



Dataset Publication Standards: Community Lessons from a Multi-Stakeholder Workshop on FAIR Implementation

COLLECTION:
INTERNATIONAL
DATA WEEK

PRACTICE PAPERS

JULIEN GOBEILL 

MATT JEFFRYES 

MELISSA HARRISON 

SCOTT EDMUNDS 

THOMAS LEMBERGER 

MARK HAHNEL 

SLAVA TYKHONOV 

PATRICK RUCH 

**Author affiliations can be found in the back matter of this article*

ubiquity press

ABSTRACT

Ninety-seven percent of biological research identifiers (accession numbers) are published in supplementary files rather than in manuscripts. These files are usually in proprietary formats, such as Excel. This makes them invisible to standard search engines and difficult for machines to process. This ‘long tail of data’ is a critical resource for reproducibility and discovery, yet it remains hidden from researchers and curators alike. Although Findable, Accessible, Interoperable and Reusable (FAIR) standards exist, they have not yet been adopted and/or coordinated.

To address this issue, we organised a multi-stakeholder workshop with international experts from data repositories, publishing houses, text mining, and standards organisations, at SciDataCon 2025 in Brisbane. The solutions presented at the workshop, along with subsequent community dialogue, revealed that, while stakeholders agree on the necessary standards, key questions remain about how to implement them. Three barriers emerged: timing (should FAIRification occur during or after peer review?), governance (who enforces standards?), and incentives (who funds what?).

Our recommendations are as follows: authors should deposit datasets early; publishers should embed FAIR standards in the peer review process; repositories should harmonise metadata to enable cross-platform discovery; funders should require and support FAIR compliance.

This paper argues that implementing FAIR standards requires coordinated action across the research ecosystem with regard to technical standards, governance, incentives, and workflows.

CORRESPONDING AUTHOR:

Julien Gobeill

SIB Text Mining Group, Swiss
Institute of Bioinformatics,
Geneva, Switzerland; BiTeM
Group, Information Sciences,
HES-SO/HEG Geneva,
Switzerland

julien.gobeill@sib.swiss

KEYWORDS:

FAIR data; Supplementary
data; Data repositories;
Metadata standards;
Biocuration; Research data
management

TO CITE THIS ARTICLE:

Gobeill, J., Jeffries, M.,
Harrison, M., Edmunds, S.,
Lemberger, T., Hahnel, M.,
Tykhonov, S. and Ruch, P 2026
Dataset Publication Standards:
Community Lessons from a
Multi-Stakeholder Workshop
on FAIR Implementation.
Data Science Journal, 25:
38, pp. 1–8. DOI: <https://doi.org/10.5334/dsj-2026-038>

Most biological research data never reaches the databases and repositories built to hold it; instead, it is published in supplementary files attached to journal articles, such as spreadsheets, documents, and images (Gobeill et al., 2025; Pop and Salzberg, 2015). This is the ‘long tail’ of research data: a vast collection of small, scattered datasets that, taken together, represent an enormous scientific resource (Ferguson et al., 2014; Heidorn, 2008). A recent analysis of Global Core Biodata Resources (Global Biodata Coalition, 2022) found that 97% of the accession numbers referencing these resources appear only in supplementary files, not in the article text (Gobeill et al., 2025). These files are typically stored as Excel or other proprietary formats, which leaves the data invisible to standard search engines and difficult for machines to read.

The FAIR principles state that data should be Findable, Accessible, Interoperable and Reusable, and offer a clear framework for addressing this issue (Wilkinson et al., 2016). A growing body of work, including community assessments and technical guidelines, has documented where supplementary data falls short against each of these principles and what could be done to improve it (Boeckhout, Zielhuis and Bredenoord, 2018; Pattyn et al., 2017; Rocca-Serra et al., 2023). Yet, a gap persists between knowing what should be done and knowing how to do it at scale. Recommendations accumulate faster than they are implemented, and the barriers are rarely technical alone: they involve workflows, incentives, governance, and the way different tools and platforms fit together.

