. The desire by some national governments for more sustainable practices and zero

[22] Stall, S.; McEwen, L.; Wyborn, L.; Hoebelheinrich, N.; Bruno, I. Growing the FAIR Community at the Intersection of the Geosciences and Pure and Applied Chemistry. *Data Intell.* 2020, 2(1−2), 139−50. https://doi.org/10.1162/dint_a_00036

[23] Klöcking, M.; Wyborn, L.; Lehnert, K. A.; Ware, B.; Prent, A. M.; Profeta, L.; Kohlmann, F.; Noble, W.; Bruno, I.; Lambart, S.; Ananuer, H.; Barber, N. D.; Becker, H.; Brodbeck, M.; Deng, H.; Deng, K.; Elger, K.; de Souza Franco, G.; Gao, Y.; Ghasera, K. M.; Hezel, D. C.; Huang, J.; Kerswell, B.; Koch, H.; Lanati, A. W.; ter Maat, G.; Martínez-Villegas, N.; Yobo, L. N.; Redaa, A.; Schäfer, W.; Swing, M. R.; Taylor, R. J. M.; Traun, M. K.; Whelan, J.; Zhou, T. Community Recommendations for Geochemical Data, Services and Analytical Capabilities in the 21st Century. *Geochim. Cosmochim. Acta*, 2023, *in press*. https://doi.org/10.1016/j.gca.2023.04.024

[24] *Virtual Atomic and Molecular Data Center (VAMDC) Consortium.* http://www.vamdc.org (accessed 20230428).

[25] Zwölf, C. M.; Moreau, N. Assessment of the FAIRness of the Virtual Atomic and Molecular Data Centre following the Research Data Alliance evaluation framework. *Eur. Phys. J. D.* 2023, 77, 70. https://doi.org/10.1140/epjd/s10053-023-00649-x

[26] *Ocean Data and Information System (ODIS).* https://oceaninfohub.org/odis (accessed 20230428).

[27] Buttegieg, P. L. WorldFAIR Project (D11.1) An assessment of the Ocean Data priority areas for development and implementation roadmap. 2023. Zenodo. https://doi.org/10.5281/zenodo.7682399

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

pollution in the chemistry industry[28] creates a need for an evolution in the way we assess chemical risk in order to protect health and the environment[29]. Being able to reliably and comprehensively capture chemical substances and associated data will be key to effectively achieving these aims.

2.2.7. Life sciences

Chemical entities are relevant to many biological processes and thus the development of new drugs. Being able to bridge chemistry and biology is vital for powering human and machine analysis, in particular exploiting AI/ML technologies. In the pharmaceutical industry, interoperability challenges are magnified when organisations merge and as partnership networks grow and become more diverse. FAIR characterisation of data is important for reporting and regulation of research outputs[30].

2.2.8. Cross-disciplinary themes

Themes that emerge from this cursory overview of chemistry-adjacent disciplines, and should feed into consideration of FAIR Chemistry guidance, include the following:

- Representation of discrete molecular entities is critical but insufficient for describing the composition of the chemical substances studied across disciplines - it is also critical to consider mixtures, impurities and agglomerations where constitution may be unknown and/or inherently heterogeneous (eg. a distribution of particle sizes or surface functionalisation).
- There is a need to look beyond the molecule to particles and surfaces and to consider the state under the specific conditions and transformations that may occur as a result of extrinsic factors.
- Connecting chemistry data to physical samples as well as their position in time and in space is absolutely key along with factors relating to their natural environment or storage conditions.
- We need to be able to distinguish atomic-level properties such as isotopic abundances and excited states.
- Some disciplines have mature standards whilst in others these are absent or emerging - there are opportunities to be informed by the former and to help address the needs of the latter.
- The need to develop better practices in data reporting and interoperability is relevant to challenges and efficiency improvements in industry as well as in academia.

[28] *EU Chemicals strategy.* https://environment.ec.europa.eu/strategy/chemicals-strategy_en (accessed 20230428).
[29] *Partnership for the Assessment of Risk from Chemicals (PARC).* https://www.eu-parc.eu (accessed 20230428).
[30] Blanke, G.; Doerner, T.; Lynch, N. Chemical Data in Life Sciences R&D and the FAIR Principles. 2020. Zenodo. https://doi.org/10.5281/ZENODO.3970745

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- Guidance should look to inform and be informed by regulatory drivers that result from national and international chemicals management and sustainability policies and ambitions.

2.3. Community-level strategies for data reporting

When starting to think about guidance for chemistry, we are not operating in a void. In this section we identify case studies from domains related to chemistry where there have been successful efforts over time to standardise on policy and practice in order to maximise the potential for data sharing and reuse in alignment with the FAIR data principles. We also look at more recent data infrastructure initiatives that are looking to achieve similar success in supporting research data management in the academic sector across the chemical and physical sciences more broadly. These case studies have been informed by ongoing collaboration with IUPAC generally, through WorldFAIR WP03, and by various meetings and workshops (for example, the RDA P20 Chemistry session; see summary in Appendix 5.2).

2.3.1. Scientific case study: Crystallography

In 2022, around 60,000 small-molecule crystal structure datasets were published, the majority associated with one of ~16,000 articles published in journals that are primarily chemistry-related. The majority will also have been deposited in a trusted data repository prior to publication. The data submitted will have been in a standard semantic file format and will commonly be passed through a community-developed validation service that checks for completeness and consistency.

The culture of data sharing in crystallography has its roots in the early publication of derived and processed data in printed form dating back to the 1940s. The digital data publication workflows that have evolved since the early 1990s have been enabled and catalysed by the adoption of common standards by stakeholders operating across the research lifecycle. At the core of these workflows is the Crystallographic Information File (CIF)[31] which can semantically represent many facets of a diffraction experiment including sample preparation and characteristics, instrument type and configuration, data collection parameters and conditions, data processing methods and constraints as well as the data generated and the final model structure derived from this.

The CIF specification was first published in 1991 and over the course of around a decade became the standard format used for publishing crystal structure data sets. This was greatly aided by the adoption of CIF as the default export format from software tools commonly used to solve and refine a structure. In the following decade there was much focus on establishing joined-up policies and workflows involving data repositories and journal publishers that enshrine guidance issued by the

[31] Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography. *Acta Crystall. Sec. A Found. of Crystall.* 1991, 47(6), 655–85. https://doi.org/10.1107/S010876739101067X

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

International Union of Crystallography (IUCr)[32] and make it easy for researchers to comply with this. Key elements of these workflows include use of checkCIF[33] to generate validation reports and an encouragement to publish processed as well as the final derived data. As with the initial adoption of CIF, default output settings in software tools has greatly influenced the extent to which processed data is included in datasets submitted for publication.

In parallel to these efforts in chemical crystallography, the structural biology community developed similar standards, policies and workflows for macromolecular crystallography. These include a standard exchange format (mmCIF)[34] based on the same semantic framework as CIF and a comprehensive data validation pipeline[35]. Journals publishing macromolecular crystal structures typically require a validation report to have been generated prior to submission of an associated manuscript. In structural biology there has been more attention paid to the publication and archival of raw data[36] with repositories emerging, designed to specifically support this[37,38]. In chemical crystallography, the value of publishing raw data is less clear but is under active consideration.[39]

Datasets deposited in crystallographic data repositories are assigned accession IDs which have the characteristics of a persistent ID in that they are unique and immutable and can also be resolved either through services maintained by the repository or via mechanisms such as identifiers.org[40]. Since 2014, chemical crystallographic data sets have also been assigned DOIs[41] and advantage taken of the metadata registration services offered by DOI registration agencies such as DataCite[42]. These services provide a foundation for interoperability enabling metadata and links between data objects

[32] Larsen, S.; Kostorz, G. Publication Standards for Crystal Structures. 2011. http://www.iucr.org/home/leading-article/2011/2011-06-02

[33] Spek, A. L. Structure Validation in Chemical Crystallography. *Acta Crystall. Sec. D Biol. Crystall.* 2009, 65(2), 148–55. https://doi.org/10.1107/S090744490804362X

[34] Bourne, P. E.; Berman, H. M.; McMahon, B.; Watenpaugh, K. D.; Westbrook, J. D.; Fitzgerald, P. M. D. 'Macromolecular Crystallographic Information Fil'. In *Methods in Enzymology*, 277:571–90. Macromolecular Crystallography Part B. Academic Press, 1997. https://doi.org/10.1016/S0076-6879(97)77032-0

[35] Gore, S.; Velankar, S.; Kleywegt, G. J. Implementing an X-Ray Validation Pipeline for the Protein Data Bank. *Acta Crystall. Sec. D, Biol. Crystall.* 2012, 68(4), 478–83. https://doi.org/10.1107/S0907444911050359

[36] Kroon-Batenburg, L. M. J.; Helliwell, J. R.; McMahon, B.; Terwilliger, T. C. Raw Diffraction Data Preservation and Reuse: Overview, Update on Practicalities and Metadata Requirements. *IUCrJ* 2017, 4(1), 87–99. https://doi.org/10.1107/S2052252516018315

[37] *Integrated Resource for Reproducibility in Macromolecular Crystallography.* https://proteindiffraction.org (accessed 20230428).

[38] *SBGrid Data Bank.* https://data.sbgrid.org (accessed 20230428).

[39] Coles, S.; Sarjeant, A. IUCr Workshop on "When Should Small Molecule Crystallographers Publish Raw Diffraction Data?" *ACA RefleXions*, 2021.

[40] Wimalaratne, S. M.; Juty, N.; Kunze, J.; Janée, G.; McMurry, J. A.; Beard, N; Jimenez, R.; Grethe, J. S.; Hermjakob, H.; Martone, M. E.; Clark, T. Uniform Resolution of Compact Identifiers for Biomedical Data. *Scient. Data* 2018, 5(1), 180029. https://doi.org/10.1038/sdata.2018.29

[41] Datasets, Data Citation and DOIs (2104) https://www.ccdc.cam.ac.uk/discover/blog/post-35 (accessed 20230428).

[42] 'DataCite: Helping You to Find Access and Reuse Research Data', n.d. http://www.datacite.org (accessed 20230428).

to be harvested by resources such as the Data Citation Index[43] and platforms implementing Scholix[44] such as ScholeXplorer[45].

There is a limit to the degree that general data services and repositories can enable and support domain-specific discovery and reuse of data and crystallographic data organisations play an important role in enabling this[46]. Crystallographic data is relevant across many domains, in particular the chemical, materials and biological sciences. Reuse of a small molecule crystal structure dataset requires a machine-actionable representation of the chemical substance studied and currently this is very rarely provided alongside the dataset. It is instead generated through automated and manual curation activities undertaken by specialist data organisations. The ability to combine rigorously described crystallographic data with a rich representation of associated chemical entities enables the aggregation of data collected across generations to provide knowledge-based solutions applicable to research challenges in industry and academia.

Guidance for implementing the FAIR data principles in chemical crystallography today can be captured simply as “put data in a trusted domain repository, in a standard format that can be validated” (see Figure 4). At the heart of this is a standard representation format backed by vocabularies that semantically define the data. Facilitating this are community-supported validation services and data repositories that support joined up data publishing workflows and undertake curation activities. Persistent Identifier (PID) registration agencies provide important services that can enable discovery and provide context at a general level; at a domain-specific level the availability and adoption of standard identifiers and community representation formats are key enablers of interoperability and reuse in a chemistry context.

Finally, it is important to recognise the range of stakeholders who contribute to FAIR data publishing success in crystallography: journal publishers and data repositories, tool providers and instrument manufacturers, scientific unions and individual researchers willing to champion best practice and contribute to the development of standards.

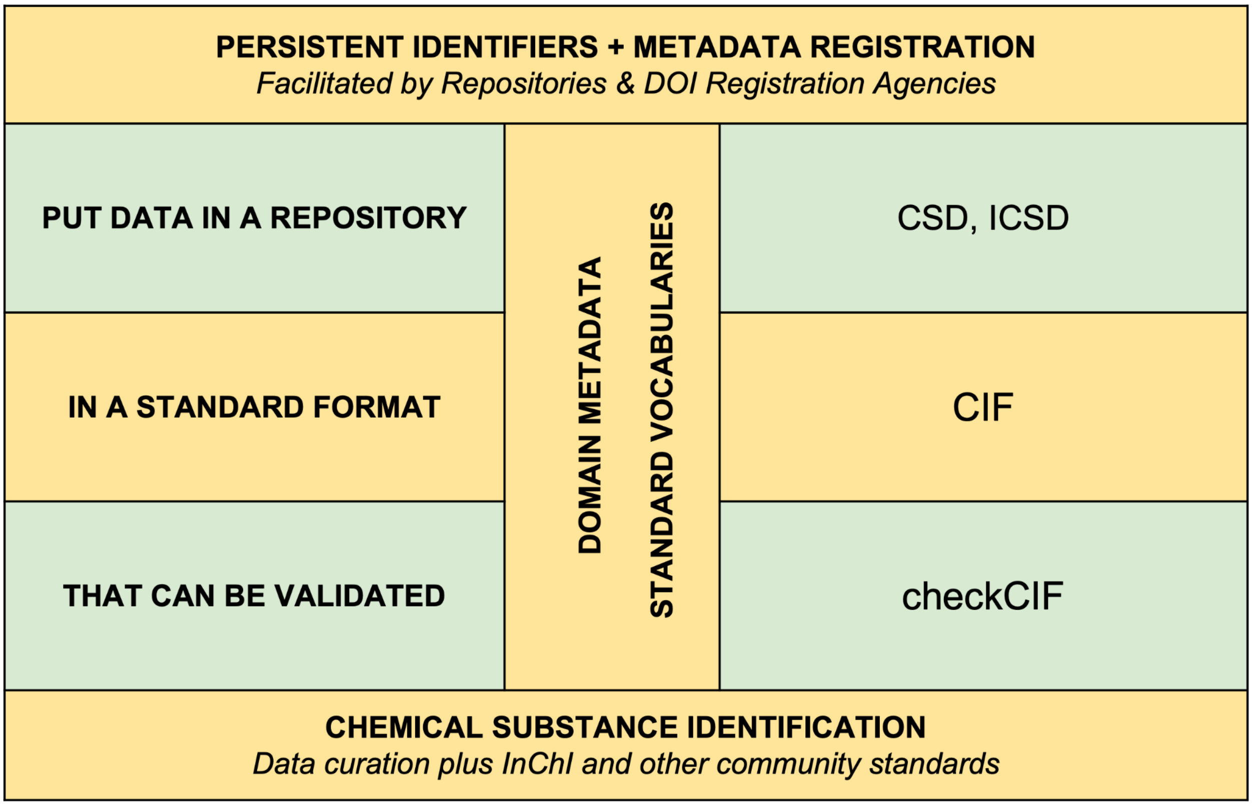
[43] *Data Citation Index.* https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/data-citation-index (accessed 20230428).

[44] Burton, A.; Aryani, A.; Koers, H.; Manghi, P.; La Bruzzo, S.; Stocker, M.; Diepenbroek, M.; Schindler, M.; Fenner, M. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Mag.* 2017 23 (1/2). https://doi.org/10.1045/january2017-burton.

[45] *ScholeXplorer.* https://scholexplorer.openaire.eu/#/ (accessed 20230428).

[46] Bruno, I.; Gražulis, S.; Helliwell, J. R.; Kabekkodu, S. N.; McMahon, B.; Westbrook, J. Crystallography and Databases. *Data Sci. J.* 2017, 16, 38. https://doi.org/10.5334/dsj-2017-038.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.



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

PUT DATA IN A REPOSITORY

IN A STANDARD FORMAT

THAT CAN BE VALIDATED

DOMAIN METADATA

STANDARD VOCABULARIES

CSD, ICSD

CIF

checkCIF

CHEMICAL SUBSTANCE IDENTIFICATION
Data curation plus InChI and other community standards

Figure 4. Schematic representing implementation of FAIR Data Principles in chemical crystallography today. At the core of this are standard formats supported by standard vocabularies that semantically capture data and metadata (Crystallographic Information File-CIF). Domain repositories (Cambridge Structural Database (CSD), Inorganic-Cambridge Structural Database (ICSD)) provide archival and curation services that take advantage of community validation services such as checkCIF. DOI registration agencies provide persistent identifiers and metadata registration that enable discovery at a general level whilst association of data with community-supported chemical substance representations enables reuse across domains.

2.3.2. Scientific case study: Thermodynamic properties

At latest count, almost 2.7 million thermodynamic properties values reported in the literature from over 123 thousand datasets on chemical systems have been captured in the ThermoML Archive[47,48]. These data have been collected in conjunction with five major journals focused on reporting thermophysical and thermochemical measurements. The ThermoML Archive is stewarded by the United States National Institute of Standards and Technology (NIST). The data are openly accessible and formatted in alignment with the ThermoML standard, co-developed by NIST and IUPAC for complete and accurate exchange of thermodynamic property data.

[47] *NIST/TRC ThermoML Archive.* https://trc.nist.gov/ThermoML (accessed 20230428).

[48] Riccardi, D.; Bazyleva, A.; Paulechka, E.; Diky, V.; Magee, J. W.; Kazakov, A. F.; Townsend, S. A.; Muzny, C. D. ThermoML/Data Archive, National Institute of Standards and Technology, 2021. https://doi.org/10.18434/mds2-2422

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The archive itself was compiled over a period spanning more than fifteen years, but the effort to define parameters and expressions for reporting thermodynamic data extends back to the mid twentieth century. These efforts were carried out by NIST and other authoritative standard bodies including IUPAC and the International Bureau of Weights and Measures (BIPM). The need for reporting standards for thermodynamic data was discussed as early as 1953 at the 8th Conference of Calorimetry[49]. Post-World War II rebuilding and economic expansion drove the need for more extensive and higher-quality data. This period coincided with many other innovations in metrology and data expression, from the ratification of the SI units in 1960[50] and into this present era of digital expression and cloud-based data exchange services.

