Research data repositories chosen by researchers across broad range of disciplines, from an analysis of 145,000 data availability statements

Serena C. Tan[1], Dave Flanagan[1], Elisha Morris[1], and Chris Graf[1]

[1]Wiley

July 8, 2020

Abstract

We analysed 145,000 data availability statements (DASs) submitted by research authors to 176 Wiley journals between 2013 and 2020, from the same dataset we previously used to identify the impact of new policies at journals on trends in the use of DASs. We looked at URLs and DOIs contained within those DASs to identify the most common repositories (and other locations) used by researchers to store and share the new research data they create. We resolved DOIs, and captured destination for DOIs (as well as URLs). We mapped destinations to research disciplines, and ranked them to show data services and repositories most often used by researchers who choose to submit to Wiley journals. We share this information as source data and in dynamic figures here, as inspiration and direction for journal teams and research authors.

Introduction

Researchers are creating data, code, and software in previously unimaginable quantities. Data sources are everywhere. Researchers use new tools, like digital notebooks, to compile and version data, to code, and to compile and summarize the methods and results of their work for sharing with others. They have ways to control access to research data, and ways to share it when they're ready to. Some of the researchers that make the most significant impact in both the human sciences and natural sciences are those whose work is data intensive, those that adopt new technology in everything they do, new practices as they go, and those that are active in sharing their data to maximize and accelerate the impact of their work.

Sharing data can, however, come with several challenges and concerns for researchers, and, on the publishing side, for authors and editors of academic research journals. Depending on the discipline, most authors have options for where and how to share their data. They may wonder which repositories are best suited for their data, where others in their community might already be sharing, and also where their data will be the most easily accessible and discoverable, and in turn, able to make the biggest impact. When it comes time to submit their work to a journal for publication, they must also take into consideration which repositories and methods are compliant with data sharing policies at those journals. Journal editors themselves also have concerns. Some wonder whether expectations around data sharing will be difficult for authors to meet and if that will that in turn affect submissions to their journal. Others are eager to promote data sharing policies but require assistance in improving their own knowledge and expertise on the matter to be able to better guide and assist their authors.

Research publishers that don't recognize the fundamental changes towards data intensive research in the human and natural sciences, that don't appreciate the challenges researchers and journal editors face while

trying to make an impact in this changing environment, and that don't offer the support researchers look to publishers for, are taking a significant risk. They miss the opportunity to be part of the research data revolution, the opportunity to support quality and best practices in data sharing, and in turn, they risk becoming increasingly less relevant. Most researchers need support to thrive in this new data intensive world. There is an opportunity right now for publishers to provide support around data sharing for researchers and authors and to drive change through their journals.

In recognition of the challenges faced by authors and journal editors alike in navigating the changing landscape around data, Wiley is developing resources that support researchers who want or need to share the new data they create, including those shared in this preprint (Wu et al., 2019). By doing that we can help lead positive change. To this end, we assessed and analyzed the data sharing behaviors of communities of research authors that publish in Wiley journals. We present the methodology, results, and our analyses here. The results presented here will inform future efforts and plans for developing useful and valuable resources for our author and editor communities. It is our hope that these resources will help facilitate a deeper understanding of best practices around data sharing, reporting standards, and data repositories.

Methods

In our previous report (Graf et al., n.d.), we extracted 124,000 Data Availability Statements (DASs) from the custom questions included for 176 journals in Wiley's electronic editorial office systems between 2013 and 2019. We limited answers to journals that asked custom questions containing the terms "data availability" or "data accessibility".

We used SpaCy's Matcher to identify and extract URLs from the answers to the custom questions for each submission, and tldextract to extract the domains and subdomains from each URL.

In total for this study we used the same data source and method to analyze about 145,000 data availability statements, in which we found about 28,000 with URLs, and of those we were able to resolve about 19,500 to repositories.

For DOI links, we used requests and the doi.org API to resolve the DOI links, and attempted to follow the link to the original source to retrieve the page domain and source. In cases where we were unable to resolve a link (for example, automatically extracted but badly-formed links, or if the connection to the source page timed out), we labelled the domain as "doi.org (unresolved)". It is worth noting that prominent repositories, for example those provided by ebi.ac.uk and nih.gov, use accession numbers not DOIs, and so any accession numbers present in DASs were not included.

Finally, we grouped and filtered domains by journal using the Wiley Online Library hierarchy of subject classifications (Table 1, and data in the associated CSV file), and plotted the results using Plotly. Table 1 shows the WOL Level 1 and Level 2 subject areas that reported the largest number of identified domains for each WOL Level 1 category. "WOL Level 1" is the top-level of that classification hierarchy. "WOL Level 2" is the second-level, and is used to provide more specific information by more discrete description of disciplines and subject areas through our analysis. Each WOL Level 1 group contains multiple WOL Level 2 subject groups (for example, the Life Sciences WOL Level 1 category encompasses subject areas such as Cell Biology, Microbiology, Genetics, and others). Figures 1 and 2 provide benchmarks, showing the numbers of DASs from submitted articles and repositories for WOL Level 1 and 2 categories.

Source code for all analyses is available on GitHub.

WOL Level 1	WOL Level 2
LIFE SCIENCES	ECOLOGY
EARTH & ENVIRONMENTAL SCIENCES	EARTH SCIENCES
MATHEMATICS & STATISTICS	STATISTICS
BUSINESS, ECONOMICS, FINANCE & ACCOUNTING	ECONOMICS
MEDICINE	ONCOLOGY & RADIOTHERAPY
SOCIAL & BEHAVIOURAL SCIENCES	PSYCHOLOGY
AGRICULTURE, AQUACULTURE & FOOD SCIENCE	ACQUACULTURE, FISHERIES & FISH SCIENCE
VETERINARY MEDICINE	VETERINARY MEDICINE
NURSING, DENTISTRY & HEALTHCARE	HEALTH & SOCIAL CARE
PHYSICAL SCIENCES & ENGINEERING	CIVIL ENGINEERING & CONSTRUCTION

Table 1: WOL Level 1 subject groups and the corresponding WOL Level 2 subject with the largest number of identified domains

Figure 1: Number of unique domains (blue) and submitted articles (red) per WOL Level 1 category. (Note that the count is on a logarithmic scale for comparisons.)

Figure 2: Number of unique domains (blue) and submitted articles (red) per WOL Level 2 category. (Note that the count is on a logarithmic scale for comparisons.)

Results

Figure 3 shows the top domains found within DASs analyzed; for each domain, WOL Level 1 categories are shown. Figures 4 through to 8 show domains found for top WOL Level 2 subject categories taken from within the top five WOL Level 1 categories (the data for other categories is in the data files associated with Table 1). Figures 9, 10 and 11 resolve subdomains from the nih.gov URLs.

Figure 3: Top domains found within DASs analyzed

Figure 4: Top domains for Ecology

Figure 5: Top domains for Earth Sciences

Figure 6: Top domains for Statistics

Figure 7: Top domains for Economics

Figure 8: Top domains for Oncology and Radiotherapy

Figure 9: Number of submitted articles per NIH subdomain (e.g., “nda” corresponds to “nda.nih.gov”) per WOL Level 1 category.

Figure 10: Number of submitted articles per NIH subdomain per WOL Level 2 category.

Figure 11: Number of submitted articles per NIH subdomain for Oncology and Radiotherapy.

Analysis and Discussion

In general, our results mirror the results of surveys like the Wiley Open Research Survey 2019 (*What Do Researchers Think About Open Data?*, n.d.), where researchers in the Life Sciences most often report that they share data, with Medicine closely behind.

The source data associated and shared with Table 1 includes two files with domains and repositories for: higher-level WOL Level 1 categories; and for more detailed WOL Level 2 categories. This data can be used to identify commonly-used repositories, as inspiration and direction for journal teams and researchers. Dynamic Figures 3 to 8, discussed below, plot the source data for WOL Level 1 categories and for a selection of WOL Level 2 categories. Figures 9, 10, and 11 offer further detail about DOIs that resolve to the nih.gov domain.

Figure 3 shows the large number of unresolvable doi.org links in submitted DASs. Our sample for this analysis only included DASs that indentified datasets (excluding DASs that report, for example, that data are available on request from the authors and similar). It is reasonable to expect submitted manuscripts to contain errors, including in links provided by authors. These errors may be the reason why so many links that appear in our analysis to be doi.org links do not resolve. Publishers and research authors correct errors and oversights through the publishing process. Providing clear information when it is needed, to support better data citation practices by research authors, is important (*Data Sharing & Citation — Wiley*, n.d.).

The large proportion of unresolvable doi.org links in our data set may have implications for the observations we make below, which may change if we were able to resolve those unresolvable doi.org links.

The largest number of domains identified in Figure 3 come from Life Sciences journals. Among the most highly referenced resolvable domains, Dryad (datadryad.org), the collection of repositories at the NIH (nih.gov), GitHub (github.com), and the general use data repository Figshare (figshare.com) represent the top four highly-used data sharing and storage locations. The top-referenced data sharing domains and resources were referenced largely by researchers from the Life Sciences. It is worth noting that GitHub, Research Gate, and Google appear in our results. These organisations do not currently offer the same commitments to data preservation, access, and citation that specialised repository services offer for research data. It is also noted that the high number of referenced domains from the Life Sciences journals, particularly in Ecology (see Figure 4 above) may also reflect that a large portion of Wiley's Life Sciences portfolio has required data archiving since as far back as 2011.

In Figure 3, the Mathematics & Statistics category is the only WOL Level 1 category where doi.org (unresolved) is not the highest. Here most responses are in github.com. Similar patterns are also seen in the WOL Level 2 categories discussed below, Statistics and Oncology & Radiotherapy. As well as the most DASs, Life sciences also has the largest range of repositories used (22). Physical Sciences & Engineering uses the least, at only 2 repositories.It is clear from Figure 3 that github.com is not only used by computer science researchers but by researchers publishing in a diverse range of disciplines. We speculate that this reflects the increasingly important role across the research spectrum of computation and data science to handle the analysis of big data (*Responsible Data Science is the New Frontier*, n.d.), and the increasingly interdisciplinary nature of research. Figure 3 also shows that in Social & Behavioral Sciences data sharing is dominated by osf.io, the Open Science Framework from the Center for Open Science. Interestingly, osf.io does not describe itself as specifically for social and behavioral scientists; it seems the Center for Open Science's roots in psychology run deep. Other results for top WOL Level 1 categories (Life Sciences; Social & Behavioral Sciences; Earth & Environmental Sciences; Medicine; and Business, Economics, Finance & Accounting) show the success of repositories or domains that exclusively (or predominantly) serve researchers from a particular discipline or subject category; some of these are discussed below.

Figure 4 shows the dominance of datadryad.org in Ecology, as is the case across the Life Sciences: DASs that link to datadryad.org are almost exclusively in the Life Sciences (Figure 3). This speaks to datadryad.org's origins in evolutionary biology and ecology, and to its future across Life Sciences and perhaps beyond: datadryad.org only relatively recently recast itself as a general-purpose data repository.

Figure 5 shows specialist domains serving data sharing in Earth Sciences, namely usgs.gov, noaa.gov, and pangaea.de. Generalist repositories zenodo.org and figshare.com also feature highly, and github.com features when we broaden out from Earth Sciences WOL Level 2 category to consider the Earth & Environmental Sciences WOL Level 1 category in general (Figure 3).

Figure 6 shows that github.com dominates the links found in DASs in Statistics. In fact, github.com scores higher than unresolvable doi.org links for Statistics DASs. This tells us something about precision and familiarity with data sharing in Statistics, where more doi.org links resolve than do not (unlike in Ecology, Earth Sciences, or Economics, Figures 4, Figure 5, or Figure 7).

Figure 7 shows that in Economics worldbank.org is the most common domain. For the first time in this analysis researchgate.net, most commonly referred to as a professional network for science rather than a data sharing service, makes an appearance. As mentioned above, Research Gate does not currently offer the same commitments to data preservation, access, and citation that specialised repository services offer for research data.

