Why is getting credit for your data so hard?

Wouter Haak[1,*], *Alberto* Zigoni[1], and *Helen* Kardinaal-de Mooij[1], *Elena* Zudilova-Seinstra[1]

[1]Elsevier, Mendeley Data, Radarweg 29, 1043NX Amsterdam, The Netherlands

Abstract

Institutions, funding bodies, and national research organizations are pushing for more data sharing and FAIR data. Institutions typically implement data policies, frequently supported by an institutional data repository. Funders typically mandate data sharing. So where does this leave the researcher? How can researchers benefit from doing the additional work to share their data?

In order to make sure that researchers and institutions get credit for sharing their data, the data needs to be tracked and attributed first. In this paper we investigated where the research data ended up for 11 research institutions, and how this data is currently tracked and attributed. Furthermore, we also analysed the gap between the research data that is currently in institutional repositories, and where their researchers truly share their data.

We found that 10 out of 11 institutions have most of their public research data hosted outside of their own institution. Combined, they have 12% of their institutional research data published in the institutional data repositories. According to our data, the typical institution had 5% of their research data (median) published in the institutional repository, but there were 4 universities for which it was 10% or higher.

By combining existing data-to-article graphs with existing article-to-researcher and article-to-institution graphs it becomes possible to increase tracking of public research data and therefore the visibility of researchers sharing their data typically by 17x. The tracking algorithm that was used to perform analysis and report on potential improvements has subsequently been implemented as a standard method in the Mendeley Data Monitor product. The improvement is most likely an under-estimate because, while the recall for datasets in institutional repositories is 100%, that is not the case for datasets published outside the institutions, so there are even more datasets still to be discovered.

Keywords: research data; data metrics; Scholix; RDM; researcher incentives; institutional data repository; data policy; Mendeley Data; Scopus; data monitor

* Corresponding author: w.haak@Elsevier.com

1 Sharing research data: it starts with tracking

The value of sharing data in the world of research is well known and requires no further elaboration [1, 2]. How data should be shared has subsequently be structured in the FAIR data principles: Findable, Accessible, Interoperable, and Re-useable data [3]. Funding bodies [4] and research communities, like for example the American Geophysical Union (AGU) [5], have subsequently agreed and adopted standard data sharing mandates. While this has led to an overall increase in data sharing [6 and figure 1], there remain a lot of obstacles in rewarding researchers and institutions for sharing their data [7, 8]. In short: there is a lot of 'stick' but very little 'carrot' to incentivise researchers to share their data. We see an increased willingness from researchers to share their data [9], we surely see the effect of data policies but all this is hampered by the inability to track datasets as effectively as we can do with publications, including ensuring that researchers are credited for sharing their data.

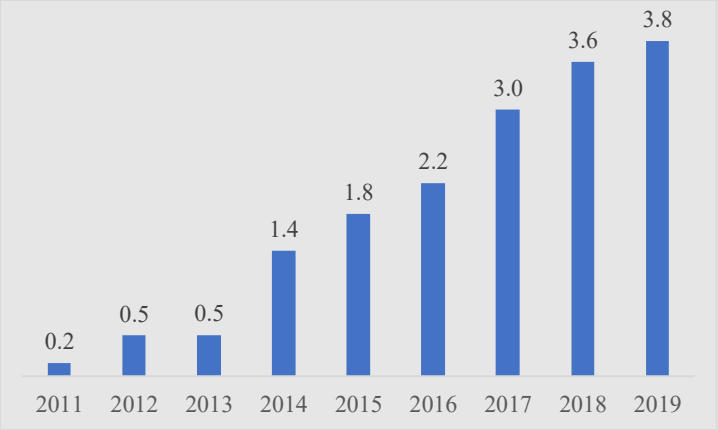

Fig. 1. DataCite DOI registrations [6]

The scientific community has a long-standing track record of using scientometrics as a basis to help policy makers and research administrators to incentivise researchers and to ensure credit is given to researchers. If it can be measured and tracked in a consistent and persistent way, the reward system can be expanded to incorporate metrics around data sharing [10, 11, 12]. One of the most basic fundamentals of tracking data is to be able to count the number of datasets, and be able to attribute them to the right author and institution. A major step forward to solve this puzzle happened when publishers and data repositories joined forces through the Research Data Alliance and World Data System Working Group on the Scholarly Link Exchange (Scholix) iniative. Both publishers and repositories started contributing article-dataset links to central Scholix hubs, like for example DataCite, CrossRef and OpenAire [13, 14, 15].

In this article we have investigated how much benefit research institutions have from combining Scholix information (article-dataset linking) with author and affiliation information (article-author and article-affiliation) from Scopus. [16, 17]

The aim is to compare the existing method of tracking research datasets, typically from an in-house institutional data repository with the broader datasets that have been shared elsewhere. In short: how much better can institutions track research data if they could

combine article-dataset relations such as those provided by Scholix hubs with author and affiliation metadata obtained from bibliographic databases such as Scopus?
Because it is our belief that most datasets end up outside the institution, where they should be, in domain specific repositories or data centers. If they are not tracked, they are not counted and neither the researcher nor the institution will get rewarded.

2. An analysis of research data for 11 institutions

For this analysis, we selected 11 universities, many of which have been our evaluation partners and hence we had an opportunity to check the correctness of the numbers that we collected. Please note that, we don't claim that this analysis is complete and it can be further expanded by adding more universities worldwide. The authors would be happy to expand their analysis to additional universities upon request. Table 1 lists the universities that have been selected.

Table 1. Universities that have been selected for this analysis

Sample:
University of Barcelona
École des Ponts ParisTech
Monash University
Nanyang Technological University
Polytechnic University of Milan
Rensselaer Polytechnic Institute
Università Cattolica del Sacro Cuore
University of Hull
University of Manchester
University of Milan at Bicocca
University of Virginia

For each institution we have counted the number of (public) datasets by counting the number of datasets on their institutional data repository. 10 out of 11 universities had some datasets on their repositories; one university didn't have an institutional data repository.

Table 2. Sample size – counts from March 2020

Sample:	**size**
Nr of institutions	11
# datasets in Institutional Data repositories	1896
# datasets in external data repositories	33,137
Total datasets tracked	**35,033**

In order to derive the total number of datasets for an institution, the following graph analysis was done:

1) We took 14.6 mln dataset corpus available in the Mendeley Data Monitor search index in May 2020 and checked for each dataset if it has an associated article using the Scholxplorer API [16]

2) All articles that belong to a specific institution then have been selected, using the institutional Scival/Scopus IDs [17]
3) Using "isSupplementOf" article-dataset relation, the datasets got attributed to the corresponding institutions.
4) The enriched set of affiliated datasets have been stored in the Mendeley Data Monitor search index [18]

This data is a snapshot taken on May 2020; later snapshots will return larger counts due to increased data sharing. The total number of datasets associated with a researcher or institution grew from 1896 datasets to 35033 datasets.

The result of this analysis for these 11 institutions indicates that they can get credit for their data a lot better if they also track external data repositories where their researchers publish their data. As can be seen in table 3, the improvement is typically 17x versus only focusing on the tracking of research data in their institutional repository.

This is an approximate outcome, where three factors can further refine the analysis. These factors are:

a) The recall for the institutional repository is 100%. Recall for the outside repositories is known to be lower than 100%, because not all data repositories are indexed in Scholix. (there are more datasets out there than we could find or affiliate). Therefore, the number of 26x improvement is expected to be higher still when this recall improves over time in the future.
b) The precision for the institutional repository is 100%. Precision for the outside repositories has three known potential errors. First of all, the relationship of the article to the dataset may be represented incorrectly in the Scholix database. This error is unknown but manual checks have shown it to be negligible for the "IsSupplementTo" relationship that we used in the OpenAire Scholexplorer database. Secondly, the institutional affiliation of the article in the Scopus database may contain errors; again this error is known to be very small (between 1% and 5% of affiliations). Lastly, it is possible that entities that are counted as a 'dataset' are actually not a dataset at all. We have chosen to classify an entity as a dataset if it is classified as a dataset in the Scholix database (OpenAire Scholexplorer); manual checks showed that this error cannot be excluded as some datasets pointed to documents or tombstones rather than actual datasets.
c) We investigated 11 institutions. In order to be representative of the world of research, the sample needs to be larger.

Table 3. Analysis

	Total	Average per institution	Median per institution
% datasets published in Institutional Data repositories	5%	12%	5%
% of datasets published in external data repositories	95%	88%	95%
Improvement in tracking and attribution when all datasets are tracked		41x	17x

The underlying data how the analyses were done can be found in a public dataset on Mendeley Data [19].

3. Conclusion

Our analysis shows that most research data is shared outside research institutions in generalist or domain-specific repositories. The data for 11 institutions that we analysed indicates that institutions can improve their data sharing compliance 14x, if they expand their dataset tracking beyond their institutional repositories. Note that some institutions listed above have already expanded their reporting on datasets to include the external repositories as well.

We also found that automated methods are possible, so the burden on researchers to increase reporting and administrative overhead to manually track and report on their data sharing is not needed.

The tracking algorithm that has been used as the basis for analysis presented in this article, has subsequently been built at scale within the (commercial) Mendeley Data Research Data Management system from Elsevier, under the name of "Data Monitor".

Acknowledgements

We would like to thank the co-chairs of the Research Data Alliance Scholarly Link Exchange working group (Scholix WG): Adrian Burton, Hylke Koers, Martin Fenner, Wouter Haak, Paolo Manghi, and Rachael Lamney

Competing Interests

The authors all work for Elsevier BV, the parent company of Mendeley Data and Mendeley Data Monitor. All metrics and analyses in this article can be independently reproduced by using the following resources:
The Scholix hub from OpenAire: https://www.openaire.eu/scholexplorer
The Scopus API from Elsevier.com: https://dev.elsevier.com/

References

1. Borgman, C. 2012. Journal of the American Society for Information Science and Technology, **63**: 1059–1078. (2012) DOI: https://doi.org/10.1002/asi.22634
2. Piwowar, H and Vision, T. PeerJ, **1** (e175). (2013) DOI: https://doi.org/10.7717/peerj.175
3. Wilkinson, M, et al. Scientific Data **3** 3:160018 (2016) DOI : https://doi:10.1038/sdata.2016.18
4. Digital Curation Centre (dcc.ac.uk): Overview of funders' data policies: http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
5. Stall, S. Copdess.org – Enabling FAIR data project (AGU) (2018): http://www.copdess.org/enabling-fair-data-project/
6. Sample taken from DataCite (May 1, 2020) https://stats.datacite.org/
7. Bierer, B, et al. N Engl J Med, **376**. (2017) DOI: https://doi.org/10.1056/NEJMsb1616595
8. Mongeon, P, et al. Aslib Journal of Information Management, **69**: 545–556. (2017) DOI: https://doi.org/10.1108/AJIM-01-2017-0024
9. Fane, B, What is the State of Open Data in 2019? (2019) DOI: https://dx.doi.org/10.6084/ m9.figshare.10011788

10. Cousijn, H, et al. Data Science Journal, **18: 9**, pp. 1–7. (2019) DOI: https://doi.org/10.5334/dsj-2019-009
11. Cantu-Ortiz, F 2017. Research Analytics: Boosting University Productivity and Competitiveness Through Scientometrics. Boca Raton: CRC Press. DOI: https://doi.org/10.1201/9781315155890
12. Kratz, J and Strasser, C. Scientific Data, **2.** (2015) DOI: https://doi.org/10.1038/sdata.2015.39
13. Burton, A, et al. D-Lib Magazine, **23** (1/2) (2017) DOI: https://doi.org/10.1045/january2017-burton
14. Burton, A, et al. Scholix Metadata Schema for Exchange of Scholarly Communication Links. Zenodo. (2017) DOI: https://doi.org/10.5281/zenodo.1120265
15. Burton, A, et al. Program, **51(1)**: 75–100. (2017) DOI: https://doi.org/10.1108/PROG-06-2016-0048
16. The Scholix hub from OpenAire: https://www.openaire.eu/scholexplorer - sample taken April 2020
17. The Scopus API from Elsevier.com: https://dev.elsevier.com/ - sample taken April 2020
18. https://data.mendeley.com/research-data/ Mendeley Data Monitor search index; available through API from https://datasearch.elsevier.com/api/docs#/search/search_1
19. Zudilova-Seinstra, Elena; Zigoni, Alberto; Haak, Wouter (2020), Mendeley Data, **V2**, doi: http://dx.doi.org/10.17632/k5p45z33kb.2