Geoff Krause
Department of Information Science, Dalhousie University, Halifax, NS, Canada

Timothy Bowman
Wayne State University, Detroit, MI, United States

Domenic Rosati
Scite, Brooklyn, NY, United States

Philippe Mongeon
Department of Information Science, Dalhousie University, Halifax, NS, Canada

Michael Smit
Department of Information Science, Dalhousie University, Halifax, NS, Canada

MEASURING DATA RE-USE IN OPENALEX BY RESEARCHERS, INSTITUTIONS, AND COUNTRIES (Paper)

Abstract:

Open Data is a concept that is receiving increased attention and support in academic environments, with one justification being that shared data may be reused in further research. But what evidence exists for such reuse, and what is the relationship between the producers of shared datasets and researchers making use of them? This work in progress makes use of dataset citations in the OpenAlex bibliometric database to analyze the relationship between the creators of datasets and authors who cite them, at individual, institutional, and national levels.

1. Introduction

The sharing of data used in and produced by research is an increasing trend, and one that supports both the research community and the visibility of researchers' own work (McKiernan et al., 2016; Drachen et al., 2016). It is increasingly common for publishers and funders to require data be made open and available (Jones et al., 2019; Neylon, 2017). When researchers use other authors' findings to support their own, they provide citations to enhance credibility and accountability, and to acknowledge others' work. It would seem to make sense to do the same when re-using data produced by others, yet data citation practices have only recently started to be formalized (Peters et al., 2016) and are not common across all fields (Robinson-García et al., 2016). Nevertheless, some such citations have been recorded by large bibliometric databases such as OpenAlex (Priem et al., 2022). OpenAlex is one of the largest databases of scholarly works' metadata and citations, and makes its data available in a free and open manner; it contains data across a multitude of disciplines, publications, and geographic locales.

Though research data is increasingly being made available, it remains to be seen whether the goal of re-usability is being actualized, and if so, under what circumstances. Through citations, we can compare the authorship of papers citing data to that of the datasets themselves, to determine whether re-use is limited to a dataset's creators, through self-citation, or has spread to

other researchers, institutions, or nations. The particular questions we will seek to address in this project are:

1. How are citations to datasets in OpenAlex distributed, and what proportion are self-citations?
2. How frequently are datasets cited by authors working in the same institution and country as the original dataset creators?
3. What sort of patterns emerge in the citing of datasets across different institutions or countries?

2. Data & Methods

The data used is drawn from OpenAlex, a bibliometric database replacing and building upon the Microsoft Academic Graph (Priem et al., 2022). This data was accessed through a snapshot, circa May 2022, hosted on a PostgreSQL server by the Maritime Institute of Science, Technology and Society. It contains records for over 211 million works and 2.4 billion citations. Processing and analysis of the data were done through custom SQL and R scripts.

Of works available in OpenAlex, 531,299 were identified as datasets. However, over 117,000 of these, more than 20% of all dataset records, had a single author/source[i]; as these likely represent an artefact of OpenAlex's data sources, and no citations to them were present, these were excluded from further analysis, leaving 414,022 datasets. In the data provided through OpenAlex, we find 124,132 citations to the remaining datasets by non-dataset works.[ii]

Self-citations were identified by comparing all authors on both the dataset and citing work, using data extracted from OpenAlex. This includes matches across the OpenAlex AuthorID, ORCID, exact names, and tokenized name groups.

Institution data (including country) is tied to individual authorships rather than works themselves. The completeness of this information is dependent on metadata provided in data sources used by OpenAlex. For the list of citing papers, this is relatively complete, with around 76% of papers having this information available for first authors. However, dataset records are comparatively lacking: only around 30% of all datasets have this information for first-listed creators, while for datasets with citations, institutional links are available for fewer than 3%.

Additional institution data was linked by matching all authorship records (using OpenAlex AuthorIDs and ORCIDs) for works in OpenAlex tied to the same author over a five-year period surrounding creation of datasets. Finally, institutions were added from the 'last_known_institution' field on the author record, where available.

3. Results

Citation data from OpenAlex shows 117,115 non-dataset works citing 14,499 datasets (3.5% of all datasets being examined), around 8.6 citations per dataset. Just over half of these, 7,490, are cited only once, and these single citations account for less than 7% of all citation pairs. The most-cited dataset, the Facial Action Coding System, has 3,346 citations, around 2.7% of all pairs. The top 10% of cited datasets accounts for nearly 75% of all citations. Figure 1 demonstrates the extreme skewedness of citations across datasets.

