Embedding Research Data Management Support in the Scholarly Publishing Workflow

Authors:

Iain Hrynaszkiwicz and Rebecca Grant

Correspondence to iain.hrynaszkiwicz@nature.com

Springer Nature, The Campus, Trematon Walk, Wharfedale Road, London N1 9FN

Abstract

In 2016 the publisher Springer Nature introduced four standard research data policies for its journals, enabling more journals to adopt a data policy appropriate for their discipline and community. These standard policies have been adopted by more than 1,500 journals, and similar initiatives to standardise journal research data policies have since been introduced by other large publishers. To support researchers and editors Springer Nature launched a Research Data Helpdesk, which to October 2017 had received more than 300 enquiries. A large survey of researchers with more than 7000 respondents in 2017 revealed that many researchers need support with data management and curation tasks. In 2017 Springer Nature introduced a pilot service to provide additional support to researchers who wish to make their data available alongside their published articles. This Research Data Support service provides hands-on assistance to researchers in uploading their data to a repository, selecting an appropriate licence, enhancing metadata, and cross-referencing the data and its associated publication. The data curation standards were subject to blinded testing by professional editors, and curated datasets scored much higher for metadata quality and completeness on average. We describe the implementation of – and lessons learned from providing – a third party data deposition and curation service at a large scholarly publisher, which has been used by authors publishing in journals including *Nature*, and *BMC Ecology*. We conclude with current and future developments, which extend the Research Data Support service to any published researcher, and to research institutions and conferences, providing opportunities to embed research data management support earlier in the scholarly publishing workflow. This paper was presented at the PV 2018 Conference: Adding value and preserving data.

Full text

Research data policies

In 2016 the publisher Springer Nature introduced four standard research data policies for its journals (Iain Hrynaszkiwicz, 2016), enabling more journals to adopt a data policy – one

which is appropriate for their discipline and community. This was in response to evidence (Naughton & Kernohan, 2016) that the research data policies of journals and publishers were found to be confusing for researchers and support staff and were in need of standardisation. The four policies aim to support researchers in following good practice in the sharing and archiving of research data in accordance with community expectations; facilitate compliance with institutional and research funder requirements to share data; and simplify the journal policy landscape for researchers and support staff. As of April 2018, these standard policies have been adopted by more than 1,500 journals at Springer Nature (the second largest scholarly publisher), and similar initiatives to standardise journal research data policies have since been introduced by other large publishers: Wiley, Elsevier and Taylor & Francis in 2017-18.

Several major publishers are now participating in an international effort to standardise journal and publisher data policies via the global Research Data Alliance (RDA), chaired by representatives from Jisc in the UK, the Australian National Data Service, Springer Nature, and Wiley (Iain Hrynaszkiewicz, Simons, Goudie, & Hussain, n.d.). This interest group in 2018 released for public comment a draft master policy framework, which is ultimately intended to be adopted by all journals and publishers. Policy is known to have an impact on researchers' willingness to share data (Schmidt, Gemeinholzer, & Treloar, 2016) but there is a need to provide ensure policy is implemented effectively (Vines et al., 2013).

Providing support with a Research Data helpdesk

To support the introduction of its standard policies, Springer Nature introduced in July 2016 a Research Data Helpdesk. The helpdesk provides free email-based advice to editors, authors (researchers) and support staff to help them to implement and comply with journal research data policies. The helpdesk is accessible at any point in the peer-review and publication process, and outside of it, although the it is not intended to become part of the peer-review process. To October 2017, the helpdesk had received more than 300 enquiries. More than half (53%) of people contacting the helpdesk were researchers (authors); 43% were editors, (both professional and academic), and the remaining 4% were other stakeholders such as librarians and repository managers. The most common types of queries related to policy implementation (122 queries); data repositories (53 queries); policy compliance (46 queries) and on writing data availability statements (31 queries). A survey of users of the helpdesk found a high level of satisfaction, with 84% of respondents being satisfied or very satisfied with the advice they received (Astell, Hrynaszkiewicz, Grant, Smith, & Salter, n.d.).

Understanding costs

As well as promoting the benefits of data sharing it is important to understand the costs, if research data management support is to be available to a high proportion of researchers. Funding agencies have begun to more explicitly acknowledge and provide funding for research data management in grant awards and guidelines ("Frequently Asked Questions (FAQs) for Public Access (nsf18041) | NSF - National Science Foundation," 2018). There can be increased costs in the publishing process when research data policies are enhanced. Understanding these costs is important for journals and publishers who routinely monitor

manuscript processing time and costs, and typically strive to peer review papers rigorously and rapidly. Some types of data policy require additional checks on submitted or accepted manuscripts, which can have associated costs – while also providing benefits (Piwowar & Vision, 2013) – and are important to quantify, particularly for large publishing operations.

