

Reusing Chemical Data Across Disciplines: Initiatives and Common Challenges

by Fatima Mustafa, Iseult Lynch, Jan Theunis, Anjana Elapavalore, Hiba Mohammed Taha, Jeremy Frey, Felix Bach, Christian Bonatto Minella, and Leah McEwen

This work discusses reuse of chemical data across disciplines and the role of various data initiatives and projects including PARC, NORMAN-SLE, MassBank, WorldFAIR, PSDI and NFDI4Chem to facilitate increased data sharing. Improved machine-readable chemical data supports global research and interdisciplinary methodologies crucial for sustainable development and achievement of UNESCO's Open Science priorities and the UN Sustainability Development Goals. Examples of success and ongoing approaches include integrating toxicology and chemical exposure data using ontologies, linking specialised chemical data collections with larger repositories such as PubChem, and developing IUPAC International Chemicals Identifier (InChI) extensions for nanomaterials and mixtures. National data infrastructure projects in the UK and Germany focus on digitising and standardising chemical research data management workflows, aiding scientists in data collection, storage, processing, analysis, disclosure, and reuse. These global initiatives aim to enhance chemical data interoperability to solve real-world problems, foster collaboration, and promote innovation while considering sustainable data resources beyond individual projects.

Chemistry underlies many critical worldwide issues including climate [Gür, 2022], health [Kenny and Anushree, 2021], and food availability [Jansen *et al.*, 2020]. Linking chemical exposure to adverse health outcomes requires integrating chemical data with environmental, biological, and toxicological data. Making chemical data more FAIR (Findable, Accessible, Interoperable, and Reusable) across disciplines enables easier integration, identification of connections (causation and correlation), and incorporation into models to address global challenges more effectively.

Standard scientific criteria for describing chemicals and chemical properties are defined by the International Union of Pure and Applied Chemistry (IUPAC). IUPAC led the WorldFAIR case study in Chemistry to align these scientific standards with the FAIR principles and facilitate implementation into data systems, tools and workflows [McEwen and Bruno, 2023, Thiessen *et al.*, 2023, Chalk *et al.*, 2024]. The WorldFAIR initiative [WorldFAIR Project, 2024], led by CODATA and the

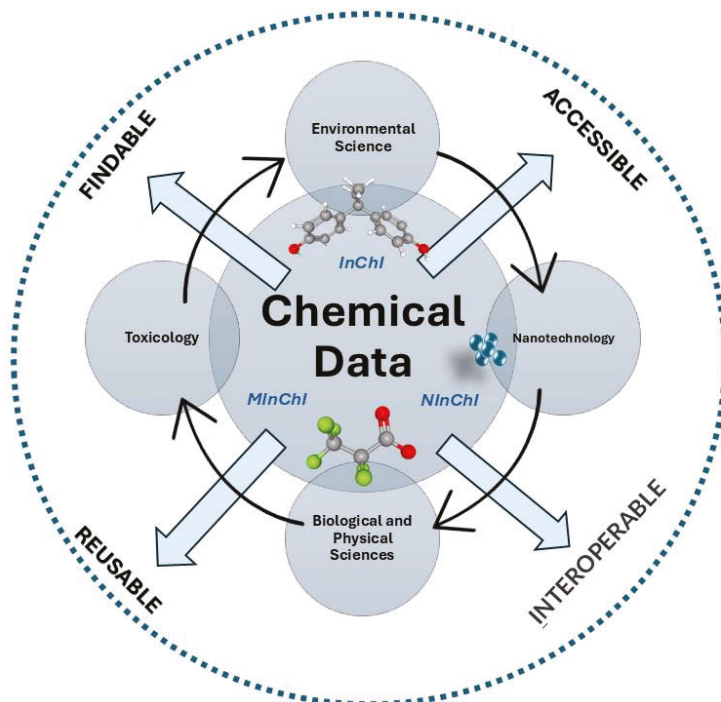
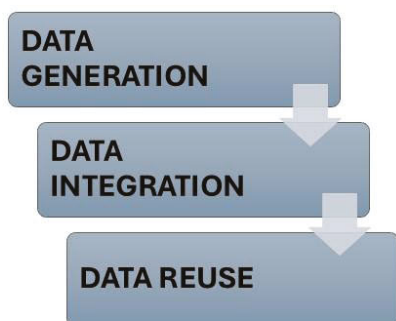
Research Data Alliance (RDA), is facilitating implementation of the FAIR Data Principles [Wilkinson *et al.*, 2016] across disciplines through a global framework for interoperability and collaboration.

