

FAIR and Open Data in Science: The Opportunity for IUPAC

by Ian Bruno, Simon Coles, Wolfram Koch, Leah McEwen, Fabienne Meyers, and Shelley Stall

At the start of 2020, IUPAC's Committee on Publications and Cheminformatics Data Standards (CPCDS) formed a Task Force to propose guidelines for the dissemination and sharing of machine-readable chemical data. These guidelines would be for IUPAC to adopt internally and recommend to the wider chemistry community.

The FAIR Data Principles were to be central to these guidelines with the aim to ensure that scientific data management and stewardship will make chemical research data findable, accessible, interoperable, and reusable by **both** humans and machines [1].

In a digital context, FAIR means describing data outputs such that they are *machine-actionable*, for example in applications such as modelling and machine learning. Data outputs are **findable** through unique and persistent identifiers that are associated with basic machine-actionable meta-data to distinguish different outputs. Data outputs are **accessible** through systems that employ universal Internet protocols and allow for authentication to access sensitive content. Data outputs are **interoperable** through formats that incorporate authoritative and referable domain vocabularies that are syntactically and semantically parsable. Data are **reusable** through rich metadata that enable linking and compiling, provenance trails that establish authority and open licenses [2].

The membership of the task force includes high level experts supporting chemistry research endeavours and those with leadership positions in scientific organisations active in FAIR initiatives. A key focus was on IUPAC's own data outputs and the opportunity there is for these to be managed in a FAIR manner so they become more readily available for consumption by digital technologies and provide an exemplar in best practice for the wider community.

Whilst the FAIR Principles are central to many discussions pertaining to policy and practices surrounding research data, the extent to which they penetrate the horizons of most chemists in industry and academia may be limited. One of the first questions the Task Force thus contemplated was why the FAIR Principles might matter to chemists and to IUPAC. To answer this,

we turned to a range of broader principles, provocations, and reports that help frame the practical utility of the FAIR Principles, and to initiatives that demonstrate their practical application in other disciplines. The following section highlights some of the landmarks identified as we undertook our tour of the landscape surrounding the FAIR Principles.

Taking in the Scenery

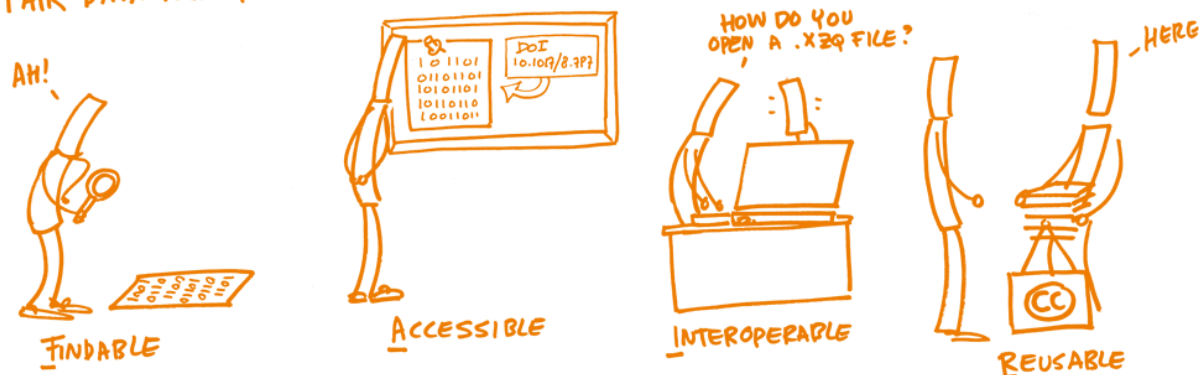
We began our review with "Science as an Open Enterprise," a report published by the Royal Society in 2012 which highlights the enormous potential modern digital technologies have for exploring massive amounts of data to address matters of public policy and business [3]. The report identifies six key areas for action that are very much focussed on data. These include publishing data in reusable form, cultivating experts in the management of digital data, and establishing common standards for sharing information—an area of activity very much aligned with IUPAC's core mission.