At the Nature journals, in 2016, an analysis of the impact of adding data availability statements (DASs) to every article, as part of enhancements to their policies, found that on average it increased manuscript processing time by professional editors by 10 minutes (Grant & Hrynaskiewicz, 2018). The publisher deemed the increased time and its associated cost reasonable, given the importance and benefits of including a DAS, and this cost did not limit implementation of mandatory DASs at all Nature journals.

Where researchers need more support

Seeking to understand more about the practical challenges researchers face in sharing research data, in 2018 Springer Nature published a survey of researchers which had received more than 7,700 responses (Stuart et al., 2018). The main barrier to data sharing was identified by respondents as 'Organising data in a presentable and useful way' (46%). Other common challenges were 'Unsure about copyright and licensing' (37%); 'Not knowing which repository to use' (33%); 'Lack of time to deposit data' (26%) and 'Costs of sharing data' (19%). The most commonly reported barrier suggests there may be a lack of skills, support or resources available to researchers who wish to curate, share, manage and archive data. Three quarters of researchers also expressed the significance of discoverability (find-ability) of their research data, rating it highly for importance. Interviews with funding agencies (n = 13) and higher education institutions (n = 15) have also revealed that metadata skills and expertise are a commonly perceived barrier to data sharing.

In 2017, Springer Nature introduced a pilot service (Iain Hrynaskiewicz, 2017) to provide additional support to researchers who wish to make their data available. This Research Data Support service (<https://www.springernature.com/gp/authors/research-data-policy>) provides hands-on assistance to researchers in uploading their data to a repository, selecting an appropriate licence, enhancing metadata, and cross-referencing the data and any associated publication. Data are hosted and preserved on Springer Nature's figshare repository. The service was initially available to authors submitting manuscripts to select BMC journals before being expanded to nearly 100 journals and, from January 2018, being made available to any published researcher and to institutions.

The editorial standards and data curation criteria were developed by, and in consultation with, data curation and data publishing experts and journal editors and were subjected to blinded testing by professional journal editors - many of whom are former researchers. Ten professional editors representing life, social and physical sciences were randomly assigned four datasets each to assess, half of which (20) had been curated according to the standards of the Research Data Support service and half (20) which had not. Responses were collected via an online survey and the curated datasets scored much higher for metadata quality and completeness on average (rated 4.1 out of 5.0 for overall quality, compared to 2.8 for unedited; Figure 1).

Figure 1: Metadata quality and completeness scores for edited (curated) compared to unedited datasets

The service focuses on curation of metadata rather than directly editing the contents of files, but also intends to guard against the publication of sensitive data and inappropriate file formats, and the use of inappropriate licence terms.

Users of the service include authors publishing in *Nature* (Giles, Xu, Near, & Friedman, 2017a, 2017b), and *BMC Ecology* (J. Rakotoniaina et al., 2017; J. H. Rakotoniaina et al., 2017). Publication and sharing of sensitive data remains a challenge, and is the most common reason for submissions to the Research Data Support service being rejected. Authors' submissions are checked against established guidelines on data anonymisation (I. Hrynaszkiewicz, Norton, Vickers, & Altman, 2010), and authors are asked for details of participant consent for data publication where appropriate, although anonymisation of the data themselves ultimately remains the authors' responsibility. Authors are advised of alternatives to publicly sharing non-anonymised data, such as repositories that provide managed or safeguarded access - such as the UK Data Archive.

Integration of research data support services into the traditional scholarly publishing workflow and systems provides opportunities to increase data sharing, and the visibility of available support to researchers, but also presents challenges. Research Data Support is currently available as an integrated service to selected journals using the Editorial Manager manuscript submission system. Testing and development of the integration is on-going, as with other manuscript submission systems, and has revealed that visibility and engagement of users with the service via this route could be improved. Supporting authors whose associated manuscript is rejected from their initial target journal were also an important workflow consideration. The service provides three options where authors' associated manuscript has been rejected (and will often be sent to another journal for consideration): The curated data can be transferred to a personal figshare account; the data can be deleted; or the data can be held privately for a fixed period. Key to provision of the level of curation offered (details at <https://www.springernature.com/gp/authors/research-data-policy/help-faqs/15369356> 'What is involved in data curation?') is access to the associated manuscript or documentation, to provide a record for the dataset that links to the published article but which is also comprehensible without the article.

Current and future developments

While the Research Data Support service ended its pilot phase in early 2018 and introduced fees to ensure sustainability of the service, different use-cases and workflows continue to be tested and developed. These include the provision of research data support to conferences, such as the International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS; <https://springernature.figshare.com/tacas>). This collaboration led to a higher level of uptake by authors compared to current journal-based workflows.