Focus on the precision and accuracy of reporting of thermodynamic property data became a hard push in the 1950s and 1960s[51]. Property data are directly relevant to development of many materials necessary for building infrastructure and ascertainment of data quality and uncertainty is an absolute requirement for engineering applications. Several data services emerged around this time to compile and analyse thermodynamic data from the previous century of literature. These efforts made apparent the importance of complete data reporting so that users and compilers could assess the quality of published property information, and this was emphasised in a 1972 IUPAC guide[52]. Recognition of the need to compile reported property data into different services with maximum precision and a minimum intervention for a wide variety of applications was a motivation for the subsequent development of the NIST ThermoML Archive, originally released in the early 2000s[53].

A key requirement for the ThermoML Archive was the opportunity to exchange information across organisations and systems accurately over the Internet. Critical parameters for expressing the data needed to be defined and agreed on for consistency and to maximise the potential for automated processes. Thus, the XML-based ThermoML storage and exchange format was developed by IUPAC

[49] McCullough, J. P.; Westrum, Jr., E. F; Evans, W. H. Calorimetry Group Adopts Revised Resolution on Data Publication. *Science*. 1960, 132(3440), 1658-1659. https://doi.org/10.1126/science.132.3440

[50] 11th General Conference on Weights and Measures. *Meas. Tech.* 1960, 3, 909–912. https://doi.org/10.1007/BF00977503

[51] Ku, H. H. Precision Measurement and Calibration: Selected NBS Papers on Statistical Concepts and Procedures. In NBS Special Publications, 300v1, Commerce Department; National Institute of Standards and Technology (NIST), 1969. https://www.govinfo.gov/app/details/GOVPUB-C13-6769ef50616bc8a1ca657841ccc19c92 (accessed 20230428).

[52] Physical Chemistry Division, Commission on Thermodynamics and Thermochemistry; International Union of Pure and Applied Chemistry. A Guide to Procedures for the Publication of Thermodynamic Data. *Pure App. Chem.* 1972, 29(1-3), 395-408. https://doi.org/10.1351/pac197229010395

[53] Frenkel, M.; Chirico, R. D.; Diky, V.; Muzny, C.; Dong, Q.; Marsh, K. N.; Dymond, J. H.; Wakeham, W. A.; Stein, S. E.; Königsberger, E.; Goodwin, A. R. H.; Magee, J. W.; Thijssen, M.; Haynes, W. M.; Watanasiri, S.; Satyro, M.; Schmidt, M.; Johns, A. I.; Hardin, G. R. New Global Communication Process in Thermodynamics: Impact on Quality of Published Experimental Data. J. Chem. Inf. Model. 2006, 46(6) 2487–2493. https://doi.org/10.1021/ci600208f

and NIST and deployed in the Archive[54]. The metadata and data structure build on other standard recommendations relevant to well described thermodynamic properties, including definitions for Quantities, Units and Symbols in Physical Chemistry (the IUPAC 'Green Book')[55] and the Guide for Estimation of Uncertainty in Measurement ('GUM')[56]. The ThermoML schema that describes these parameters covers over 120 thermochemical and thermophysical properties, including uncertainties and related metadata.

Also critical for the Archive was the development of a regular pipeline for the ingestion of newly published data from the literature[57]. Five journals focusing primarily on related topics were targeted to pilot a workflow in conjunction with NIST to capture thermodynamic data values and associated necessary information (metadata) into the ThermoML format for incorporation into the Archive (see Figure 5). Templates were developed to guide the capture and review of crucial information by authors, journal reviewers and database editors. Semi-automated mechanisms for data capture to lower the barrier for authors as well as additional steps for assessing completeness and further curation necessitated several iterations of the workflow. Analysis of the data capture indicated that even with specific guidance, around one third of all reported values were missing crucial information about measurement parameters or had other issues that made the data unusable. These issues were partially mitigated by improved review processes to identify where information may be missing and to circle back to authors for correction. This is an indicator of how significant the challenge of incomplete data reporting can be in any field, particularly if manual transfer of information is involved, and how important the process of checking for provenance and completeness is at each exchange.

[54] Chirico, R. D.; Frenkel, M.; Diky, V. V.; Marsh, K. K.; Wilhoit, R. C. ThermoML; An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 2. Uncertainties. *J. Chem. Eng. Data.* 2003, 48(5) 1344–1359. https://doi.org/10.1021/je034088i

[55] Renner, T. *Quantities, units and symbols in physical chemistry*; International Union of Pure and Applied Chemistry. Royal Society of Chemistry, 2007. https://doi.org/10.1039/9781847557889

[56] *GUM: Guide to the Expression of Uncertainty in Measurement*; Joint Committee for Guides in Metrology, International Bureau of Weights and Measures, 2008. https://www.bipm.org/en/committees/jc/jcgm/publications (accessed 20230428).

[57] Chirico, R. D.; Frenkel, M.; Magee, J. W.; Diky, V.; Muzny, C. D.; Kazakov, A. F.; Kroenlein, K.; Abdulagatov, I.; Hardin, G. R.; Acree Jr., W. E.; Brenneke, J. F.; Brown, P. L.; Cummings, P. T.; de Loos, T. W.; Friend, D. G.; Goodwin, A. R. H.; Hansen, L. D.; Haynes, W. M.; Koga, N.; Mandelis, A.; Marsh, K. N.; Mathias, P. M.; McCabe, C.; O'Connell, J. P.; Pádua, A.; Rives, V.; Schick, C.; Trusler, J. P. M.; Vyazovkin, S.; Weir, R. D.; Wu, J. Improvement of Quality in Publication of Experimental Thermophysical Property Data: Challenges, Assessment Tools, Global Implementation, and Online Support. *J. Chem. Eng. Data.* 2013, 58(10), 2699–2716. https://doi.org/10.1021/je400569s

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

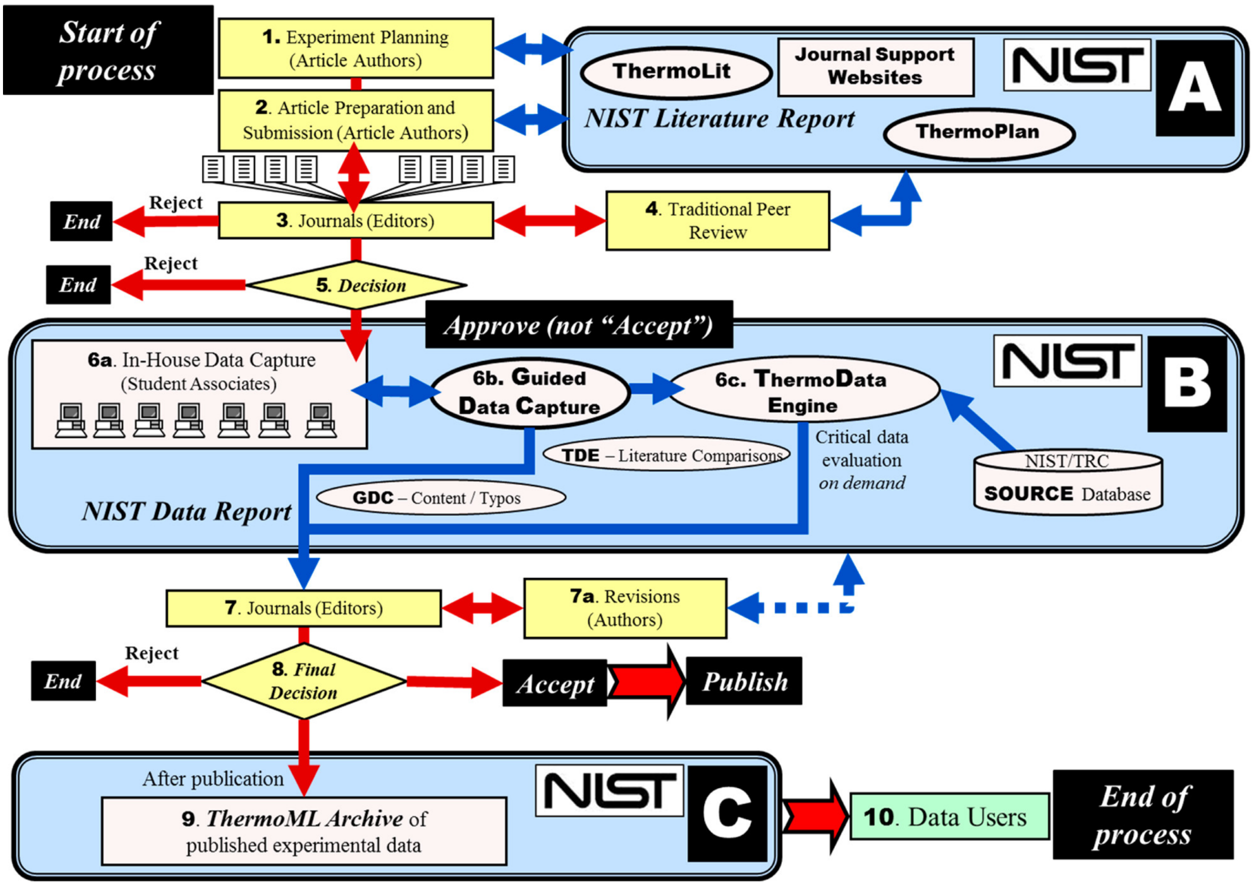


Figure 5. Workflow used for the NIST-Journal cooperation to compile reported thermophysical property data. (Source: J. Chem. Eng. Data. 2013, 58(10), 2699-2716. Used with permission from ACS Publications.)

In addition to making the ThermoML Archive available for re-use by the community, NIST also used the corpus to formulate and test rigorous automated methods for critical evaluation, developed as part of their ThermoData Engine (TDE) project[58]. At the time of this writing, the ThermoML Archive includes data reported up to and including the year 2020. It is not known if the reporting process from journals is currently active and whether the workflow was considered sustainable by the parties involved. The dissemination infrastructure of the existing corpus has recently been upgraded in alignment with other NIST databases, including additional machine-readable export options (e.g. JSON-LD[59]) and other FAIR-enabling enhancements, such as an API[60]. The entire content remains

[58] Frenkel, M.; Chirico, R. D.; Diky, V.; Yan, X.; Dong, Q.; Muzny, C. ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept. *J. Chem. Inf. Model.* 2005, 45(4) 816−838. https://doi.org/10.1021/ci050067b

[59] JSON-LD: Javascript Object Notation – Linked Data.

[60] Riccardi, D.; Trautt, Z.; Bazyleva, A.; Paulechka, E.; Diky, V.; Magee, J. W.; Kazakov, A. F.; Townsend, S. A.; Muzny, C. Towards improved FAIRness of the ThermoML Archive. *J. Comp Chem.* 2022, 43(12), 879-887. https://doi.org/10.1002/jcc.26842

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

openly accessible and a worked example of how it can be reused is illustrated in Appendix 5.5 of this report.

The ThermoML Archive filled a vacuum for programmatically accessible, compiled and structured, machine-processable thermodynamic property data. The process stressed the importance of ensuring these data are well enough described and expressed according to authoritative scientific standards to be compiled and assessed. IUPAC has subsequently refined formal guidance on reporting thermodynamic property data in 2012[61] and most recently in 2021[62], framed around key principles for property data that are broadly applicable across chemistry and other measurement-based fields. These guidelines emphasise complete and unambiguous description and form the basis for Section 3 of this report on articulating criteria for FAIR machine-readable data expression and exchange. An important goal will be to improve on previous efforts to enable more regular and higher quality sharing of research data as a more sustainable pipeline for further compiling, curation and re-use.

2.3.3. Infrastructure case study: National data initiative in Germany

Germany recently launched a major research funding initiative to develop a National Data Infrastructure (NFDI) to facilitate FAIR data sharing across the research community. The NFDI Initiative comprises 26 domain-based consortia, including NFDI4Chem, NFDI4Cat, NFDI-MatWerk, FAIRmat and others with intersections to chemistry[63]. NFDI4Chem is organised around the implementation of technologies to support the full data life cycle in chemistry research and provides a real-time case study for enabling FAIR chemical data in practice[64]. An overall objective is to advance Research Data Management (RDM) in chemistry research labs and all resources developed in the project are openly available to encourage broader re-use.

NFDI4Chem is anchored on key deployments including a cloud-based Electronic Laboratory Notebook (ELN), a suite of data-specific and general repositories and an ontology mapping service. Other task areas are focusing on facilitating interoperability and community engagement. The project has produced a number of landscape reviews considering current availability and status of community standards and resources relevant to data sharing in chemistry, including ontologies, data

[61] Chirico, R. D.; de Loos, T. W.; Gmehling, J.; Goodwin, A. R. H.; Gupta, S.; Haynes, W. M.; Marsh, K. N.; Rives, V.; Olson, J. D.; Spencer, C.; Brennecke, J. F.; Trusler, J. P. M. Guidelines for reporting of phase equilibrium measurements (IUPAC Recommendations 2012). *Pure App. Chem.* 2012, 84(8) 1785-1813. https://doi.org/10.1351/PAC-REC-11-05-02
[62] Bazyleva, A,; Abildskov, J.; Anderko, A.;, Baudouin, O.; Chernyak, Y.; de Hemptinne, J.; Diky, V.; Dohrn, R.; Elliott, J. R.; Jacquemin, J.; Jaubert, J.; Joback, K. G.; Kattner, U. R.; Kontogeorgis, G. M.; Loria, H.; Mathias, P. M.; O'Connell, J. P.; Schröer, W.; Smith, G. J.; Soto, A.; Wang, S.; Weir, R. D. Good reporting practice for thermophysical and thermochemical property measurements (IUPAC Technical Report). *Pure App. Chem.* 2021, 93(2), 253-272. https://doi.org/10.1515/pac-2020-0403
[63] *Nationale Forschungsdaten Infrastruktur (NFDI).* https://www.nfdi.de (accessed 20230429).
[64] *Nationale Forschungsdaten Infrastruktur, Chemistry Consortium (NFDI4Chem).* https://www.nfdi4chem.de (accessed 20230429).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

formats and journal guidelines[65,66,67]. Generally, many resources appear to exist, but at various levels of accessibility and suitability for supporting open and FAIR data infrastructure. Key to enabling digital workflows is the ability to transfer metadata seamlessly between systems and federation of core services will depend on trust and agreement around standards. Implementable standards involve a number of supporting components, including published guidelines, checklists, data models, ontology terms and reference implementations.

NFDI4Chem and IUPAC have a common goal: to increase the pipeline of chemistry data that are FAIR and available for re-use. Towards this goal, NFDI4Chem and IUPAC are collaborating directly on several joint activities on the above themes, including expansion of digital standards development, implementation of standard notations and terms into metadata schema and ontologies and development of guidance for chemistry data sharing in chemistry journals. These collaborations provide IUPAC with the opportunity to perceive strengths and limitations of current standards to support interconnected chemistry data systems in the context of an RDM workflow.

2.3.4. Infrastructure case study: National data initiative in the UK

The Physical Data Sciences Infrastructure project (PSDI)[68] is aiming to provide data infrastructure that brings together and builds upon the various data systems researchers currently use in the physical sciences across the United Kingdom. This includes integrating data from experiment, simulation and theory and making data from various resources accessible to support analysis workflows and reuse in AI/ML activities. A pilot phase[69] was conducted including a series of case studies to assess the infrastructure needs, challenges, and opportunities in different contexts and to evaluate the potential for development of an integrated resource.

The pilot study[70] ran during 2021 and surfaced the need to reduce barriers for researchers caused by a fragmented data landscape. The PSDI project aims to focus on connecting existing resources, enabling automated workflows and ensuring curation of data, software and tools beyond project lifetimes. Professionalising data skills, and achieving international community consensus around standards are also considered to be key to achieving these aims. PSDI and IUPAC are collaborating in a number of areas, including implementation of standards and resources for skills development.

[65] Strömert, P.; Hunold, J.; Castro, A.; Neumann, S.; Koepler, O. Ontologies4Chem: the landscape of ontologies in chemistry. *Pure App. Chem.* 2022, 94(6), 605-622. https://doi.org/10.1515/pac-2021-2007

[66] Rauh, D.; Blankenburg, C.; Fischer, T. G.; Jung, N.; Kuhn, S.; Schatzschneider, U.;, Schulze, T.; Neumann, S. Data format standards in analytical chemistry. *Pure App. Chem.* 2022, 94(6), 725-736. https://doi.org/10.1515/pac-2021-3101

[67] Parks, N. A.; Fischer, T. G.; Blankenburg, C.; Scalfani, V. F.; McEwen, L. R.; Herres-Pawlis, S.; Neumann, S. The current landscape of author guidelines in chemistry through the lens of research data sharing. *Pure App. Chem.* 2023, *Ahead of Print*. https://doi.org/10.1515/pac-2022-1001

[68] *Physical Sciences Data Infrastructure (PSDI).* https://psdi.ac.uk (accessed 20230428).

[69] *PSDI Pilot.* https://www.psdi.ac.uk/the-pilot (accessed 20230428).

[70] *PSDI Pilot Report: Connecting Digital Research Infrastructures for the Physical Sciences.* https://www.psdi.ac.uk/the-pilot/report (accessed 20230428).

