Information Services & Use 42 (2022) 441–445
DOI 10.3233/ISU-220167
IOS Press

Supporting Open Science with frictionless publication workflows: The Tree of Life project at Wellcome Open Research

Rebecca Grant*
Head of Data and Software Publishing, F1000, 240 Blackfriars Road, London, UK

Abstract. This paper is based on a presentation delivered as part of the NISO Plus 2022 panel discussion titled "Open Science: catch phrase, or a better way of doing research?" that focused on the workflows of Open Science and opportunities for collaboration by stakeholders including publishers, repository infrastructure providers, and the wider research community.
While the aims and outputs of Open Science are well-defined, this paper explores the workflows that are necessary to support the production of "open scientific knowledge", as defined by UNESCO. Producing research outputs as open scientific knowledge is an activity that is undertaken alongside traditional research practices and must be planned for from the beginning of the research process.
This paper explores the challenges and opportunities associated with Open Science workflows, focusing on an innovative new automated publishing pipeline on the *Wellcome Open Research* publishing platform.

Keywords: Tree of life project, wellcome open research, F1000, open data, open science

1. Introduction

The publication of the UNESCO Recommendation on Open Science inspired a panel discussion at the NISO Plus 2022 conference, "Open Science: catchphrase, or a better way of doing research?" During the panel the speakers, who included representatives from the American Geophysical Union, the Shanghai Information Center for Life Science (CAS), the repository Dryad, and academic publisher F1000, considered not only the beneficial outputs of Open Science practices, but also the methodologies and workflows involved in achieving Open Science.

As an academic publisher, F1000 provides research publishing solutions and services to organisations including the European Commission, Wellcome, and the Bill & Melinda Gates Foundation, as well as directly to researchers through the F1000 Research publishing platform. F1000 publishing platforms are fully Open Access, support open peer review and the open publication of new versions of articles, and have strong Open Data policies that require that authors share all of the research data underlying their articles.

This paper considers the challenges associated with Open Science workflows and the publication of open research outputs and describes a project on the *Wellcome Open Research* publishing platform which is exploring how automated workflows can assist researchers to publish their outputs openly, at scale.

*E-mail: Rebecca.grant@f1000.com.

2. The challenges of Open Science

The UNESCO Recommendation on Open Science [1] states that Open Science "[…] aims to make multilingual scientific knowledge openly-available, accessible, and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation, and communication to societal actors beyond the traditional scientific community". The knowledge produced by Open Science practices may take the form of Open Access publications, research data, metadata, educational resources, software, source code, and hardware; it also refers to the possibility of opening research methodologies and evaluation processes.

The production of open scientific knowledge requires investment of time and resources on the part of the researcher, as well as an awareness and understanding of the activities involved. Additional motivators or credit may also be necessary to encourage the adoption of such practices by researchers who have not previously worked in "Open" ways. Hagger notes the move towards Open Science practice requires a complete shift in mindset on the part of the researcher, towards an assumption that every part of a research project will be subject to examination by others [2].

Not only is it necessary for researchers to develop an Open Science mindset, but it is important that they do so early in their research lifecycle if the intention is to share their research outputs openly. A study by Gownaris et al. identified that researchers' awareness of the practices associated with Open Science is particularly low in relation to early phases of research (e.g., the study design stage). This lack of awareness can cause "path dependencies" that prevent the open sharing of outputs in later stages of the research [3]. As an example, such path dependencies can impact on the ways that researchers share additional outputs openly when publishing a research article. Many academic publishers, including F1000, enforce strict Open Data sharing policies for their authors [4]. If researchers who have conducted their research with human research participants do not consider Open Science practices at the planning stage of their research, they may fail to gain participants' permission to share their research data openly. This can then cause issues when the researcher attempts to publish their research article, as they cannot comply with the journal's Open Data policy.