Research data support provision by scholarly publishers is a new development and traditionally researchers may seek support via dedicated teams within their institutions; by seeking advice from their funders; from external expert agencies and services such as Jisc or the Digital Curation Centre; or by working directly with specialist data repositories – where and if they exist for a given a community. Not all researchers have access to these services however, and researcher support for data sharing will likely remain a mixed economy for the foreseeable future. The best way to support a researcher on their specific situation varies from case to case and may need to consider applicable funder and institution research data policies. Research data support from publishers can potentially complement institutional services, for example where institutional services are not able to scale or provide sufficient flexibility for the volume of research data outputs of their researchers. Working directly with institutions and funding agencies, and with researchers earlier in the research life-cycle than the point of publication in a peer-reviewed journal, also provide opportunities to further embed research data management support in the scholarly workflow.

Data availability

Data underlying the analysis of helpdesk queries are not publicly available due to the inability to suitably anonymise these customer data, and as consent was not obtained to publish the queries. The data are available on reasonable request from the authors and via researchdata@springernature.com. The data supporting the plots in figure 1 are available publicly in figshare with the identifier <https://doi.org/10.6084/m9.figshare.6200357> (Smith, Grant, & Hrynaskiewicz, n.d.).

Acknowledgement

The authors thank Mr Graham Smith, Springer Nature, for support with production of figure 1 and preparation of the data supporting the figure.

Conflict of interest statement

Both authors are employees of Springer Nature.

References

- Astell, M., Hrynaskiewicz, I., Grant, R., Smith, G., & Salter, J. (n.d.). Have questions about research data? Ask the Springer Nature Helpdesk.
<https://doi.org/10.6084/m9.figshare.5890432.v2>
- Frequently Asked Questions (FAQs) for Public Access (nsf18041) | NSF - National Science Foundation. (2018). Retrieved April 30, 2018, from
<https://www.nsf.gov/pubs/2018/nsf18041/nsf18041.jsp#q46>
- Giles, S., Xu, G.-H., Near, T., & Friedman, M. (2017a). Fukangichthys: CT scan data and surface files from middle Triassic fossil scanilepiform fish. figshare.
<https://doi.org/10.6084/m9.figshare.c.3814360>
- Giles, S., Xu, G.-H., Near, T. J., & Friedman, M. (2017b). Early members of “living fossil” lineage imply later origin of modern ray-finned fishes. *Nature*, 549(7671), 265–268.
<https://doi.org/10.1038/nature23654>
- Grant, R., & Hrynaskiewicz, I. (2018). The impact on authors and editors of introducing Data Availability Statements at Nature journals. *bioRxiv*, 264929.
<https://doi.org/10.1101/264929>
- Hrynaskiewicz, I. (2016). Promoting research data sharing at Springer Nature. Retrieved from <http://blogs.nature.com/ofschemesandmemes/2016/07/05/promoting-research-data-sharing-at-springer-nature>
- Hrynaskiewicz, I. (2017). New services to support open research. Retrieved April 30, 2018, from <https://blogs.biomedcentral.com/bmcblog/2017/04/27/new-services-to-support-open-research/>
- Hrynaskiewicz, I., Norton, M. L., Vickers, A. J., & Altman, D. G. (2010). Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ*, 340(jan28 1), c181–c181. <https://doi.org/10.1136/bmj.c181>
- Hrynaskiewicz, I., Simons, N., Goudie, S., & Hussain, A. (n.d.). Research Data Alliance Interest Group: Data policy standardisation and implementation. Retrieved April 30, 2018, from <https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation>
- Naughton, L., & Kernohan, D. (2016). Making sense of journal research data policies. *Insights the UKSG Journal*, 29(1), 84–89. <https://doi.org/10.1629/uksg.284>
- Piowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, e175. <https://doi.org/10.7717/peerj.175>
- Rakotoniaina, J. H., Kappeler, P. M., Kaesler, E., Hämäläinen, A. M., Kirschbaum, C., & Kraus, C. (2017). Hair cortisol concentrations correlate negatively with survival in a wild primate population. *BMC Ecology*, 17(1), 30. <https://doi.org/10.1186/s12898-017-0140-1>
- Rakotoniaina, J., Kappeler, P., Kaesler, E., Hämäläinen, A., Kirschbaum, C., & Kraus, C. (2017). Capture-mark-recapture data modelling survival rates of *Microcebus murinus* in relation to glucocorticoid level, parasite infection and body condition. figshare.
<https://doi.org/10.6084/m9.figshare.5259415>
- Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum’s Open Data Survey. *PloS One*, 11(1), e0146695.
<https://doi.org/10.1371/journal.pone.0146695>
- Smith, G., Grant, R., & Hrynaskiewicz, I. (n.d.). Quality and completeness scores for curated and non-curated datasets. figshare.
<https://doi.org/10.6084/m9.figshare.6200357>
- Stuart, D., Baynes, G., Hrynaskiewicz, I., Allin, K., Penny, D., Lucraft, M., & Astell, M. (2018). *Whitepaper: Practical challenges for researchers in data sharing*.
<https://doi.org/10.6084/m9.figshare.5975011.v1>
- Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., ... Yeaman, S. (2013). Mandated data archiving greatly improves access to research data. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, fj.12-218164-. <https://doi.org/10.1096/fj.12-218164>