FAIR Data: What It Is and How We Can Support Its Principles

MODERATOR:
Jenna Daenzer
Peer Review Coordinator
GENETICS & G3

SPEAKERS:
Taunton Paine
National Institutes of Health

Gabriele Hayden
University of Oregon Libraries

Karen Yook
WormBase and microPublication Biology

Matt Giampoala
American Geophysical Union

REPORTER:
Tony Alves
HighWire Press

The FAIR (Findable, Accessible, Interoperable, and Reusable) Data principles have become essential guidelines for modern data management and scholarly publishing. These principles are designed to enhance the quality and impact of research by ensuring data is managed and shared in a way that maximizes its utility and accessibility. The CSE 2024 Annual Meeting session on FAIR Data brought together experts from various fields to discuss the challenges and opportunities associated with implementing these principles. Each speaker provided unique insights into the practicalities and benefits of FAIR Data.

NIH Data Management and Sharing Policy

The National Institutes of Health (NIH) Data Management and Sharing Policy underscores the importance of data sharing to advance rigorous and reproducible research. Taunton Paine, Director in the Scientific Data Sharing Policy Division of the Office of Science Policy at the NIH, emphasized that sharing data enables the validation of research results, makes high-value datasets accessible, and accelerates future research directions. The policy also aims to increase opportunities for citation and collaboration, promoting public trust and transparency in research. The NIH has a long history of encouraging data sharing, with policies dating back to 2003 for data sharing plans and more specific policies for genomic data and clinical trials introduced in subsequent years. Despite these efforts, data accessibility remains a challenge, with studies showing low rates of data availability and sharing across various disciplines (Table).

The policy requires that all NIH-funded research include a data management and sharing plan (DMSP). This plan must outline how data will be managed and shared and emphasizes the use of established repositories to ensure data quality and accessibility. The DMSP is assessed by NIH staff and can be updated to reflect changes during the research project. The NIH also provides resources to help researchers comply with these requirements, including guidance on selecting appropriate repositories and protecting participant privacy. The policy acknowledges that not all data generated during research will be suitable for sharing and provides criteria for determining which data should be shared. Factors such as informed consent, privacy concerns, and legal or ethical restrictions are considered valid reasons for limiting data sharing.

https://doi.org/10.36591/SE-4703-04

Role of Data Management Librarians

Gabriele Hayden, Research Data Management and Reproducibility Librarian at the University of Oregon Libraries, highlighted the crucial role of data management librarians in supporting FAIR Data principles. Librarians offer regular workshops on data management tools such as R, Python, and GitHub, and provide consultations on programming, statistical methods, and data lifecycle management. They assist researchers with writing data management plans, identifying appropriate repositories, and developing metadata to ensure data is findable and usable. These services are essential for helping researchers navigate the complexities of data management and sharing, especially in disciplines with diverse data norms and structures.

However, data management librarians face significant challenges. They cannot enforce data-sharing policies or ensure compliance across hundreds of disciplines. This enforcement gap often results in low rates of data sharing, even when authors have committed to sharing their data. A study published in PLOS ONE found that fewer than 21% of authors who included data-sharing plans in their articles provided links to repositories storing the data. The session underscored the need for enforceable data-sharing policies to ensure that the benefits of data sharing, such as increased citations and greater research impact, are realized.

Publishing FAIR Data

Karen Yook from WormBase and microPublication Biology discussed the importance of curating published data to ensure it meets FAIR principles. Curation involves both entity identification and fact extraction, which ensures data is correctly annotated and linked to relevant metadata. This process helps make data findable, accessible, interoperable, and reusable. Yook highlighted the challenges of ensuring

CONTINUED

Table. List of studies showing low rates of data availability and sharing across various disciplines.

Author	Finding	Year
Tedersoo et al.[1]	• Evaluated data availability in 875 papers across nine disciplines published 2000–2019 • Data requests successful 39.4% on average; ranged 27.9%–56.1% per field, 19.4% of requests declined after repeated follow-up	2021
Errington et al.[2]	• Attempted to repeat 193 experiments from 53 high-impact cancer biology papers; able to obtain data for 32% of experiments	2021
Gabelica et al.[3]	• Requested data from 1,792 papers published January 2019 with data availability statements; 6.8% of authors provided the requested data	2022
Narang et al.[4]	• Evaluated data availability for 213 NIH*-funded pediatric clinical trial publications • Individual-level participant data available for 3.3% of publications	2023
Hussey[5]	• Requested data from 52 papers employing Implicit Relational Assessment Procedure over previous 5 years; 26.9% of authors provided the requested data	2023 (preprint)
Ioannidis et al.[6]	• Reviewed 5,340 papers on COVID-19 in 9 infectious disease journals in 2019 and 2021; 9% of papers made data available (rates by journal ranged 5%–25%)	2023
Hamilton et al.[7]	• Reviewed 105 meta-analyses of data sharing in 2,121,580 papers published 2016–2021 • Found declared and actual public data availability of 8% and 2%, respectively • Success in privately obtaining data from authors ranged between 0% and 37%	2023

*NIH, National Institutes of Health.