This collaborative paper arose from a session held at the International Data Week 2023, in Salzburg, Austria, titled "Beyond FAIR: Reusing Chemical Data Across-disciplines with CARE, TRUST, and Openness", exploring the integration of chemical data across disciplines in several case studies in WorldFAIR and partner initiatives to demonstrate how FAIR chemical and chemistry data can drive innovation and address real-world problems.

Working with Chemical data in different contexts

Currently, multiple **active projects** are developing methodologies, processes and partnerships for advancing cross-disciplinary research areas that utilise chemical data. Next-generation chemical risk assessment approaches are being developed through the Partnership for the Assessment of Risks of Chemicals (PARC) [Marx-Stoelting *et al.* 2023]. PARC has prioritised sets of chemicals for ongoing evaluation, building on the Human BioMonitoring project (HBM4EU) and extending these based on needs identified by regulators and stakeholders. Chemical data must be linked to environmental, population, biological, and toxicological data to assess risk (risk = exposure x hazard) and manage identified risks effectively [Doe, 2023]. All PARC chemicals have unique identifiers using the IUPAC International Chemical Identifier (InChI) [Heller *et al.*, 2013] to enable linking of metadata with datasets on their environmental occurrence, toxicity and adverse outcomes [Willighagen *et al.*, 2013; Watford *et al.*, 2019]. PARC will also serve as a use case for practical implementation of the extension of InChI to cover mixtures (MInChI) [Clark *et al.*, 2019], as humans and the environment are exposed to numerous chemicals simultaneously and in parallel.

The NORMAN Network of reference laboratories, research centres and related organisations developed a Suspect List Exchange (NORMAN-SLE) [Mohammed Taha *et al.*, 2022] to support monitoring of emerging environmental substances. The environmental analytical chemistry community uses suspect screening with high-resolution mass spectrometry (HRMS) to detect chemicals in samples based on suspect lists [Hollender *et al.*, 2023]. The NORMAN-SLE provides open access to 119 suspect list collections from over 80 contributors, totalling over 120,000 unique substances. Suspect lists from collaborators, including metadata and



transformation information, are further annotated with identifiers and notations, including chemical structures, InChI, InChIKey, and IUPAC name and archived on Zenodo. This content is further mapped and integrated into the PubChem repository, hosted by the National Institutes of Health (NIH) [Kim *et al.*, 2023], through an automated workflow [Mohammed Taha *et al.*, 2022].

Non-targeted environmental cheminformatics studies rely on diverse resources and tools, particularly spectral and chemical compound databases such as MassBank and PubChem, to identify unknown compounds [Elapavalore *et al.*, 2023]. MassBank (an open mass spectral library) and PubChem both exemplify the FAIR principles for accessibility and interoperability and facilitate programmatic data reuse [Schymanski and Bolton, 2021], including integration of MassBank records into PubChem to enrich spectral data content [Elapavalore *et al.*, 2023]. Information is extracted from MassBank, validated and adapted to the PubChem record format. Fields such as Authors and Instrument are displayed, with the complete MassBank summary file forming the basis for a substance file in PubChem. Cross-links back to the original MassBank data are provided by SPLASH, an unambiguous, database-independent spectral identifier inspired by InChI [Wohlgemuth, *et al.*, 2015].