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

A case study undertaken to explore the importance of chemical structure for physical sciences infrastructure[71] noted the criticality of consistent and reliable structure representation to support data discovery and reuse across the research lifecycle. Infrastructure should support a range of structure representations to reflect the differing contexts in which data are generated and reused. Standard chemical identifiers should be widely adopted to enable interoperability across resources. Validated digital representations of substances should be considered as essential metadata and connected to sample metadata where appropriate.

2.4. Guidance for open science and data sharing

Guidance can provide direction, purpose and motivation; it can be prescriptive, aspirational or visionary. Here we review categories of guidance that already impact on the chemistry data landscape ranging from visionary reports that outline overarching goals for science generally through aspirational perspectives from within the chemistry community to specific guidance and policies that practically move the community towards a desired end state.

2.4.1. Framing: Overarching goals

A U.K. Royal Society report on *Science as an Open Enterprise*[72] published in 2012 highlighted the enormous potential that massive amounts of data created by modern technologies have for advancing science and its application in public policy and business. It articulated areas for action necessary to realise this potential that included greater openness, the need for data to be published in reusable forms and the importance of standards in achieving this.

A U.S. National Academies report on *Open Science By Design*[73] published in 2018 put forward the view that openness and sharing of information are fundamental to the progress of science and to the effective functioning of the research enterprise. It specifically called out the importance of sharing and preserving research results under the FAIR principles to help facilitate reproducibility and ensure the reliability of knowledge. Themes of integrity, reproducibility and replicability in research and science are explored in other National Academies reports[74,75] which further reinforce the importance of access to well-managed research data.

[71] Bruno, I.; Ward, S. PSDI Case Study 8: The Role of Structure in Physical Sciences Data Management, 2022. https://www.psdi.ac.uk/case-study-8 (accessed 20230428).

[72] The Royal Society. *Science as an Open Enterprise*, 2012. https://royalsociety.org/~/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf (accessed 20230428).

[73] The National Academies. *Open Science by Design*. National Academies Press, 2018. https://doi.org/10.17226/25116.

[74] The National Academies. *Fostering Integrity in Research*. *Fostering Integrity in Research*. National Academies Press, 2017. https://doi.org/10.17226/21896

[75] The National Academies. *Reproducibility and Replicability in Science*. *Reproducibility and Replicability in Science*. National Academies Press, 2019. https://doi.org/10.17226/25303

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

More recently, the UNESCO *Recommendation on Open Science*[76] published in 2021 provides an international framework for open science policy that aims to recognise the urgency and complexity of current global challenges and the transformative potential of open science practices to address these equitably and transparently. The recommendation highlights the importance of machine-actionable open research data managed in accordance with FAIR principles as a key enabler of broader aims.

Overarching goals that emerge from such reports and recommendations include addressing global sustainability challenges, enriched economic benefits, increased transparency and reproducibility, enabling inter-disciplinary collaboration and ensuring universal equitable access to knowledge.

2.4.2. Perspectives: Desired end state

Whilst formal guidance is typically considered to come from established organisations or societies, encouragement and motivation can also be drawn from the perspectives of individuals, research groups, and communities of practice. The following summarises key themes from recent perspective articles that explore how current practices in chemistry and related areas could and should improve and how barriers to achieving a desired end state might be overcome[77,78,79,80,81,82,83]:

- Improved management and reporting of chemistry data is crucial for improving the reproducibility and usability of chemistry research outputs. Realising the predictive power of data-driven computational methods is hindered by poor data management practices and competing standards.
- There is a significant emphasis on the role that tools and infrastructure must play in enabling best practice, in particular electronic lab notebooks (ELNs) and laboratory information management systems (LIMS). For these to be effective however, they need to implement common standards.

[76] UNESCO *Recommendation on Open Science*, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000379949 (accessed 20230428)

[77] Willoughby, C.; Frey, J. G. Data Management Matters. *Dig. Disc.* 2022, 1(3), 183–94. https://doi.org/10.1039/D1DD00046B.

[78] Rzepa, H. S. The Long and Winding Road towards FAIR Data as an Integral Component of the Computational Modelling and Dissemination of Chemistry. *Israel J. Chem.* 2022, 62(1–2). https://doi.org/10.1002/ijch.202100034

[79] Mroz, A. M.; Posligua, V.; Tarzia, A.; Wolpert, E. H.; Jelfs, K. E. Into the Unknown: How Computation Can Help Explore Uncharted Material Space. *J. Am. Chem. Soc.* 2022, 144(41), 18730–18743. https://doi.org/10.1021/jacs.2c06833

[80] Jablonka, K. M.; Patiny, L.; Smit, B. Making the Collective Knowledge of Chemistry Open and Machine Actionable. *Nat. Chem.* 2022, 14(4), 365–76. https://doi.org/10.1038/s41557-022-00910-7

[81] Yano, J.; Gaffney, K. J.; Gregoire, J.; Hung, L.; Ourmazd, A.; Schrier, J.; Sethian, J. A.; Toma, F. M. The Case for Data Science in Experimental Chemistry: Examples and Recommendations. *Nat, Rev. Chem.* 2022, 6(5), 357–70. https://doi.org/10.1038/s41570-022-00382-w

[82] Herres-Pawlis, S.; Bach, F.; Bruno, I. J.; Chalk, S. J.; Jung, N.; Liermann, J. C.; McEwen, L. R.; Neumann, S.; Steinbeck, C.; Razum, M.; Koepler, O. Minimum Information Standards in Chemistry: A Call for Better Research Data Management Practices. *Ange. Chem. Int'l. Ed.* 2022, 61(51), e202203038. https://doi.org/10.1002/anie.202203038

[83] Agrawal, A.; Choudhary, A. Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mat.* 2016, 4(5), 053208. https://doi.org/10.1063/1.4946894

- There needs to be consistency in policies and practice and some call for there to be mandatory data deposition accompanying published results. This might extend to reporting of failed experiments as well as successful ones, recognising the importance of balancing positive and negative data when training machine learning models.
- We have standards and solutions that can be adopted now and should take advantage of what we have to identify gaps and needs for improvement. Adoption barriers are cultural as much as technical and efforts need to be made to cultivate a change in mindset, not least by incorporating best practices into undergraduate and graduate teaching curricula.

2.4.3. Practical: Specific and targeted

When it comes to putting in place best practices and implementing solutions, then more specific guidance is required to drive activities. This may come from individual researchers - champions in their field identifying practical solutions and communicating these to colleagues within an institution (e.g. Cambridge University Open Data FAQs for chemists)[84], from experts collaborating across institutions to establish best practice (e.g. the Data Curation Network)[85], or from consortia looking to develop infrastructure and best practice for different user groups across a region (e.g. NFDI4Chem Knowledge Base)[86]. The *ACS Guide to Scholarly Communication* provides general guidance for the specific task of communicating results in the chemical sciences, including research data, and is an example of a professional society drawing on the expertise amongst its members[87].

Scientific Unions may issue specific guidance to a particular stakeholder group (e.g. IUCr to publishers of crystal structure data)[88]. Scientific Unions also establish the recommendations and standards needed to reliably report specific data types in a field. In the case of IUPAC this includes terminologies, representation formats and recommendations needed to ensure the precise communication of specific chemistry data types and objects[89].

2.4.4. Policies: Enforcing community norms

Driving adoption of specific guidance are overarching policies issued by influential bodies typically in a position to monitor and enforce compliance. These policies may be issued by regulators who set

[84] University of Cambridge, Chemistry Library. *Open Data FAQs for chemists.* https://www-library.ch.cam.ac.uk/open-data-faqs-chemists (accessed 20230429).

[85] *Data Curation Network.* https://datacurationnetwork.org (accessed 20230429).

[86] *NFDI4Chem Knowledge Base.* https://knowledgebase.nfdi4chem.de/knowledge_base (accessed 20230429).

[87] *ACS Guide to Scholarly Communication*; Banik, G. M., Baysinger, G., Kamat, P. V., Pienta, N. J., Eds; American Chemical Society, 2020. https://doi.org/10.1021/acsguide

[88] Larsen, S.; Kostorz, G. Publication Standards for Crystal Structures, 2011. http://www.iucr.org/home/leading-article/2011/2011-06-02 (accessed 20230429).

[89] International Union of Pure and Applied Chemistry. *Recommendations and Technical Reports.* https://iupac.org/what-we-do/recommendations (accessed 20230429).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

out the rules of doing business in the chemical sector (e.g. the U. S. Food and Drug Administration[90], the European Chemicals Agency[91], etc.), governments issuing edicts that federal funders most abide by (e.g. the U.S. Office of Science and Technology[92]), funders who attach conditions to the grants they issue (e.g., the European Commission[93], U.S. National Institutes of Health[94], the Wellcome Trust[95]), and journal publishers who specify criteria that researchers must satisfy in order to submit their work for publication[96].

Enforcing norms requires that these have been agreed to by a community and that tools and enablers are available to ensure there can be no reasonable excuse for not abiding by these norms. Different disciplines and different sub-domains will be at differing levels of maturity when it comes to having a clear sense of community consensus and the standards needed to shape this. Guidance must thus allow for a community to approach compliance with FAIR data expectations through stages of a journey, something that the RDA *Research Data Policy Framework for Journals and Publishers*[97] recognises by setting out a path that allows progression from suggesting to checking and enforcing an increasing number of policy features.

3. Implementation frameworks: Provisioning guidance

In this section we describe requirements that must be satisfied to achieve FAIR sharing of chemistry data as well as IUPAC FAIR-enabling resources (FERs) that are essential to implementation of FAIR in chemistry. Drawing from a range of existing and emerging standards and practices from IUPAC and the broader community, together these begin to define a framework that can be used to guide the generation of practical policy for reporting FAIR chemistry data. We highlight approaches that are sufficiently concrete to be implemented, initiatives that are in progress, and areas where there is a need to initiate efforts to establish cross-community consensus.

[90] *For example, the U.S. Federal Food, Drug, and Cosmetic Act.* https://www.fda.gov/regulatory-information/laws-enforced-fda/federal-food-drug-and-cosmetic-act-fdc-act (accessed 20230430).

[91] *For example, the E. C. Registration, Evaluation, Authorisation and Restriction of Chemicals Regulation.* https://echa.europa.eu/regulations/reach/legislation (accessed 20230430).

[92] *For example, the 2022 U.S. Office of Science & Technology Public Access Memo.* https://www.osti.gov/2022-OSTP-Public-Access-Memo-Release (accessed 20230430).

[93] *For example, the E.C. Open Research Europe Data Availability Policy.* https://open-research-europe.ec.europa.eu/about/policies#dataavail (accessed 20230430).

[94] *For example, the U.S. National Institutes of Health Data Management and Sharing Policy.* https://sharing.nih.gov/data-management-and-sharing-policy (accessed 20230430).

[95] *For example, the Wellcome Trust Data, software and materials management and sharing policy.* https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy (accessed 20230430).

[96] Parks, N. A.; Fischer, T. G.; Blankenburg, C.; Scalfani, V. F.; McEwen, L. R.; Herres-Pawlis, S.; Neumann, S. The current landscape of author guidelines in chemistry through the lens of research data sharing. *Pure App. Chem.* 2023, Ahead of Print. https://doi.org/10.1515/pac-2022-1001

[97] Hrynaszkiewicz, I.; Simons, N.; Hussain, A.; Grant, R.; Goudie, S. Developing a Research Data Policy Framework for All Journals and Publishers. *Data Sci. J.* 2020, 19(1), 5. https://doi.org/10.5334/dsj-2020-005

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

3.1. RIPE for sharing

The FAIR data principles are framed from the perspective of data reuse. Process engineers, researchers, and an increasing number of automated processes need complete and unambiguous description of research results in convenient forms that are easy to find, retrieve and compile. To get more FAIR data out in consumable forms, we also need to consider the other side of the equation – critical parameters for documenting data during the lifecycle upstream of sharing to ensure that meaning and quality can be assessed and reassessed appropriately. Developments of this kind will greatly enhance collaboration within and between research groups. It is not enough for data objects to be accessible; data need to be reusable scientifically and to enable reuse must be accompanied by contextual information for interpretation. In order to fully enable the Interoperable and Reusable facets of FAIR for machines, we need a framework for the provision of rich metadata within domains bounded by formal research methodology.

Multiple dimensions of functional data need to be considered and articulated collectively among those who are ultimately sharing data to enable networked technical mechanisms downstream to kick in, including: information content (scope of metadata); expression of information (structure of metadata); and where this information appears for technical mechanisms to exploit (exposure of metadata). These may be framed as a series of questions regarding the completeness of data description in scope, form and function, as follows:

- **Reliability**: is all the information required for unambiguous positioning of the data relative to their scientific context present?
- **Interpretability**: are data and metadata expressed in a way that is scientifically interpretable and agnostic to any systems (and/or can be converted with provided information and definitions)?
- **Processability**: are data and metadata in forms that are processable by common protocols, architectures and infrastructure?
- **Exchangeability**: are the metadata necessary for finding, accessing, retrieving and processing exposed to APIs via registries, repositories and other information systems?

The acronym **RIPE** can help to focus on information requirements, resources and practices to facilitate provision of data that are mature for sharing - truly usable and fully AI-ready across a broad range of needs. In distinguishing these aspects, the documentation process can be more readily managed and verified for completeness. A range of existing standards and resources provide definitions that support these requirements. IUPAC is collaborating with the community to instantiate these into reusable digital motifs that can be deployed across different systems and formats to enable the broadest possible use (see Table 1).

Table 1. Standardised definitions for describing chemical data (items in italics are in progress)

RIPE for sharing	Chemical data	Standard definitions (examples)
Reliable *context 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*

3.1.1. Reliable

The reporting principles for thermodynamic data outlined by IUPAC in 2021[98] can readily be applied to chemistry and materials data more generally and these provide an excellent foundation for applying the FAIR data principles to maximise the potential to derive meaning from the data. The 2021 guide emphasises that unambiguous description is essential for reliable reuse of data and provides the inspiration and starting point for the RIPE framework.

To be reliable, published data need to be well-defined and for measurements on chemical systems, this implies inclusion of metadata that fully describe samples, measurements, uncertainty and provenance (see Table 2). Collectively, this description allows the user to ascertain how much the studied system deviates from an idealised one - for example, a pure compound.

Table 2. Critical metadata components needed to describe chemical measurements

Concept	Critical metadata
Samples	Identification of the substance or material, including chemical nature and composition Description of the studied samples, including source, temporal information, and information about purity

[98] Bazyleva, A,; Abildskov, J.; Anderko, A.;, Baudouin, O.; Chernyak, Y.; de Hemptinne, J.; Diky, V.; Dohrn, R.; Elliott, J. R.; Jacquemin, J.; Jaubert, J.; Joback, K. G.; Kattner, U. R.; Kontogeorgis, G. M.; Loria, H.; Mathias, P. M.; O'Connell, J. P.; Schröer, W.; Smith, G. J.; Soto, A.; Wang, S.; Weir, R. D. Good reporting practice for thermophysical and thermochemical property measurements (IUPAC Technical Report). *Pure App. Chem.* 2021, 93(2), 253-272. https://doi.org/10.1515/pac-2020-0403

Concept	Critical metadata
	Identification of the properties and states (phases) to which they are related Specification of the conditions at which the properties were measured Unambiguous statement of units for every property and state variable
Measurements	Experimental methods, computational procedures Reference substances, instruments, calibration procedures and results
Uncertainty	Purity of the samples and reference materials Control of measurement conditions Uncertainties of measuring devices and calibration data Method validation and validation criteria Comparison with previously published values
Provenance	Data origin: experiment, calculation, derivation Process: observation, interpretation (metrological traceability) Documentation: when/who/etc., standard formats, versions, dates, connected parts Use case in which data are generated, as part of the scope and if/how future users consider if the data are fit-for-purpose

Data quality and integrity matter. Data will likely be reused in different contexts than the original measurement or observation and it will not be known what the requirements for accuracy or the tolerance for error might be for any given use[99]. Documenting a full description of the measurement, including reagent purity, conditions, instrumentation, quality controls and other factors potentially contributing to systematic error will provide those reusing data with information to assess the fitness for purpose for their specific use case. Critical evaluation of chemical and physical property data from reported values utilises this information to develop uncertainty budgets[100,101] and assess best measurements that are convenient to reuse. However, the critical

[99] Eisenhart, C. Expression of the uncertainties of final results. *Science.* 1968, 160(3833), 1201-1204. DOI: https://doi.org/10.1126/science.160.3833.1201

[100] Joint Committee for Guides in Metrology. JCGM 200:2012, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM), BIPM, Sèvres, France (2012). https://www.bipm.org/en/publications/guides (accessed 20230524).

[101] Joint Committee for Guides in Metrology. JCGM 100:2008, Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement (GUM), BIPM, Sèvres, France (2008). https://www.bipm.org/en/publications/guides (accessed 20230524).

evaluation process is dependent on full reporting of measurement parameters, which can be found to be lacking for many reported studies. Forthcoming guidance from IUPAC[102] reviews this process.

3.1.2. Interpretable

Reporting data with well articulated metadata provides the opportunity to fully describe the context of the measurement and sample as outlined in Section 3.1.1. There are some fundamental requirements for how data values are expressed that align with scientific conventions and allow data to be unambiguously interpreted. IUPAC and other authoritative bodies define expressions for many different types of physical quantities, including their equations, units and symbols, as described in the Green Book (and the Concise Guide[103]). Systems that collect and process data should adhere to these conventions and digital infrastructure also needs to incorporate them as these definitions are essential for data to be computationally analysed.