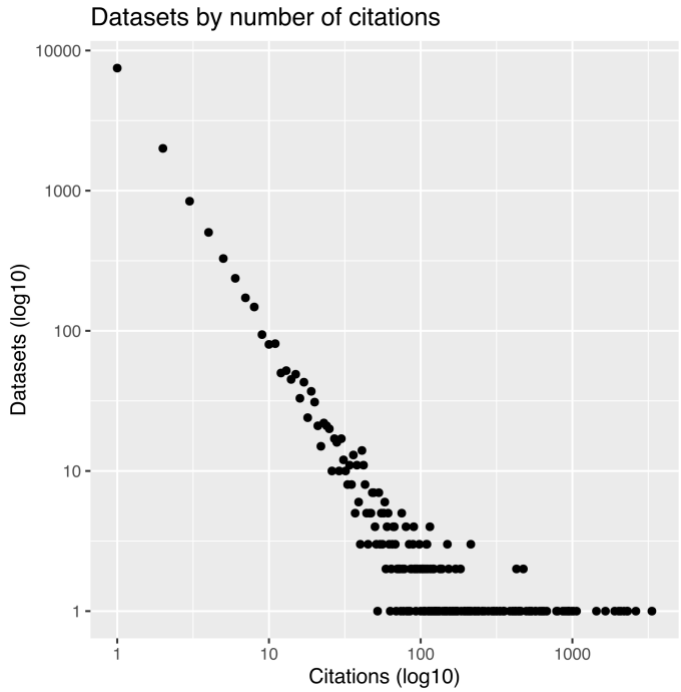

Figure 1: Datasets by number of citations

Author matching resulted in 7,911 identified matches, or just over 7% of all citation pairs. This is far lower than found in certain targeted studies (Dudek et al., 2019). Datasets with self-citations accounted for only 31% of all cited datasets. Self-citations were highly clustered around the less-cited datasets, with much lower proportions in the more-cited datasets, compared to the distribution of overall citations (see Figure 2).

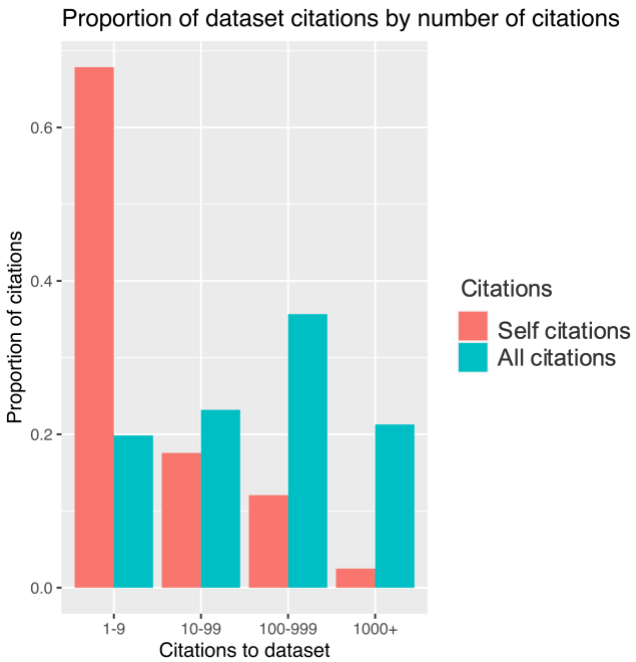

Figure 2: Proportion of dataset citations by number of citations

Of datasets examined, approximately one third have institution data available for the first creator. There are 7,155 creator institutions listed, across 144 countries. US-based institutions are predominant, accounting for 42% of datasets. Germany follows, with 13%, and the UK with 9%. Between them, the top 10 countries produced more than 83% of all datasets.

Position	Institution	Country	Datasets
1	NamesforLife	US	8291
2	Alfred Wegener Institute for Polar and Marine Research	Germany	4836
3	University of Bremen	Germany	2985
4	University of Cambridge	UK	1772
5	Kiel University	Germany	1156
6	Center for Advanced Study in the Behavioral Sciences	US	1101
7	University of Southampton	UK	943
8	University of Washington	US	938
9	French National Centre for Scientific Research	France	915
10	University of Michigan - Ann Arbor	US	865

Table 1: Top 10 producers of datasets overall

Position	Country	Datasets	%
1	United States	57832	42.14
2	Germany	18189	13.25
3	United Kingdom	12458	9.08
4	Australia	4669	3.40
5	Canada	4564	3.33
6	China	3574	2.60
7	France	3548	2.59
8	Netherlands	3368	2.45
9	Spain	2846	2.07
10	Italy	1997	1.46

Table 2: Top 10 dataset-producing countries

Institution data is available for 8,188 cited datasets in OpenAlex, representing 1,915 institutions in 93 countries. There is some overlap with top data-producing institutions and countries overall, and producers of cited datasets.

Position	Institution	Country	Datasets
1	Alfred Wegener Institute for Polar and Marine Research	Germany	293
2	University of Cambridge	UK	132
3	University of Bremen	Germany	104
4	University of Edinburgh	UK	96
5	University of Minnesota	US	89

6	University of Queensland	Australia	81
7	Max Planck Society	Germany	76
8	University of Southampton	UK	74
9	GEOMAR Helmholtz Centre for Ocean Research Kiel	Germany	61
10	Harvard University	US	55

Table 3: Top 10 producers of cited datasets

Position	Country	Datasets	%
1	United States	3408	41.74
2	Germany	1063	13.02
3	United Kingdom	904	11.07
4	Australia	320	3.92
5	Canada	320	3.92
6	Netherlands	225	2.76
7	France	197	2.41
8	Spain	196	2.40
9	Italy	130	1.59
10	Switzerland	120	1.47

Table 4: Top 10 cited dataset-producing countries

There are a total of 69,551 citation pairs where both citing work and dataset have institution information for the first author, representing 56% of citations. Of these, 3,494, or about 5%, involve citing works and datasets produced at the same institution. This is a lower rate than for self-citation. Possible causes may include authors or data creators not having institution data associated with them, authors collaborating with data creators across institutions, or researcher mobility between institutions. Institutions outside the US are well-represented here, as are institutions other than universities.

Overall, 743 institutions have matched citation pairings, though only 432 have more than one such citation. The top ten institutions represent nearly 19% of such pairings.

Position	Institution	Country	Citation matches	% (matched institutions)
1	Alfred Wegener Institute for Polar and Marine Research	Germany	148	4.24
2	Max Planck Society	Germany	87	2.49
3	University of Cambridge	UK	64	1.83
4	University of Queensland	Australia	57	1.63
5	University of Minnesota	US	56	1.60
6	University of New Mexico	US	55	1.57
7	National Institutes of Health	US	54	1.55
8	University of Bremen	Germany	48	1.37
9	University of Edinburgh	UK	46	1.32

10	University of Pennsylvania	US	44	1.26

Table 5: Top 10 citation-matched institutions

Of the 69.372 pairs with country information available for both dataset and citing work, 30,751, or 44%, produced matches. The vast majority of these, about 82%, are US-to-US citations. Other countries in the top 10 for matched-country citation pairings are similar to the results for data production. The United States' dominant position points to a higher average rate of citation to locally produced datasets.

Position	Country	Citation matches	% (matched countries)
1	United States	25402	82.61
2	Germany	1300	4.23
3	United Kingdom	1248	4.06
4	Australia	919	2.99
5	Canada	319	1.04
6	France	148	0.48
7	Netherlands	139	0.45
8	Spain	129	0.42
9	Italy	106	0.34
10	Sweden	105	0.34

Table 6: Top 10 citation-matched countries

There are 37,357 pairings of different citing/cited institutions in the data. No single pairing rises to the level of even 0.1% of citations. Pairings involving data produced by governmental or non-profit research organizations appear prominently.

Position	Citing institution	Dataset-producing institution	Citations
1	Johns Hopkins University (US)	National Institutes of Health (US)	63
2	Northwestern University (US)	National Institutes of Health (US)	62
3	University of North Carolina at Chapel Hill (US)	RTI International (US)	61
4	Centers for Disease Control and Prevention (US)	Center For Policy Research (US)	54
5	University of California, San Diego (US)	University of Iowa (US)	51
6	University of Pittsburgh (US)	Linköping University (Swe.)	49
7	King's College London (UK)	University of Iowa (US)	47
8	University of California, Davis (US)	Cornell University (US)	47
9	University of North Carolina at Chapel Hill (US)	Center For Policy Research (US)	46
10	Chinese Academy of Sciences (China)	German Meteorological Service (Ger.)	44

Table 7: Top 10 mixed-institution citation pairs

There were 1,533 pairings of different countries, of which 916 occurred more than once. The United States again shows up frequently, and the top four pairings involve the UK, Canada, Australia, and Germany citing US-produced datasets, and doing so more often than they do their own. The US also appears to cite datasets produced by other countries, including Germany, Sweden, and Canada, more often than those countries cite locally produced datasets.

