https://doi.org/10.1038/s41591-024-02879-x

Understanding the provenance and quality of methods is essential for responsible reuse of FAIR data

Check for updates

Data availability and reusability are critical to open research. The FAIR principles provide a minimal set of guiding principles for making data findable, accessible, interoperable and reusable[1]. Open data are not necessarily FAIR, and FAIR data are not necessarily open. Since their publication in 2016[1], the FAIR principles have accelerated the open data movement by inspiring activities and infrastructure development[2–4]. The principles are also being adapted for other research outputs, such as software[5]. As funders increasingly demand FAIR practices and researchers work to implement the FAIR principles, additional actions should be taken for responsible data use and reuse.

The FAIR principles indirectly outline the responsibilities of the data depositor by identifying dataset properties that facilitate reuse. However, the data provenance and the quality of the methods and procedures used to generate and validate data are often overlooked. This information is essential for responsible data reuse. FAIR data evaluations typically focus on the question: “Can I reuse these data?” We argue that it is time to also ask, “Should I re-use these data?” and “How should I reuse these data responsibly?”. These questions allocate responsibilities between the data depositor and the prospective data user. This shift should include several elements.

Although FAIR data are necessary for reusability, this does not guarantee scientific rigor, trustworthiness or research quality. In addition to determining whether data are FAIR, prospective data users should consider whether the data are appropriate to answer their research question. Furthermore, data users must consider the rigor and quality of the study design and procedures used to generate the data, and whether reuse is likely to yield trustworthy results. Sharing FAIR data may encourage others to uncritically reuse those data. Reuse of data from poorly designed experiments may yield valuable insights if users address design limitations when analyzing and interpreting the data[6,7]. However, uncritical reuse of problematic data to generate new, untrustworthy findings may be harmful. We believe that comprehensive FAIR data-sharing evaluations (such as EOSC call HORIZON-INFRA-2022-EOSC-01) should consider the quality of the data provenance.

We propose that creating guiding principles that outline the responsibilities of the data user will facilitate and enhance responsible data reuse. These responsibilities might include performing a systematic search to identify datasets relevant to the new research question, assessing and describing the scientific rigor of the methods and procedures used to generate or collect those data, and determining whether identified datasets are appropriate to answer the research question. If the underlying study has a high risk of bias, users should develop an analysis plan to address this bias or avoid using that dataset. Researchers who aim to combine datasets should determine whether any datasets have properties that preclude combination. Pre-registration of secondary analyses is important, as data reuse might make it possible to test many exploratory hypotheses, at low cost, and then publish only those supporting a particular view or reaching a particular evidence threshold. Additionally, users could share products resulting from studies that reused data, including protocols, modified data, code, software or tools. Further discussion is needed to clearly define guiding principles.

Responsible reuse of data requires knowing how those data were generated. However, progress on open and reusable methods and procedures lags far behind progress toward open data. Sharing detailed methods descriptions may facilitate a broader spectrum of data reuse, including for purposes not anticipated by the data depositors. FAIR highlights the importance of metadata in helping potential data users to understand the dataset[1], and several groups have set domain-specific metadata standards. The term ‘metadata’, however, is poorly understood by many researchers. Furthermore, the importance of metadata is often disregarded, as sharing of high-quality metadata and data are typically not incentivized or rewarded. Data-sharing requirements are often ‘unfunded mandates’, introduced without adequate training. Many depositors provide little metadata, or simply cite a paper describing the study. This is often inadequate, as publications regularly lack essential methodological details[8].

The research community can take several steps to solve these problems. When using the technical term ‘metadata’, researchers should state that metadata include detailed methods. Data management plans should include sharing of high-quality information about methods. Method descriptions contextualizing datasets should include detailed information about the study aim, study design, methods used, any additional measures taken to reduce the risk of bias, and study limitations. Data depositors should also share guidance for responsible dataset reuse. Research assessment systems must reward and incentivize sharing of methods, data and code as separate research outputs. The academic assessment system primarily values papers, and so researchers who share methods, data and code are doing more work without additional reward. This must change.

Data repositories can contribute by providing fields enabling data depositors to link detailed methods shared in methods repositories. This could include pre-registrations, study design protocols, reusable step-by-step protocols and data validation or analysis plans. Many researchers use generalist repositories, which allow unstructured depositing of data, methods and other materials. Further research is needed to determine whether the presence of structured fields in methods repositories improve reporting. Generalist repositories should have machine-readable systems for determining what materials (such as methods, data and code) an entry contains.

Methods are crucial to scientific advancement; they are not simply a tool to contextualize datasets. If properly shared, methods could be more widely reused than data. Open and reusable methods should be shared as separate, essential research products. A vibrant open methods community is needed to champion this, such as exists for open data and open code.