As described in the 2012 IUPAC guide for reporting phase equilibria, the rules that define these quantities can provide an “unequivocal accounting basis” for complete and precise description of measurements, including all variables, constraints and dependencies. These expressions of rules can form the basis for metadata reporting standards, as with the IUPAC ThermoML schema which is based on the Gibbs phase rule, a general principle governing thermodynamic systems[104]. Different digital approaches to representing units and quantities may lend themselves to different use cases, for example use of LaTex for presentations to a human audience and more semantic motifs for ontologies or databases[105]. It is also critical to maintain the integrity of units and quantities across formats, workflows and data processing generally. IUPAC is involved in a number of projects to coordinate development of digital representations that encapsulate key quantity definitions and expressions[106,107].

Chemical composition is an object of study in many fields and will be an essential component of most sample descriptions. There are many modes of expression, based on how the physical size and state of the sample is described, which may vary depending on how samples are handled in a given

[102] Shaw, D. G.; Bruno, I.; Chalk, S.; Hefter, G.; Hibbert, D. B.; Hutchinson, R. A.; Magalhães, M. C. F.; Magee, J.; McEwen, L. R.; Rumble, J.; Russell, G. T.; Waghorne, E.; Walczyk, T.; Wallington, T. J. Chemical data evaluation: general considerations and approaches for IUPAC projects and the chemistry community (IUPAC Technical Report). *Pure App. Chem.* 2023, *in press.*

[103] Renner, T. Quantities, units and symbols in physical chemistry; International Union of Pure and Applied Chemistry. Royal Society of Chemistry, 2007, 3rd ed. https://doi.org/10.1039/9781847557889. (Summary sheet of 4th edition at https://publications.iupac.org/ci/2011/3304/July11_green-sup-4p.pdf)

[104] Chirico, R. D.; de Loos, T. W.; Gmehling, J.; Goodwin, A. R. H.; Gupta, S.; Haynes, W. M.; Marsh, K. N.; Rives, V.; Olson, J. D.; Spencer, C.; Brennecke, J. F.; Trusler, J. P. M. Guidelines for reporting of phase equilibrium measurements (IUPAC Recommendations 2012). *Pure App. Chem.* 2012, 84(8) 1785-1813. https://doi.org/10.1351/PAC-REC-11-05-02

[105] CODATA. *Units of measure representations inventory.* https://codata.org/wp-content/uploads/2022/12/DRUM_Units_Inventory_120522.pdf (accessed 20230524).

[106] CODATA Task Group. *Digital Representation of Units of Measurement (DRUM).* https://codata.org/initiatives/task-groups/drum (accessed 20230501).

[107] IUPAC Project. *Preparation of the 5th Edition of the IUPAC Green Book.* https://iupac.org/project/2019-001-2-100 (accessed 20230501).

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

field. Which quantity (e.g. mass, volume, number or amount of chemical entities) is used as the basis reflects the macro or molecular level focus of the original context and can impact which quantities can be more readily reused for different purposes, such as comparing field measurements or theoretical predictions. Thus, re-users may frequently need to convert between different expressions of composition when compiling heterogeneous data, and the level of accuracy at which this can be done will depend on unambiguous expression of units and other auxiliary information for condition dependent quantities such as volume[108,109]. Complete documentation and full understanding of the many nuances of units and quantity expressions is a high bar for busy researchers and others curating data. Some harmony around digital expression of composition will be critical for the ability to use data more broadly for interdisciplinary use cases.

3.1.3. Processable

For a data object to be processable it needs to be in a format that is systematically structured so that a machine can parse and operate on it. This requires not just a formalised syntax but also formalised terminologies that clearly describe the parameters and scope of specific data items. We also need data models and schema that can capture relationships between core components of the data object and descriptions of related objects. In chemistry, core components include specification of the specific sample studied, representations of the chemical entities known to be in the sample, and expressions of the values arising from a measurement or analysis as described in the previous sections.

3.1.3.1. Digital representation of chemical entities

Fundamental to the processability equation for chemical data is digital representation of the chemical entities to which a given data set relates. The availability and limitations of formal and community conventions for representing a chemical structure have been covered in Section 2.1.1. Here we note that in order to maximise the processability potential of a data object, it should ideally be accompanied by a representation that delineates a chemical structure as precisely as possible as well as standard identifiers that will enable discovery and provide connections across resources. Critical to advancing FAIR will be community consensus on the strengths, weaknesses and limitations of the variety of representation formats and identifiers currently available relative to the breadth of needs for characterising samples and materials across the disciplines.

[108] Cvita˘s. T. Quantities Describing Compositions of Mixtures. *Metrologia*. 1996, 33, 35-39. https://doi.org/10.1088/0026-1394/33/1/5

[109] Horvath, A. L.; Getzen, F. W.; Eds. Carbon Dioxide in Water and Aqueous Electrolyte Solutions. Solubility Data Series, International Union of Pure and Applied Chemistry, Vol. 62, pp vi-xv; Oxford University Press, 1996. https://iupac.github.io/SolubilityDataSeries/volumes/SDS-62.pdf (accessed 20230501).

3.1.3.2. Systematic representation formats for chemical data

Some standard representation formats exist for representing some chemical measurement data outputs, including CIF (chemical crystallography)[110], ThermoML (thermodynamic data)[111], JCAMP-DX (various spectra types)[112] and mzML (mass spectra)[113], among others. Domain ontologies and terminologies such as the IUPAC Gold Book[114] of formally defined chemical concepts can be used to underpin standard formats and enable semantic representation of data items. CIF[115] and ThermoML[116,117] are both accompanied by dictionaries that define data items relevant to their fields in the context of the format. Projects are also underway to define key metadata elements for several chemical properties that can be incorporated into native formats and data models, including isotopic abundances, atomic weights, solubility and mixture composition[118,119, 120].

3.1.3.3. Semantic terminologies

Established terminologies can be manifest into frameworks that are broadly used across domains - for example, a project to convert CIF data dictionaries into the Elementary Multiperspective Material Ontology (EMMO) framework[121] and reference to the IUPAC Gold Book from various ontologies for chemical methods (CHMO), reactions (RXNO), chemical entities (ChEBI, ChemOnt), and various others that have been developed for the life sciences[122]. These ontologies have the potential to further enrich data models with representations of the multi-faceted relationships inherent in chemical systems under study. Granularity of classification will be important for the disambiguation of various concepts in different fields. Ongoing work in IUPAC and the broader

[110] IUCr. *Crystallographic Information Framework*. https://www.iucr.org/resources/cif (accessed 20230501).

[111] IUPAC. *ThermoML – An XML-based IUPAC Standard for Thermodynamic Property Data.* https://iupac.org/what-we-do/digital-standards/thermoml (accessed 20230501).

[112] IUPAC. *JCAMP-DX - Joint Committee on Atomic and Molecular Physical Data Data Exchange format.* https://iupac.org/what-we-do/digital-standards/jcamp-dx (accessed 20230501).

[113] Martens L., Chambers M., Sturm M., Kessner D., Levander F., Shofstahl J., Tang W.H., Römpp A., Neumann S., Pizarro A.D., Montecchi-Palazzi L., Tasman N., Coleman M., Reisinger F., Souda P., Hermjakob H., Binz P.A., Deutsch E.W..mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics.* 2011 Jan;10(1):R110.000133. https://github.com/HUPO-PSI/mzML (accessed 20230501).

[114] IUPAC Gold Book. *Compendium of Chemical Terminology.* https://goldbook.iupac.org (accessed 20230524).

[115] IUCr. *CIF dictionaries.* https://www.iucr.org/resources/cif/dictionaries (accessed 20230501).

[116] Frenkel, M.; Chiroco, R.; Diky, V.; Dong, Q.; Marsh, K.; Dymond, J.; Wakeham, W.; Stein, S.; Königsberger, E.; Goodwin, A. XML-based IUPAC standard for experimental, predicted, and critically evaluated thermodynamic property data storage and capture (ThermoML) (IUPAC Recommendations 2006). *Pure App. Chem.* 2006, 78(3), 541-612. https://doi.org/10.1351/pac200678030541

[117] Frenkel, M.; Chirico, R.; Diky, V.; Brown, P.; Dymond, J.; Goldberg, R.; Goodwin, A.; Heerklotz, H.; Königsberger, E.; Ladbury, J.; Marsh, K.; Remeta, D.; Stein, S.; Wakeham, W.; Williams, P. Extension of ThermoML: The IUPAC standard for thermodynamic data communications (IUPAC Recommendations 2011). *Pure App. Chem.* 2011, 83(10), 1937-1969. https://doi.org/10.1351/PAC-REC-11-05-01

[118] IUPAC Project. *Machine-Accessible Periodic Table.* https://iupac.org/project/2019-020-2-024 (accessed 20230501).

[119] IUPAC Project. *Development of a metadata schema for critically evaluated solubility measurement data.* https://iupac.org/project/2020-018-1-024 (accessed 20230501).

[120] IUPAC *Project*. *InChI extension for mixture composition*. https://iupac.org/project/2015-025-4-800 (accessed 20230501)

[121] EMMO. *CIF ontology.* https://github.com/emmo-repo/CIF-ontology (accessed 20230501).

[122] NFDI4Chem. *Terminology Service.* https://terminology.nfdi4chem.de/ts (accessed 20230501).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

community is focusing on bridging concepts across sub-disciplines as well as formulating high-level concepts in chemistry and across neighbouring disciplines[123].

3.1.3.4. Data models

The IUPAC FAIRSpec project is addressing the broader challenge of linking data objects associated with measurement outputs, molecular representations and sample descriptions[124]. Experimental results will involve collections of data sets that variously map on to different chemical entities occurring in different samples. FAIRSpec aims to define a data model that can capture the objects involved and the relationships between them, based on the association of characterization data assessed through various spectroscopic techniques to features of chemical structures. The documentation strategy under development in FAIRSpec exposes key metadata that will allow a reuser to locate relevant datasets and quickly ascertain whether additional scrutiny of the data collection is warranted. This concept, of publishing a digital finding aid, will offer a mechanism for working around the limitations of proprietary formats, whilst still enabling a degree of processability[125].

3.1.4. Exchangeable

Exposure of domain metadata through APIs and widely adopted metadata registries such as DataCite can facilitate discovery and exchange of data between systems. Collectively these mechanisms function as an index of networked data resources, enabling search and retrieval across many disparate datasets. For chemistry data to be exchangeable across domains, key data items represented by the formats and terminologies needed to make data processable must be fed into these frameworks. Particularly important are indications of relevant chemical entities, as many data resources are organised around chemical substances.

As a canonical structure based identifier, InChI supports a key function in this environment – to the extent that a chemical substance can be represented, any associated datasets where InChIs have been generated can be identified. This provides a point of connection that supports findability, accessibility, integration and validation. Many chemistry data resources currently utilise connections based on InChI to federate across sources (e.g., UniChem[126]; VAMDC[127]) or aggregate data from a

[123] IUPAC. *Joint Subcommittee on the IUPAC Gold Book.* https://iupac.org/body/039 (accessed 20230524).

[124] IUPAC project. *Development of a Standard for FAIR Data Management of Spectroscopic Data.* https://iupac.org/project/2019-031-1-024 (accessed 20230524).

[125] Hanson, R. M.; Jeannerat, D.; Archibald, M,; Bruno, I. J.; Chalk, S. J.; Davies, A. N.; Lancashire, R. J.; Lang, J.; Rzepa, H. S. 'IUPAC Specification for the FAIR Management of Spectroscopic Data in Chemistry (IUPAC FAIRSpec) – Guiding Principles'. *Pure App. Chem.* 2022, 94(6), 623-636. https://doi.org/10.1515/pac-2021-2009.

[126] https://chembl.gitbook.io/unichem/api (accessed 20230505).

[127] Zwölf, C. M.; Moreau, N. Assessment of the FAIRness of the Virtual Atomic and Molecular Data Centre following the Research Data Alliance evaluation framework. *Eur. Phys. J. D.* 2023, 77, 70. https://doi.org/10.1140/epjd/s10053-023-00649-x

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

range of sources (e.g., PubChem[128]; ChemSpider[129]). Other linear notations such as SMILES can also support automated queries, although different conventions for these notations may yield inconsistent retrieval. IUPAC is currently developing open formalised reference documentation for SMILES that articulates standard interpretation to facilitate interoperability of reading the SMILES language[130].

A key challenge for metadata exchange is navigating variation between data systems around the degree of specificity of structure characteristics, the degree of knowability about the associated molecular structure from the data to hand, and the data model used for representation. As different resources provide different information about different compounds, an equivalent of a telephone exchange is needed that can connect these together in a generic way. To enable this at a level more specific than general frameworks but sufficiently broad from a chemical perspective, WorldFAIR D3.3 is developing a protocol for general API specification that can be adopted by resources wishing to enable exchange of chemistry data based on chemical structure.

3.2. IUPAC standards as FAIR Enabling Resources

IUPAC formulates standard definitions and notations across the sub-disciplines of chemistry around three major pillars - i) chemical representation, ii) chemical and physical properties, and iii) chemical concepts[131]. These are made available to the community to consult in the context of their own use cases and business scenarios through a variety of formal publications and other outputs. Many of these resources lend themselves to providing useful reference material for describing and managing chemical data, as discussed in previous sections. However, very few are presently available in fully digitalised forms that could be readily incorporated into workflows and tools that support FAIR data.

The WorldFAIR project is reviewing the efficacy of the FAIR Implementation Profile (FIP) methodology and tools to assess how the FAIR attributes are enabled in various research data resources[132]. Central to this approach is the utilisation of FAIR Enabling Resources (FERs), technical

[128] PubChem. *Programmatic Access documentation.* https://pubchem.ncbi.nlm.nih.gov/docs/programmatic-access (accessed 20230505)

[129] *Royal Society of Chemistry APIs.* https://developer.rsc.org (accessed 20230505).

[130] IUPAC project. *IUPAC SMILES+ Specification.* https://iupac.org/project/2019-002-2-024 (accessed 20230524).

[131] *International Union of Pure and Applied Chemistry.* https://iupac.org/what-we-do (accessed 20230524).

[132] Schultes, E.; Magagna, B.; Hettne, K. M.; Pergl, R.; Suchánek, M.; Kuhn, T. Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Grossmann, G., Ram, S. (eds). Advances in Conceptual Modeling. 2020. In, *Lecture Notes in Computer Science*, vol. 12584. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_13

motifs of various types that provide mechanisms for implementing aspects of FAIR[133,134]. The FIP method defines a typology of FERs mapped to the FAIR attributes that provides also a useful framework for assessing the potential for existing chemistry standards to enable FAIR. Working from the premise that standards from IUPAC and other authoritative bodies provide the scientific backbone for accurate chemistry data exchange, we can consider two questions relative to the FIP analysis: 1) how do these standards fit into the FER typology to enable FAIR re-use of chemistry data?; and 2) are these standards themselves available in forms that are FAIR and accessible for programmatic re-use?

The WP03 case study has thus far generated two FIPs exploring these two facets, including a general assessment of IUPAC digital standards and their respective functions as FERs, and a more specific assessment of the IUPAC Gold Book Compendium of Chemical Terminology as a FAIR vocabulary. The goal is to provide some indication of *FAIR application* of IUPAC standards in supporting chemistry data exchange and the *FAIR status* of IUPAC standards for those who need to use them. Further details on the FIP and FER analysis to date are summarised in Appendix 5.4.

Generally, IUPAC standards have great potential to support Interoperability. IUPAC defines key concepts and rulesets and provides some reference implementations that can be used directly as FERs. Standard parameters for chemical representation and various chemical properties can be incorporated into file formats, data structures and semantic models. These parameters can also be used in criteria to validate implementations for Re-usability. Chemical identity and property types also play a high level role in how chemistry data are organised and indexed in different data systems, and key notations can also be used in metadata schema to support Findability and in API specifications to support Accessibility. Further consideration on use of these standards in enabling FAIR is addressed earlier (Section 3.1) in this report.

These FIP assessments of IUPAC standards have already been reviewed by other WorldFAIR case studies considering how to work with chemistry related data, including Nanomaterials (WP04)[135]; Geochemistry (WP05)[136]; and Oceanography (WP11)[137]. Additional FIPs assessing other IUPAC

[133] *FIP Wizard 3.0 User Guide.* https://osf.io/5ygzx (accessed 20230429).

[134] Jacobsen, A.; de Miranda Azevedo, R.; Juty, N.; Batista, D.; Coles, S.; Cornet, R.; Courtot, M.; Crosas, M.; Dumontier, M.; Evelo, C. T.; Goble, C.; Guizzardi, G.; Hansen, K. K.; Hasnain, A.; Hettne, K.; Heringa, J.; Hooft, R. W. W.; Imming, M.; Jeffery, K. G.; Kaliyaperumal, F.; Kersloot, M. G.; Kirkpatrick, C. R.; Kuhn, T.; Labastida, I.; Magagna, B.; McQuilton, P. Meyers, N.; Montesanti, A.; van Reisen, M.; Rocca-Serra, P.; Pergl, R.; Sansone, S.; da Silva Santos, L. O. B.; Schneider, J.; Strawn, G.; Thompson, M.; Waagmeester, A.; Weigel, T.; Wilkinson, M. D.; Willighagen, E. L.; Wittenburg, P.; Roos, M.; Mons, B.; Schultes, E. FAIR Principles: Interpretations and Implementation Considerations. *Data Intell.* 2020, 2(1-2), 10–29. https://doi.org/10.1162/dint_r_00024

[135] McEwen, L; Exner, T. Example FIPs – WorldFAIR project: Chemistry & Nanomaterials. PARC FAIR Data and Tools webinar series, 2023-04-20. https://nanocommons.github.io/user-handbook/training-courses/PARC-training (accessed 20230428).

[136] Prent, A. WorldFAIR Project (D5.1) Formalisation of OneGeochemistry (1.0). 2022. Zenodo. https://doi.org/10.5281/zenodo.7380947

[137] Buttegieg, P. L. WorldFAIR Project (D11.1) An assessment of the Ocean Data priority areas for development and implementation roadmap (Version 1). 2023. Zenodo. https://doi.org/10.5281/zenodo.7682399

standards will be developed later in the project and will feed into a roadmap for IUPAC to target areas where further standards development is needed for chemistry data and outline steps to improve FAIR access to IUPAC standards. A further opportunity would be to engage with other organisations responsible for standards relevant for chemistry data to profile these as additional FERs - for example, the CIF format stewarded by IUCr.

3.3. Aligning across interoperability frameworks

Also under review in the WorldFAIR project is an effort to establish a cross-disciplinary interoperability framework (CDIF). Some initial scoping work has been done to identify existing broadly used technologies that resources in different domains can incorporate and thereby to establish some common areas of technical alignment for exchanging FAIR data[138]. Many technologies exist for enabling Findability and Accessibility, as common functions for users; how IUPAC can enable these in its standards as a means to facilitate their broader use will be considered in further assessments.

Interoperability across domains will necessitate further collective engagement to identify the common concepts needed for integration of heterogeneous data. As discussed above in Section 3.1.1, well-described chemistry data is essential for assessing quality in various fields. A broadly common need is for digital motifs that describe a multitude of facets pertaining to observations and measurements, including representation of quantities, conditions, methods and other constraints and dependencies defined in metrology. A number of semantic models for measurement description have recently emerged[139,140] and considering these along with various existing formats for specific measurement types relevant for chemistry, such as ThermoML, provide opportunities for further exploration.

Broadly applicable approaches to describing samples is another common need. Critical in these models will be the relationship between the various chemical species involved, along with other characteristics of the material and physical state, conditions, provence and other facets. Articulating technical motifs for these requirements will enable modelling of chemical data around many cross-disciplinary challenges - for example, identifying high-risk, persistent chemical pollutants in different environments. IUPAC would also be able to utilise such models for samples and measurement to report critically evaluated chemical and physical properties for broader access and reuse.

[138] Gregory, A.; Hodson, S. Cross-Domain Interoperability Framework (CDIF) Working Documents. 2023. Zenodo. https://doi.org/10.5281/zenodo.7652742

[139] RDA. I-ADOPT Framework. https://www.rd-alliance.org/group/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg/wiki/i-adopt (accessed 20230524).

[140] Open Geospatial Consortium. Observations and Measurements ontology. https://www.ogc.org/standard/om (accessed 20230524).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

3.4. Ecosystems of implementation

Implementing FAIR chemical data reporting practices depends on critical activities throughout the full lifecycle of data. A point where FAIR requirements are commonly considered is, at the time of publication, when data are made available to the broader community, but the usability of these data will be impacted by factors both upstream and downstream. Whilst FAIR data sharing is a useful target for funder and journal policies, realisation of FAIR needs to be approached from the very beginning of the research process to reduce the number of ambiguities in measurements and documentation and ensure that results will be RIPE for reuse across domains[141].

Data Management Plans are required by many funders but do not provide researchers with direction on the practical frameworks and infrastructure that ensure effective data management. Organisations and professionals involved in building and managing information systems and services need to consider what is required for data to be 'born FAIR' at the point they are generated and how FAIR attributes can be added and enhanced from conception through analysis and reuse, throughout the research data lifecycle[142]. Similar considerations have been identified by those working on FAIR initiatives in industry where the need for prospective FAIRification combined with quality assessment has been identified as essential for maximising data value in biopharmaceutical research and development (R&D)[143].

Vital to facilitating FAIR management of chemical data over time will be the deployment and adoption of information systems both upstream and downstream that incorporate digital standards and manifest the principles of RIPE and FAIR. The panoply of Electronic Lab Notebooks (ELNs), Digital Research Notebooks and Laboratory Information Management Systems (LIMS) available to researchers need to provide output that are RIPE for reuse and interoperable with other systems[144,145,146]. Concerted efforts to address barriers are being made through regional initiatives:

[141] Bazyleva, A,; Abildskov, J.; Anderko, A.;, Baudouin, O.; Chernyak, Y.; de Hemptinne, J.; Diky, V.; Dohrn, R.; Elliott, J. R.; Jacquemin, J.; Jaubert, J.; Joback, K. G.; Kattner, U. R.; Kontogeorgis, G. M.; Loria, H.; Mathias, P. M.; O'Connell, J. P.; Schröer, W.; Smith, G. J.; Soto, A.; Wang, S.; Weir, R. D. Good reporting practice for thermophysical and thermochemical property measurements (IUPAC Technical Report). *Pure App. Chem.* 2021, 93(2), 253-272. https://doi.org/10.1515/pac-2020-0403

[142] NIST. *Research Data Framework (RDaF).* https://www.nist.gov/programs-projects/research-data-framework-rdaf (accessed 20230524).

[143] Harrow, I.; Balakrishnan, R.; Küçük McGinty, H.; Plasterer, T.; Romacker, M. 'Maximizing Data Value for Biopharma through FAIR and Quality Implementation: FAIR plus Q'. *Drug Disc. Today.* 2022, 27(5), 1441–47. https://doi.org/10.1016/j.drudis.2022.01.006

[144] Kanza, S.; Willoughby, C.; Gibbins, N.; Whitby, R.; Frey, J. G.; Erjavec, J.; Zupančič, K.; Hren, M; Kovač, K. Electronic lab notebooks: can they replace paper? *J. Cheminform.* 2017, 9(1). https://doi.org/10.1186/s13321-017-0221-3

[145] Harvard Longwood Medical Area Research Data Management Working Group (2021) "Electronic Lab Notebook Comparison Matrix". https://doi.org/10.5281/ZENODO.4723752

[146] PSDI. *Case Study 6: Process Recording and Digital Research Notebooks.* https://www.psdi.ac.uk/case-study-6 - Process Recording and Digital Research Notebooks (accessed 20230505).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

NFDI4Chem is looking to encourage adoption of the Chemotion ELN across institutions in Germany; a PSDI pathfinder project[147] is exploring process recording approaches that can work across a range of solutions.

Success will be further maximised if the software tools that mediate generation of data from instruments, analyse and refine data, and perform subsequent calculations also adopt standardised data formats. Critically, outputs need to capture all the data and metadata needed to enable discovery, reuse and interoperability. This extends to commonly used open source and proprietary toolkits used to process, analyse and inter-convert chemistry data and structure representations.

Whilst lab-based, real-time FAIR data management approaches are a critical precursor to FAIR implementation, the role of those involved in facilitating publication of results is not diminished. This traditionally includes those involved in the publication of articles - Publishers, Editors, Authors, and Reviewers (PEARs). All of these groups need to engage in the specification and operation of policies and processes that ensure data underpinning a paper are appropriately described, validated and reviewed.

Another key enabler of FAIR data reporting are trusted data resources including repositories and databases, in particular those that specialise in a domain. For example, resources focused on a specific analytical method (CSD - crystallography[148]; Massbank - mass spec[149,150,151]; nmrXiv - spectra[152]) or specific properties (ChEMBL - bioactivities[153]; CompTox - toxicology[154]). These play a critical role in bridging communities by providing workflows that channel researcher outputs through to publication outlets, enriching data objects so they are broadly discoverable and reusable across domains. Further, they can facilitate links to other research outputs, and ensure preservation and availability of data to future generations of researchers in ways that can satisfy evolving scientific and technological needs.

Technical capabilities are only one consideration. Willingness and ability to support such systems and adapt workflows can impact effective adoption. There is an ongoing need to educate current researchers in the expectations they should have for how their data are managed and reported. Librarians, data stewards and information professionals can provide guidance and services to help researchers achieve the best for their data within the limitations of the current ecosystem. Finally,

[147] PSDI. *WP4 – First round Pathfinders.* https://www.psdi.ac.uk/current-work/wp4-pathfinders (accessed 20230505).
[148] CCDC. *The Cambridge Structural Database (CSD).* https://www.ccdc.cam.ac.uk/solutions/software/csd (accessed 20230505).
[149] *MassBank Europe.* https://massbank.eu/MassBank/About (accessed 20230505).
[150] *MassBank Japan.* http://www.massbank.jp/About (accessed 20230505).
[151] *MassBank of North America.* https://mona.fiehnlab.ucdavis.edu (accessed 20230505).
[152] *nmrXiv homepage.* https://nmrxiv.org (accessed 20230505).
[153] EMBL-EBI. *ChEMBL Database.* https://www.ebi.ac.uk/chembl (accessed 20230505).
[154] EPA. *CompTox Chemicals Dashboard.* https://comptox.epa.gov/dashboard (accessed 20230505).

through these interactions, we need to encourage champions to catalyse the development of further digital standards, tools and infrastructure needed to fully implement FAIR in the chemistry ecosystem.

Effectively communicating scholarship and sharing quality data broadly across disciplines and communities of practice is a collective effort. There is a need for actors across the research ecosystem to rise to the promise of FAIR and to work to optimise practices, processes, and systems to enable FAIR. We emphasise 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. Resources developed by these stakeholders further enable the research community, and represent a powerful mechanism for coordinating broad engagement with FAIR data sharing. Table 3 summarises a number of suggestions arising from community workshops to support active implementation of FAIR data practices[155,156,157,158,159,160].

[155] RDA-IUPAC-EPA workshop. *Prioritising Digital Data Challenges in Chemistry: Road-mapping Technical Opportunities and Business Cases.* Rally, 2016. https://iupac.org/event/prioritizing-digital-data-challenges-in-chemistry (accessed 20230524).

[156] IUPAC-CODATA workshop. *Supporting FAIR Exchange of Chemical Data through Standards Development.* Amsterdam, 2018. https://iupac.org/event/supporting-fair-exchange-chemical-data-standards-development (accessed 20230524).

[157] McEwen, L.; Martinsen, D.; Lancashire, R.; Lampen, P.; Davies, A. Are your spectroscopic data FAIR?. *Spectroscopy Europe*. 1 August 2018. https://doi.org/10.1255/sew.2018.a2

[158] NSF-IUPAC-CINF workshop. *FAIRPublishing Guidelines for Spectral Data and Chemical Structures.* Orlando, 2019. https://iupac.org/event/fair-publishing-guidelines-for-spectral-data-and-chemical-structures (accessed 20230524).

[159] Scalfani, V.; McEwen, L. *NSF FAIR Chemical Data Publishing Guidelines Workshop on Chemical Structures and Spectra: Major Outcomes and Outlooks for the Chemistry Community.* 2019. https://doi.org/10.7298/fs2d-hx95

[160] WorldFAIR-IUPAC-CINF workshop. *Advancing FAIR Chemistry: Developing New Services for Sharing Chemical Data.* Indianapolis, 2023. https://iupac.org/event/advancing-fair-chemistry-developing-new-services-for-sharing-chemical-data (accessed 20230524).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Table 3. Examples of FAIR-enabling functions provided by different services and roles

	PIDs & Registered Metadata	**Domain Repositories**	**Open Standard Formats**	**Data are 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	✶save data files in standard formats ✶prepare ReadMe

4. Guidance summary

Well-defined chemical data are broadly applicable across a number of disciplines and use cases. To facilitate the reporting of chemical data that are FAIR and align with existing scientific and domain standards, chemical data need to be RIPE (Reliable, Interpretable, Processable, Exchangeable) for sharing. RIPE may be defined by the following criteria, which were discussed in detail in Section 3.1:

1) All critical representative information is at hand and **reliable**;
2) Information is **interpretable** via scientifically robust standard conventions;
3) Data and information objects are **processable** in a lossless manner and aligned with community standards;
4) Metadata are appropriately exposed and **exchangeable** for discovery and retrieval by programmatic means.

To meet these criteria, stakeholders involved in scientific communication must incorporate established digital chemistry notations into their tools and workflows. A number of examples, listed in Table 4, can already be used to support functions that enable FAIR, including file formats, ontologies, and formal terminologies to use in metadata. Further examples and needs for community effort to provide missing enablers were discussed throughout the report.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Table 4. Examples of standard notations in chemistry and related fields that enable FAIR

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

Of particular importance for reusing chemical data is the association of values from the measurement output with characterisation of the sample and the chemical entities involved, as represented in Figure 6. These associations are critical to capture as part of data reporting as multiple data objects will likely be involved. Metadata must describe components in sufficient detail to enable validation of these relationships as well as the components themselves.

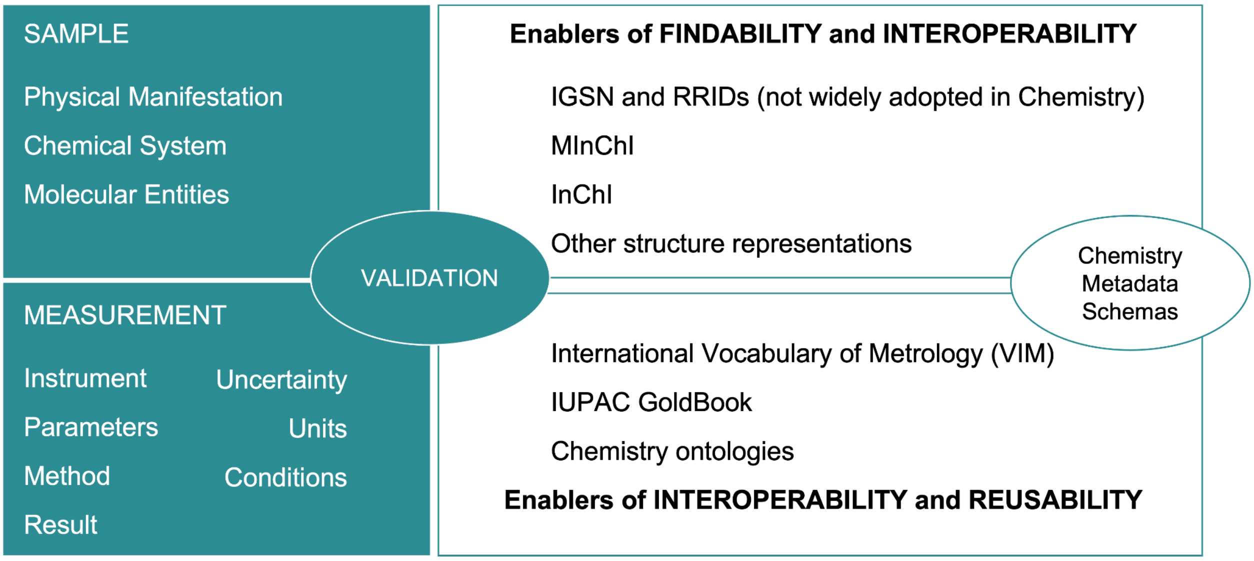

SAMPLE
Physical Manifestation
Chemical System
Molecular Entities
MEASUREMENT
Instrument
Uncertainty
Parameters
Units
Method
Conditions
Result
VALIDATION
Enablers of FINDABILITY and INTEROPERABILITY
IGSN and RRIDs (not widely adopted in Chemistry)
MInChI
InChI
Other structure representations
Chemistry Metadata Schemas
International Vocabulary of Metrology (VIM)
IUPAC GoldBook
Chemistry ontologies
Enablers of INTEROPERABILITY and REUSABILITY

Figure 6. Schematic that positions attributes of samples and measurements and their relationships in the context of FAIR

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Putting data in trusted chemistry-enabled repositories in standard formats that can be validated is an aspirational target for data sharing in chemistry. There are opportunities to apply existing digital standards more broadly in current and emerging workflows for data sharing that will immediately help to improve the accessibility of data and advance our understanding of FAIR. Table 5 provides a brief summary of the present availability of chemistry motifs that can support key functionalities and improve the RIPEness of chemical data.

Table 5. Present availability of standard chemistry motifs to support RIPE and FAIR

Metadata to include with DOI registration to facilitate discovery		
Available now	Domain motifs including IUPAC Gold Book terms and InChI for chemical identity can be incorporated into metadata now.	**Exchangeable**
In progress	FAIRSpec guidance for registered metadata being developed.	
File formats and repositories for various chemistry data types		
Available now	Standard formats for some data types available (see Tables 1 & 4). Some domain repositories support these formats (see discussion in Sections 3.1 and 3.4).	**Processable**
In progress	Additional formats and metadata specifications under development in IUPAC. Additional community domain repositories under development.	
Unit representation for different use cases		
Available now	IUPAC standard definitions for quantities available via the IUPAC Gold Book and IUPAC Green Book Various digital motifs available lend themselves to different scenarios.	**Interpretable**
In progress	Representation of additional types of units under review in the DRUM Task Group.	
In progress	Active projects addressing how to maintain integrity of units across formats, workflows and processes.	**Processable**

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Use of chemical terminology in descriptive metadata and ontologies		
Available now	Referrable terms in any metadata scenario (via Gold Book).	**Interpretable**
In progress	Definition of high level concepts in chemistry.	
Available now	Referrable terms in ontologies.	**Processable**
In progress	Classification of chemical concepts used in different sub-disciplines of chemistry; *would also be interesting to track-back linkages from other fields.*	

As with the research community at large, the chemistry domain is still very much progressing towards widely enabling FAIR in practice. Whilst some communities of practice such as crystallography have mature workflows and digital components in place to support specific data types[161], practices around data sharing generally are nascent and there are still many gaps in digital infrastructure and standard notation. Priorities for action by communities involved in chemistry data and information infrastructure include:

1. Take advantage of what is already available

- Incorporation of existing digital chemistry motifs into current and emerging tools and workflows.