Stakeholders' Open Science policies are increasing however, emerging from funding agencies [5,6] institutions [7], and academic publishers [8]. These policies tend to focus on two elements of Open Science: Open Access publishing and Open Data sharing. There is evidence that such policy mandates are effective in changing researcher behaviour, for example BioMed Central and the Public Library of Science (PLoS) journals demonstrated increased data sharing when strong research data policies were introduced in 2015 and 2014 respectively [9]. While mandates increase, the motivators or rewards for researchers who practice Open Science are not always clear. The ON-MERRIT project has identified discrepancies between researchers' attitudes towards Open Science practices, and the extent to which they are rewarded through institutions' policies on promotion, review and tenure. For example, 70.4% of surveyed researchers believed that sharing research data openly should be considered as an important or very important factor in promotion decisions, but only 26.6% of institutional policies designate it to be so [10]. In parallel, 65% of surveyed researchers have stated that they have never received credit (for example in the form of a citation) because they had shared their data openly [11]. Initiatives such as the UKRN (UK Reproducibility Network)'s five-year programme of work with its consortium of institutional members [12], and the Montreal Neurological Institute (MNI)'s announcement of their full commitment to Open Science demonstrate the ways that institutions can address the new challenges that Open Science presents [13].

Nevertheless, many challenges associated with Open Science remain: to produce open outputs like articles, research data, software, hardware, or education materials, it is necessary to plan for "openness" from the beginning of the research project. To consider a new "open" way of working, a change of mindset may be needed, as researchers have not traditionally shared all project outputs in this way. While funder and publisher policies are increasingly mandating some facets of Open Science, the tangible benefits for researchers are not yet clear.

3. Automating publishing workflows: The Tree of Life project at Wellcome Open Research

While stakeholders continue to balance mandates and motivators, efforts are underway to improve the workflows of Open Science and to make processes more frictionless. For example, academic journals with Open Data policies often require authors to deposit their research data openly before submitting a manuscript. Authors may not be aware of this requirement until they are midway through their submission to a journal and will then need to discontinue their submission while they deposit their research data appropriately elsewhere. A 2022 pilot on *Nature* journals has addressed this by embedding data deposition functionality into the manuscript submission system, allowing authors to quickly deposit their data into the figshare repository as part of the manuscript submission workflow [14]. The publisher eLife provides similar integration with the Dryad data repository, prompting authors to upload data during the manuscript submission process [15]. For readers, integrations such as the Code Ocean widget on F1000 allow immediate replication of analyses with an interface embedded into published papers [16]. The Center for Open Science (COS) also provides digital Open Science badges, allowing readers to easily identify articles where Open Science practices have been undertaken, for example at selected Taylor & Francis journals [17].

An innovative project on the publishing platform *Wellcome Open Research* has addressed not only frictionless data submission, but a pipeline that allows the submission of the entire manuscript via an API (Application Programming Interface). Once submitted, automated benchmarking reports are used to support the peer review of the articles and to allow the peer reviewers to make their decision more easily and quickly. This process was developed by F1000, which powers the *Wellcome Open Research* publishing platform, and the Wellcome Sanger Institute in order to rapidly publish the genome sequences of thousands of animals, plants, fungi, and micro-organisms that live on and around Britain and Ireland as part of the Tree of Life project. This approach is intended to allow high volumes of genome sequences to be published in the Tree of Life Gateway on *Wellcome Open Research*, in the form of Genome Note articles [18]. In a standard workflow, the genome is sequenced in the lab and deposited at an appropriate data repository. The scientist then drafts a Genome Note article and submits it via the manuscript submission system on *Wellcome Open Research*. Peer review takes place, and any revisions are made by the author, with Genome Note being revised and resubmitted as appropriate. An example of a published Genome Note which has been published using this standard workflow is 'The genome sequence of the northern goshawk, *Accipiter gentilis* (Linnaeus, 1758) [19]'.