To explore this gap, we convened a multi-stakeholder workshop, ‘Bridging the FAIR gap’, at SciDataCon 2025, part of International Data Week. Rather than presenting a single solution, the workshop brought together experts from different perspectives: data repositories, publishers, text-mining groups, and standards organisations. The aim was deliberately practical: to test recommendations against the realities faced by the people who would have to implement them. Our examples are taken from the life sciences, which have a long-standing tradition of open literature and biocuration. However, the barriers we have identified are organisational rather than domain-specific, and are likely to apply wherever data is published as a supplement to a narrative.

This paper synthesises what emerged. After describing the workshop format, we summarise the solutions presented, each addressing a different facet of the FAIR challenge. We then draw together the recurring tensions that surfaced in discussion, and close with a set of recommendations aimed at the different actors in the research ecosystem, grounded not in principle but in what the community judged workable.

METHODS

This paper presents an expert-informed perspective, derived from a workshop rather than a formal qualitative study. Entitled ‘Bridging the FAIR gap’, the workshop was held at SciDataCon 2025 (International Data Week, Brisbane, 14 October 2025) and gathered around 20 participants, including representatives from data repository providers, publishers, text-mining groups, and standards organisations. The session comprised a series of short presentations, each outlining a practical solution to part of the problem, followed by an open panel discussion with the audience.

The questions raised during the session and the responses given were recorded in a shared online document. The recommendations presented in this paper were not transcribed verbatim during the discussion; instead, they were reconstructed by the authors based on the discussions and subsequent exchanges. Similarly, the three barriers (timing, governance, and incentives) used to structure the discussion were not derived through a formal coding or consensus procedure, but identified by the authors as recurring themes throughout the session. A draft of this paper, including the recommendations, was subsequently shared with workshop participants who had registered on the shared document for comment and endorsement.

Therefore, these observations should be read as the considered synthesis of an engaged group of practitioners rather than as the output of a systematic empirical study.

BIODIVERSITY PMC: MAKING SUPPLEMENTARY DATA SEARCHABLE

Standard search engines such as MEDLINE and PubMed Central (PMC) only provide a partial view of the literature, because they do not index supplementary file content ([Gobeill et al., 2025](#)). Our service continuously collects all PMC supplementary data files and extracts text from most of them (85%), including about 24 million manuscript figures processed by OCR (Optical Character Recognition), producing a total of 35 million searchable documents. These documents are available via the SIB Literature Services (SIBiLS) collection and are enriched with automatically mapped biomedical entities drawn from a growing set of controlled vocabularies ([Gobeill et al., 2020](#)).

The collection is openly accessible via a web interface, Biodiversity PMC ([Pasche et al., 2023](#)), which also includes Plazi's taxonomic treatments and a subset of Zenodo contents. To assess the value of supplementary files, we searched for mentions of the 52 Global Core Biodata Resources by both name and accession number. Indexing supplementary files yielded a 6% gain in recall when searching by resource name. More strikingly, 97% of accession numbers identified appeared only in supplementary files. This confirms that these resources are mainly used to annotate results tables and datasets rather than the narrative.

However, the same analysis revealed a limitation: 87% of these accession numbers are published in Excel format (mostly xlsx, with 6% in the older, deprecated xls binary format). These formats favour human readability, but severely restrict machine reuse and interoperability. Beyond parsing difficulties (merged cells, multiple sheets), metadata are typically sub-optimal (often absent, or scattered in a separate sheet or in free-text cells).

EUROPE PMC: TEXT-MINING AND DEPOSITING SUPPLEMENTARY DATA

Europe PMC is a life science literature database, bringing together abstracts from PubMed, Agricola, and 34 life science preprint servers, and full-text articles from PMC ([Rosonovski et al., 2024](#)). Through collaboration with the BioStudies database, we have made supplementary material more findable and accessible—supplementary files from PMC automatically receive a BioStudies entry and DOI ([Sarkans et al., 2018](#)).

We have extended this collaboration to integrate with our annotation platform. Text-based supplementary files are processed through our text-mining pipeline to extract and ground entities. The following entity types are extracted: chemicals (ChEBI), diseases (UMLS), experimental methods (Experimental Factors Ontology), genes/proteins (UniProt), Gene Ontology terms, organisms (NCBI taxonomy), accession numbers, and resource names. These annotations are available through our Annotations Application Programming Interface (API) ([Europe PMC, no date](#)), and through a listing on the relevant BioStudies entry.

