Information Services & Use 35 (2015) 185–188
DOI 10.3233/ISU-150780
IOS Press

Data accessibility and reproducibility: Moving to transparent publishing in the biosciences

Bernd Pulverer
Chief Editor, EMBO Journal, and Head of Scientific Publications, EMBO
E-mail: bernd.pulverer@embo.org

1. Introduction

As the volume of archived scientific data and information increases exponentially, the issue of the reliability and reproducibility of research results has come to the fore. The published output per year now exceeds one million peer-reviewed papers in more than twenty-five thousand journals with an estimated growth of up to 5% per annum. In addition to the published literature, open science policies encourage posting of a vast amount of data in structured and unstructured repositories. We must find ways of handling, validating, and accessing data and information more effectively and more efficiently or, else, the sheer volume threatens to obstruct scientific progress.

Scientific journals continue to be the dominant channel for sharing vetted data. This is because in the value chain of establishing research quality, selective journals are most effective. They remain deeply embedded in the research ecosystem: from filtering based on peer review, editorial governance, and revision cycles; to broad dissemination, discoverability, and reliable archiving; on to usage/citation metrics, and reputation based value assignment – which together inform research assessment.

Journal policy influences researcher behaviour. Hence, if journals develop and agree consistent new standards, this is an excellent opportunity for establishing better ways of handling data and information. In particular this is the case if standards are set in coordination with research funders and institutions.

In what follows, firstly, I will highlight the shift to transparent publishing at EMBO Press, for which open science provides context. Second, I will be reviewing the problem of research reliability and reproducibility, focussing on adapting the editorial processes for enhanced quality assurance. Transparent peer review and data transparency are highlighted as key elements. I will argue that the scientific paper needs to be reinvented: data packaged into the long established information unit of the research paper is optimized for the human reader, but largely opaque to computational access. Navigating the information wave necessitates that research publications are optimized for discoverability. Finally, I summarize how current challenges and changes could catalyse agreement on new standards.

2. The move to transparent publishing in the context of open science

Transparent publishing means that data, methods, and protocols must be accessible easily and efficiently for interpretation and re-use. Transparent publishing is consistent with open data policies, such as requests by the European Commission for the deposition of data underlying publications, and advocating a move towards machine-readable versions of datasets and publications. Transparent publishing does not have to change or undermine existing business models, and can be complementary to or, indeed, an alternative to Open Access.

Transparent publishing is also highly compatible with more open forms of peer review, which add accountability for referees, and valuable information for authors and readers. Transparency better promotes and recognizes the value of referees, the peer review system, and the whole value chain.

3. The problem: Reliability and reproducibility

The reliability and reproducibility of research results published in journal articles has come under scrutiny, including public claims that much research is not reproducible, at least not from the information provided. Concurrently, in policy and practice, we have two major trends in open science:

(1) Depositing data in structured databases, which is increasingly required by journal editors, and mandated by research funders;
(2) Storage for essentially unlimited volumes of primary, unstructured data.

More systematically, we observe the following key trends and issues:

(a) Increasing specialization, coupled with rapid technological development and information overload, making it difficult for peers to stay abreast of cutting-edge research results;
(b) The sheer volume of data and the lack of suitable platforms leading to only limited transparency and granularity in reporting research results;
(c) Too few incentives for sharing data, and for establishing best practice for authors, referees, and journals;
(d) The dominance of publication metrics based on research assessment skews the behaviour of researchers and journals towards rapid publication and excessive claims.

4. In search of scalable solutions

To address the key issues identified above, we need to adapt investment, incentives, and policies.

(1) Investment: How much do publishers, research funders, and the scholarly community want to invest in data curation, review, and publishing?
(2) Incentives: Can we create academic incentives for publishing and valuing negative, confirmatory or refuting data? What will be the economic benefits for journals that value reproduction, correction, and retraction of data? How do we make this process efficient?
(3) Policies and governance: Is there agreement for requiring authors to share and correct data? Can refereeing be improved with formal training and incentives? Will referees be held accountable for their judgement?

One way of addressing the issue of reproducibility would be to require that research results (i.e. data) be validated prior to publication, as suggested by some in the wake of widely reported 'reproducibility scandals'. Data validation would imply that another researcher replicate the results or that they generate data that lead to the same conclusions. There are logistical and practical hurdles associated with reproducing a specific experiment in another lab, especially before publication. Research assessment does not reward repeating experiments and intermediaries may emerge to offer this type of service (for example, at Science Exchange[1] researchers can already order experiments to be conducted in another lab). Reproducing experiments systematically would be a very expensive solution, and in any case is unlikely to be scaleable. More principally, lack of reproducibility does not necessarily imply that a given dataset is flawed.

We must acknowledge that we have not yet developed a scalable solution for addressing the reliability and reproducibility of published research results systematically. Any more focus on reviewing, validation, and reproducing data would only increase workload. Mandated by funders and publishers or not, more refereeing, commenting, accountability, collaboration, and so on would make the workload unbearable, especially for those in high demand as reviewers, that is the top scholars doing the best research.

5. Adapting the editorial process

Nevertheless, we can optimize the current publication process. To be effective, optimization has to involve authors, referees, readers, journals, and funders.

For example, at EMBO Press, we now have the following enhanced editorial process:

- Explicit editorial policies and a mandatory author checklist on data presentation and ethics.
- Transparent review process: publication of anonymous referee reports, editorial decision letters, and author responses in full – with open referee identities optional (i.e. open peer review).
- Cross-referee commenting, whereby referees are encouraged to respond to each other's criticisms.
- Prepublication ethics screening for plagiarism and image manipulation.
- Source data: publishing minimally processed data underlying figures as linkout from figures on a CC-0 license to support re-use.

We have found that the transparent review process and cross-referee commenting are an excellent teaching tool that systematically supports the improvement of peer review, leading to better decision-making. Systematic prepublication analysis for data integrity uncovers that 20% of otherwise publishable manuscripts have data problems that require correction. Problems can usually be corrected successfully, and less than 1% have to be rejected. Tracking such rejects for subsequent publication elsewhere remains a challenge.

We are working on further enhancements, including:

(a) Data oriented search functionality that directly points to figures and data (http://sourcedata.embo.org);
(b) Enhanced recording of methods and protocols linked to specific experiments;
(c) Unambiguous reagent identifiers;
(d) Supporting the versioning of papers to encourage correction and addition.

[1] https://www.scienceexchange.com/how-it-works.

6. Agreeing a new standard

This article has reviewed how the sheer quantity of research publications in conjunction with the emergence of data-driven science and pressures to publish from research assessment has led to research reliability and reproducibility becoming problematic and key issues. While solutions have to emerge at every level – including researchers, research institutions, funders, databases, and journals, most promising in the interim is adapting the editorial process. A number of largely isolated initiatives for scalable solutions at this level have emerged, but these will only lead to global change if they are widely adopted in a consistent manner.

The scholarly community, and those serving it, must urgently develop and investigate (more) scalable solutions for increasing reliability and reproducibility. For if we do not, this very likely will erode trust in the scientific enterprise. As the problem becomes deeper and more public, it may also lead to solutions being imposed from the outside. Hence, new standards should emerge without delay and the publishing community would do well to play a fundamental part in developing solutions.