The WorldFAIR nanomaterials case study focused on implementation of “on the fly” approaches to meta-data annotation from the perspectives of the data

provider and data users to bridge the gap between data generation and management [Exner *et al.*, 2023]. Open-source platforms KoNstanz Information MinEr (KNIME) can be used to automate FAIRification workflows, annotate datasets with identifiers such as InChI and support transparent, reproducible workflows and integrate with platforms providing computationally-ready datasets like NanoPharos [Lynch *et al.*, 2023]. The utility of KNIME-linked nanoinformatics models is critical in evaluating model precision and reliability, and the automated mechanism for channelling modelling results back into NanoPharos augments the database with new research findings and enhances the reusability and overall value of the datasets. A critical step in integrating and identifying nanomaterials datasets is formalising an extension of the IUPAC InChI standard for nanomaterials (NInChI) that covers over composition, size, shape, crystal phase, surface ligand composition, binding modality [Lynch *et al.*, 2020].

Common needs in using and exchanging chemical data

The chemical sciences, as a very broad discipline, presents an interesting landscape to appreciate the needs and opportunities for sharing and reusing data across domains. Every tangible material in the natural and human-created environment has a chemical nature which impacts its utility and behaviour in the environment. The persistence and distribution of chemicals

of concern in the natural environment is truly a global challenge and numerous disciplines need to work with chemical data that is interoperable across these domains. Molecular entities associated with these chemicals are fundamental to our understanding of biological and material properties and underlie the configuration of many property data models and resources. Standards and protocols are critical to enable these chemical representations and to tie them to broader contexts, including for example managing and documenting data:

- associated with complex mixtures of multiple chemical components at real-world scale and under real world conditions.
- collected on the behaviour and transformations of these chemicals under changing conditions and their impact on biological, environmental, and materials formulation pathways.

A substantial barrier to the exchange of chemical data is the lack of standardised system-to-system interoperability across data resources and analysis tools. Each resource may use different motifs and models for chemical representation and interpretation, which impacts reuse of associated data. Challenges may also arise through inconsistent syntax, lack of adherence to rule-sets, and variables querying and exporting. Confirming the identity of chemical substances is an important part of tracking provenance and reusability of chemical data. IUPAC is developing a consistent approach for chemistry resources to expose information about the chemical representations used in their system through a common application programming interface (API) protocol that would facilitate navigation across these resources. The IUPAC InChI standard chemical identifier discussed in several of the case studies highlighted in this paper is a critical component for finding and matching related chemical records, and providing the links between InChIs and records in individual data sources through a common protocol could foster data hubs and one-stop shops for cross-links [Thiessen *et al.*, 2023].

As chemical data and chemical principles are increasingly applied to cross-disciplinary use cases, the contexts for describing and analysing chemical substances and properties becomes more diverse. Structured chemical data are necessary for enhancing clarity and interdisciplinary collaboration, however the range of practices in data management and representation lead to inconsistency, even if overall more data are shared in more open repositories. There is a common need for broadly shared practices to enable better

discovery and integration of data across resources, starting with more active collection and sharing of chemical data.

Practices for managing and curating chemical data

Despite advances in digitalisation, chemistry data often remain underutilised and many chemists still use traditional paper laboratory notebooks [Steinbeck *et al.*, 2020 and Herres-Pawlis *et al.* 2020]. In the Physical Sciences, many research bodies have their own data infrastructures, which limits the sharing, integration, and reuse of data across systems. Large scale infrastructure projects are emerging to develop open source tools and services that incorporate chemical data standards and aid scientists in collecting, storing, processing, analysing, disclosing, and reusing data, including two national initiatives in Germany [Nationale Forschungsdateninfrastruktur Chemistry Consortium, NFDI4Chem] and the UK [Physical Sciences Data Infrastructure, PSDI].