In this new publishing workflow, the Sanger Institute sequences each genome, and the sequence is deposited into the ENA (European Nucleotide Archive). Using information from the sequencing equipment and contextual information written by Sanger-affiliated researchers, an XML file for the Genome Note is then compiled. During the sequencing process, the Sanger Institute also collates information and metrics about the quality of each genome assembly. The metrics include Base pair QV, Scaffold N50/NG50, and BUSCO completeness, which are added to the Genome Note XML, and published alongside the article as an automated benchmarking report to support peer review. Some metrics are also

represented as figures which are made available in the body of the article. A package containing the XML file (including the Genome Note article text, and the automated benchmarking report) plus the figures is sent directly to the *Wellcome Open Research* platform via the API, skipping the manual article submission process. The file is checked by the F1000 editorial team, and the Genome Note is published. At the same time, the automated benchmarking report is published as part of the Open Peer Review information panel which is visible alongside the Genome Note. The automated benchmark report will then be verified by human peer reviewers when they assess the Genome Note, as well as any readers of the published article.

As the Sanger Institute intends to publish up to seventy thousand Genome Notes as part of the Tree of Life project, this workflow allows publication volumes which would not be possible in a traditional publishing workflow. As the automated benchmarking reduces the manual effort that would be required to generate and review these sequences via the traditional publishing process, lower Article Processing Charges (APCs) can also be charged for each publication.

This automated publishing workflow is a unique across publishers and could represent a new publishing approach where information flows directly from lab equipment to the publishing platform or journal. The automated benchmarking report can also be used as a support mechanism to reduce the burden on peer reviewers, as peer reviewing can be time consuming and labour intensive [20]. As an additional benefit, the sequenced data is shared both in an appropriate data repository and through the published Genome Note which is indexed and citable, allowing the author to gain maximum credit for their outputs. There is further potential to apply elements of this workflow to other data types which are published at scale or require rapid dissemination, for example health data gathered during the Covid-19 pandemic.

4. Conclusion

Open Science practices can represent a challenge to researchers, as they may represent new ways of working with unclear reward structures. As stakeholders in Open Science, it is beneficial for publishers to consider accompanying Open Science policy mandates with new approaches to facilitate easier publication of open research outputs. Technical solutions such as integrated data deposition or automated publishing pipelines can help to reduce friction in Open Science publishing, reducing the burden on the researcher. As the workflows of Open Science become easier and more standardised, they will balance with policy mandates, rewards, and incentives to create a way of doing research that is more accessible, more straightforward and more beneficial than traditional, closed methods.

Acknowledgements

The author would like to thank her co-panellists at NISO, Shelley Stall (AGU), Jennifer Gibson (Dryad) and Yongjuan Zhang (CAS); and her colleagues at F1000 and the Sanger Institute who have developed the automated publishing workflow for the Tree of Life gateway on *Wellcome Open Research*.

About the Author

Dr. Rebecca Grant is Head of Data and Software Publishing at F1000. She was previously a Research Data manager at Springer Nature and has worked on policy development and served as a digital archivist at the National Library of Ireland, and at the Digital Repository of Ireland. She received her PhD in 2020 from the University College Dublin. Her thesis explored the connections between archival theory and practice and the management of research data.

References

[1] UNESCO Recommendation on Open Science. 2021 [cited 2022 May 31]. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en, accessed August 18. 2022.

[2] M.S. Hagger, Developing an open science 'mindset', *Health Psychology and Behavioral Medicine* **10**(1) (2022), 1–21. doi:10.1080/21642850.2021.2012474, accessed August 18, 2022.

[3] N.J. Gownaris, K. Vermeir, M.-I. Bittner, L. Gunawardena, S. Kaur-Ghumaan, R. Lepenies et al., Barriers to full participation in the open science life cycle among early career researchers, *Data Science Journal* **21**(1) (2022), 2. doi:10.5334/dsj-2022-002, accessed August 18, 2022.

