Title: How to enhance data FAIRness

Authors: Zegni Triki[1*] and Redouan Bshary[2]

Affiliation:

[1] Department of Zoology, Stockholm University, Svante Arrheniusväg 18 B, Stockholm, Sweden

[2] Institute of Biology, University of Neuchâtel, Emile-Argand 11, 2000 Neuchâtel, Switzerland

***Correspondence to:** Zegni Triki, email: zegni.triki@gmail.com

Abstract

In recent years, we witnessed an increasing number of funding agencies, scientific journals and scientists agreeing that society and science benefit from open access to research data. Benefits derive mainly from increased access to knowledge for all and improved transparency in academia. However, despite the advances in open science and open data, three significant aspects still need considerable policing: data quality, the accompanying summaries with basic information of the data files (i.e., metadata), and codes used to generate the research outcomes. Only by having these three components together, we can achieve efficient data sharing and reuse, and hence higher transparency. Here, we propose two complementary approaches that potentially can help with shared data quality: (i) data file(s) sharing should be guided step-by-step in public archives with mandatory metadata, and (ii) journals creating assistant data editor positions at editorial boards with a leading role in data quality and computational reproducibility. Forty-four editors-in-chief in the field of behaviour, ecology and evolution shared their opinion with us regarding these two approaches. Although most of the views were divided, the majority estimated that their current editorial board members do not have the necessary skills to assess the quality of shared data. Since data is the core of research studies, we should consider not only data presence but also quality as a requirement for publication.

Main text

With the open data movement, the practice of data sharing is expanding among biologists. The issue, as it stands, is the overall insufficient quality of the archive files [1,2], in terms of Findability, Accessibility, Interoperability, and Reusability, or the FAIR data principles [3]. It might be intuitive to think that sharing good quality data depends on the scientist's benevolence in contributing to public goods. Yet, recent research found that even highly cooperative individuals fail to archive "better" quality data (Green et al., *unpublished*). Thus, it appears that one major problem is a lack of scientists training in data management and data archiving [4,5]. After all, the archiving of reusable data should not depend on scientists' benevolence. Hence, the question is how to ensure high quality FAIR data archiving without constantly relying on the authors. Therefore, we propose implementing two complementary measures that we believe will strongly improve the quality of shared data and increase reuse. The first measure uses computer macroinstructions that assist authors in archiving their data following a set of step-by-step instructions. Such policy can provide a simple, automatised and standardised quality check. For example, once the author uploads their data file, the macro can create a list with the dataset's column headings while providing information on the data type in each column (e.g., numerical or categorical). Most importantly, the macro will add a task of requesting mandatory field entries for each column heading in turn. This will ensure minimum metadata presence for every data file uploaded to public archives.

Alone, computer macroinstructions will have limitations in assessing a dataset's completeness. No computer program can determine data quality at this stage, yet a human can. At this stage comes our second proposed measure: journals could create a dedicated assistant editor position to check data quality and FAIRness of submitted papers. The data assistant editor could further review the research outcomes of a given paper by reusing authors' data and code (computational reproducibility) to reproduce their findings if this were esteemed necessary by referees. An expert in the journal's discipline should fulfil the position, as this will facilitate the task of the data assistant editor in evaluating the quality and FAIRness of the data, as well as the validity and relevance of the statistical approach. We think that such a task will be beyond a statistician’s skills. Only a biologist will appreciate and understand potential data collection constraints that prevent “perfect” datasets in our field. Subsequently, the authors will deliberately invest in data sharing to avoid publication delays.

To get feedback on to what extent these two approaches can be a promising solution to improve the quality and FAIRness of shared data, we asked 160 editors-in-chief in the field of behaviour, ecology and evolution for their opinion. We sent out the survey, to the 160 editors, on three occasions between 28th April and 25th October 2021. Forty-four editors filled in the survey anonymously (see Supplementary Material for a copy of the survey and the shared data for detailed replies). Thirty-four among the 44 were editors of scientific society journals.

Overall, there was no general agreement on who is responsible for the quality check of shared data: 32% of the participants suggested referees as the ones responsible for such task, while 25% suggested the editorial staff, 18% the public data repositories, and 14% suggested the authors. Despite this disagreement, up to 66% of the editors considered a dedicated data editor with the help of macroinstructions in data archiving as a useful measure to improve the quality and FAIRness of shared data.

Regarding implementing macroinstructions in data sharing, 57% of editors identified public repositories as key responsible, while 27% said it is the publishers. Fifty-two per cent of the editors (strongly) agreed that macroinstructions could improve the quality and FAIRness of shared data, while 41% were rather neutral, with 7% disagreeing.

From the editors' replies, the main potential issues of macroinstructions would be complicating the submission procedure (70%), the unwillingness of data repositories to implement macroinstructions (54%), and that such measure will not appeal to authors (45%) (multiple choices were possible).

Sixty-six per cent of the editors viewed the addition of a dedicated data editor to their editorial board as an asset, compared to 16% disagreeing. It became clear that such opinions stem from the trust of the editors in their editorial team skills when it comes to dealing with data checks. When asked whether their editorial board can assess the quality of shared data, 57% (strongly) disagreed, while 29% (strongly) agreed. Although 61% of editors thought that a data editor should check the reuse of shared data, such as completeness of data files, accompanying metadata and codes, they agreed that this could be a way too labouring task if one person checks all manuscripts. The main potential issues with the implementation of a dedicated data editor would be the additional costs (93.2%), a delay in the speed of publications (61.4%), the low attractiveness of assuming such a role (50%) and the lack of appeal from the authors (34.1%) (multiple choices were possible). Nevertheless, in a scenario where the editorial board does have a data editor, we asked the participants when the data editor should intervene. Again, the opinions were quite divided, with 41% proposing that the assessment should take place before the manuscript is sent out to review, 32% advocated assessment in parallel with reviews, while 9% chose the moment when the manuscript is pending acceptance would be best.

