

Article

Data2paper: Giving Researchers Credit for Their Data

Neil Jefferies ^{1,*} , Fiona Murphy ², Anusha Ranganathan ³ and Hollydawn Murray ⁴¹ Bodleian Libraries, University of Oxford, Oxford OX2 0EW, UK² Department of Meteorology, University of Reading, Reading RG6 6AS, UK; f.murphy@reading.ac.uk³ Digital Nest Ltd., Oxford OX1 3LE, UK; anusha@digitalnest.co.uk⁴ F1000Research, London W1T 4LB, UK; hollydawn.murray@f1000.com

* Correspondence: neil.jefferies@bodleyan.ox.ac.uk

Received: 1 March 2019; Accepted: 23 May 2019; Published: 27 May 2019



Abstract: Initially funded as part of the Jisc Data Spring Initiative, a team of stakeholders (publishers, data repository managers, coders) has developed a simple workflow to streamline data paper submission. Metadata about a dataset in a data repository is combined with ORCID metadata about the author to automate and thus greatly reduce the friction of the submission process. Funders are becoming more interested in good data management practice, and institutions are developing repositories to hold the data outputs of their researchers, reducing the individual burden of data archiving. However, to date only a subset of the data produced is associated with publications and thus reliably archived, shared and re-used. This represents a loss of knowledge, leading to the repetition of research (especially in the case of negative observations) and wastes resources. It is laborious for time-poor researchers to fully describe their data via an associated article to maximise its utility to others, and there is little incentive for them to do so. Filling out diverse submission forms, for the repository and journal(s), makes things even lengthier. The app makes the process of associating and publishing data with a detailed description easier, with corresponding citation potential and credit benefits.

Keywords: data papers; publication; open data; credit

1. Introduction

Data papers (Newman and Corke [1]) are overwhelmingly not only open themselves but also based on open data held in repositories, unlike more conventional publications. However, they currently represent a relatively small proportion of research outputs, even though they are arguably often disproportionately more useful to the community in terms of reproducibility of work, quality control of data (through peer review), dissemination of techniques and capture of negative results to prevent repetition (see Kratz & Strasser, and references therein [2]).

At the same time, data papers and data journals are becoming more interesting to the establishment. In June 2016, Earth System Science Data became the first data journal to achieve an impact factor and at 8.286, it already ranks 2nd in Meteorology & Atmospheric Sciences and 3rd in Geosciences, Multidisciplinary. While there are well-articulated arguments for distrusting the Journal Impact Factor as a measurement of worth [3] (which have given rise to a number of responses such as The Royal Society's rationale for a more diverse, nuanced set of metrics [4]), this result still has some significance as it clearly illustrates that data papers are cited by primary research articles in considerable numbers, and so form part of the overall knowledge canon. This has been accompanied by increased publisher interest—possibly in part due to increased understanding of the need to publish the 'whole research story' and partly as data papers' potential revenue-raising opportunities begin to emerge as a result.

Having observed this situation, the project team aimed to drive the deposit of data in repositories and encourage the growth of data papers by simplifying the process through the removal of redundant

metadata entry and streamlining publisher submissions into a single consistent workflow. Thus the goal was to both increase the amount of content held by open data repositories and increase the prevalence of open data papers to comprise a more significant fraction of the research output mix.

This paper describes the project to date, outlines key outcomes and sets out a framework for further development work.

2. Project History

Jisc's 'Research data spring' programme [5] aimed to 'find new technical tools, software and service solutions, which will improve researchers' workflows and the use and management of their data' by using a project model which supported innovative partnerships between stakeholders through a series of small funding calls and pitching sessions. From an initial call for ideas to the final showcase for projects that ran throughout its three phases, it ran from November 2014 to the end of October 2016.

Initially titled 'Giving Researchers Credit for their Data' (now 'Data2Paper'), the idea for the app originated with the Bodleian Library at Oxford (representing a data repository) and F1000Research (representing a publisher), with input at the initial pitching session from an independent data publishing consultant.

For Phase 1, an initial feasibility study was carried out based on a report commissioned by the WDS-RDA Publishing Data Workflows Working Group [6]. This report analysed data paper publication workflows and indicated that mapping DataCite and ORCID metadata to journal submission metadata requirements was a viable approach. The author of the report (Murphy) joined the project team.

