



International Journal of Bioinformatics and Cheminformatics

Journal homepage: <https://trpub.org/journal.php?id=MGxgAxAuO>

TRPub

IJBC

International Journal of
Bioinformatics and
Cheminformatics

Vol. 1 Issue 1 2026

Perspective

Open Access

Why FAIR Data Principles Still Struggle to Reach Everyday Cheminformatics Workflows

Ramesh Raghavan^{1*}, Vahan Tran², Pedro Fortin³

¹Department of Life Sciences, Elbe Academy of Life Sciences, Czech Republic

²Department of Life Sciences, Patagonia Graduate Institute of Life Sciences, Chile

³Department of Life Sciences, Sharavati Institute for Advanced Life Sciences, Hungary

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijbc.2026.v1.i1.pp26-31](https://doi.org/10.62715/ijbc.2026.v1.i1.pp26-31)

Received: 14 January 2026

Accepted: 12 March 2026

Published: 12 May 2026

Abstract

The FAIR principles have been endorsed by funders, publishers and most large pharmaceutical organisations, and the chemical sciences have standards of their own that should make compliance straightforward: a line notation for structures, an international chemical identifier, open databases of bioactivity and public repositories that issue persistent identifiers. Yet the data produced in an ordinary medicinal chemistry or cheminformatics project still largely fail the principles, and the reasons have less to do with ignorance than with the structure of the work. This article argues that the gap persists because FAIR was written for datasets while chemistry produces decisions, because the hard part of interoperability in chemistry is semantic rather than syntactic, because licensing and competitive concerns bite hardest exactly where the data are most useful, and because the effort of depositing data falls on people who are not rewarded for it. Counter-arguments are considered, including the view that current infrastructure is adequate and that mandates will close the gap on their own. The article suggests where modest changes by journals, funders and research groups would do most good.

Keywords: FAIR principles, Research data management, Cheminformatics, Chemical data curation, Interoperability, Open science

© 2026 Ramesh Raghavan et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

The FAIR principles, published by [Wilkinson et al. \(2016\)](#), asked that research data be findable, accessible, interoperable and reusable, with emphasis on machines as well as people doing the finding and reusing. They were adopted with unusual speed. Funders wrote them into data management plans, publishers into policies, and the pharmaceutical industry into internal programmes described in surveys such as that of [Wise et al. \(2019\)](#). Guidance on what the principles require in practice followed ([Mons et al., 2017](#); [Jacobsen et al., 2020](#)).

Chemistry ought to have found this easy. It has a compact line notation for structures ([Weininger, 1988](#)), a standard identifier designed precisely so that two laboratories can agree on what a compound is ([Heller et al., 2015](#)), open toolkits for reading and writing chemical formats ([O Boyle et al., 2011](#)), and

* Corresponding author: Ramesh Raghavan, Department of Life Sciences, Elbe Academy of Life Sciences, Czech Republic. E-mail: ramesh.raghavan@elbe.edu

© 2026 Ramesh Raghavan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

large public archives of structures and bioactivity ([Kim et al., 2023](#); [Mendez et al., 2019](#)). A discipline with that infrastructure should be an exemplar.

It is not. Most data generated in an ordinary synthesis and screening project never leave the group in machine-readable form, and much of what is published remains locked in page images, in supplementary documents designed for reading and in spreadsheets whose columns mean whatever their author intended. This article sets out why, and argues that the obstacles are structural rather than technical.

2. What FAIR Asks of a Chemistry Dataset

Stated concretely for a cheminformatics project, the principles ask for a few specific things. Each compound should carry an identifier that resolves and that another group can match to its own records. Each measurement should travel with the conditions that produced it. The files should use formats that software can parse without guesswork, and the terms in them should come from a shared vocabulary. The whole should carry a licence that says what may be done with it, and enough provenance that a reader can judge its quality. Table 1 lists these requirements against the point at which each commonly fails.

3. Four Reasons the Gap Persists

Four arguments explain the distance between policy and practice better than any appeal to the conservatism of chemists. They are summarised in Figure 1 and taken in turn.