DOME AND GIGASCIENCE: INTEGRATING MACHINE LEARNING STANDARDS IN PUBLICATION PROCESS

There has been a substantial increase in scientific publications utilising machine learning (ML), and as a publisher of data-intensive computational research GigaScience journal has been at the forefront of dealing with these papers. The 'black box' nature of ML research has made understanding the intelligibility and reproducibility of these studies more challenging than traditional computational research. Being focussed on assessing research based on reproducibility and transparency, to address these challenges, GigaScience has experimented with integrating ML standards and annotation into the publication process, eventually deciding upon the DOME (Data, Optimisation, Model, Evaluation) guidelines ([Walsh et al., 2021](#)) developed by the ELIXIR-ML Focus Group ([Edmunds et al., 2026](#)).

On top of using the DOME standard to help screen which ML publications are suitable to send out for review, the journal also developed a workflow that integrated associated tools into the publishing process. The DOME Data Stewardship Wizard (DOME-DSW) streamlines the process of metadata capture and presents it in a form to share with peer reviewers, and then facilitates the submission of DOME annotations to the DOME Registry ([Attafi et al., 2024](#)). If submissions pass review, the DOME Registry ID is cited in the references of the published paper, providing additional discoverability, transparency, and provenance for the readers. While this process slows the review process down by a few hours, combining the curation and peer review is potentially more efficient than doing these steps separately. A new DOME Copilot tool automates and

greatly speeds up the annotation process through large language model (LLM)-based metadata extraction, removing further barriers in this FAIRification process (Farrell et al., 2026).

SOURCEDATA: CURATING FIGURE SOURCE DATA DURING PUBLISHING

To bridge open-access publication with the deposition and curation of open data, EMBO Press has developed the SourceData framework (Liechti et al., 2017), which makes the source data behind scientific publications openly accessible and citeable. The SourceData workflow is fully integrated into the publication process at EMBO Press: the research data underlying the figures of every paper are curated and systematically deposited in the EMBL-EBI database BioStudies. Accordingly, the Data Availability Section of papers published by EMBO Press includes the accession number of the associated BioStudies record, providing a convenient single point of entry to the source data underlying the published paper. The SourceData curation workflow has also enabled the distribution of a large machine-learning training dataset (Abreu-Vicente et al., 2026).

Curation and quality checks of source data are costly and time-consuming for authors and editorial teams, creating a high barrier to adopting open science practices. To scale up and improve the efficiency and transparency of these steps, EMBO Press is developing an open library of AI editorial skills, multimodal AI quality checks (mmQC), that leverage the multimodal capabilities of AI models to jointly analyse images, captions, and the underlying data. Every skill in mmQC is benchmarked against an appropriate test dataset. This allows us to verify the reliability of AI-assisted quality checks and, importantly, to avoid ambiguities in the underlying editorial policies.

Making AI-executable curation and QC skills openly available will enhance transparency in the editorial process by clearly defining policies. This will allow communities to contribute well-defined standards and checklists. Additionally, incorporating these open, transparent AI skills into authoring tools will enable researchers to verify their manuscripts prior to submission.

GREI: FEDERATED DISCOVERY AND STANDARDISED INFRASTRUCTURE

While domain-specific databases are the ideal home for biological data, much of the ‘long tail’ of research data does not fit neatly into these resources. Consequently, researchers frequently rely on generalist repositories to share their datasets. However, a lack of standardisation across these platforms has historically hindered the Interoperability and Reusability of the deposited data.

To address this infrastructural fragmentation, the US National Institutes of Health (NIH) launched the Generalist Repository Ecosystem Initiative (GREI). Rather than creating a new, monolithic platform, GREI fosters ‘coopetition’ among seven established generalist repositories, including Figshare, Dataverse, Mendeley Data, Vivli, Zenodo, Dryad, and the Open Science Framework.