A straw-man workflow proposal (see Figure 1) was developed, along with a set of metadata requirements and the resulting survey was circulated amongst data repositories and publishers. Of the repositories and publishers surveyed, 91% and 94% respectively were interested in implementing this tool, establishing both the demand for such a service and the practicality thereof with overwhelmingly positive responses from both camps. The study also provided a useful technical insight to publisher and data repository platforms. As a result of the study a number of respondents indicated a willingness to contribute resources to the next phase. A full report, including the survey results, can be found at Murphy, Jefferies and Ingraham, 2015 [7].

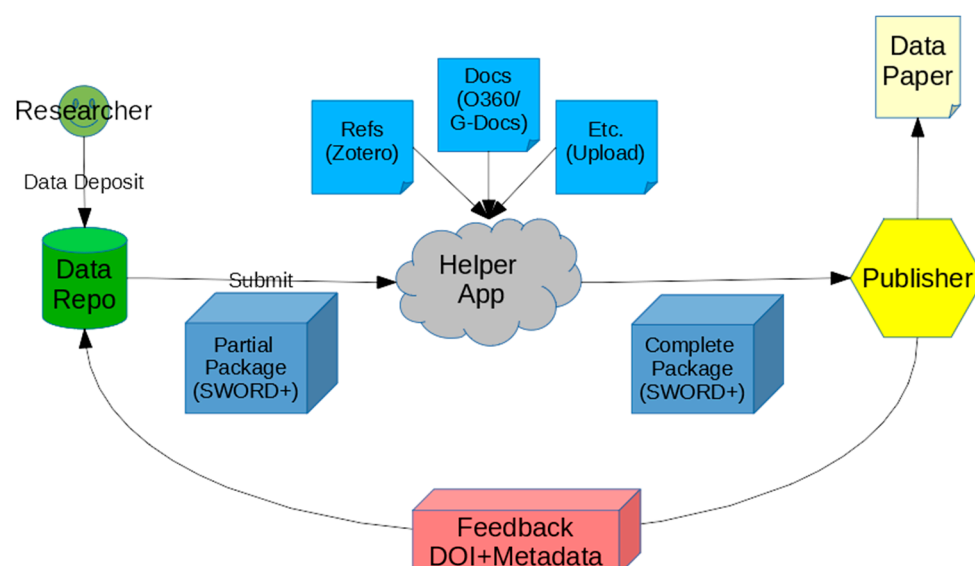


Figure 1. Straw-man Workflow for Helper App (as at Phase 1 pitching session in February 2015).

Following the July 2015 assessment process, a second phase was funded to develop a prototype of the helper app and to demonstrate the linking of a data repository to a publisher's data paper submission system. A more formal specification of the cloud-based helper app was produced along with

a specification for the API for moving metadata packages between systems—for which SWORD V2 [8] was selected both for its functionality and for the fact that it has libraries for a number of languages and repository systems already. In order to participate, a data repository should support DataCite DOI's and metadata and ORCID. Crucially, ORCID provides an identification and authentication system which bridges multiple publishers and repositories.

The prototype helper app was based on Fedora 4/Hydra (and in particular the Sufia Hydra-head) since it essentially acts like a short term repository for material in transit between data repositories and publishers. It is also a scalable platform suitable for more widespread deployment. Development work by Anusha Ranagathan of Digital Nest with code and documentation stored in GitHub and hosting via AWS. Repositories taking part in this proof-of-concept were Oxford's ORA-Data, Figshare and Mendeley Data. Publishers taking part were F1000Research, and Elsevier's Data-in-Brief.

The prototype was presented at the December 2015 Jisc Data Spring event and further work towards a sustainable production service was funded as a Phase 3.

Phase 3 (running from March to October 2016) consisted of building the end-to-end solution, publishing the documentation, demonstrating live publication capability, putting preliminary governance in place and beginning work on a long-term sustainable service model. A wide range of publishers and repositories had expressed interest in the project and its capabilities, but were not in a position to take positive action within the time scale required for Phase 3. Partners and potential partners at this stage included Elsevier/Mendeley Data, University of Manchester, Earth Science Information Partners (ESIP), Data Science Journal, CERN/Zenodo, World Data Center for Climate (WDCC) and University of Edinburgh. During this time hosting migrated from AWS, which was proving somewhat problematic, to an Azure instance provided by Microsoft Research.