The potential for fully virtual documentation and data management environments that support the entire research data cycle, including instrument integration, electronic laboratory notebooks and interlinked FAIR data repositories are being realised through implementation pilots. Work on a zinc complex for bioplastic production was successfully replicated in China using deposited data in the NFDI4Chem Chemotion repository [Hermann *et al.*, 2020]; however, while the published paper was cited, the data itself was not, highlighting the challenge of tracking and incentivizing data reuse. In another example, current papers in biomolecular simulations lack sufficient details of how to repeat calculations. Piloting a workflow system in PSDI that captures all simulation phases and metadata detailing inputs, outputs, calculations, computers used, and key files resulted in a complex provenance map, even for simple simulations. This underlines the need for computer-readable and processable, FAIR provenance information. One potential approach is adoption of the Modelling Data (MODA) metadata reporting templates, as promoted by the WorldFAIR Nanomaterials case study via which an online tool to guide users in computing their model metadata has been developed to enhance update and reduce the risk of error; <https://www.enaloscoud.novamechanics.com/insight/moda/> [Kolokathis *et al.*, 2024].

A common goal across these national infrastructure projects is to provide open source tools that are broadly accessible to data stewards as well as researchers supporting chemical data sharing and management. Lack

Initiatives and Common Challenges

of standardised interfaces and metadata across repositories hinder seamless data integration and tracking. Standardised terminologies and ontologies can structure and harmonise chemical data, and enhance clarity and interoperability. The NFDI4Chem Terminology Service (TS) indexes terminologies most pertinent to chemistry to support data curators in selecting appropriate terms for various scientific contexts from a wide range of disparate terminologies and ontologies connected to chemistry and other domains.

Collaborative curation and management will be necessary to maximise the benefit of these community resources and sustain data resources beyond the funded lifetime of individual projects. Working with existing infrastructures and domain experts to join up systems still involves bespoke approaches to meet needs across these use cases. As more FAIR data standards are developed in chemistry and other domains, it will be important to further align with broader cross-domain practices such as described in the (WorldFAIR initiated) Cross-Domain Interoperability Framework (CDIF) to integrate general FAIR principles, like discovery metadata and data structure, with domain-specific elements such as controlled vocabularies and ontologies [Cross-Domain Interoperability Framework (CDIF) Working Group, 2023].

Conclusions

Central to the effectiveness of many cross-domain studies involving chemical analysis is the strategic utilisation of various cheminformatics tools that can operate across open-source chemical compound databases. The collective services for FAIR data curation and compilation from these and other resources enable integration of data from many sources across these platforms. Filling the gaps in chemical databases with expert knowledge from many domains maximises the potential of FAIR chemical data to transcend boundaries, facilitate revolution of fields such as environmental and health monitoring, and illuminate pathways towards a sustainable future.

Community standards to facilitate common languages for chemical data description are key to the ability to align data with the FAIR principles and successfully exchange chemical data across systems. Further standardisation and implementation of terminologies and ontologies is needed to enable virtual federation for discovery and interoperability across the chemical, material, life and other data sciences to enable reuse in combination with datasets from many domains. Fostering the development of expertise in FAIR data curation and standards development, and

provision of user-friendly tools to support implementation of these standards, will be critical for realising and sustaining the potential of linked data.


Investment in developing fit for purpose open-source tools and infrastructure through large-scale initiatives such as NFDI4Chem, PSDI and PARC can actively germinate the transition for researchers towards using FAIR, support reproducible and reusable processes, and contribute to broader implementation of the FAIR data principles globally. Emphasising data citations alongside traditional paper citations is critical to ensure proper attribution and traceability, promoting reproducibility and scientific integrity. Reliable open-source data enriched with metadata encourages reuse, fosters collaboration, accelerates research, and minimises redundant efforts. However, cross community issues are persistent and collaborative approaches are needed for implementing interoperability and reuse in practice.

Acknowledgments and Funding

The session underpinning this paper was funded via the Horizon Europe WorldFAIR (Grant Agreement No. 101058393) project in which the UoB's participation is funded by UKRI / Innovate UK via the Horizon Europe guarantee fund (Grant No. 1831977).