GREI, and its subsequent repository of best practices, provides a working model for how repositories can practically coordinate to enhance FAIR implementation. GREI’s primary mission is to establish a cohesive set of capabilities, consistent metadata standards, and unified social infrastructure across its participating platforms. By harmonising metadata and implementing federated discovery through DataCite, GREI ensures that a single query can surface relevant datasets regardless of which specific repository hosts them.

GREI focuses on optimising the primary destination where researchers actively deposit long-tail data. It directly answers the community’s call for data repositories to standardise and coordinate rather than develop parallel, isolated solutions. As dataset publishing becomes common in academia, the standards to which dissemination platforms are held have a guidepost in GREI.

SYNTHESIS

Taken together, these initiatives (Table 1) show that tools for making the long tail of data FAIR already exist. However, they operate at different stages and address different facets of the problem. Some act at the source by embedding curation and standards within the publication workflow (DOME/GigaScience and SourceData). Others recover value after publication by mining and indexing supplementary files that are already in the literature (Biodiversity PMC and Europe PMC). One initiative operates at the level of the repositories where the data is ultimately stored, harmonising metadata and enabling discovery across platforms (GREI). No single initiative is designed or able to solve the problem alone. The following discussion examines what they collectively reveal: where progress is being made and where the harder obstacles remain.

INITIATIVE	MAIN FAIR CHALLENGE	INTERVENTION POINT	APPROACH
Biodiversity PMC	Findability	Post-publication	Text-mines and indexes PMC supplementary files, enriched with mapped biomedical entities
Europe PMC	Findability, Accessibility	Post-publication	Deposits supplementary files in BioStudies (entry + DOI) and exposes text-mined entities via an Annotations API
DOME/ GigaScience	Interoperability	Pre-publication (peer review)	Embeds DOME metadata capture and curation into the editorial workflow
SourceData	Interoperability, Reusability	Pre-publication (curation)	Curates figure source data into BioStudies; develops multimodal AI quality checks (mmQC)
GREI	Interoperability, Reusability	Repository level	Harmonises metadata and enables federated discovery across generalist repositories

Table 1 Overview of the solutions presented at the workshop.

DISCUSSION

Taken together, the solutions point to an encouraging conclusion and a harder problem. Findability and Accessibility are largely within reach: text-mining approaches such as Biodiversity PMC and Europe PMC index supplementary file content after publication, and make these files accessible via structured records and APIs. Harder problems remain once data is found and retrieved; data must still be understood and reused. Here, Interoperability and Reusability remain the real bottleneck.

The discussion revealed a general consensus on the desired outcome, but not on the approach. Participants agreed on the important standards: DOME, Croissant, Dublin Core, BioC, and others. The open issue is how to implement them. Three barriers emerged during the discussion.

The first is timing: when should FAIRification happen? Two philosophies sat side by side. Acting before publication, as GigaScience does by embedding DOME curation within the peer-review process, makes data FAIR by design. But it requires new infrastructure, curator time, and changes to editorial workflows. The second approach, taken by text-mining pipelines, involves acting after publication by recovering structure from files already in the literature. This method scales easily and requires nothing of authors, but can only approximate what earlier intervention would have captured cleanly. These two approaches are complementary rather than exclusive: one improves future data, while the other rescues backlog data.

The second barrier concerns governance: who should enforce the standards, and which standards should prevail? Currently, adoption is largely voluntary. There is no single body with the authority to mandate a common metadata schema. The proliferation of standards is part of the problem; without coordination, each new format risks contributing to the very fragmentation it was intended to solve.

The third barrier concerns incentives: who funds the work, and who receives credit for it? FAIRification is not free. Activities such as embedding curators in peer review, maintaining text-mining pipelines, and harmonising repository metadata all require sustained resources. Furthermore, the individuals responsible for preparing high-quality datasets rarely receive recognition for their efforts. Without a funding model and a credit system, good practice depends on goodwill. Initiatives such as APICURON, which tracks and attributes curation activities to give credit to individual contributors, point towards a solution ([Hatos et al., 2021](#)).

Within the cases considered in the workshop, participants identified organisational, governance and incentive barriers as more consequential than the absence of technical capabilities. The necessary tools already exist and, in several cases, are capable of working at scale. What is missing is coordination between the various stakeholders who must use them together: authors, publishers, repositories, funders, and standards organisations. Coordination connects timing, governance, and incentives, and it is where the workshop's recommendations are directed.