Following on from Phase 3, further small grants from Jisc and the Sloan Foundation have kept the project moving forward, if somewhat more slowly than originally planned. Two major factors have contributed to this change of pace:

1. While a number of repositories notionally support DataCite and ORCID, most have not been in a position to update existing content to the new standards, or are only progressing slowly. As such, only newly deposited material meets the requirements for Data2paper, which represents a significantly smaller pool of source material than originally anticipated.
2. Initial research indicated that most publishers used one of a small number of submission management platforms which represented the integration targets for Data2paper. In practice, it transpires that most data journals, and many others, do not make use of these systems and have their own submission workflows, frequently with a significant human element. In some cases, those that *did* use the standard platforms desired to move away from them because of their poor support for data-oriented workflows. This represented a significant barrier to adoption.

Over time, it is expected that both of these situations will improve. DataCite and ORCID support in repositories and publishers is growing steadily and more modern publishing platforms such as F1000Research, Ubiquity Press' Open Journal System and the Collaborative Knowledge Foundation's publishing suite are all gaining industry traction. Data2paper is engaging with all these platforms, so that they should integrate out-of-the-box but team members are also working with the community to advance the state of repositories and publication tools so that Data2paper become more widely deployable.

More tactically, the project has pivoted towards a focus on the BioInformatics community, where the prevalent repository and publishing platforms are better aligned with Data2paper. In particular, BioStudies and GigaScience at the repository end with F1000Research, Wellcome and GigaScience as publishers. This focus provides us with a scholarly community with repositories and publishing platforms that are in a suitable state to be readily integrated with Data2Paper, and will be used to user test and subsequently tune the application, before we seek wider deployment.

3. Outcomes

3.1. Service Vision

The initial aim was to enable researchers to produce accreditable research outputs quickly and seamlessly, incentivising both data paper publication and data deposit in repositories. As we worked with the community, a broader range of stakeholder benefits emerged (Figure 2):

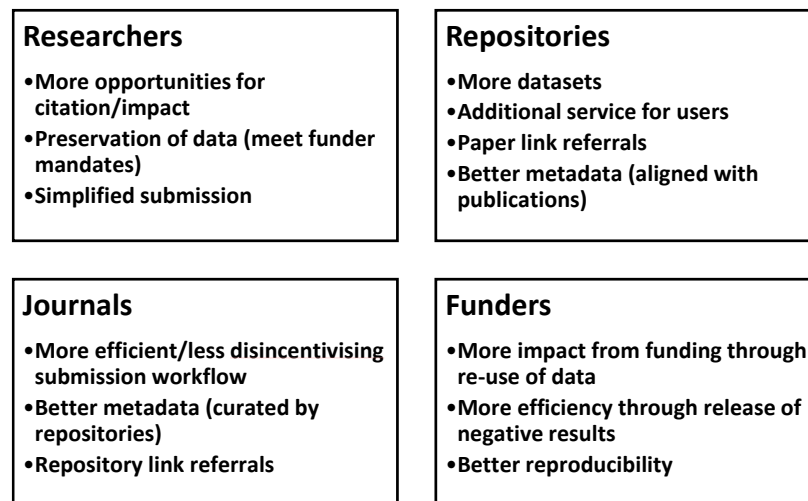


Figure 2. Data2paper Stakeholder benefits.

The service comprises the helper app that resides ‘in the cloud’, providing a consistent, streamlined data paper submission workflow connecting multiple data repositories with multiple publishers. Repositories that wish to connect to the service are provided with an API key and publishers must provide a suitable endpoint. The current workflow is shown below (Figure 3) and represents a simplification and streamlining compared to the initial proposal (Figure 1). SCHOLIX [9] integration as added in 2018 as a result of an OpenAIRE grant and allows the paper details to be fed back to Data2paper and, optionally, the source data repository via the SCHOLIX API. Initially, we had anticipated implementing such functionality within the app but saw no need to re-invent the wheel when SCHOLIX emerged.