3.1. FAIR was written for datasets, chemistry produces decisions

A genome assembly or a climate record is a dataset in the sense the principles assume: a coherent object, produced once, with an obvious boundary. A medicinal chemistry programme produces a stream of small, heterogeneous observations whose value lies in their relation to one another. A single assay result means little on its own; the series it belongs to means a great deal. There is no natural unit to deposit, no obvious moment at which the data are finished, and the object that would be most useful to others, the whole programme with its failures, is the one nobody is preparing to release.

3.2. The hard part of interoperability is semantic

Chemistry solved syntactic interoperability decades ago. Two groups can exchange structures with confidence, and the early argument for marking up chemical information in a structured way ([Murray-Rust and Rzepa, 1999](#)) has largely been won at the level of the file format. What they cannot exchange reliably is meaning. An inhibition value depends on the enzyme construct, the substrate concentration, the buffer, the preincubation time and the detection method, and two numbers labelled identically may not be comparable. Aggregation across sources therefore requires judgement, which is why curation remains skilled work ([Fourches et al., 2010](#)) and why integration projects such as the one described by [Williams et al. \(2012\)](#) invested so heavily in mapping vocabularies. Making data interoperable in this sense is a scientific task, not a formatting one.

Table 1: FAIR requirements for a chemistry dataset and where each tends to fail in practice

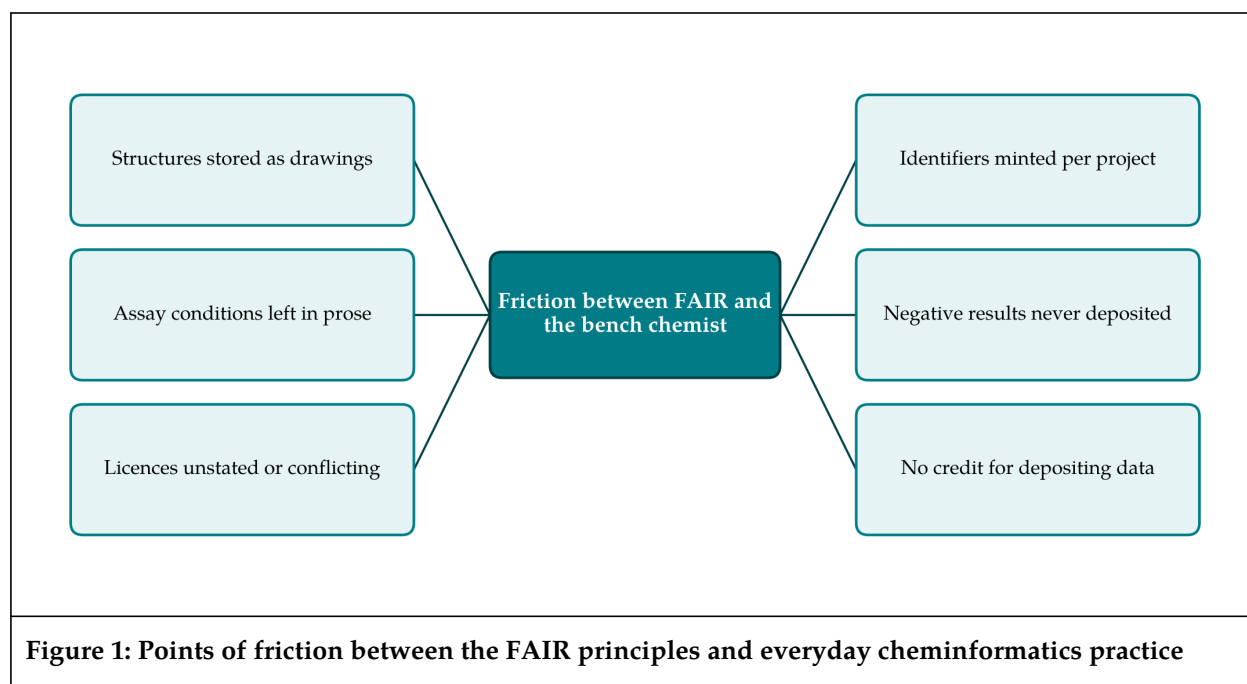
Requirement	Where it breaks in everyday practice	What would ease it
Resolvable compound identifier	Structures recorded as drawings or internal codes	Standard InChI alongside the drawing (Heller et al., 2015)
Machine-readable measurement	Values and units embedded in prose or figures	A tabular deposit required with the manuscript
Shared vocabulary for entities	Assay and endpoint named differently by every group	Ontology terms where they exist (Hastings et al., 2016)
Explicit licence	No licence, or terms that forbid redistribution	A default open licence stated by the journal