2. Develop guidance and services that facilitate best practices

- Adoption of standard formats
- Validation of data in standard formats
- Deposition of data into domain-enabled repositories.

3. Collaborate to establish cross-community consensus

- Clarification and communication of best practice
- Criteria for assessing whether standards are appropriately adopted in accordance with guidelines
- Reasonable adaptations in lieu of absent standards
- Approaches to innovation for standards development.

Our landscape review indicates that there are activities and initiatives that provide a foundation for addressing these priorities but that more work is needed to accommodate the digital needs of

[161] Bruno, I., Gražulis, S., Helliwell, J. R., Kabekkodu, S. N., McMahon, B., & Westbrook, J. (2017). Crystallography and Databases. *Data Sci. J.* 16(0), 38.DOI: https://doi.org/10.5334/dsj-2017-038

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

chemistry more broadly and to satisfy requirements from other disciplines. We have identified implementation frameworks that we believe can provide direction to those developing the standards, tools and services needed to advance these priorities as well as foster alignment and consensus across communities engaged in establishing best practices and policies for the reporting of FAIR chemical data.

To enact these frameworks, we need to establish axes of engagement that align with existing community practices of consensus building. IUPAC contributes through the provision of objective scientific expertise and consensus based communication standards in chemistry. New methods of dissemination and collaborative development are needed to facilitate consistent and comprehensive communication of our current understanding and to advance more complex needs in the digital context. Ultimately, the success of this endeavour will depend on effective curation - fostering roles and expertise to ensure a fully integrated cycle of data collection, reporting and re-use.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

5. Appendices

5.1. WorldFAIR Chemistry webinar series

Summary of the Webinar Series: What is a Chemical? [162]

September 2022-February 2023. Virtual
Organisers: Leah McEwen, Ian Bruno, Stuart Chalk and Fatima Mustafa

1. Description

Chemical substances touch on all areas of laboratory science and chemistry underlies many critical worldwide issues, including climate, health, food availability and sustainable development. Increased reporting of machine-readable chemical data will support active research in chemistry and related sciences worldwide, and will be essential to the development of the interdisciplinary science critical to address the UN Sustainable Development Goals and UNESCO's priorities around Open Science. IUPAC is the world authority on chemical nomenclature, terminology, and standardised methods of measurement, and is engaging in a concerted effort through collaboration with the broader chemistry and data science communities to translate a range of assets and activities into the digital domain. IUPAC is leading the chemistry WorldFAIR project-WP03, an EU funded program, aiming to align standards development and implementation with the FAIR data principles. This will facilitate development of guidelines, tools and validation services that support scientists to share and store chemical data in a FAIR manner and support the ability to compile and interpret data across scientific disciplines. In this webinar series, we aimed to:

1. Understand the chemical substance notations used within multiple disciplines *(geochemistry, nanochemistry, atmospheric chemistry, environmental chemistry, oceanography, crystallography, etc.)*.

2. Explore the data resources that are used in these applied areas *(life sciences and pharmaceuticals, agriculture and crop protection, dyes and pigments, and machine learning)*, and understand the current ways of communication and accessing data by other groups.

3. Investigate various digital and machine-readable depictions or notations of chemical substances, reactions and datasets (*InChI, HELM, SMILES, Graphical Representation, Systematic Representation, Media Types, ChEBI, Ontologies4Chem, and notations of Mixtures-Molecules and Complex Substance Schema)*.

[162] https://doi.org/10.5281/zenodo.7903683 (accessed 20230505)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Overall, the webinar series highlighted the current status of working with chemical notations, development of digital tools to transform chemical notations into digital entities and ways to implement the FAIR data principles across the chemical enterprise.

2. Overview of Webinars

1. "What is a chemical? Handling Chemical Data Across Disciplines" [163]

   In this webinar, seven experts from diverse backgrounds were invited: Emma Schymanski (Environmental Sciences), Iseult Lynch (Nanomaterials), Lesley Wyborn (Geochemistry), Kerstin Lehnert (Astromaterials), Marie-Lise Dubernet (Astrophysics), Ken Kroenlein (Material Science), and Dylan Walsh (Polymers). The panellists discussed the status of working with machine-readable formats of chemicals within their field and indicated challenges and future needs. The conversation revealed areas of emphasis that users and developers of machine-readable formats should work to advance and to which IUPAC can potentially contribute.

   - **Representation and finding chemical data** (If we can't represent materials, how are we going to be able to find them?) **:** Identifiers for small molecule chemicals such as InChI and InChIKey are widely used. Similiarialy for polymers, BigSmiles is in use; however stochasticity is a challenge to translate into machine-readable format. Polymer properties are based on assemblies, and assembling and processing needs to be known as they affect properties. Moreover, a notation for nanoparticles that builds on InChI is still in development. Linking as-synthesised material to various transformed forms (during storage, upon dispersion, in the environment, body etc.) to correlate actual form with effect is needed. This could help geochemists in identifying isotope ratios / crystal phases or domains etc.
   - **Real sample complexity** (Can we have chemical identifiers to reflect sample complexity?) **:** There are a variety of different types of chemical samples (cosmochemical, ceramic, alloys, nanomaterials, mixtures, etc.) with multiple factors (type, age, source, storage, uniformity, history, etc.) that play a role in sample representation.

2. “What is a chemical? Applying Chemical Data to Industrial Challenges” [164]

   In this webinar, six experts from multiple applied chemical areas were invited: Lutz Weber (AI & ML), Teodoro Laino (AI & ML), Yannick Djoumbou Feunang (Agir-Chem), Nelson Vinueza Benitez (Dyes and Pigments), Nick Lynch (Pharmaceuticals and life Science), Gunther Schadow (Pharmaceutical-Regulatory). The panellists discussed the challenges and forthcoming regulation requirements for using chemical data in machine-readable formats and provided many recommendations:

[163] https://doi.org/10.5281/zenodo.7259101 (accessed 20230505)

[164] https://doi.org/10.5281/zenodo.7259727 (accessed 20230505)

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- Need an easy to use more expressive molecular representation, across all nominal identifiers, extending InChI for bigger structures, and making molecules open for everyone!
- Explore and adopt technologies to aid chemists as they generate data.
- Define rules for structures, organising and sharing physical property data with structures.

3. "What is a chemical? User' Perspectives on Digital Machine Readable Depictions" [165]

   In this webinar, experts in five various digital machine-readable depictions were hosted: Greg Landrum (InChI), Dana Vanderwall (HELM), Jonathan Goodman (Graphical Representation), Michelle Rogers (Systematic Nomenclature), and Vincent Scalfani (SMILES). The panellists identified some gaps in chemical representation/identifiers such as:

   - Graphical representation is not standardised for small molecules including organometallics, and isomers.
   - The need to define/determine uniqueness, capturing process of macromolecules including glycans, and lipids.
   - Representation of created mixture of unique chemicals and differentiating these to a "natural product" such as canola oil.
   - Collecting critical information to make determinations on consistent representation is challenging.
   - Standardisation of description for non-standard structures.
   - Variability in interpretation of structures in different jurisdictions.

   The conversation invited users to be aware of common depictions, to use them in different applications and to understand what can and cannot pass from sketching tools into standard formats.

4. "What is a chemical? Innovations in Chemical Descriptions" [166]

   Experts of five creative descriptions were featured in this webinar: Henry Rzepa (Media Types), Adnan Malik (ChEBI), Oliver Koepler (Ontolgies4Chem), Alex Clark (Mixtures-Molecules), and Ken Kroenlein (Complex Substance Schemas). The panellists identified technical challenges that need to be focused on including:

   - Solubility: need for better solubility models;
   - Data capture: FAIR pushing forward, balance with informatics community building connections in parallel;
   - Data quality: particularly auto-validating before publication;
   - Community consensus: IUPAC chemical classification;

[165] https://doi.org/10.5281/zenodo.7435258 (accessed 20230505)
[166] https://doi.org/10.5281/zenodo.7683138 (accessed 20230505)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- Many nuanced details arise in real world challenges:
  - What are critical variables in different use cases?
  - How many layers need to be captured in standard data models for interoperability?
  - How many layers arise empirically? How many can be computed?

3. IUPAC Potential Contribution:
Overall, the four conversations reflected what the IUPAC can potentially contribute to. Below are some of the suggestions:

- Representations of challenging mixtures;
- How to capture distributions around samples e.g polymer length, phases,etc?
- Education of what machine-readable is and what FAIR is?
- Modularity of things: the sample space is incredibly complex. IUPAC can identify what the core commonalities are in sample space.
- Having a high level controlled vocabulary/machine actionable for the chemical space
- Unify data models that can be broadly used;
- There is a need for coordination of activities in enabling the community to not reinvent the wheel e.g vocabulary;
- IUPAC FAIR Cookbook! reach the base communities at the bench level;
- Facilitate Interoperability based on sharing across the communities and representations;
- Supporting interoperability of FAIR, similar to approach with SMILES+;
- Expanding InChI to other areas (e.g., organometallics);
- Education: creating tables and guides of what can be done with different representations, and documenting common use cases;
- Determining what is critical about different classes of entities (small well characterised, what do we need to know about proteins, other macromolecules, nanomaterials);
- What are the key parameters to capture in representations? May need different approaches for different applications;
- Classification is another area that intersects with describing chemical moieties;
- Coordinate common classification schemas for molecular classes and build active chemical and cross-disciplinary ontology community exchange and practice.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

5.2. Research Data Alliance (RDA) P20 Plenary summary

Summary of Symposium: Describing diverse chemistry datasets across distributed data resources[167]

March 23, 2023. RDA 20th Plenary Meeting - Gothenburg (Hybrid)
Name of session organisers: Ian Bruno, Stuart Chalk, Leah McEwen
Group(s) organising the session: Chemistry Research Data Interest Group (CRDIG)
Topics: Chemistry interoperability

1. **Speakers**
   - National Research Data Infrastructure for Chemistry: Johannes Hunold
   - NFDI4Chem-Repositories: Felix Bach
   - Implementing standards for sharing FAIR chemical data, Steffen Neumann, NFDI4Chem and Leah McEwen, IUPAC
   - PSDI: Brian Mathews
   - Catalysis Hub PathFinder: Abraham Nieva de la Hidalga
   - Geochemistry Data Interoperability: Kerstin Lehnert

2. **Session summary**

   This session provided updates and perspectives from regional and disciplinary initiatives relevant to chemistry, focussing on the challenge of describing chemistry data sets to enable interoperability and reuse across resources and domains. This was followed by a discussion that aimed to identify cross-community challenges that might be addressed through activities within the RDA. The discussion identified areas of focus that the group will aim to take forward in collaboration with other RDA groups and community initiatives.

3. **Key outcomes/actions/takeaways**

   1. Opportunity for cross-community collaboration on agreeing standard approaches for describing samples in chemistry that align with wider initiatives.
   2. Desire for a common approach to describing analytical techniques and a catalogue of available standards for representing analytical results in chemistry.
   3. Value in Identifying use cases that establish what is needed for discovery and interoperability across chemistry resources.

[167] https://www.rd-alliance.org/plenaries/rda-20th-plenary-meeting-gothenburg-hybrid/describing-diverse-chemistry-datasets-across (accessed 20230505)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

5.3. American Chemical Society (ACS) 2023 Spring Meeting workshop summary

Summary of Workshop: Advancing FAIR Chemistry: Developing New Services for Sharing Chemical Data [168]

March 27, 2023. ACS Spring Meeting. Indianapolis. USA
Organisers: Leah McEwen, Ian Bruno, Stuart Chalk, Evan Bolton and Fatima Mustafa

*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.

1. Speakers

- Cornell University: Leah McEwen
- Cambridge Crystallographic Data Center: Ian Bruno
- University of North Florida: Stuart Chalk
- U.S. National Center for BioInformatics: Evan Bolton
- St. Olaf College: Robert Hanson
- University of Alabama: Vincent Scalfani
- IUPAC: Fatima Mustafa

2. Workshop summary

The goal of WorldFAIR Chemistry is to support the use of chemical data standards in research workflows to enable downstream data reuse through practical direction and resources. The aim of this workshop was to engage the input and expertise of stakeholders across the chemistry community on prototype services and other IUPAC standards activities in progress. We presented early work on each of these resources and invited the community to share feedback on what will help them to implement these in your workflows - What works well? What needs further refinement? What is missing?

WorldFAIR Chemistry prototype deliverables:

- **Guidance** [169]: recommendations for managing and sharing FAIR chemical data for various stakeholders;

[168] https://doi.org/10.5281/zenodo.7903727 (accessed 20230505)
[169] https://iupac.org/project/2022-027-1-024 (accessed 20230505)

- **Cookbook** [170]: recipes (Cookbook) for preparing and depositing FAIR machine-enabled chemical data;
- **Protocols** [171]: universal protocol for browser-based validation and lookup services.

IUPAC standards projects:

- **Doc-a-thon: Chemical representation best practices for humans and machines**. This session focused on reviewing the existing IUPAC graphical representation standards for chemical structure diagrams and stereochemical representation with considerations for machine-readability.
- **IUPAC FAIRSpec** [172]**-ready aggregations: Recommendations for researchers, authors, and publishers.** IUPAC is developing specifications for aggregating spectroscopy data and chemical structures for reporting spectroscopy measurements. The resulting data collections can be created either manually by a researcher or automatically by an electronic laboratory notebook (ELN) or laboratory instrument management system (LIMS) and the accompanying metadata can also be used as a finding aid.

3. Key discussion points

1. Ideally, data management resources should be capturing metadata from the point of sample identification, through experimental techniques, data collection, measurement parameters (including units), instruments, analysis, etc. through publication and re-use.
2. IUPAC standards need models for digital representation, including chemical systems, quantities, measurements and criteria for interoperability.
3. Ambiguity may not always be apparent until formats are moved from system to system, round-trip testing can help with review of specifications.
4. Consider data curators and others managing data who don’t necessarily have extensive domain knowledge.

5.4. WorldFAIR Chemistry WP03 FAIR Implementation Profiles

5.4.1. FAIR Implementation Profile of IUPAC digital standards

This table summarises a FAIR Implementation Profile (FIP) of IUPAC digital standards generally. Note, individual standards will have more specific profiles; see example for the IUPAC Gold Book FIP given in section 5.4.2 below.

[170] https://iupac.org/project/2022-028-1-024 (accessed 20230505)
[171] https://iupac.org/project/2022-029-1-024 (accessed 20230505)
[172] https://iupac.org/project/2019-031-1-024 (accessed 20230505)

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FAIR Principle name	Referring to MetaData or Data	FIP question	FAIR Implementation Profile of IUPAC standards generally
F1	MD	What globally unique, persistent, resolvable identifier service do you use for metadata records?	Crossref DOIs (primarily for formal narrative descriptions of standards).
F1	D	What globally unique, persistent, resolvable identifier service do you use for datasets?	Crossref DOIs (datasets currently encapsulated inline in text).
F2	MD	What metadata schemas do you use for findability?	Crossref DOIs (for bibliographic metadata of standards publications). *NB: domain level notations for chemical entities and other concepts can also be incorporated into F2 registered metadata schema for indexing and searching.*
F3	D	What is the schema that links the persistent identifiers of your data to the metadata description?	Crossref DOIs (usual approach to journal type publications).
F4	MD	Which service do you use to publish your metadata records?	Crossref (via publishing partner for IUPAC formal standard publications).
F4	D	Which service do you use to publish your datasets?	Crossref (via publishing partner for IUPAC formal standard publications).
A1.1	MD	Which standardised communication protocol do you use for metadata records?	HTTPS (via publishing partner for IUPAC formal standard publications). *NB: domain level notations for chemical entities and other concepts can also be incorporated into A1.1 API protocols for record retrieval.*
A1.1	D	Which standardised communication protocol do you use for datasets?	HTTPS (via publishing partner for IUPAC formal standard publications).
A1.2	MD	Which authentication & authorisation service do you use for metadata records?	N/A (IUPAC formal standard publications are open to read.)
A1.2	D	Which authentication & authorisation service do you use for datasets?	N/A (IUPAC formal standard publications are open to read.)
A2	MD	What metadata preservation policy do you use?	Crossref

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FAIR Principle name	Referring to MetaData or Data	FIP question	FAIR Implementation Profile of IUPAC standards generally
I1	MD	What knowledge representation language (allowing machine interoperation) do you use for metadata records?	Crossref (various schemas for journal articles, book chapters. etc.)
			InChI (implemented in metadata records for chemical datasets.)
			SMILES+ (implemented in metadata records for chemical datasets.)
			HELM (primarily used in local systems, could be implemented in metadata records for chemical datasets.)
			RInChI (primarily used in local systems, could be implemented in metadata records for chemical datasets.)
			MInChI (primarily used in local systems, could be implemented in metadata records for chemical datasets.)
			NInChI (under development, could be implemented in metadata records for chemical datasets.)
I1	D	What knowledge representation language (allowing machine interoperation) do you use for datasets?	ThermoML (XML schema associated with some journals and compiled datasets.)
			AIF (metadata schema under development, current work to secure adoption by relevant journals.)
			MAPT (metadata schema under development; to be recommended for reporting and handling data in journals and compiled datasets.)
			Solubility Schema (metadata schema under development; to be recommended for reporting and handling data in journals and compiled datasets.)
I2	MD	What structured vocabulary do you use to annotate your metadata records?	IUPAC Gold Book (reference to formal chemical concepts in standard specifications.) *NB: IUPAC also endorses formal concepts from the VIM (BIPM Metrology Vocabulary).*
I2	D	What structured vocabulary do you use to encode your datasets?	IUPAC Gold Book (reference to formal chemical concepts in standard specifications.) *NB: IUPAC also endorses formal concepts from the VIM (BIPM Metrology Vocabulary).*

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FAIR Principle name	Referring to MetaData or Data	FIP question	FAIR Implementation Profile of IUPAC standards generally
I3	MD	What semantic model do you use for your metadata records?	FAIRSpec (under development; format agnostic specification for association of metadata and data components to enable interoperability between many existing data formats.)
I3	D	What semantic model do you use for your datasets?	JCAMP-DX (implemented by instrument vendors, analysis software, ELNs, repositories.)
R1.1	MD	Which usage licence do you use for your metadata records?	CC-BY-NC-ND 4.0 (default licence for reuse of IUPAC standard specifications.)
R1.1	D	Which usage licence do you use for your datasets?	CC-BY-NC-ND 4.0 (default licence for reuse of IUPAC standard specifications.)
R1.2	MD	What metadata schema do you use for describing the provenance of your metadata records?	Crossref DOIs (formal copy of record published IUPAC standard specifications.)
R1.2	D	What metadata schema do you use for describing the provenance of your datasets?	Crossref DOIs (formal copy of record published IUPAC standard specifications.)
R1.3	MD	What validation service do you use to verify consistency of your metadata records with standards?	*NB: this declaration is not presently available in the FIP assessment, but deemed critical for verifying successful representation of metadata in standard schema. Presently used in crystallography with checkCIF, approaches under development in several IUPAC projects.*
R1.3	D	What validation service do you use to verify consistency of your dataset records with standards?	*NB: this declaration is not presently available in the FIP assessment, but deemed critical for verifying successful representation of metadata in standard schema. Presently used in crystallography with checkCIF, approaches under development in several IUPAC projects.*

5.4.2. FAIR Implementation Profile for the IUPAC Gold Book

This table summarises the initial FAIR Implementation Profile (FIP) of the IUPAC Gold book.