JT, IL, AE and HMT acknowledge support from PARC, funded by the European Union Horizon Europe Research and Innovation Programme [Grant Agreement number 101057014] in which UoB's participation is funded by UKRI / Innovate UK via the Horizon Europe guarantee fund (Grant No. 1752317).

NFDI4Chem is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure (NFDI4/1). NFDI4Chem – Chemistry Consortium in the NFDI (Project number 441958208).

PSDI acknowledges the funding support by the UKRI Digital Research Infrastructure (DRI) scheme through EPSRC grants EP/X032701/1, EP/X032663/1 and EP/W032252/1. 

References

- Cross-Domain Interoperability Framework (CDIF) Working Group (2023) Cross Domain Interoperability Framework (CDIF): Discovery Module (v01 draft for public consultation). <https://doi.org/10.5281/zenodo.10252563>
- Chalk, S., Munday, S., et al. (2024). WorldFAIR (D3.2) Training Package: FAIR Chemistry Cookbook (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.10711950>
- Clark, A.M., McEwen, L.R., et al. Capturing mixture composition: an open machine-readable format for representing mixed substances. *J Cheminform* 11, 33 (2019). <https://doi.org/10.1186/s13321-019-0357-4>

- Doe J. The risk assessment paradigm for chemicals: a critical review of current and emerging approaches, In: Present Knowledge in Food Safety. Editors: Knowles ME, Anelich LE, et al. Academic Press, 2023, 568-574. <https://doi.org/10.1016/B978-0-12-819470-6.00015-9>
- Elapavalore, A., Kondić, T., et al. (2023). Adding open spectral data to MassBank and PubChem using open source tools to support non-targeted exposomics of mixtures. *Environmental Science: Processes & Impacts*, 25(11), pp. 1788-1801. <https://doi.org/10.1039/D3EM00181D>
- Exner, T.E., Papadiamantis, A.G., et al. Metadata stewardship in nanosafety research: learning from the past, preparing for an "on-the-fly" FAIR future. *Frontiers in Physics*, 2023, 11, <https://doi.org/10.3389/fphy.2023.1233879>
- Gür, T.M., 2022. Carbon dioxide emissions, capture, storage and utilization: Review of materials, processes and technologies. *Progress in Energy and Combustion Science*, 89, p. 100965. <https://doi.org/10.1016/j.pecs.2021.100965>
- Heller, S., McNaught, A., et al. (2013). InChI-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5, pp. 1-9. <https://doi.org/10.1186/1758-2946-5-7>
- Herres-Pawlus S., Liermann J. C., Koepler O. (2020). Research Data in Chemistry – Results of the first NFDI4Chem Community Survey. *Z. Anorg. Allg. Chem.*, 646, 1748–1757 <https://doi.org/10.1002/zaac.202000339>
- Hermann, A., Hill, S., et al. (2020). Next generation of zinc bisguanidine polymerization catalysts towards highly crystalline, biodegradable polyesters. *Angewandte Chemie International Edition*, 59(48), pp. 21778-21784. <https://doi.org/10.1002/anie.202008473>
- Hollender, J., Schymanski, E.L., et al. (2023). NORMAN guidance on suspect and non-target screening in environmental monitoring. *Environmental Sciences Europe*, 35(1), p. 75. <https://doi.org/10.1186/s12302-023-00779-4>
- Jansen, T., Claassen, L., et al. (2020). 'All chemical substances are harmful.' public appraisal of uncertain risks of food additives and contaminants. *Food and Chemical Toxicology*, 136, p. 110959. <https://doi.org/10.1016/j.fct.2019.110959>
- Kenny, C. and Priyadarshini, A.. "Review of current healthcare waste management methods and their effect on global health." *Healthcare*. Vol. 9. No. 3. MDPI, 2021. <https://doi.org/10.3390/healthcare9030284>
- Kim, S., Chen, J., C et al. (2023). PubChem 2023 update. *Nucleic acids research*, 51(D1), pp. D1373-D1380. <https://doi.org/10.1093/nar/gkac956>