[4] Data availability. [Cited 2022 May 31]. Available from: https://f1000research.com/about/policiesdataavail, accessed August 18, 2022.

[5] Overview of funders' data policies. [Cited 2022 May 31]. Available from: https://www.dcc.ac.uk/guidance/policy/overview-funders-data-policies, accessed August 18, 2022.

[6] Research Funders' Open Access Policies. [Cited 2022 May 31]. Available from: https://v2.sherpa.ac.uk/juliet/, accessed August 18, 2022.

[7] UK Institutional data policies. [Version 6 updated 2016, cited 2022 May 31]. Available from: https://www.dcc.ac.uk/guidance/policy/institutional-data-policies.

[8] Publisher Data Availability Policies Index. [Cited 2022 May 2022]. Available from: https://www.chorusaccess.org/resources/chorus-for-publishers/publisher-data-availability-policies-index/, accessed August 18, 2022.

[9] G. Colavizza, I. Hrynaszkiewicz, I. Staden, K. Whitaker and B. McGillivray, The citation advantage of linking publications to research data, *PLoS ONE* **15**(4) (2020), e0230416. doi:10.1371/journal.pone.0230416, accessed August 18, 2022.

[10] N. Pontika, T. Klebel, D. Pride, P. Knoth, S. Reichmann, H. Metzler et al., ON-MERRIT D6.1 Investigating Institutional Structures of Reward & Recognition in Open Science & RRI, 2021. doi:10.5281/zenodo.5552197, accessed August 18, 2022.

[11] Science, Digital; N. Simons, G. Goodey, M. Hardeman, C. Clare, S. Gonzales et al., *The State of Open Data 2021*. Digital Science: 2021. doi:10.6084/m9.figshare.17061347.v1, accessed August 18, 2022.

[12] Open Research Programme. [Cited 2022 May 31]. Available from: https://www.ukrn.org/open-research-programme/, accessed August 18, 2022.

[13] G. Rouleau, Open Science at an institutional level: An interview with Guy Rouleau, *Genome Biol* **18** (2017), 14. doi:10.1186/s13059-017-1152-z, accessed August 18, 2022.

[14] Springer Nature and Figshare announce pilot to improve data sharing. [Cited 2022 May 31]. Available from: https://group.springernature.com/gp/group/media/press-releases/springer-nature-and-figshare-announce-data-sharing-pilot/20301098, accessed August 18, 2022.

[15] Announcing eLife and Dryad's seamless data publishing integration. [Cited 2022 May 31]. Available from: https://elifesciences.org/inside-elife/0e9d5c4c/announcing-elife-and-dryad-s-seamless-data-publishing-integration, accessed August 18, 2022.

[16] Reanalyse(a)s: making reproducibility easier with Code Ocean widgets. [2017 April 20, Cited 2022 May 31]. Available from: https://blog.f1000.com/2017/04/20/reanaly-seas-making-reproducibility-easier-with-code-ocean-widgets/, accessed August 18, 2022.

[17] Open Science Badges. [Cited 2022 May 31]. Available from: https://authorservices.taylorandfrancis.com/open-science-badges/, accessed August 18, 2022.

[18] J. Threlfall and M. Blaxter, Launching the tree of life gateway, *Wellcome Open Res* **6** (2021), 125. doi:10.12688/wellcomeopenres.16913.1, accessed August 18, 2022.

[19] K. August, M. Davison, C. Bortoluzzi et al., The genome sequence of the northern goshawk, *Accipiter gentilis* (Linnaeus, 1758), *Wellcome Open Res* **7** (2022), 122. doi:10.12688/wellcomeopenres.17821.1, accessed August 18, 2022.

[20] A. Severin and J. Chataway, Overburdening of peer reviewers: A multi-stakeholder perspective on causes and effects, *Learned Publishing* **34**(4) (2021), 537–546. doi:10.1002/leap.1392, accessed August 18, 2022.