FAIR Principle name	Referring to MetaData or Data	FIP question	FAIR Implementation Profile of the IUPAC Gold Book resource
F1	MD	What globally unique, persistent, resolvable identifier service do you use for metadata records?	Crossref DOIs (each term registered as a book chapter, overall resource registered as a book).
F1	D	What globally unique, persistent, resolvable identifier service do you use for datasets?	Crossref DOIs (each term registered as a book chapter, overall resource registered as a book).
F2	MD	What metadata schemas do you use for findability?	Crossref DOIs (each term registered as a book chapter, overall resource registered as a book).
F3	D	What is the schema that links the persistent identifiers of your data to the metadata description?	Crossref DOIs (each term registered as a book chapter, overall resource registered as a book).
F4	MD	Which service do you use to publish your metadata records?	Crossref DOIs (each term registered as a book chapter, overall resource registered as a book).
F4	D	Which service do you use to publish your datasets?	GoldBook (locally developed and supported by IUPAC.).
A1.1	MD	Which standardised communication protocol do you use for metadata records?	HTTPS (CrossRef metadata retrieval).
A1.1	D	Which standardised communication protocol do you use for datasets?	HTTPS (GoldBook API). *NB: specifications for retrieval defined in the IUPAC Gold Book project.*
A1.2	MD	Which authentication & authorisation service do you use for metadata records?	HTTPS (CrossRef metadata retrieval).
A1.2	D	Which authentication & authorisation service do you use for datasets?	N/A (IUPAC formal standard specifications are open access.)
A2	MD	What metadata preservation policy do you use?	Crossref
I1	MD	What knowledge representation language (allowing machine interoperation) do you use for metadata records?	Crossref (schema book chapters.)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FAIR Principle name	**Referring to MetaData or Data**	**FIP question**	**FAIR Implementation Profile of the IUPAC Gold Book resource**
I1	D	What knowledge representation language (allowing machine interoperation) do you use for datasets?	SKOS (under review for implementation.) InChI (implemented in metadata records for chemical datasets.) MathJax (under review for possible implementation for equations.) DRUM (digital representation for units of measurement, under review for possible implementation.) *NB: other digital approaches to notation of quantities and symbols under study.*
I2	MD	What structured vocabulary do you use to annotate your metadata records?	Crossref (Dublin Core.)
I2	D	What structured vocabulary do you use to encode your datasets?	IUPAC terminology, VIM, formal terms from other authoritative bodies
I3	MD	What semantic model do you use for your metadata records?	Crossref (events data.)
I3	D	What semantic model do you use for your datasets?	*None. NB: chemistry-based classification approaches under study.*
R1.1	MD	Which usage licence do you use for your metadata records?	CCO 1.0 (Crossref open metadata.)
R1.1	D	Which usage licence do you use for your datasets?	CC-BY-NC-ND 4.0 (default licence for reuse of IUPAC standard specifications.)

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FAIR Principle name	Referring to MetaData or Data	FIP question	FAIR Implementation Profile of the IUPAC Gold Book resource
R1.2	MD	What metadata schema do you use for describing the provenance of your metadata records?	Crossref DOIs (each term registered as a book chapter, overall resource registered as a book).
R1.2	D	What metadata schema do you use for describing the provenance of your datasets?	Crossref DOIs (each term registered as a book chapter, overall resource registered as a book).
R1.3	MD	What validation service do you use to verify consistency of your metadata records with standards?	*NB: this declaration is not presently available in the FIP assessment, but deemed critical for verifying successful representation of metadata in standard schema. Presently used in crystallography with checkCIF, approaches under development in several IUPAC projects.*
R1.3	D	What validation service do you use to verify consistency of your dataset records with standards?	*NB: this declaration is not presently available in the FIP assessment, but deemed critical for verifying successful representation of metadata in standard schema. Presently used in crystallography with checkCIF, approaches under development in several IUPAC projects.*

5.4.3. IUPAC FAIR Enabling Resources

This table summarises IUPAC digital standards that can be used as FAIR Enabling Resources (FERs).

FAIR Principle name	**FAIR Enabling Resource Type**	**FAIR Enabling Resource**	**Resource status**	**Notes**
F2	metadata schema	InChI (IUPAC International Chemical Identifier)	existing FER	Linear notation can be implemented into metadata schema for indexing and searching
F2	metadata schema	SMILES+ (IUPAC SMILES standard reader)	FER in development	Linear notation can be implemented into metadata schema for indexing and searching

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

F2	metadata schema	HELM (Hierarchical Editing Language for Macromolecules)	existing FER	Linear notation can be implemented into metadata schema for indexing and searching
F2	metadata schema	RInChI (Reaction InChI application)	existing FER	Linear notation can be implemented into metadata schema for indexing and searching
F2	metadata schema	MInChI (Mixture InChI application)	FER in development	Linear notation can be implemented into metadata schema for indexing and searching
F2	metadata schema	NInChI (Nanomaterials InChI application)	FER in development	Linear notation can be implemented into metadata schema for indexing and searching
F2	metadata schema	Gold Book (IUPAC Compendium of Chemical Terminology	existing FER	Referrable terminology can be implemented into metadata schema for indexing and searching.
A1.1	communication protocol	IUPAC chemical structure API protocol	FER in development	API protocol specification for verifying and exchanging chemical notation
I1	knowledge representation language	InChI (IUPAC International Chemical Identifier)	existing FER	Can be implemented in metadata records for chemical datasets.
I1	knowledge representation language	SMILES+ (IUPAC SMILES standard reader)	FER in development	Can be implemented in metadata records for chemical datasets.
I1	knowledge representation language	HELM (Hierarchical Editing Language for Macromolecules)	existing FER	Can be implemented in metadata records for chemical datasets.
I1	knowledge representation language	RInChI (Reaction InChI application)	existing FER	Can be implemented in metadata records for chemical datasets.
I1	knowledge representation language	MInChI (Mixture InChI application)	FER in development	Can be implemented in metadata records for chemical datasets.
I1	knowledge representation language	NInChI (Nanomaterials InChI application)	FER in development	Can be implemented in metadata records for chemical datasets.

I1	knowledge representation language	ThermoML (IUPAC thermodynamic data specification)	existing FER	Can be used for reporting measurement data and in compiled datasets.)
I1	knowledge representation language	AIF (IUPAC adsorption metadata schema)	FER in development	Can be used for reporting measurement data and in compiled datasets.)
I1	knowledge representation language	MAPT (IUPAC machine-accessible periodic table metadata schema)	FER in development	Can be used for reporting measurement data and in compiled datasets.)
I1	knowledge representation language	Solubility Schema (IUPAC solubility metadata schema)	FER in development	Can be used for reporting measurement data and in compiled datasets.)
I2	structured vocabularies	Gold Book (IUPAC Compendium of Chemical Terminology	existing FER	Referrable terminology can be implemented into metadata schema and referenced in ontologies.
I3	semantic model	FAIRSpec (specification for collections of spectroscopic data and chemical structures)	FER in development	Can be used for associating data types in data systems and ontologies.
I3	semantic model	JCAMP-DX (exchange format for spectroscopic data types	existing FER	Used for representing spectroscopic data.
R1.3	validation specification	IUPAC validation criteria for chemical structure notations	FER in development	Can be used for verifying chemical structure notations.
R1.3	validation specification	IUPAC validation criteria for chemical metadata	FER in development	Can be used for verifying successful representation of metadata in standard schema and formats. Existing example is checkCIF used in crystallography. Additional validation criteria being defined inIUPAC digital projects.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

5.5. Worked Example: Reuse of the NIST TRC ThermoML Dataset

The NIST ThermoML Archive dataset (https://doi.org/10.18434/mds2-2422) contains nearly 12,000 ThermoML XML files collected from researchers publishing papers in five journals from 2003-2019. This dataset is high quality due to the application of the ThermoML schema and the efforts by NIST to validate the data entered.

The ThermoML schema (v4) (http://trc.nist.gov/ThermoML.xsd) is extremely detailed and the structure and layout the schema file, as well as the data structure it defines, helps the user appreciate how the data and metadata in the files are structured. This is possible because naming conventions are used to define XML elements, names are descriptive allowing human and machine understanding of content, useful documentation has been added and categorization of thermophysical properties helps classification of techniques. All of this made it possible to fully understand how the data was structured and thus informed the appropriate way in which it should be translated into a MySQL database.

As an example of reuse of this data and making it more available for reuse the following have been developed in the Chalk Research Group at the University of North Florida.[173]

1. The data has been converted to a MySQL database (https://github.com/ChalkLab/Dataset-NIST-TRC-MySQL), using scripts in the PHP hypertext preprocessor language, to allow the dataset to be more easily searched and digitally accessed. In addition, the database has been constructed to match the ThermoML schema and in each database table entries have been deduplicated (e.g., keywords, references, substances) to keep the database DRY (Don't Repeat Yourself). For chemical substances, CAS Registry Numbers, IUPAC names and PubChem IDs, in addition to the InChI strings/keys available in the ThermoML files, have been added to make substances more findable.
2. This database has been used as the basis for a website created using the CakePHP framework and is available online (https://trc.stuchalk.domains.unf.edu) and has been made available for download (https://github.com/ChalkLab/CakePHP_TRC). The site allows users to search and retrieve data based on substances, mixtures, references, and keywords and to view individual datasets with graph representation. The data is also downloadable as JSON-LD in the SciData Framework (see below) for conversion to RDF, and storage in a graph database.
3. A set of JSON-LD (122403) files, one for each dataset has been generated using a PHP script and mode available on GitHub (https://github.com/ChalkLab/Dataset-NIST-TRC-JSONLD). This version of the data is ready for ingesting into a graph database, where the SciData

[173] Stuart Chalk, ORCID: 0000-0002-0703-7776. https://orcid.org/0000-0002-0703-7776 (accessed 20230524).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

ontology provides a framework (as RDF triples) by which the data can be searched using SPARQL queries.

4. The JSON-LD dataset in (3) has been ingested into a graph database (Apache Jena) and made available as an Apache Fuseki SPARQL endpoint (https://scidata.unf.edu/sparql/#/dataset/nisttrc/query). This allows searching of the content and linking of this data with data available on other SPARQL endpoints.

The three GitHub repositories above have been assigned DOIs using Zenodo, making them findable and citable.

NOTE: It should be noted that the original versions of the source dataset used here were lacking some consistency in content. Specifically, the XML files did not consistently include InChIStrings and did not have InChIKeys, but rather proprietary CAS Registry Numbers. These have since been incorporated into the ThermoML Archive by NIST.[174]

[174] Riccardi, D.; Trautt, Z.; Bazyleva, A.; Paulechka, E.; Diky, V.; Magee, J. W.; Kazakov, A. F.; Townsend, S. A.; Muzny, C. Towards improved FAIRness of the ThermoML Archive. *J. Comp Chem.* 2022, 43(12), 879-887. https://doi.org/10.1002/jcc.26842

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6. Bibliography

Battino, R.; Clever, H. L.; Fogg, P. G. T.; Young, C. L. Introduction to the Solubility Data Series; Solubility of Gases in Liquids. In, *Carbon Dioxide in Water and Aqueous Electrolyte Solutions*; Scharlin, P., Ed.; Solubility Data Series, Vol. 62; International Union of Pure and Applied Chemistry, Oxford University Press, 1996; pp vi-xiv.

Bazyleva, A,; Abildskov, J.; Anderko, A.;, Baudouin, O.; Chernyak, Y.; de Hemptinne, J.; Diky, V.; Dohrn, R.; Elliott, J. R.; Jacquemin, J.; Jaubert, J.; Joback, K. G.; Kattner, U. R.; Kontogeorgis, G. M.; Loria, H.; Mathias, P. M.; O'Connell, J. P.; Schröer, W.; Smith, G. J.; Soto, A.; Wang, S.; Weir, R. D. Good reporting practice for thermophysical and thermochemical property measurements (IUPAC Technical Report). *Pure App. Chem.* 2021, 93(2), 253-272. https://doi.org/10.1515/pac-2020-0403

Beilstein, F. R. *Handbuch der Organischen Chemie*, 1st ed.; Leopold Voss, 1881; *(data from 1771).*

Blanke, G.; Doerner, T.; Lynch, N. Chemical Data in Life Sciences R&D and the FAIR Principles. 2020. Zenodo. https://doi.org/10.5281/ZENODO.3970745

Bourne, P. E.; Berman, H. M.; McMahon, B.; Watenpaugh, K. D.; Westbrook, J. D.; Fitzgerald, P. M. D. 'Macromolecular Crystallographic Information Fil'. In *Methods in Enzymology*, 277:571–90. Macromolecular Crystallography Part B. Academic Press, 1997. https://doi.org/10.1016/S0076-6879(97)77032-0

Burton, A.; Aryani, A.; Koers, H.; Manghi, P.; La Bruzzo, S.; Stocker, M.; Diepenbroek, M.; Schindler, M.; Fenner, M. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Mag.* 2017 23 (1/2). https://doi.org/10.1045/january2017-burton.

Bruno, I.; GraŽulis, S.; Helliwell, J. R.; Kabekkodu, S. N.; McMahon, B.; Westbrook, J. Crystallography and Databases. *Data Sci. J.* 2017, 16, 38. https://doi.org/10.5334/dsj-2017-038.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Bruno, I.; Ward, S. PSDI Case Study 8: The Role of Structure in Physical Sciences Data Management, 2022. https://www.psdi.ac.uk/case-study-8 (accessed 20230428).

Buttegieg, P. L. WorldFAIR Project (D11.1) An assessment of the Ocean Data priority areas for development and implementation roadmap. 2023. Zenodo. https://doi.org/10.5281/zenodo.7682399

Chirico, R. D.; de Loos, T. W.; Gmehling, J.; Goodwin, A. R. H.; Gupta, S.; Haynes, W. M.; Marsh, K. N.; Rives, V.; Olson, J. D.; Spencer, C.; Brennecke, J. F.; Trusler, J. P. M. Guidelines for reporting of phase equilibrium measurements (IUPAC Recommendations 2012). *Pure App. Chem.* 2012, 84(8) 1785-1813. https://doi.org/10.1351/PAC-REC-11-05-02

Chirico, R. D.; Frenkel, M.; Diky, V. V.; Marsh, K. K.; Wilhoit, R. C. ThermoML; An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 2. Uncertainties. *J. Chem. Eng. Data.* 2003, 48(5) 1344–1359. https://doi.org/10.1021/je034088i

Chirico, R. D.; Frenkel, M.; Magee, J. W.; Diky, V.; Muzny, C. D.; Kazakov, A. F.; Kroenlein, K.; Abdulagatov, I.; Hardin, G. R.; Acree Jr., W. E.; Brenneke, J. F.; Brown, P. L.; Cummings, P. T.; de Loos, T. W.; Friend, D. G.; Goodwin, A. R. H.; Hansen, L. D.; Haynes, W. M.; Koga, N.; Mandelis, A.; Marsh, K. N.; Mathias, P. M.; McCabe, C.; O'Connell, J. P.; Pádua, A.; Rives, V.; Schick, C.; Trusler, J. P. M.; Vyazovkin, S.; Weir, R. D.; Wu, J. Improvement of Quality in Publication of Experimental Thermophysical Property Data: Challenges, Assessment Tools, Global Implementation, and Online Support. *J. Chem. Eng. Data.* 2013, 58(10), 2699–2716. https://doi.org/10.1021/je400569s

Coles, S.; Sarjeant, A. IUCr Workshop on "When Should Small Molecule Crystallographers Publish Raw Diffraction Data?" *ACA RefleXions*, 2021.