- Kolokathis, P.D., Sidiropoulos, N.K., et al. Easy-MODA: Simplifying Standardized Registration of Scientific Simulation Workflows through MODA Template Guidelines powered by the Enalos Cloud Platform. *Computational and Structural Biotechnology Journal*, 2024.
- Lynch, I., Afantitis, A., et al. Can an InChI for Nano Address the Need for a Simplified Representation of Complex Nanomaterials across Experimental and Nanoinformatics Studies? *Nanomaterials* 2020, 10, 2493. <https://doi.org/10.3390/nano10122493>
- Lynch, I., Afantitis, A. (2023). WorldFAIR Project (D4.1) Nanomaterials domain-specific FAIRification mapping. <https://doi.org/10.5281/zenodo.7887340> (Accessed 23 July 2024).
- Marx-Stoelting, P., Rivière, G., et al. (2023). A walk in the PARC: developing and implementing 21st century chemical risk assessment in Europe. *Archives of Toxicology*, 97(3), pp.893-908. <https://doi.org/10.1007/s00204-022-03435-7>
- McEwen, L., & Bruno, I. (2023). WorldFAIR Project (D3.1) Digital recommendations for Chemistry FAIR data policy and practice (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.7887283>
- Mohammed Taha, H., Aalizadeh, R., et al. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. *Environ Sci Eur* 34, 104 (2022). <https://doi.org/10.1186/s12302-022-00680-6>
- PSDI. Available at: <https://www.psd.ac.uk>. (Accessed 24 July 2024).
- Schymanski, E.L. and Bolton, E.E. (2021). FAIR chemical structures in the Journal of Cheminformatics. *Journal of cheminformatics*, 13(1), p.50. <https://doi.org/10.1186/s13321-021-00520-4>
- Steinbeck, C., Koepler, O. (2020) NFDI4Chem - Towards a National Research Data Infrastructure for Chemistry in Germany. Research Ideas and Outcomes 6: e55852. <https://doi.org/10.3897/rio.6.e55852>
- Thiessen, P., Bolton, E., and McEwen, L. R. (2023). WorldFAIR (D3.3) Utility services for Chemistry Standards (1.1). Zenodo. <https://doi.org/10.5281/zenodo.10514901>
- Watford, S., Edwards, S., et al. Progress in data interoperability to support computational toxicology and chemical safety evaluation. *Toxicol Appl Pharmacol*. 2019, 1;380: 114707. <https://doi.org/10.1016/j.taap.2019.114707>
- Wilkinson, M., Dumontier, M., et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- Willighagen, E.L., Waagmeester, A., et al. (2013). The ChEMBL database as linked open data. *Journal of cheminformatics*, 5, pp.1-12. <https://doi.org/10.1186/1758-2946-5-23>
- Wohlgenuth, G., Mehta, S.S., et al. (2016). SPLASH, a hashed identifier for mass spectra. *Nature biotechnology*, 34(11), pp.1099-1101. <https://doi.org/10.1038/nbt.3689>
- WorldFAIR Project. Available at: <https://worldfair-project.eu/> (Accessed 23 July 2024).
- Fatima Mustafa¹, Iseult Lynch^{2,3}, Jan Theunis⁴, Anjana Elapavalore⁵, Hiba Mohammed Taha⁶, Jeremy Frey⁶, Felix Bach⁷, Christian Bonatto Minella⁷, Leah McEwen⁸
- ¹ Department of Chemistry, The University of Texas, San Antonio, USA
- ² School of Geography, Earth and Environment Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom
- ⁴ VITO HEALTH, Flemish Institute for Technological Research (VITO), Mol, Belgium
- ³ Centre for Environmental Research and Justice, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom
- ⁵ Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg. ORCIDs: HMT: 0000-0001-7820-4335; AE: 0000-0002-0295-6618.
- ⁶ Computational Systems Chemistry, School of Chemistry and Chemical Engineering, University of Southampton, UK
- ⁷ FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
- ⁸ Materials Science and Engineering, Cornell University, USA