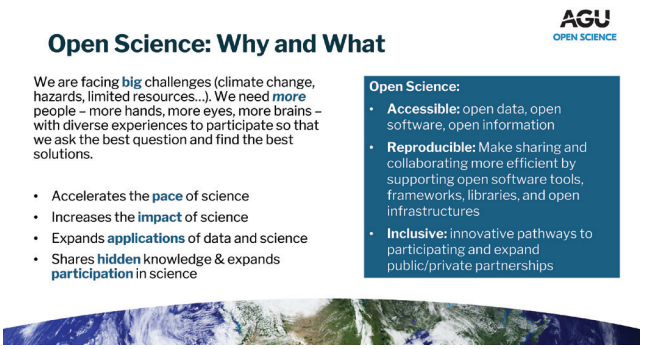
data quality and completeness, noting that peer review alone is insufficient to guarantee good data. She advocated for databases and journals to work together to validate data, append FAIR metadata, and maintain data accessibility long after publication.

An example of effective data curation was provided through the work done by microPublication Biology. The journal publishes single experiment results with DOIs and ensures these are discoverable on platforms like PubMed and Google Scholar. Data from these publications are curated directly into community databases such as WormBase, SGD, and FlyBase, which enhances their visibility and utility. The curation process involves verifying entities, annotating facts, and correcting any incorrect or missing information, which is vital for maintaining the integrity and usability of published data.

AGU's Open Science and Data Strategy

Matthew Giampoala, Vice President for Publications for the American Geophysical Union (AGU), discussed the AGU's commitment to open science and FAIR Data. AGU's strategy includes requiring data and software sharing in published outputs, integrating data into peer review, and connecting articles to curated repositories. The AGU has been actively promoting open science through initiatives like the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS) and the Enabling FAIR Data project. These efforts aim to ensure that all data supporting publications are preserved in trusted repositories and properly cited.

AGU's Open Science strategy is designed to accelerate the pace of science, increase its impact, and expand applications of data and science (Figure). The organization emphasizes the importance of making research accessible, reproducible, and inclusive. By supporting open data, open software, and open information, AGU aims to foster collaboration and innovation across the scientific community. The strategy includes establishing methodologies for measuring the adoption of FAIR Data policies and increasing the usage and citation of data sets.



CONTINUED

Conclusion

The CSE 2024 Annual Meeting session on FAIR Data highlighted the critical importance of adopting FAIR principles in scholarly publishing. Despite the challenges in implementation and enforcement, the benefits of making data findable, accessible, interoperable, and reusable are clear. By promoting rigorous data management practices and fostering collaboration between researchers, librarians, publishers, and repositories, the scientific community can enhance the transparency, reproducibility, and impact of research. The ongoing efforts by organizations like the NIH, AGU, and WormBase serve as exemplary models for how to integrate FAIR principles into the fabric of scientific research and publishing.

References and Links

1. Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen Ä, Pedaste M, Raju M, Astapova A, Lukner H, et al. Data sharing practices and data availability upon request differ across scientific disciplines. Sci Data. 2021;8:192. https://doi.org/10.1038/s41597-021-00981-0.
2. Errington TM, Denis A, Perfito N, Iorns E, Nosek BA. Reproducibility in cancer biology: challenges for assessing replicability in preclinical cancer biology. eLife. 2021;10:e67995. https://doi.org/10.7554/eLife.67995.
3. Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. J Clin Epidemiol. 2022;150:33-41. https://doi.org/10.1016/j.jclinepi.2022.05.019.
4. Narang C, Ouvina M, Rees CA, Bourgeois FT. Data sharing for pediatric clinical trials funded by the US National Institutes of Health. JAMA Netw Open. 2023;6:e2325342. https://doi.org/10.1001/jamanetworkopen.2023.25342.
5. Hussey I. Data is not available upon request. PsyArXiv. 2023. https://doi.org/10.31234/osf.io/jbu9r.
6. Zavalis EA, Contopoulos-Ioannidis DG, Ioannidis JPA. Transparency in infectious disease research: meta-research survey of specialty journals. J Infect Dis. 2023;228:227–234. https://doi.org/10.1093/infdis/jiad130.
7. Hamilton DG, Hong K, Fraser H, Rowhani-Farid A, Fidler F, Page MJ. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. BMJ. 2023;382:e075767. https://doi.org/10.1136/bmj-2023-075767.