A parallel report on "Open Science By Design," published by the U.S. National Academies in 2018, specifically calls out the sharing and preservation of research results under FAIR principles whilst highlighting the need to prepare data and tools to support reproducibility, replicability, and reuse [4]. The importance of the availability of well-managed research data to support reproducibility and replicability in science, and the role that the FAIR Principles have in enabling this, is further emphasised in other National Academies reports specifically focused on this area [5,6].

A policy driver for adoption of the FAIR Principles comes from research funders across regions and domains [7-11]. Their policies recognise research data as a public good and increasingly expect data management plans that reflect elements of the FAIR Principles to ensure data are adequately identified, described, licensed and attributed. Publication policies are also key and a recent International Science Council report articulating principles for publishing in the digital era offers a comprehensive argument for open publication of data in support of claims in articles [12]. It further notes that datasets need to "intelligently open" through procedures formalised by the FAIR Principles.

The economic case for supporting the FAIR Principles is laid out in a report published by the European Commission which estimates the cost to the EU member states of not having FAIR research data to be a minimum of € 10.2 billion per year with a significant portion of this (around 96 %) being due to duplication of effort and storage [13].

FAIR DATA PRINCIPLES



(Reproduced from <https://book.fosteropenscience.eu/en/02OpenScienceBasics/02OpenResearchDataAndMaterials.html>)

Whilst the FAIR Principles have their roots in academic communities, they have also been adopted within the life sciences industry to help shape strategies that support digital transformation and accelerate the application of AI and deep learning to industry data [14]. The opportunity to apply FAIR principles to chemical data within industry to enable interoperability outside of localised silos has further been highlighted [15]. The principles espoused by FAIR become critical when we envisage future laboratories where experimental data are automatically captured and analysed, conceivably in real time, to generate actionable insights that inform subsequent experimental choices [16–18].

The FAIR Principles are central to various European infrastructure activities supporting research in chemistry and beyond. NFDI4Chem for example is a German initiative that aims to build an open and FAIR infrastructure for research data management in Chemistry [19]. More broadly, the EU-sponsored FAIRsFAIR project is producing a range of recommendations in support of FAIR data policies, practices and services [20]. The output of FAIRsFAIR will contribute to the rules of participation in and governance of the European Open Science Cloud, a virtual environment being developed with funding from the European Commission to underpin data-driven science in Europe [21].

Internationally, the FAIR Data Principles are found at the heart of the CODATA Beijing Declaration, a statement of core principles to encourage global cooperation around research data [22]. They also feature in the recently published UNESCO draft Recommendation on Open Science [23] which, in common with the Beijing Declaration, has at its heart the UN Sustainable Development Goals [24] and the reliable and efficient exchange of information and data needed to address the complex and interconnected

global challenges facing the world today.

Successful implementation of the FAIR Principles requires input from stakeholders and service providers across and around the research life cycle. A Commitment Statement adopted by the Coalition for Publishing Data in the Earth and Space Sciences outlines commitments for researchers, funding agencies, publishers, repositories and institutions in support of open and FAIR data [25, 26]. The Royal Society report [3] extends a similar list to include business investors in research, government and regulators. Both highlight a role for societies, academies, and other professional bodies, such as scientific unions, that includes:

- Exploring how enhanced data management could benefit their constituency and how habits, credit and recognition might need to change to achieve this;
- Supporting the promulgation and development of FAIR data principles and educating their communities about these;
- Participating in the development of community standards, infrastructure, tools, and services to enable open and FAIR data practices.

These and other key themes that emerged align well with IUPAC's mission to develop tools that enable the application and communication of chemical knowledge, including:

- The relevance of FAIR for enabling robust reproducibility;
- The importance of establishing trust in research data across sectors;
- The need for infrastructure that enables the long term stewardship of research data;
- Opportunities FAIR offers to underpin improvements in the publication of research data;
- The opportunity for collaboration around FAIR across academia and industry.



Cover of the recently published UNESCO draft Recommendation on Open Science [23] which, in common with the Beijing Declaration, has the FAIR Data Principles at the heart, along with the UN Sustainable Development Goals.

