Assessing scientific merit through data quality in a domain repository

JANEZ ŠTEBE

University of Ljubljana, Faculty of Social Sciences; Social Sciences Data Archives

ABSTRACT

A widespread open science publishing culture means the researchers involved should be rewarded for the merit contained in the research output they produce and share. A recent survey shows that "Open and FAIR data management and sharing" is the most deserving of reward. (Grattarola et al., 2024). Research assessment should prioritize quality, encompassing the entire spectrum of research outputs, including research data, computer code, and other. Evaluating the scientific quality of research data is crucial for including it as a legitimate scientific output, akin to the way literature is evaluated in the editorial and peer review publishing selection.

The data review, traditionally forming part of the pre-ingest 'appraisal and selection' stage within a repository, has gained prominence since the variety of data published has expanded. Transparent documentation and information concerning data quality can help in determining the reusability of data (Sharma, 2024; Kindling and Strecker, 2022).

Authors themselves are the first to critically assess the quality of their data and its overall significance. They can then choose to publish the data in the most suitable repository, considering factors like acquisition criteria and processing intensity. Domain-specific repositories are more appropriate for assessing the data characteristics that are important for the research community.

Repositories are also distinguished by levels of curation. The most intensive "Data-level curation" is more than simply checking the quality, consistency and completeness of (meta)data. Such curation actively enhances these elements in collaboration with the author, similar to the traditional literature publishing process.

While depositing data 'as is' is preferable to not sharing at all and should be duly recognised, research evaluation should differentiate such data from reviewed data whose scientific quality has been established. Only the latter can be evaluated similarly to peer-reviewed literature.

Assessing research data quality in a domain repository

The way research data quality is assessed varies depending on the intended re-use purpose. In the European Statistical System (ESS), data quality reporting utilises a standard scheme featuring 18 criteria like relevance, accuracy and comparability. Some journals have already introduced the 'data editor' role (Muench, 2023) that focuses on the overall data and computer code consistency to assess the "computational reproducibility", provided that the material is openly accessible, well documented, and cited in an article. Although these criteria suit the specific purpose they are meant for, research data repositories need to more broadly examine various quality aspects.

The consideration is the reuse potential of future data weighted against the cost of digital curation activities. Not all data can be processed to the highest curation standards. Further, data re-use potential entails a prediction that respects the context that determines the value, not taking any criterion in isolation. The value depends on whether the content of data is rare or it duplicates data already in the collection. There are no absolute measures. Combinations of criteria are also considered (Gutmann et al., 2004; Whyte and Wilson, 2010). During the COVID-19 period, it was important to collect timely data and share it widely, even if some methodological quality factors may have been compromised.

ADP Template for Evaluating Research Data Quality

To determine the re-use potential, processing, long-term curation costs, and scientific data quality, the Slovenian social science data archives (ADP) are testing a "Template for the evaluation of research data quality" it produced itself. Elements from the above-mentioned sources, supplemented with general social science research reports on aspects of ratings (e.g., Miller, 1991: 642-644) for research data, are incorporated in the template and tailored to the repository's mission and designated user community.

The template's criteria are categorised in different sections distinguishing which role can assess them. In the first section, the Data archivist assesses the completeness of data, formal aspects of metadata, adherence to the minimum set established by the CESSDA community, the format, and whether the data are sufficiently clean and documented on a granular level to facilitate informed re-use. Legal and ethical conditions for data sharing are also verified. This provides an elementary data transparency review that is thus performed. The basic scientific relevance is assessed regarding its further research re-use potential: if the data cover the research topic's multifaceted nature; if the sample represents the complete or an important population; and if methodological relevance and research design complexity are demonstrated.

The relevance of a study and associated data for the repository collection is assessed in the second section and performed by the Head of acquisition. The historical and cultural relevance and uniqueness of data are considered, along with suitability for use in teaching or citizen science, etc.

Thus far, the assessment already evaluates a few scientific qualities. In the remaining section, the Domain specialist (from either the repository team or outside) assesses some broader aspects of scientific fitness for use for studying a wide range of theoretically or practically inspired problems. This includes methodological quality assurance and the study's significance for various research areas or for addressing important societal problems.

The written assessment is then presented orally to the acquisition commission, which decides on the study's category (self-deposit or long-term curation) and produces a summary. Data published in the long-term curation system that meets most criteria of scientific relevance also receives a score, which is entered in the Slovenian bibliographic system where it counts as scientific output for the researcher's promotion.

Benefits and Challenges of the System

The comprehensive evaluation system aims to minimise subjective and arbitrary evaluations. Nonetheless, it requires additional effort from data repository staff already in the pre-ingest phase. The observations made are noted in internal documentation and in the long-term curation regime and communicated to the author for further formal data, metadata, and documentation quality assurance on a higher level.

Although the scientific quality evaluation cannot be completely objective, as there is always a certain arbitrariness in weighting the importance of different criteria, like with a literature peer review, attracting established researchers for a role that brings little reward is hard. A compromise is to rely on experienced researchers among the repository staff. The discussion in the acquisition commission also helps overcome the limitations of the primary evaluation: it is usually enough for the final decision on scientific merit to establish that at least some of the scientific quality criteria have been met.

The scores associated with the data publication incentivise researchers to maintain high quality throughout the data lifecycle, including the effort involved in preparing the data for publishing, and discourage the intentional reduction of data quality to gain a competitive advantage. The data itself is already recognised as a scientific product in its own right, and its secondary re-use can further enhance the researcher's citation reputation.

KEYWORDS

data repository; data curation; research data quality; research assessment; re-use

REFERENCES

1. Grattarola, F., Shmagun, A.; Erdmann, C., Cambon-Thomsen, A., Thomsen, M. and Mabile, L. (2024). Gaps Between Open Science Activities and Actual Recognition

Systems: Insights from an International Survey. SocArXiv. https://doi.org/10.31235/osf.io/hru2x.

2. Gutmann, M., Schürer, K., Donakowski, D. and Beedham, H. (2004). The selection, appraisal, and retention of social science data. Data Science Journal, 3(0):209-221. https://doi.org/10.2481/dsj.3.209.

3. Miller, D. C. (1991). Handbook of Research Design and Social Measurement. (5th ed., str. XIV, 704). Los Angeles: Sage Publications.

4. Muench A. (2023). The roles of data editors in astronomy. Sci Ed. 2023,46:8-10.

5. https://doi.org/10.36591/SE-D-4601-04

6. Kindling, M., & Strecker, D. (2022). Data Quality Assurance at Research Data Repositories. Data Science Journal, 21(1), 18. https://doi.org/10.5334/dsj-2022-018

7. Sharma, S. (2024). Peer Reviewing Data and Software: A Pilot Project (1.0). Zenodo.

8. Whyte, A. & Wilson, A. (2010). How to Appraise and Select Research Data for Curation. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides