How are data repositories used to share research data? Preliminary evidence from the Public Library of Science (PLoS) data availability statements

Chenyue Jiao
University of Illinois Urbana-Champaign, United States. cjiao4@illinois.edu

Kai Li
Renmin University of China, China. kai.li@ruc.edu.cn

The 21st century marks the rise of data-driven science, i.e., a new research mode where data plays a central role in the production of scientific knowledge (Hey et al., 2009). This change brings a new requirement that data should be effectively shared to facilitate its reuse and enhance the reproducibility and transparency of sciences (Borgman, 2015). During the past decade, data availability statements have been increasingly embraced by academic communities as a venue to share research data, which is a statement offered by the authors in the publication to specify whether the data underlying the research is shared and if so, how it can be accessed. The Public Library of Science (PLoS) is one of the first major publishers implementing this policy from the beginning of 2014 (Bloom et al., 2014), which was reported to have greatly improved the availability of data in their publications (Byrne, 2017).

As an emerging data source, very few empirical studies have been conducted to examine what information is included in data availability statements and their efficiencies. One notable exception is Federer and colleagues' work (2018) that examined the extent to which PLoS' new data policy was compliant by all PLoS ONE research articles published between March 2014 and May 2016, based on the information from data availability statements. They reported that only 18.2% of publications depositing their data in external data repositories, which is the journal's recommended data sharing method. However, given the short publication window examined by this research, little is known about whether and how this policy was implemented differently over time.

In the present project, we strive to investigate how data repositories are used in PLoS publications, per the data availability statements, and particularly how the use of data repositories is influenced by various factors related to the research and authors. In this presentation, we aim to present preliminary findings of what data repositories are mentioned in PLoS data availability statements from 2014 to 2020 and how the pattern changes over the years.

In this project, we retrieved all 262,895 English-language research articles published in nine PLoS journals up to the end of 2020 by using the *rplos* package of the R programming language (Chamberlain, 2021). Among these publications, 145,717 articles (55.4% of the sample) include a data availability statement, which are analyzed in this research. From all these statements, we identified all 89 data repositories that are: 1) among the twenty most frequently mentioned repositories identified in Federer and colleagues' study (2018) and 2) recommended by PLoS[1]. We search the name variations of all targeted repositories in each sentence of the statement by using regular expression patterns, to identify these repositories from statement texts. To validate our results, one coder manually inspected 500 randomly selected data availability statements (evenly split between those with and without any identified repository). With the information, we updated our regular expression patterns and repeated the procedure. In the updated results, we found that all repositories were correctly identified based on another random selection of 500 statements.

We found that only 27,504 statements (18.87% of all publications with a data availability statement) used external data repositories, a percentage number that is similar to the Federer paper (i.e., 18.2%). However, there is a steady growth of repository mentions in our sample over time. As shown in Table 1, the percentage of publications mentioning external data repositories rises from 13.7% in 2014 to 23.3% in 2020. This indicates that however slow the change is, data repositories are playing increasingly important roles in data sharing in PLoS publications.

[1] https://journals.plos.org/plosone/s/recommended-repositories

Among the 89 targeted data repositories, only 74 of them were found in our corpus. Table 2 shows the times in which each of the top 10 data repositories are mentioned. This list is composed of general-purpose data repositories, such as Figshare and GitHub, and those dedicated to biological and life sciences, such as Gene Expression Omnibus (GEO) and GenBank. Moreover, Figure 1 shows how these 10 data repositories are used over time, whose y-axis represents the percentage of publications in which a specific repository is mentioned in the data availability statement each year. Despite individual differences, there is a general pattern that repositories across knowledge domains, especially GitHub and Zenodo, are more frequently used over the years, as compared to those dedicated to biological and life sciences.

Year	Statements with repository	Total statements	Percentage
2014	1,420	10,385	13.67%
2015	4,650	30,101	15.45%
2016	4,169	24,274	17.17%
2017	4,369	23,124	18.89%
2018	4,220	20,548	20.54%
2019	4,216	18,122	23.26%
2020	4,460	19,163	23.27%
Total	27,504	145,717	18.87%

Table 1: Percentages of statements mentioning data repositories over time

Repository	Statements with repository
Figshare	5,088
GEO	4,106
Genbank	3,202
Dryad	3,139
GitHub	3,117
SRA	2,672
OSF	2,134
Bioproject	1,281
Zenodo	1,133
Dataverse	1,080

Table 2: Numbers of statements mentioning top 10 repositories

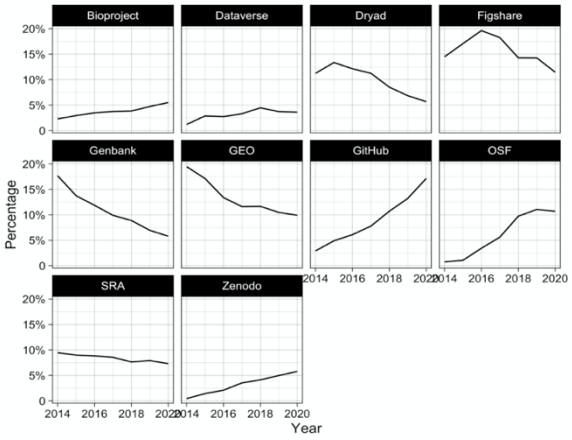

Figure 1: Trends of top 10 data repositories over time

Our project bridges an important gap between quantitative science studies and data studies, by using data availability statements that have been rarely used in empirical research. In our future work, we intend to explore the disciplinary distributions of data repositories by integrating the paper-level subject field classification of PLoS publications, to better understand how data sharing activities are rooted in different disciplinary norms. We also plan to examine various factors that may affect the selection of data repositories by researchers, such as costs, scopes, and services. These efforts will offer up-to-date evidence about how data is shared in scientific publications, a topic that has not been fully investigated in quantitative science studies. More importantly, results from this project will also help researchers to select data repositories for their datasets more effectively and support publishers and funders to develop research policies regarding data sharing and reuse.

References

Bloom, T., Ganley, E., & Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLoS Biology*, *12*(2), e1001797. https://doi.org/10.1371/journal.pbio.1001797

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Byrne, M. (2017). *Making Progress Toward Open Data: Reflections on Data Sharing at PLOS ONE*. Retrieve from: http://blogs.plos.org/everyone/2017/05/08/making-progress-toward-open-data/

Chamberlain, S., Boettiger, C., & Ram, K. (2016). rplos: Interface to the search API for PLoS journals. R package version 0.6.4.

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WC: Microsoft Research.