FAIR data should increase the opportunities for secondary analysis. Previously, these analyses have been conducted by, or in close collaboration with, the researchers who collected the data, or involved large, well-documented, publicly available datasets, such as population studies or government registries. FAIR data sharing can further expand the number and types of available datasets, while reducing the need for collaboration between the data depositor and the data user or re-user, but this will require changes to data deposition and reuse strategies.

We encourage those with relevant expertise who are interested in contributing to principles for responsible data reuse to contact us.

Tracey L. Weissgerber[1]✉, **Małgorzata Anna Gazda**[2], **Gustav Nilsonne**[1,3,4], **Gerben ter Riet**[5,6], **Kelly D. Cobey**[7], **Julia Prieß-Buchheit**[8], **Jorge Noro**[9], **Robert Schulz**[1], **Joeri K. Tijdink**[10], **Evgeny Bobrov**[1], **Alexandra Bannach-Brown**[1], **Delwen L. Franzen**[1], **Ugo Moschini**[11], **Florian Naudet**[12,13], **Ulrich Mansmann**[14], **Maia Salholz-Hillel**[1], **Anita Bandrowski**[15,16] & **Malcolm R. Macleod**[17]

[1]QUEST Center for Responsible Research, Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany. [2]Department of Biological Sciences, University de Montréal, Montreal, Quebec, Canada. [3]Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden. [4]Swedish National Data Service, University of Gothenburg, Gothenburg, Sweden. [5]Center of Expertise Urban Vitality, Faculty of Health, Amsterdam University of Applied Sciences, Amsterdam, the Netherlands. [6]Department of Cardiology, Amsterdam University Medical Center, Amsterdam, the Netherlands. [7]Meta-Research and Open Science Program, University of Ottawa Heart Institute, Ottawa, Ontario, Canada. [8]Institut für Pädagogik, Kiel University, Kiel, Germany. [9]Institute for Interdisciplinary Research, Center for Business and Economics Research (CeBER), University of Coimbra, Coimbra, Portugal. [10]AmsterdamUMC, location VUmc, Department of Ethics, Law and Humanities, Amsterdam, the Netherlands. [11]Data Analysis Office, Istituto Italiano di Tecnologia, Genoa, Italy. [12]University of Rennes, CHU Rennes, Inserm, Irset (Institut de recherche en santé, environnement et travail)–UMR_S 1085, CIC 1414 (Center of Clinical Investigation of Rennes), Rennes, France. [13]Institut Universitaire de France (IUF), Paris, France. [14]Department of Medical Information Sciences, Biometry, and Epidemiology, Medical Faculty, Ludwig-Maximilians Universität München, Munich, Germany. [15]Department of Neuroscience, University of California, San Diego, San Diego, CA, USA. [16]BIH Visiting Professor (funded by Stiftung Charité), Berlin Institute of Health at Charité–Universitätsmedizin Berlin, Berlin, Germany. [17]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK.

✉e-mail: tracey.weissgerber@bih-charite.de

Published online: 21 March 2024

References

1. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
2. Gaiarin, S. P. *Zenodo* https://doi.org/10.5281/zenodo.4486280 (2020).
3. Devaraju, A. & Huber, R. *Patterns* **2**, 100370 (2021).
4. Patel, B. & Soundarajan, S. *F1000Res.* **11**, 836 (2022).
5. Chue Hong, N. P. et al. *Zenodo* https://doi.org/10.15497/RDA00068 (2022).
6. Nielson, J. L. et al. *Nat. Commun.* **6**, 8581 (2015).
7. Torres-Espín, A. et al. *eLife* **10**, e68015 (2021).
8. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. *eLife* **10**, e67995 (2021).

Competing interests

T.L.W. leads PRO-MaP (Promoting Reusable and Open Methods and Protocols), which aims to improve the quality and reporting of methods and protocols in the life sciences. A.B. is founder and CEO of SciCrunch Inc., a company devoted to improving the scientific literature. U. Mansmann is coordinating the doctoral network MSCA-DN SHARE-CTD (HORIZON-MSCA-2022-DN-01 101120360), funded by the EU, and is a member of the national working group on Data Sharing of, and is supported by, the German Medical Informatics Initiative (BMBF 01ZZ23048, 01ZZ2316). F.N. received funding from the French National Research Agency (ANR-17-CE36-0010), the French Ministry of Health and the French Ministry of Research; is a work package leader in the OSIRIS project (Open Science to Increase Reproducibility in Science), which has received funding from the EU's Horizon Europe research and innovation program under grant agreement no. 101094725; and is a work package leader for the doctoral network MSCA-DN SHARE-CTD (HORIZON-MSCA-2022-DN-01 101120360), funded by the EU. K.D.C. is co-chair of DORA (Declaration of Research Assessment).