Despite the complexity of the current issue of shared data quality and FAIRness and the diverse opinions we received from the editors, the most frequent conclusion was that both macroinstructions and data editor would be useful (36%), while 20% preferred coupling macroinstructions with data quality check by the reviewers. On the other hand, only 9% of editors thought the current system was satisfying.

Trying to imagine the perspective of our colleagues, we are optimistic that a data editor will be an asset for editorial boards. If there is an issue with data quality/reuse, it is tremendously advantageous for the authors and the editorial staff to find out before the study is published. When a third party detects mistakes after publication, it often converges into battles over thorough data check and potential retraction, eventually leading to implicit or explicit accusations of defection. We see it as an encouraging sign that most editors (70.5%) would allow revision and resubmission in the scenario where the data editor would detect problems with the shared data. Also, as journals nowadays are moving slowly away from the hunt for significant results, a change in the findings and conclusions should become a less important criterion for acceptance as long as the scientific questions and methods remain sound [6].

Public repositories and journals/publishers have much to gain from improving the quality and FAIRness of shared data. The resulting positive reputation in this regard will be attractive for authors, which translates into more usage of repositories and more submitted manuscripts to choose from for the journals. Consequently, both public repositories and journals/publishers should invest in providing macroinstructions for data sharing. Public repositories might be more resourceful than journals in developing and adjusting data sharing macroinstructions for every scientific discipline. Some editors brought our attention to an already existing tool somewhat close to our suggestion for macroinstructions, DataSeer. DataSeer is a tool that verifies whether shared data match the text describing data collection in the manuscript. It is a platform that uses artificial intelligence to help authors tailor their data sharing to the journal's requirements following best practices (see: https://dataseer.ai). Thus, some starting points already exist.

On the other hand, creating data editor positions is a task for scientific journals. By checking the files, data editors may detect mistakes in either data files, analyses, or codes that warrant corrections, which might affect the study conclusions. This in itself will help journals reduce the number of published errata and possible retractions. Furthermore, we anticipate that data editors will promote the establishment of macroinstructions at the journal level to improve own efficiency. Without such improved efficiency, we agree with the editors-in-chief's opinion that the data editor's duties might become overwhelming. Nevertheless, being a data editor can be highly attractive to junior colleagues. The new generation of postdocs and newly graduated PhDs in the field of behaviour, ecology and evolution, for instance, have tremendous statistical knowledge and skills to offer. They can be an excellent asset for journals, and in return, they enrich their CVs with a demonstrated key competence. Furthermore, we do not think the financial argument against a data editor holds well. In many journals, (associate) editors, for example, are unpaid or receive very little compensation. In the scenario where having a data editor on the editorial board adds costs for journals, maybe the latter can consider having this as part of the publication fees. After all, the funding agencies can play an active role by paying the extra costs and even encouraging grantees to publish in journals that ensure quality control and FAIRness of shared data. Overall, we believe that the returned benefits over time should overcome any costs, providing journals with data editors a competitive edge.

When asked about the best practice for future data sharing, several editors reached out and greeted the initiative as interesting and timely. Although, one editor commented: "*I think you are all very naïve…*". We think two important players can make our proposal work; authors interested in extra support to make their study as accessible and flawless as possible and funding agencies that are increasingly caring about open and FAIR data. For example, the funding agency, the Swiss National Science Foundation (SNSF), has a policy to push toward open access. It exclusively covers the publication fees for 100% open access journals, but not hybrid ones. For good or bad, this policy clearly selects against publishing in hybrid journals. One can imagine that if a similar approach applies to increase sharing high quality data, it will select for better data FAIRness. As soon as key players team up, we can increase the quality of publications by improving data FAIRness. This next level of transparency will eventually also reinforce the credibility of science.

Ethics statement:

The Ethics committee of the University of Neuchatel in Switzerland approved this study under permit number #80-2021. We have asked all participants for their consent to participate in the survey. We did not collect any personal data on the participants, and participation was anonymous. As a thank you for participation, we donated 10 € for every filled survey to The Nature Conservancy charity.

Acknowledgements

We kindly thank all editors-in-chief who took the time to participate in this survey and share their opinions with us. We also thank Dominique Roches for his valuable input on an earlier version of the paper.

Funding:

The Swiss National Science Foundation supported this study (https://www.snf.ch/en). Grant number: P400PB_199286 to ZT and 173334 to RB. The funder had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

Author contributions:

The authors contributed equally.

Competing interests: Authors declare no competing interests.

Data and materials availability: data is accessible on Figshare: https://figshare.com/s/4ed6ba1ce5ff0af4c4d8 [7].

References:

1. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biol. 2015;13: e1002295. doi:10.1371/journal.pbio.1002295

2. Culina A, van den Berg I, Evans S, Sánchez-Tójar A. Low availability of code in ecology: A call for urgent action. PLoS Biol. 2020;18: e3000763.

3. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. doi:10.1038/sdata.2016.18

4. Roche D, Berberi I, Dhane F, Soeharjono S, Dakin R, Binning S. The quality of open datasets shared by researchers in ecology and evolution is moderately repeatable and slow to change. 2021.

5. Strasser CA, Hampton SE. The fractured lab notebook: undergraduates and ecological data management training in the United States. Ecosphere. 2012;3: 1–18.

6. Brembs B. Reliable novelty: New should not trump true. PLOS Biol. 2019;17: e3000117. doi:10.1371/journal.pbio.3000117

7. Triki Z, Bshary R. Data from: How to enhance data FAIRness. 2021. doi:10.6084/m9.figshare.17091572