Cvita˘s, T. Quantities Describing Compositions of Mixtures. *Metrologia.* 1996, 33, 35–39. https://doi.org/10.1088/0026-1394/33/1/5

Dhaked, D. K.; Ihlenfeldt, W.; Patel, H.; Delannée, V.; Nicklaus, M. C. 'Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2'. *J. Chem. Info. Model.* 2020, 60(3), 1253–75. https://doi.org/10.1021/acs.jcim.9b01080

Eisenhart, C. Expression of the uncertainties of final results. *Science.* 1968, 160(3833), 1201-1204. DOI: https://doi.org/10.1126/science.160.3833.1201

Frenkel, M.; Chirico, R.; Diky, V.; Brown, P.; Dymond, J.; Goldberg, R.; Goodwin, A.; Heerklotz, H.; Königsberger, E.; Ladbury, J.; Marsh, K.; Remeta, D.; Stein, S.; Wakeham, W.; Williams, P. Extension of ThermoML: The IUPAC standard for thermodynamic data communications (IUPAC Recommendations 2011). *Pure App. Chem.* 2011, 83(10), 1937-1969. https://doi.org/10.1351/PAC-REC-11-05-01

Frenkel, M.; Chiroco, R.; Diky, V.; Dong, Q.; Marsh, K.; Dymond, J.; Wakeham, W.; Stein, S.; Königsberger, E.; Goodwin, A. XML-based IUPAC standard for experimental, predicted, and critically evaluated thermodynamic property data storage and capture (ThermoML) (IUPAC

Recommendations 2006). *Pure App. Chem.* 2006, 78(3), 541-612. https://doi.org/10.1351/pac200678030541

Frenkel, M.; Chirico, R. D.; Diky, V.; Muzny, C.; Dong, Q.; Marsh, K. N.; Dymond, J. H.; Wakeham, W. A.; Stein, S. E.; Königsberger, E.; Goodwin, A. R. H.; Magee, J. W.; Thijssen, M.; Haynes, W. M.; Watanasiri, S.; Satyro, M.; Schmidt, M.; Johns, A. I.; Hardin, G. R. New Global Communication Process in Thermodynamics: Impact on Quality of Published Experimental Data. *J. Chem. Inf. Model.* 2006, 46(6) 2487–2493. https://doi.org/10.1021/ci600208f

Frenkel, M.; Chirico, R. D.; Diky, V.; Yan, X.; Dong, Q.; Muzny, C. ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept. *J. Chem. Inf. Model.* 2005, 45(4) 816–838. https://doi.org/10.1021/ci050067b

Gmelin, L. *Handbuch der anorganischen Chemie*, 1st ed.; Franz Varrentrapp, 1817; *(data from 1772).*

Gregory, A.; Hodson, S. Cross-Domain Interoperability Framework (CDIF) Working Documents. 2023. Zenodo. https://doi.org/10.5281/zenodo.7652742

Gore, S.; Velankar, S.; Kleywegt, G. J. Implementing an X-Ray Validation Pipeline for the Protein Data Bank. *Acta Crystall. Sec. D, Biol. Crystall.* 2012, 68(4), 478–83. https://doi.org/10.1107/S0907444911050359

Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography. *Acta Crystall. Sec. A Found. of Crystall.* 1991, 47(6), 655–85. https://doi.org/10.1107/S010876739101067X

Hanson, R. M.; Jeannerat, D.; Archibald, M,; Bruno, I. J.; Chalk, S. J.; Davies, A. N.; Lancashire, R. J.; Lang, J.; Rzepa, H. S. 'IUPAC Specification for the FAIR Management of Spectroscopic Data in Chemistry (IUPAC FAIRSpec) – Guiding Principles'. *Pure App. Chem.* 2022, 94(6), 623-636. https://doi.org/10.1515/pac-2021-2009.

Harrow, I.; Balakrishnan, R.; Küçük McGinty, H.; Plasterer, T.; Romacker, M. 'Maximizing Data Value for Biopharma through FAIR and Quality Implementation: FAIR plus Q'. *Drug Disc. Today.* 2022, 27(5), 1441–47. https://doi.org/10.1016/j.drudis.2022.01.006

Horvath, A. L.; Getzen, F. W.; Eds. *Carbon Dioxide in Water and Aqueous Electrolyte Solutions.* Solubility Data Series, International Union of Pure and Applied Chemistry, Vol. 62, pp vi-xv; Oxford University Press, 1996.

Hrynaszkiewicz, I.; Simons, N.; Hussain, A.; Grant, R.; Goudie, S. Developing a Research Data Policy Framework for All Journals and Publishers. *Data Sci. J.* 2020, 19(1), 5. https://doi.org/10.5334/dsj-2020-005

Jacobsen, A.; de Miranda Azevedo, R.; Juty, N.; Batista, D.; Coles, S.; Cornet, R.; Courtot, M.; Crosas, M.; Dumontier, M.; Evelo, C. T.; Goble, C.; Guizzardi, G.; Hansen, K. K.; Hasnain, A.; Hettne, K.; Heringa, J.; Hooft, R. W. W.; Imming, M.; Jeffery, K. G.; Kaliyaperumal, F.; Kersloot, M. G.; Kirkpatrick, C. R.; Kuhn, T.; Labastida, I.; Magagna, B.; McQuilton, P. Meyers, N.; Montesanti, A.; van Reisen, M.; Rocca-Serra, P.; Pergl, R.; Sansone, S.; da Silva Santos, L. O. B.; Schneider, J.; Strawn, G.; Thompson, M.; Waagmeester, A.; Weigel, T.; Wilkinson, M. D.; Willighagen, E. L.; Wittenburg, P.; Roos, M.; Mons, B.; Schultes, E. FAIR Principles: Interpretations and

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Implementation Considerations. *Data Intell.* 2020, 2(1-2), 10–29. https://doi.org/10.1162/dint_r_00024

Kanza, S.; Willoughby, C.; Gibbins, N.; Whitby, R.; Frey, J. G.; Erjavec, J.; Zupančič, K.; Hren, M; Kovač, K. Electronic lab notebooks: can they replace paper? *J. Cheminform.* 2017, 9(1). https://doi.org/10.1186/s13321-017-0221-3

Klöcking, M.; Wyborn, L.; Lehnert, K. A.; Ware, B.; Prent, A. M.; Profeta, L.; Kohlmann, F.; Noble, W.; Bruno, I.; Lambart, S.; Ananuer, H.; Barber, N. D.; Becker, H.; Brodbeck, M.; Deng, H.; Deng, K.; Elger, K.; de Souza Franco, G.; Gao, Y.; Ghasera, K. M.; Hezel, D. C.; Huang, J.; Kerswell, B.; Koch, H.; Lanati, A. W.; ter Maat, G.; Martínez-Villegas, N.; Yobo, L. N.; Redaa, A.; Schäfer, W.; Swing, M. R.; Taylor, R. J. M.; Traun, M. K.; Whelan, J.; Zhou, T. Community Recommendations for Geochemical Data, Services and Analytical Capabilities in the 21st Century. *Geochim. Cosmochim. Acta*, 2023, *in press*. https://doi.org/10.1016/j.gca.2023.04.024

Kroon-Batenburg, L. M. J.; Helliwell, J. R.; McMahon, B.; Terwilliger, T. C. Raw Diffraction Data Preservation and Reuse: Overview, Update on Practicalities and Metadata Requirements. *IUCr J.* 2017, 4(1), 87–99. https://doi.org/10.1107/S2052252516018315

Ku, H. H. Precision Measurement and Calibration: Selected NBS Papers on Statistical Concepts and Procedures. In NBS Special Publications, 300v1, Commerce Department; National Institute of Standards and Technology (NIST), 1969. https://www.govinfo.gov/app/details/GOVPUB-C13-6769ef50616bc8a1ca657841ccc19c92 (accessed 20230428).

Lai, A.; Clark, A. M.; Escher, B. I.; Fernandez, M.; McEwen, L. R.; Tian, Z.; Wang, Z.; Schymanski, E. L. *Environ. Sci. Technol.* 2022, 56(12) 7448–7466. https://doi.org/10.1021/acs.est.2c00321

Larsen, S.; Kostorz, G. Publication Standards for Crystal Structures. 2011. http://www.iucr.org/home/leading-article/2011/2011-06-02

Lynch, I.; Afantitis, A.; Exner, T.; Himly, M.; Lobaskin, V.; Doganis, P.; Maier, D.; Sanabria, N.; Papadiamantis, A. G.; Rybinska-Fryca, A.; Gromelski M.; Puzyn, T.; Willighagen, E.; Johnston, B. D.; Gulumian, M.; Matzke, M.; Etxabe, A. G.; Bossa, N.; Serra, A.; Liampa, I.; Harper, S.; Tämm, K.; Jensen, A.; Kohonen, P.; Slater, L.; Tsoumanis, A.; Greco, D.; Winkler, D. A.; Sarimveis, H.; Melagraki, G. Can an InChI for Nano Address the Need for a Simplified Representation of Complex Nanomaterials across Experimental and Nanoinformatics Studies? *Nanomat.* 2020, 10(12) 2493. https://doi.org/10.3390/nano10122493

Martens L., Chambers M., Sturm M., Kessner D., Levander F., Shofstahl J., Tang W.H., Römpp A., Neumann S., Pizarro A.D., Montecchi-Palazzi L., Tasman N., Coleman M., Reisinger F., Souda P., Hermjakob H., Binz P.A., Deutsch E.W..mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics.* 2011 Jan;10(1):R110.000133

McCullough, J. P.; Westrum, Jr., E. F; Evans, W. H. Calorimetry Group Adopts Revised Resolution on Data Publication. *Science.* 1960, 132(3440), 1658-1659. https://doi.org/10.1126/science.132.3440

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

McEwen, L; Exner, T. Example FIPs – WorldFAIR project: Chemistry & Nanomaterials. PARC FAIR Data and Tools webinar series, 2023-04-20. https://nanocommons.github.io/user-handbook/training-courses/PARC-training (accessed 20230428).

McEwen, L.; Martinsen, D.; Lancashire, R.; Lampen, P.; Davies, A. Are your spectroscopic data FAIR?. Spectroscopy Europe. 1 August 2018. https://doi.org/10.1255/sew.2018.a2

O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* 2011, 3, 33. https://doi.org/10.1186/1758-2946-3-33

Parks, N. A.; Fischer, T. G.; Blankenburg, C.; Scalfani, V. F.; McEwen, L. R.; Herres-Pawlis, S.; Neumann, S. The current landscape of author guidelines in chemistry through the lens of research data sharing. *Pure App. Chem.* 2023, Ahead of Print. https://doi.org/10.1515/pac-2022-1001

Perrin, C. L.; Agranat, I.; Bagno, A.; Braslavsky, S. E.; Fernandes, P. A.; Gal, J.; Lloyd-Jones, G. C.; Mayr, H.; Murdoch, J. R.; Nudelman, N. S.; Radom, L.; Rappoport, Z.; Ruasse, M.; Siehl, H.; Takeuchi, Y.; Tidwell, T. T.; Uggerud, E.; Williams, I. H. Glossary of terms used in physical organic chemistry (IUPAC Recommendations 2021). *Pure App. Chem.* 2022, 94(4), 353-534. https://doi.org/10.1515/pac-2018-1010

Prent, A. WorldFAIR Project (D5.1) Formalisation of OneGeochemistry (1.0). 2022. Zenodo. https://doi.org/10.5281/zenodo.7380947

Rauh, D.; Blankenburg, C.; Fischer, T., T. G.; Jung, N.; Kuhn, S.; Schatzschneider, U.; Schulze, T.; Neumann, S. Data format standards in analytical chemistry. *Pure App. Chem.* 2022, 94(6), 725-736. https://doi.org/10.1515/pac-2021-3101

Renner, T. *Quantities, units and symbols in physical chemistry*; International Union of Pure and Applied Chemistry. Royal Society of Chemistry, 2007. https://doi.org/10.1039/9781847557889

Riccardi, D.; Bazyleva, A.; Paulechka, E.; Diky, V.; Magee, J. W.; Kazakov, A. F.; Townsend, S. A.; Muzny, C. D. ThermoML/Data Archive, National Institute of Standards and Technology, 2021. https://doi.org/10.18434/mds2-2422

Riccardi, D.; Trautt, Z.; Bazyleva, A.; Paulechka, E.; Diky, V.; Magee, J. W.; Kazakov, A. F.; Townsend, S. A.; Muzny, C. Towards improved FAIRness of the ThermoML Archive. *J. Comp Chem.* 2022, 43(12), 879-887. https://doi.org/10.1002/jcc.26842

Scalfani, V.; McEwen, L. NSF FAIR Chemical Data Publishing Guidelines Workshop on Chemical Structures and Spectra: Major Outcomes and Outlooks for the Chemistry Community. 2019. https://doi.org/10.7298/fs2d-hx95

Schultes, E.; Magagna, B.; Hettne, K. M.; Pergl, R.; Suchánek, M.; Kuhn, T. Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Grossmann, G., Ram, S. (eds). Advances in Conceptual Modeling. 2020. In, *Lecture Notes in Computer Science*, vol. 12584. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_13

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Shaw, D. G.; Bruno, I.; Chalk, S.; Hefter, G.; Hibbert, D. B.; Hutchinson, R. A.; Magalhães, M. C. F.; Magee, J.; McEwen, L. R.; Rumble, J.; Russell, G. T.; Waghorne, E.; Walczyk, T.; Wallington, T. J. Chemical data evaluation: general considerations and approaches for IUPAC projects and the chemistry community (IUPAC Technical Report). *Pure App. Chem.* 2023, *in press.*

Spek, A. L. Structure Validation in Chemical Crystallography. *Acta Crystall. Sec. D Biol. Crystall.* 2009, 65(2), 148–55. https://doi.org/10.1107/S090744490804362X

Stall, S.; McEwen, L.; Wyborn, L.; Hoebelheinrich, N.; Bruno, I. Growing the FAIR Community at the Intersection of the Geosciences and Pure and Applied Chemistry. *Data Intell.* 2020, 2(1–2), 139–50. https://doi.org/10.1162/dint_a_00036

Stohner, J.; Quack, M. *A Concise Summary of Quantities, Units and Symbols in Physical Chemistry.* 2011, International Union of Pure and Applied Chemistry. https://publications.iupac.org/ci/2011/3304/July11_green-sup-4p.pdf (accessed 20230429).

Strömert, P.; Hunold, J.; Castro, A.; Neumann, S.; Koepler, O. Ontologies4Chem: the landscape of ontologies in chemistry. *Pure App. Chem.* 2022, 94(6), 605-622. https://doi.org/10.1515/pac-2021-2007

Wimalaratne, S. M.; Juty, N.; Kunze, J.; Janée, G.; McMurry, J. A.; Beard, N; Jimenez, R.; Grethe, J. S.; Hermjakob, H.; Martone, M. E.; Clark, T. Uniform Resolution of Compact Identifiers for Biomedical Data. *Scient. Data* 2018, 5(1), 180029. https://doi.org/10.1038/sdata.2018.29

Zwölf, C. M.; Moreau, N. Assessment of the FAIRness of the Virtual Atomic and Molecular Data Centre following the Research Data Alliance evaluation framework. *Eur. Phys. J. D.* 2023, 77, 70. https://doi.org/10.1140/epjd/s10053-023-00649-x

Other reports

11th General Conference on Weights and Measures. *Meas. Tech.* 1960, 3, 909–912. https://doi.org/10.1007/BF00977503

ACS Guide to Scholarly Communication; Banik, G. M., Baysinger, G., Kamat, P. V., Pienta, N. J., Eds; American Chemical Society, 2020. https://doi.org/10.1021/acsguide

Harvard Longwood Medical Area Research Data Management Working Group (2021) "Electronic Lab Notebook Comparison Matrix". https://doi.org/10.5281/ZENODO.4723752

Joint Committee for Guides in Metrology. JCGM 100:2008, Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement (GUM), BIPM, Sèvres, France (2008). https://www.bipm.org/en/publications/guides (accessed 20230524).

Joint Committee for Guides in Metrology. JCGM 200:2012, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM), BIPM, Sèvres, France (2012). https://www.bipm.org/en/publications/guides (accessed 20230524).

The National Academies. *Fostering Integrity in Research*. *Fostering Integrity in Research*. National Academies Press, 2017. https://doi.org/10.17226/21896

The National Academies. *Open Science by Design*. *Open Science by Design*. National Academies Press, 2018. https://doi.org/10.17226/25116

The National Academies. *Reproducibility and Replicability in Science*. *Reproducibility and Replicability in Science*. National Academies Press, 2019. https://doi.org/10.17226/25303

Physical Chemistry Division, Commission on Thermodynamics and Thermochemistry; International Union of Pure and Applied Chemistry. A Guide to Procedures for the Publication of Thermodynamic Data. *Pure App. Chem.* 1972, 29(1-3), 395-408. https://doi.org/10.1351/pac197229010395

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.