Table 1: FAIR requirements for a chemistry dataset and where each tends to fail in practice

Requirement	Where it breaks in everyday practice	What would ease it
Provenance of the record	Curation steps undocumented	A published curation sequence (Bento et al., 2020)
Discoverable deposit	Data in supplementary files, not indexed	Registered repositories and standards (Sansone et al., 2019)

3.3. Openness is least available where data are most valuable

The bioactivity measurements that would most improve public models sit inside companies, and the reasons they stay there are commercial rather than technical. Academic groups face a milder version of the same pressure: data held back until the next paper, or until a patent position is clear. The FAIR principles were careful to say that accessible does not mean open, but in cheminformatics the distinction often resolves into metadata released and measurements withheld, which does little for anyone building models. Precompetitive consortia have shown that some of this can be negotiated, and their limited scope shows how hard the negotiation is.



3.4. The costs and the credit fall on different people

Depositing data properly takes a day or more per paper, and the person who does it is usually a research student with other priorities. The benefit accrues to strangers, later. Nothing in hiring, promotion or grant review currently compensates for that asymmetry, and exhortation does not change an incentive. The reproducibility discussion in the sciences generally ([Baker, 2016](#)) and in modelling specifically ([Walters, 2013](#)) has converged on the same conclusion: practices that cost individuals and benefit the field spread only when institutions make them cheap, mandatory or rewarded.

4. Counter-Arguments

Three objections deserve a hearing. The first is that the infrastructure is already adequate and the problem is simply adoption. There is something in this: a group that deposits standard identifiers, a tabular data file and a licence has met most of what the principles ask, using tools that cost nothing. But adoption is itself the thing to be explained, and attributing it to inertia stops the analysis exactly where it should begin. The more interesting version of the objection is that the remaining effort, small as it looks from

outside, lands on the busiest week of a project.

The second is that mandates will close the gap. Journal and funder requirements have certainly moved practice, and the data availability statement is now routine. What they have produced, though, is often compliance in form: a statement that data are available on reasonable request, or a spreadsheet deposited without a vocabulary or a licence. Mandates raise the floor and do not by themselves produce reusable data.

The third is that large public databases have already solved the problem, since a curator will eventually extract published values into a structured archive. This underestimates what is lost in transit. Extraction recovers the headline numbers and not the conditions, the failed compounds or the assay history, and it places the burden on a small number of curation teams. A discipline that relies on others to make its literature machine-readable has accepted a lossy pipeline as permanent.

5. What Would Actually Help

Three changes would do more than further advocacy. Journals in this field could require a structured data deposit, not a statement, with compound identifiers, measured values, units and conditions in tabular form, and could treat the absence of a licence as an incomplete submission. The cost to authors is modest once it is expected, and the effect on the usable literature would be substantial.

Funders could pay for curation explicitly rather than assuming it as an unfunded obligation, including support for the registries and standards infrastructure that makes deposits discoverable (Sansone *et al.*, 2019). And research groups could adopt a modest internal convention: a standard identifier recorded for every compound at the moment it is made or bought, a fixed vocabulary for assays and endpoints, and curation steps recorded as an ordered list rather than as the word curated. A group that does this finds deposition nearly free at publication, because the work was done as the project ran.