IUPAC Opportunities

IUPAC's more than a century of work on projects and initiatives to define a common language for chemistry position it well to embrace FAIR and other related principles that facilitate exchange of scientific information. Implementing FAIR is an opportunity for the wider community to benefit from prior work undertaken over many years to develop digital representation formats such as JCAMP-DX [27] chemical structure identifiers such as InChI [28] and IUPAC terminologies published in *Pure and Applied Chemistry* and various colour books aggregated in digital form as the GOLD Book [29]. It is also an opportunity to identify requirements for new standards and to develop metadata profiles that provide machine-readable descriptions of standard atomic weights [30], spectroscopic data [31], and critically evaluated solubility data [32]. Incorporating FAIR practices into these efforts will enable IUPAC standards to be more readily reusable and interoperable across domains.

A key factor for the success of FAIR is engaging with communities of researchers to identify and raise awareness of solutions that can enable the FAIR publication of their data. The Chemistry Implementation Network (ChIN) of the GO FAIR initiative aims to move forward the practical implementation of FAIR within the broader chemistry community through such engagement [33]. This will firstly involve facilitating and advocating the development of FAIR standards, materials and software specific to our discipline. Having developed a critical mass of tools and approaches, the second aim of the ChIN is to use these to help drive a culture change across the discipline so that the goal of FAIR chemistry data can be achieved. Set within

the landscape of a range of GO FAIR Implementation Networks covering many disciplines, the ChIN is also well placed to reach beyond the chemistry community and support the use of chemical information in other disciplines. Critical to these activities are domain-based standards and IUPAC has demonstrated support through endorsement of the Chemistry GO FAIR Implementation Network Manifesto [34].

Engagement across stakeholder groups is also key and this was a prominent feature of a 2018 NSF-sponsored workshop organised with input from CPCDS members to establish publishing guidelines for chemical structures and spectra [35]. This brought together researchers, publishers, data organisations, software developers, librarians, and standards bodies and led to practical initiatives to encourage the FAIR publication of machine-readable representations of spectra alongside journal articles [36].

FAIR provides a focal point for bringing together communities from across disciplines to explore various aspects of the management and publication of chemistry research data. As part of International Data Week in 2018, IUPAC joined forces with the International Union of Crystallography (IUCr) to organise a symposium that specifically looked at the challenges of interoperability across disciplines including chemistry, biology, crystallography, and the earth sciences [37]. The opportunities for growing the FAIR community at the intersection of the Geosciences and Pure and Applied Chemistry have further been explored [38]. Currently, IUPAC is engaging in the delivery of a Data Sharing Seminar Series for Societies in collaboration with the American Geophysical Union and other societies [39]. This aims to help communities understand

The Opportunity for IUPAC

more about how they can help enable publication of data in line with guidelines such as FAIR.


Challenges Ahead

Over the years, IUPAC has been responsible for generating data assets that are highly valuable to the scientific community. In a digital age, the value of these assets potentially grows but so does the question of how prepared IUPAC's processes and infrastructure are to fully realise this value for its mission and role. The importance of ensuring that IUPAC is producing standards that become embedded in the core digital infrastructures used in research across the chemical sciences has been previously articulated as part of a vision for a Digital IUPAC [40]. Embracing and adopting FAIR Principles will help accelerate progress towards this essential vision.

A challenge for IUPAC currently is where to host its outputs in machine-accessible form such that they are readily discoverable now and into the future. Any data repository option would need to be one that IUPAC could envisage sustaining for a number of years, including considerations of infrastructure, storage, and resources to oversee ongoing curation. It would also need to enable access to chemical objects that are born digital and machine-actionable. Such requirements are unlikely to be provided by traditional publishing platforms or partnerships. A related challenge is identifying licensing terms that can support broad reuse of outputs whilst also ensuring that integrity is preserved and opportunities for IUPAC to develop revenue streams based on these assets are retained.

The processes adopted while defining, executing, and reporting data related to outputs of projects will also impact the FAIR availability and reuse of IUPAC's assets. Many of IUPAC's outputs to date are published in forms that did not anticipate the digital expectations of today and effort will be required to transform these so they become actionable by machines. We can avoid creating the same challenge for future assets by establishing plans for disseminating outputs digitally early in a project and evolving these throughout. CPCDS are currently in the process of developing guidance that can be applied by IUPAC to ensure that its future outputs are born FAIR-ready and prepared for a digital existence that persists beyond the lifetime of a project.