RECOMMENDATIONS AND CONCLUSION

The workshop made specific recommendations for each stakeholder, summarised in [Table 2](#) and detailed below.

STAKEHOLDER	TIMING (WHEN TO FAIR-IFY?)	GOVERNANCE (WHICH STANDARDS, WHO ENFORCES?)	INCENTIVES (WHO FUNDS/CREDITS?)
Authors	Deposit datasets at peer-review time, rather than after acceptance	Provide full metadata; publish under an open licence (CC-BY or CC0)	Institutional recognition for data deposition, as already exists for open access
Publishers	Incorporate FAIR standards into the peer review process (e.g. DOME)	Require structured metadata at submission (e.g. MIAPPE); provide curation support	Reduced article processing charges for well-deposited datasets
Repositories		Harmonise metadata via shared vocabularies (e.g. DCAT); expose machine-actionable, AI-ready formats (e.g. Croissant)	Credit curators and contributors (e.g. APICURON)
Funders		Require FAIR compliance as a condition of funding; verify it by linking data management plans to the resulting deposits	Support compliance infrastructure (curation, platforms); track and credit data contributions, including at aggregate (institutional) level

Authors should ideally deposit their datasets at the time of peer review, rather than after acceptance, so that the data can be assessed alongside the manuscript. These deposits should include full metadata and be published under an open licence (CC-BY or CC0) to enable reuse and text mining. These practices can be encouraged by institutions recognising data deposition in the same way that they already reward open-access publication.

Publishers should incorporate FAIR standards into the peer-review process itself, as GigaScience does with DOME. This involves requiring structured metadata at submission and providing curation support, so that authors are not left to manage this task alone. Community reporting standards such as MIAPPE (Minimum Information About a Plant Phenotyping Experiment) can provide guidance on the expected metadata for a given data type. Recent developments demonstrate that this does not have to be labour-intensive; a DOME co-pilot now assists authors and editors in meeting the standard, thereby reducing friction and accelerating compliance (Farrell et al., 2026). Publishers can incentivise good practice further by offering reduced article processing charges for submissions accompanied by well-structured deposited datasets.

Data repositories should harmonise their metadata standards by adopting a common vocabulary, such as DCAT (Data Catalog Vocabulary) or an equivalent such as DATS (Data Tag Suite). They should also implement federated discovery, following the GREI model, so that a single query can reach across platforms. They should also make their metadata available in machine-actionable, AI-ready formats such as Croissant to enable automated discovery and quality checks. Developing search and AI evaluation benchmarks, such as those developed within the context of the BioCaddie project (Cohen et al., 2017), can help benchmark the best search technologies. Repositories can also credit the curators and contributors who enrich their records, for instance, through systems such as APICURON.

Funders hold the strongest leverage. They should require FAIR compliance as a condition of funding, supporting the infrastructure that enables compliance, including curation services and repository platforms. Equally importantly, they should track and credit data contributions, recognising the value of producing high-quality datasets as much as publishing a paper. Compliance can be verified in practice by requiring data management plans to be updated with links to the resulting deposits, and monitored at aggregate levels, such as across funded institutions.

Two complementary cross-cutting principles emerged. The first relates to services: rather than creating a single, all-encompassing platform, developers should specialise in their area of expertise, be it long-term archiving, discovery or curation. The second principle relates to interfaces. As services are specialised, their metadata and access points must be harmonised through shared vocabularies and machine-actionable formats. This enables independent tools to interoperate. Specialisation at the level of function and coordination at the level of exchange work together rather than against each other.

This raises the question that the workshop returned to most often: if coordination is the real bottleneck, who coordinates? No single organisation can mandate standards across publishers, repositories, and funders. However, venues to convene this work already exist. The Research Data Alliance provides a neutral, cross-domain forum, and CODATA's Cross-Domain Interoperability Framework (CDIF) offers a technical basis for aligning metadata standards.

Table 2 Recommendations mapped to stakeholders and the three implementation barriers.