A fourth change is cultural and costs nothing. Supervisors can ask, at the point where a project is designed rather than at the point where it is written up, what will be deposited and in what form. That question is answerable early and almost unanswerable late, because by then the records exist in whatever shape the work happened to leave them. It also makes visible the part of the dataset that is usually lost, namely the compounds that were made and found inactive, which carry information that models need and that no curator can recover from a paper in which they never appeared.

None of this requires new technology. The notations, identifiers, repositories and toolkits exist and are open. What is missing is the alignment between who bears the cost and who receives the benefit.

6. Conclusion

FAIR data practice has not reached routine cheminformatics because the principles assume a kind of research object that this field does not often produce, because the interoperability that matters here is semantic and demands expert judgement, because the most valuable measurements are held for commercial reasons, and because the labour of deposition is unrewarded. Recognising these as structural features rather than as failures of will points to remedies that are within reach: structured deposits required at submission, curation funded as part of research, and conventions adopted inside groups while the work is in progress. Chemistry has the standards it needs. What it lacks is the arrangement of incentives that would make using them the path of least resistance.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Baker, M. (2016). [1,500 scientists lift the lid on reproducibility](#). *Nature*, 533(7604): 452-454.
- Bento, A.P., Hersey, A., Felix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L.J., De Veij, M. and Leach, A.R. (2020). [An open source chemical structure curation pipeline using RDKit](#). *Journal of Cheminformatics*, 12(1): 51.
- Fourches, D., Muratov, E. and Tropsha, A. (2010). [Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research](#). *Journal of Chemical Information and Modeling*, 50(7): 1189-1204.
- Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P. and Steinbeck, C. (2016). [ChEBI in 2016: Improved services and an expanding collection of metabolites](#). *Nucleic Acids Research*, 44(D1): D1214-D1219.
- Heller, S.R., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D. (2015). [InChI, the IUPAC International Chemical Identifier](#). *Journal of Cheminformatics*, 7: 23.
- 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, R., 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.-A., 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. and Schultes, E. (2020). [FAIR principles: Interpretations and implementation considerations](#). *Data Intelligence*, 2(1-2): 10-29.
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J. and Bolton, E.E. (2023). [PubChem 2023 update](#). *Nucleic Acids Research*, 51(D1): D1373-D1380.
- Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Felix, E., Magarinos, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., Gordillo-Maranon, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C.J., Segura-Cabrera, A., Hersey, A. and Leach, A.R. (2019). [ChEMBL: Towards direct deposition of bioassay data](#). *Nucleic Acids Research*, 47(D1): D930-D940.
- Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B. and Wilkinson, M.D. (2017). [Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud](#). *Information Services and Use*, 37(1): 49-56.
- Murray-Rust, P. and Rzepa, H.S. (1999). [Chemical markup, XML, and the worldwide web. 1. Basic principles](#). *Journal of Chemical Information and Computer Sciences*, 39(6): 928-942.
- O Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. and Hutchison, G.R. (2011). [Open Babel: An open chemical toolbox](#). *Journal of Cheminformatics*, 3: 33.
- Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L. and Thurston, M. (2019). [FAIRsharing as a community approach to standards, repositories and policies](#). *Nature Biotechnology*, 37(4): 358-367.
- Walters, W.P. (2013). [Modeling, informatics, and the quest for reproducibility](#). *Journal of Chemical Information and Modeling*, 53(7): 1529-1530.
- Weininger, D. (1988). [SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules](#). *Journal of Chemical Information and Computer Sciences*, 28(1): 31-36.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone,

- S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). [The FAIR Guiding Principles for scientific data management and stewardship](#). *Scientific Data*, 3: 160018.
- Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C. and Mons, B. (2012). [Open PHACTS: Semantic interoperability for drug discovery](#). *Drug Discovery Today*, 17(21-22): 1188-1198.
- Wise, J., de Barron, A.G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., Mellino, G., Harrow, I., Smith, I., Taubert, J., van Bochove, K., Romacker, M., Walgemoed, P., Jimenez, R.C., Winnenburg, R., Plasterer, T., Gupta, V. and Hedley, V. (2019). [Implementation and relevance of FAIR data principles in biopharmaceutical R&D](#). *Drug Discovery Today*, 24(4): 933-938.