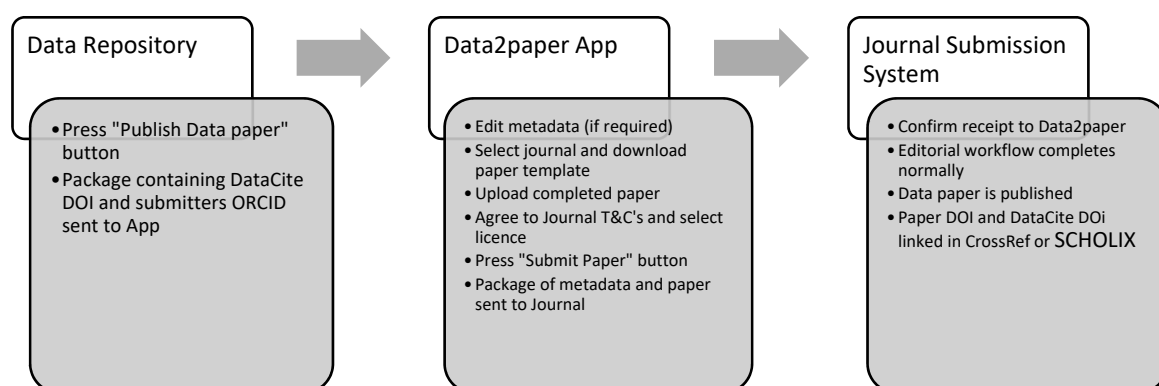


Figure 3. Data2paper Workflow Schematic.

This service is targeted at authors publishing papers that relate to their own data. Current analysis of the data paper environment shows that this workflow accounts for around 90% of data paper publishing activity. However, a logical expansion is to allow re-use of data by others, and the publication of papers that relate to multiple datasets in multiple repositories. The data exchange protocols used by the service and the helper app workflow are designed to allow this to happen.

3.2. Selecting a Journal

A key aspect of the Data2paper App is the Journal selection screen (Figure 4). This exposes Journal details in a standardised format that includes Article Guidelines and Editorial Policy but also highlights such elements as Open Access stance, APC charges and typical turnaround time. By making this information available in an accessible and systematic way, the hope is that authors can make better informed decisions about their publication routes. Maintaining these details falls to the respective journal editors and forms part of the Data2paper integration terms. It is, however, not required that any journal is Open Access.

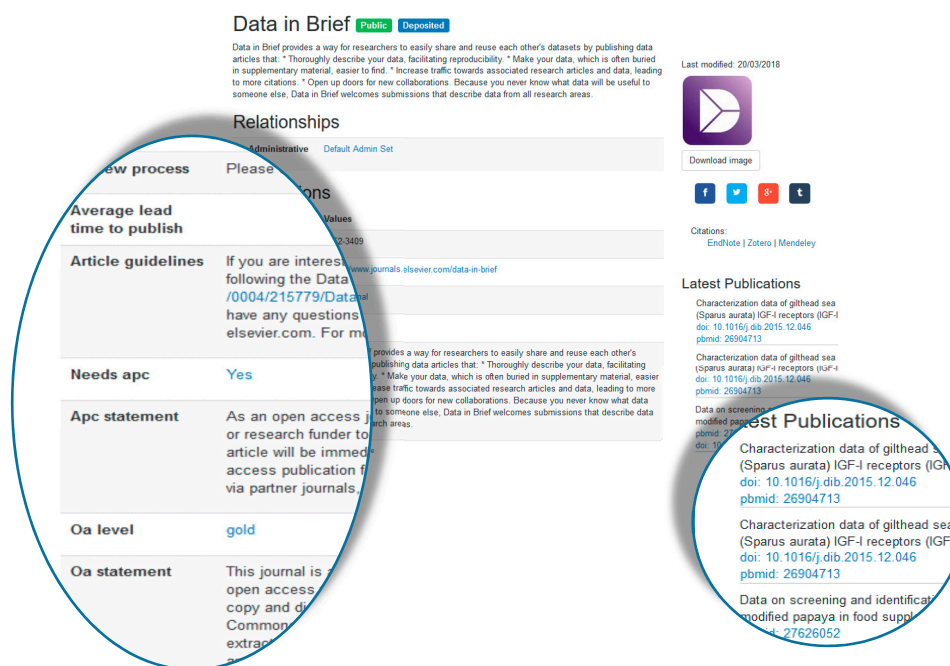


Figure 4. Journal Details in Data2paper: highlighting APC details and also recent publications (via SCHOLIX).