Central to realising FAIR Chemical Data in IUPAC is the bringing together of "standards" and "community"—two words that strike to the very core of IUPAC's essence. Indeed, bridging these is something that IUPAC is uniquely placed to do. It has the standards and it has an authoritative voice—now is the time to

combine these to advance the adoption of FAIR data principles across the global chemistry community. Doing so will unlock the opportunity of a step change in the provision and supply of quality-assured scientific data that can be used to further enable the application of the chemical sciences in tackling critical global issues facing the future of our societies and our planet. 

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):1-9. <https://doi.org/10.1038/sdata.2016.18>
2. FORCE11. Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0. FORCE11. Published September 10, 2014. <https://www.force11.org/fairprinciples>
3. The Royal Society. *Science as an Open Enterprise*; 2012. https://royalsociety.org/-/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf
4. The National Academies. *Open Science by Design*. National Academies Press; 2018. <https://doi.org/10.17226/25116>
5. The National Academies. *Fostering Integrity in Research*. National Academies Press; 2017. <https://doi.org/10.17226/21896>
6. The National Academies. *Reproducibility and Replicability in Science*. National Academies Press; 2019. <https://doi.org/10.17226/25303>
7. Common principles on data policy - UK Research and Innovation. <https://www.ukri.org/funding/information-for-award-holders/data-policy/common-principles-on-data-policy/>
8. *Division of Chemistry Advice to Principal Investigators on Data Management Plans*; 2018. <https://www.nsf.gov/bfa/dias/policy/autocompliance.jsp>
9. NOT-OD-21-013: Final NIH Policy for Data Management and Sharing. <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>
10. Data, software and materials management and sharing policy - Grant Funding. Wellcome. <https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy>
11. Schiermeier Q. How Europe's €100-billion science fund will shape 7 years of research. *Nature*. 2021;591(7848):20-21. <https://doi.org/10.1038/d41586-021-00496-z>
12. *Opening the Record of Science: Making Scholarly Publishing Work for Science in the Digital Era ISC Report*; 2021.
13. Cost-benefit analysis for FAIR research data: Cost of not having FAIR Research Data. <https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en>
14. FAIR Toolkit Launch - Pistoia Alliance. <https://www.pistoiaalliance.org/news/fair-toolkit-launch/>