Other bodies coordinate FAIR efforts across disciplines: the GO FAIR initiative on community implementation of the principles, and infrastructures such as the World Data System (WDS) and the European Open Science Cloud (EOSC) on federating trusted data services. Within the life sciences specifically, organisations such as ELIXIR and the Global Biodata Coalition's GCBR Forum are well placed to coordinate resources and set community expectations. Rather than establishing a new authority, these existing forums could agree on minimum metadata requirements, align formats, and share the cost of implementation.

These recommendations are not all equally important. At the top is coordination across the research data lifecycle, since this allows the others to function together. Next comes the early implementation of FAIR principles during publication and the use of harmonised metadata and machine-actionable formats, which allow independent tools to interoperate. Supporting all of them are incentives and credit for those who prepare data, as well as sustained funding for curation.

Ultimately, the FAIR gap will not be closed by better technology alone. As the workshop made clear, the solutions already exist, and several are operating on a large scale. However, bridging the FAIR gap is more of a collective challenge than a technical one.

ACKNOWLEDGEMENTS

We thank the organisers of SciDataCon 2025/International Data Week, and all participants of the 'Bridging the FAIR gap' workshop.

FUNDING INFORMATION

This work was supported by the ELIXIR Data Platform.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHORS CONTRIBUTIONS

All authors presented solutions at the workshop and described their respective contributions in this paper. All authors jointly wrote and edited the Discussion and Recommendations sections. All authors read and approved the final manuscript.

AUTHOR AFFILIATIONS

Julien Gobeill  orcid.org/0000-0001-9809-7741

SIB Text Mining Group, Swiss Institute of Bioinformatics, Geneva, Switzerland; BiTeM Group, Information Sciences, HES-SO/HEG Geneva, Switzerland

Matt Jeffries  orcid.org/0000-0001-9868-6271

Literature Services, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom

Melissa Harrison  orcid.org/0000-0003-3523-4408

Literature Services, European Molecular Biology Laboratory, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom

Scott Edmunds  orcid.org/0000-0001-6444-1436

Pensoft, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria

Thomas Lemberger  orcid.org/0000-0002-2499-4025

EMBO, Heidelberg, Germany

Mark Hahnel  orcid.org/0000-0003-4741-0309

Digital Science (United Kingdom), London, United Kingdom

Slava Tykhonov  orcid.org/0000-0001-9447-9830

CODATA, Paris, France

Patrick Ruch  orcid.org/0000-0002-3374-2962

SIB Text Mining Group, Swiss Institute of Bioinformatics, Geneva, Switzerland; BiTeM Group, Information Sciences, HES-SO/HEG Geneva, Switzerland