Figure 8 reports data for Oncology & Radiotherapy, where top the domains are nih.gov, cancer.gov, iarc.fr, and clinicalstudydatarequest.com. The first three of these domains score higher than unresolvable doi.org links in Oncology & Radiotherapy DASs. Again, as with Statistics above, this tells us something about precision and familiarity with data sharing in Oncology & Radiotherapy where more doi.org links resolve

than do not. Domains like nih.gov host multiple services and repositories to support data sharing; our methods did not deliver information about these.

To restate, the purpose of our study is to highlight repositories used most by researchers across disciplines, rather than to make robust comparisons about the frequency of data sharing between disciplines. Figures 1 and 2 show that our data includes DASs from many more submitted articles in Life Sciences than in any other WOL Level 1 category, and with more links to domains and repositories (28,327 DASs; 7% with links). Following Life Sciences is Medicine (4905 DASs; 11% with links), then Business, Economics, Finance & Accounting (3706; 18%), Earth & Environmental Sciences (2271; 17%) and Social & Behavioural Sciences (2146; 16%). Across all WOL Level 1 categories a mean of 17% of DASs contain links to repositories or data sharing services. The smallest number of DASs included in this study (Figure 1) is from Wiley Physical Sciences journals. The number of domains retrieved and analysed in Physical Sciences is also very small, and includes links to just two domains: nih.gov and usgs.gov. The main reason for these small numbers is the source of our data. Our source data comes from Wiley journals that use one type of editorial office software; many Wiley Physical Sciences journals use a different piece of editorial office software. The Wiley journal portfolio in the Physical Sciences is notably strong and deep, and extraction of suitable data for analysis and further investigation would be valuable.

Conclusion

Our aim with this study was to use data from DASs as a useful resource for researchers who want or need to share new data, by showing them where other researchers choose to share similar data. Similarly, we aimed to support journal editors who want to recommend repositories and data sharing services to authors based on evidence.

We have provided detailed information about frequently used domains and repositories across research disciplines and subject areas, in the Table, Figures, and in the associated data files. These can be used as directional advice and inspiration by research authors and journal editors looking for repositories to use and recommend.

What we have provided should be used in combination with other resources, for example: FAIRsharing.org (*FAIRsharing*, n.d.) and the project led by FAIRsharing.org to identify criteria that matter for repository selection (*FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter*, n.d.); CoreTrust Seal (*CoreTrust Seal*, n.d.); the ELIXIR Core Data Resources (*ELIXIR Core Data Resources — ELIXIR*, n.d.). These resources together will help facilitate a deeper understanding of best practices around data sharing, reporting standards, and data repositories.

Data availability statement

Processed source data are shared associated with Table 1. Beyond that, data from originally submitted DASs are not shared, for the same reasons described in our previous study (Graf et al., n.d.).

Disclosure of conflicts of interest

All authors are employed by Wiley and benefit from the company's success.

References

Paving the Way to Open Data. (2019). *Data Intelligence*, *1*(4), 60–72. https://doi.org/10.1162/dint_a_00021

The open data challenge: An analysis of 124,000 data availability statements, and an ironic lesson about data management plans. Authorea Inc. https://doi.org/10.22541/au.157253515.58528497

https://www.wiley.com/network/researchers/licensing-and-open-access/what-do-researchers-think-about-open-data. https://www.wiley.com/network/researchers/licensing-and-open-access/what-do-researchers-think-about-open-data

https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/index.html. https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/index.html

https://www.wiley.com/network/researchers/latest-content/responsible-data-science-is-the-new-frontier. https://www.wiley.com/network/researchers/latest-content/responsible-data-science-is-the-new-frontier

https://fairsharing.org/. https://fairsharing.org/

https://osf.io/m2bce/. https://osf.io/n9qj7/

https://www.coretrustseal.org/why-certification/certified-repositories/

https://elixir-europe.org/platforms/data/core-data-resources. https://elixir-europe.org/platforms/data/core-data-resources