15. Blanke G, Doerner T, Lynch N. Chemical Data in Life Sciences R&D and the FAIR Principles. Published online August 2020. <https://doi.org/10.5281/ZENODO.3970745>
16. Henson AB, Gromski PS, Cronin L. Designing Algorithms To Aid Discovery by Chemical Robots. *ACS Cent Sci.* 2018;4(7):793-804. <https://doi.org/10.1021/acscentsci.8b00176>
17. Szymańska E. Modern data science for analytical chemical data – A comprehensive review. *Anal Chim Acta.* 2018;1028:1-10. <https://doi.org/10.1016/j.aca.2018.05.038>
18. Shen Y, Borowski JE, Hardy MA, Sarpong R, Doyle AG, Cernak T. Automation and computer-assisted planning for chemical synthesis. *Nat Rev Methods Primer.* 2021;1(1):1-23. <https://doi.org/10.1038/s43586-021-00022-5>
19. Chemistry Consortium in the National Research Data Infrastructure. <https://nfdi4chem.de/>
20. FAIRsFAIR. <https://www.fairsfair.eu/>
21. European Open Science Cloud (EOSC) | European Commission. https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_en
22. CODATA, Committee on Data of the International Science Council, CODATA International Data Policy Committee, CODATA China High-level International Meeting on Open Research Data Policy and Practice, et al. The Beijing Declaration on Research Data. Published online November 2019. <https://doi.org/10.5281/ZENODO.3552330>
23. First draft of the UNESCO Recommendation on Open Science - UNESCO Digital Library. <https://unesdoc.unesco.org/ark:/48223/pf0000374837>
24. THE 17 GOALS | Sustainable Development. <https://sdgs.un.org/goals>
25. Stall S, Yarmey L, Boehm R, et al. Advancing FAIR Data in Earth, Space, and Environmental Science. *Eos.* 2018;99. <https://doi.org/10.1029/2018EO109301>
26. Stall S, Yarmey L, Cutcher-Gershenfeld J, et al. Make scientific data FAIR. *Nature.* 2019;570(7759):27-29. <https://doi.org/10.1038/d41586-019-01720-7>
27. Grasselli JG. JCAMP-DX, a standard format for exchange of infrared spectra in computer readable form (Recommendations 1991). *Pure Appl Chem.* 1991;63(12):1781-1792. <https://doi.org/10.1351/pac199163121781>
28. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI - the worldwide chemical structure identifier standard. *J Cheminformatics.* 2013;5(1):7-7. <https://doi.org/10.1186/1758-2946-5-7>
29. Gold V, ed. *The IUPAC Compendium of Chemical Terminology.* International Union of Pure and Applied Chemistry (IUPAC); 2019. <https://doi.org/10.1351/goldbook>
30. Machine-Accessible Periodic Table. IUPAC project. <https://iupac.org/project/2019-020-2-024>
31. Development of a Standard for FAIR Data Management of Spectroscopic Data. IUPAC project. Accessed March 26, 2021. <https://iupac.org/project/2019-031-1-024>; see also Hanson R, Jeannerat D, Archibald M, et al. FAIR enough? *Spectrosc Eur.* Published online March 16, 2021:25. <https://doi.org/10.1255/sew.2021.a9>
32. Development of a metadata schema for critically evaluated solubility measurement data. IUPAC project. <https://iupac.org/project/2020-018-1-024>
33. Coles SJ, Frey JG, Willighagen EL, Chalk SJ. Taking FAIR on the ChIN: The Chemistry Implementation Network. *Data Intell.* 2020;2(1-2):131-138. https://doi.org/10.1162/dint_a_00035
34. *Manifesto of the Chemistry GO FAIR Implementation Network.* <https://www.go-fair.org/wp-content/uploads/2019/02/ChIN-Manifesto-final-20180905.pdf>
35. Scafani VF, McEwen L. NSF OAC 2019 Workshop: FAIR Publishing Guidelines for Spectral Data and Chemical Structures. Published online 2019. <https://osf.io/psq7k/>
36. Hunter AM, Carreira EM, Miller SJ. Encouraging Submission of FAIR Data at the Journal of Organic Chemistry and Organic Letters. *J Org Chem.* 2020;85(4):1773-1774. <https://doi.org/10.1021/acs.joc.0c00248>
37. Data interoperability in chemistry, biology, and crystallography: Enabling multidisciplinary solutions to societal challenges. <https://www.scidatacon.org/IDW2018/sessions/223/>
38. 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):139-150. https://doi.org/10.1162/dint_a_00036
39. Data Sharing Seminar Series for Societies | Zenodo. <https://zenodo.org/communities/data-sharing-seminar-series-for-societies/>
40. Frey JG. Digital IUPAC. *Chem. Int.* 2014;36(1):14-16. <https://doi.org/10.1515/ci.2014.36.1.14>

* All web pages were accessed 31 March 2021

Ian Bruno is Director of Data Initiatives at the Cambridge Crystallographic Data Centre, Cambridge, UK and Secretary of the InChI Trust; <http://orcid.org/0000-0003-4901-9936>. Simon Coles is Professor of Structural Chemistry at the University of Southampton and Director of both the UK National Crystallography Service and the UK Physical Sciences Data-science Service; <http://orcid.org/0000-0001-8414-9272>. Wolfram Koch is Executive Director at the German Chemical Society (GDCh); <http://orcid.org/0000-0002-2399-8358>. Leah McEwen is Chemistry librarian at Cornell University, USA and Chair of the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS); <http://orcid.org/0000-0003-2968-1674>. Fabienne Meyers is Associate Director at the International Union of Pure and Applied Chemistry; <http://orcid.org/0000-0001-5326-8316>. Shelley Stall is Senior Director for Data Leadership at the American Geophysical Union; <http://orcid.org/0000-0003-2926-8353>.