- Abreu-Vicente, J. et al.** (2026) 'Integrating curation into scientific publishing to train AI models', *Bioinformatics*, 42(1), p. btaf685. Available at: <https://doi.org/10.1093/bioinformatics/btaf685>
- Attafi, O.A. et al.** (2024) 'DOME Registry: implementing community-wide recommendations for reporting supervised machine learning in biology', *GigaScience*, 13, p. giae094. Available at: <https://doi.org/10.1093/gigascience/giae094>
- Boeckhout, M., Zielhuis, G.A. and Bredenoord, A.L.** (2018) 'The FAIR guiding principles for data stewardship: fair enough?', *European Journal of Human Genetics*, 26(7), pp. 931–936. Available at: <https://doi.org/10.1038/s41431-018-0160-0>
- Cohen, T. et al.** (2017) 'A publicly available benchmark for biomedical dataset retrieval: the reference standard for the 2016 bioCADDIE dataset retrieval challenge', *Database*, 2017, p. bax061. Available at: <https://doi.org/10.1093/database/bax061>
- Edmunds, S.C. et al.** (2026) 'Integrating machine learning standards in disseminating machine learning research', *Data Science Journal*, 25(1), p. 1. Available at: <https://doi.org/10.5334/dsj-2026-001>
- Europe PMC.** (no date) Annotations API. Available at: <https://europepmc.org/AnnotationsApi> (Accessed: 31 July 2026).
- Farrell, G. et al.** (2026) 'DOME Copilot: A resource to automate transparent reporting of artificial intelligence methods', *bioRxiv* [Preprint]. Available at: <https://doi.org/10.64898/2026.04.16.718888>
- Ferguson, A.R. et al.** (2014) 'Big data from small data: data-sharing in the "long tail" of neuroscience', *Nature Neuroscience*, 17(11), pp. 1442–1447. Available at: <https://doi.org/10.1038/nn.3838>
- Global Biodata Coalition.** (2022) Global core biodata resources: concept and selection process. Zenodo. Available at: <https://doi.org/10.5281/zenodo.7468719>
- Gobeill, J. et al.** (2020) 'SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts', *Nucleic Acids Research*, 48(W1), pp. W12–W16. Available at: <https://doi.org/10.1093/nar/gkaa328>
- Gobeill, J. et al.** (2025) 'Unlocking the potential of PubMed Central supplementary data files', *Bioinformatics Advances*, 5(1), p. vbaf155. Available at: <https://doi.org/10.1093/bioadv/vbaf155>
- Hatos, A. et al.** (2021) 'APICURON: a database to credit and acknowledge the work of biocurators', *Database*, 2021, p. baab019. Available at: <https://doi.org/10.1093/database/baab019>
- Heidorn, P.B.** (2008) 'Shedding light on the dark data in the long tail of science', *Library Trends*, 57(2), pp. 280–299. Available at: <https://doi.org/10.1353/lib.0.0036>
- Liechti, R. et al.** (2017) 'SourceData: a semantic platform for curating and searching figures', *Nature Methods*, 14(11), pp. 1021–1022. Available at: <https://doi.org/10.1038/nmeth.4471>
- Pasche, E. et al.** (2023) 'From SIBiLS to Biodiversity PMC: foundations for the One Health Library', *Biodiversity Information Science and Standards*, 7, p. 467. Available at: <https://doi.org/10.3897/biss.7.111660>
- Pattyn, F. et al.** (2017) 'Assessment of FAIRness of open data sources in life sciences', Proceedings of the 10th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4 LS 2017). Rome, Italy, 4–7 December. CEUR Workshop Proceedings, vol. 2042. Available at: <https://doi.org/10.6084/m9.figshare.6994094>
- Pop, M. and Salzberg, S.L.** (2015) 'Use and mis-use of supplementary material in science publications', *BMC Bioinformatics*, 16(1), p. 237. Available at: <https://doi.org/10.1186/s12859-015-0668-z>
- Rocca-Serra, P. et al.** (2023) 'The FAIR Cookbook – the essential resource for and by FAIR doers', *Scientific Data*, 10(1), p. 292. Available at: <https://doi.org/10.1038/s41597-023-02166-3>
- Rosonovski, S. et al.** (2024) 'Europe PMC in 2023', *Nucleic Acids Research*, 52(D1), pp. D1668–D1676. Available at: <https://doi.org/10.1093/nar/gkad1085>
- Sarkans, U. et al.** (2018) 'The BioStudies database – one stop shop for all data supporting a life sciences study', *Nucleic Acids Research*, 46(D1), pp. D1266–D1270. Available at: <https://doi.org/10.1093/nar/gkx965>
- Walsh, I. et al.** (2021) 'DOME: recommendations for supervised machine learning validation in biology', *Nature Methods*, 18(10), pp. 1122–1127. Available at: <https://doi.org/10.1038/s41592-021-01205-4>
- Wilkinson, M.D. et al.** (2016) 'The FAIR guiding principles for scientific data management and stewardship', *Scientific Data*, 3(1), pp. 1–9. Available at: <https://doi.org/10.1038/sdata.2016.18>

TO CITE THIS ARTICLE:

Gobeill, J., Jeffries, M., Harrison, M., Edmunds, S., Lemberger, T., Hahnel, M., Tykhonov, S. and Ruch, P 2026 Dataset Publication Standards: Community Lessons from a Multi-Stakeholder Workshop on FAIR Implementation. *Data Science Journal*, 25: 38, pp. 1–8. DOI: <https://doi.org/10.5334/dsj-2026-038>

Submitted: 31 July 2026

Accepted: 05 October 2026

Published: 19 October 2026

COPYRIGHT:

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <https://creativecommons.org/licenses/by/4.0/>.

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.