Position	Citing country	Dataset-producing country	Citations	% (mixed citations)
1	United Kingdom	United States	3022	7.82
2	Canada	United States	2359	6.11
3	Australia	United States	2090	5.41
4	Germany	United States	1783	4.62
5	United States	Germany	1446	3.74
6	China	United States	1414	3.66
7	Netherlands	United States	954	2.47
8	United States	Sweden	954	2.47
9	Italy	United States	835	2.16
10	Spain	United States	785	2.03

Table 8: Top 10 mixed-country citation pairs

4. Discussion

Citations by works to datasets produced at the same institution formed a very small portion, around 5%, of all citations where institution data was available. While dataset and citing paper production were both largely dominated by institutions based in the United States, institution-matched citations appear to be more common amongst non-US institutions. Nevertheless, looking at matches at the country level, US-to-US citations were dominant, accounting for about 82% of country-matched citations. Many countries appear to cite US datasets more frequently than their own, and US-based institutions frequently make more use of other countries' datasets than researchers producing citing works within those countries.

The low rate of both self-citation and institution-matched citations for datasets suggests the possibility that authors of works utilizing data are not citing datasets upon initial use, and that formal data citation is more common when the dataset originates elsewhere, or as part of research separate from that conducted in the citing work. This may in turn indicate more re-use of data than can be captured within currently available data.

The higher rate of citations within the United States, both to datasets produced in-country and by others, may indicate a greater emphasis on data-driven research, in particular a willingness to seek out and acknowledge data from a variety of sources. This contrasts with more frequent citations within institutions, elsewhere. Another possibility is that this represents more well-developed data citation practices. A third possibility is that this represents a somewhat skewed focus on American sources within OpenAlex itself.

Understanding where and how the sharing of data between researchers, institutions, and countries takes place may help to further develop research practices and collaboration. But the effectiveness of such investigations depends strongly upon the available data sources, starting at

individual researchers' willingness to cite their use of datasets, through to bibliometric databases' ability to capture and present this information in a useful fashion.

Re-usability of datasets requires the support of both the researchers producing the data and the LIS professionals working with them. This includes ensuring that datasets are not just made available, but are findable through good metadata practices. The publication of datasets should be seen not merely as a requirement imposed on researchers, but an important research output, with citations counted towards recognition and actively encouraged. And institutions need to assist in promoting the awareness and use of datasets created by their researchers, both internally and to the outside world.

References

Drachen, T. M., Ellegaard, O., Larsen, A. V., & Fabricius Dorch, S. B. (2016). Sharing data increases citations. *LIBER quarterly*, *26*(2), 67–82. https://doi.org/10.18352/lq.10149

Dudek, J., Mongeon, P., & Bergmans, J. (2019). DataCite as a potential source for Open Data indicators. *ISSI*, *2*, 2037–2042.

Jones, L., Grant, R., & Hrynaszkiewicz, I. (2019). Implementing publisher policies that inform, support and encourage authors to share data: Two case studies. *Insights the UKSG Journal*, *32*, 11. https://doi.org/10.1629/uksg.463

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *ELife*, *5*, e16800. https://doi.org/10.7554/eLife.16800

Neylon, C. (2017). Compliance Culture or Culture Change? The role of funders in improving data management and sharing practice amongst researchers. *Research Ideas and Outcomes*, *3*, e14673. https://doi.org/10.3897/rio.3.e14673

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723–744. https://doi.org/10.1007/s11192-016-1887-4

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. https://doi.org/10.48550/ARXIV.2205.01833

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964–2975. https://doi.org/10.1002/asi.23529

[i] These are computational genetics datasets linked through the ENCODE data portal, and created by Dr. Alan Boyle at the University of Michigan – Ann Arbor.

[ii] Citations made from one dataset to another were excluded, as the nature of these is not immediately apparent, and this consists of only a small proportion of the overall citations.