3.3. Online Resources

Data2paper can be found online at <https://data2paper.org/>, which provides links to documentation, the app and many other details. The underlying code can be found on GitHub at <https://github.com/anusharanganathan/data2paper>. During Phase 3, the code was migrated from the Sufia framework to its successor, Hyrax which coincided with some additional features—such as the ability to pass papers in progress to co-authors. Additionally, much work was done to containerise the service for easy deployment across multiple cloud vendors.

4. Conclusions

The long term sustainable business model has not yet been mapped out, bearing in mind the longer path to adoption discussed earlier. In this respect, Data2paper was perhaps premature for the market and the team have modified their approach accordingly. Notwithstanding that, the potential benefits of metadata re-use and process automation in scholarly communications are too great to ignore. Reduced rekeying of information and the consequent savings in authors' time as well as downstream improvements in efficiency and accuracy are essential in an era of shrinking research budgets and increased pressure for impact.

In the purest sense, Data2paper provides a metadata reuse platform that requires a single point of integration for producers and consumers. Repositories and journals need only to maintain a single

integration with Data2paper rather than a multitude of individual point-to-point integrations. Authors see a consistent and simplified submission workflow regardless of repository or journal.

Given the somewhat uneasy relationship between institutions, libraries and publishers, it is expected that an independent not-for-profit model akin to that adopted by ORCID is the most likely business model for Data2paper. To this end, a UK legal entity, Jemura Ltd., has been created in order to run the service. It is a Company Limited by Guarantee that, thus does not have shareholders, and will ultimately be governed by the members that use the service. It is anticipated that members will be at an organisational level and that it would be free at the point of use to researchers.

Author Contributions: Conceptualisation, Funding Acquisition, Project Administration and Writing, N.J.; Investigation, Formal Analysis and Writing, F.M.; Software, A.R.; Investigation, Validation, H.M.

Funding: This research was funded by: JISC, Research Data Spring, 4807, Alfred P. Sloan Foundation, 2017-9874, OpenAIRE, OpenAIRE2020, and Microsoft Research, Azure for Research Grant.

Conflicts of Interest: Ansuha Ranaganathan and Fiona Murphy were paid consultants on this project.

References

1. Newman, P.; Corke, P. Editorial: Data papers—peer reviewed publications of high value data sets (2009). *Int. J. Robot. Res.* **2009**, *28*, 587. [CrossRef]
2. Kratz, J.; Strasser, C. Data publication consensus and controversies [version 3; referees: 3 approved]. *F1000Research* **2014**, *3*. [CrossRef] [PubMed]
3. The San Francisco Declaration on Research Assessment. Available online: <https://sfidora.org/read/> (accessed on 27 February 2019).
4. The Royal Society's Rationale for a More Diverse, Nuanced Set of Metrics. Available online: <http://rspb.royalsocietypublishing.org/citation-metrics> (accessed on 27 February 2019).
5. Jisc Data Spring. Available online: <https://www.jisc.ac.uk/rd/projects/research-data-spring> (accessed on 27 February 2019).
6. Murphy, F.; Dallmeier-Tiessen, S.; Bloom, T.; Nurnberger, A.; Austin, C.C.; Tedds, J.; Khodiyar, V. WDS-RDA-F11 Publishing Data Workflows WG Synthesis FINAL CORRECTED [Data set]. *Zenodo* **2015**. [CrossRef]
7. Murphy, F.; Jefferies, N.; Ingraham, T. Giving Researchers Credit for their Data: Jisc RDS. 2015. Available online: <https://dx.doi.org/10.6084/m9.figshare.1483297.v4> (accessed on 27 February 2019).
8. The SWORD Website. Available online: <http://swordapp.org> (accessed on 27 February 2019).
9. Scholix: A Framework for Scholarly Link eXchange. Available online: <http://www.scholix.org/> (accessed on 29 March 2019).



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).