Datasets on DataCite - an Initial Bibliometric Investigation

Anton Ninkov[1,2], Kathleen Gregory[1,2,4], Isabella Peters[3,5], & Stefanie Haustein[1,2,6]

[1] School of Information Studies, University of Ottawa, Ottawa (Canada)
[2] Scholarly Communications Lab, Ottawa/Vancouver (Canada)
[3] ZBW Leibniz Information Center for Economics, Kiel (Germany)
[4] Data Archiving & Networked Services, Royal Netherlands Academy of Arts & Sciences (Netherlands)
[5] Kiel University, Kiel (Germany)
[6] Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Montreal (Canada)

Abstract

Interest in measuring data citation and developing metrics for data is increasing. Despite this interest, basic bibliometric research investigating data sharing, data reuse and data citation practices remains relatively nascent. In this research in progress article, we use the DataCite GraphQL API to gather data for an initial investigation into dataset sharing and reuse as well as consider the current challenges. With over 8 million datasets in DataCite, we look at how datasets are dispersed by publication year, discipline, number of citations, license, institutional affiliation, and language. We find some patterns emerging, such as a recent increase in dataset publishing. However, there are still many limitations to doing this research that are discussed. As well, the future use of DataCite as a resource for doing this research and additional methods of analysis are considered.

Introduction

As data are increasingly becoming recognized scholarly outputs, funders, research managers and publishers are interested in developing data metrics to reflect the usefulness and impact of sharing data (Cousijn et al., 2019). Despite the interest in metrics for data, basic bibliometric research investigating data sharing, data reuse and data citation practices remains an underserved area.

The obstacles for conducting bibliometric research focusing on data are complex, involving decisions made in policies, practices and at the technical level (Borgman, 2016). These factors are compounded by a lack of bibliometric evidence about data sharing and reuse, particularly a lack of standardization in data citation practices. This creates a so-called "vicious circle," where bibliometricians tend not to take data as an object of study, while at the same time such research is required for developing meaningful data metrics and best practices in the field (Morissette, Peters, & Haustein, 2020).

This paper takes the first steps in addressing this vicious circle, presenting a preliminary bibliometric investigation into data sharing and citation practices, using metadata from DataCite (https://datacite.org) as a source. This research in progress article, which is part of a collaborative project involving DataCite and bibliometricians, provides an overview of the current state of the data available from DataCite. While perhaps not as large as other corpuses, DataCite is a relevant resource for bibliometric research as it is: a) not focused on a single discipline, and b) it assigns persistent identifiers (i.e. DOIs) to research data, allowing for a robust tracking of citations. Given documented disciplinary differences in data sharing, reuse and citation practices (Borgman, 2015; Tenopir et al., 2015), and the importance of accounting for disciplinary differences in data metric development (Lowenberg et al., 2019), we pay special attention to the presence (or absence) of information in DataCite about disciplinary domains in our analysis. We conclude our analysis by identifying gaps in the available data from DataCite and highlighting future areas for data-centric bibliometric research.

Background

Data sharing and citation

Data citation, and calls for its standardization, are not new matters of concern (Parsons et al., 2019). Milestones in the development of standards include the Bermuda Principles in 1996, the formation of CrossRef in 1999, and the founding of DataCite in 2009 (Lowenberg et al., 2019). Silvello (2018) extensively analyzes the extant literature on the development of such standards, as well as motivations for data citation current technical systems. The MDC initiative and DataCite have particularly contributed to efforts on the standardization of data citations in recent years (i.e. Fenner et al., 2019), as has the Scholix Framework for Interoperability in Data-Literature Information Exchange (Burton et al., 2017) and recommendations developed within the Research Data Alliance (Rauber et al., 2015). However, these projects focus on data citation infrastructure, not bibliometric research on data citation practices.

At a more granular level, other work analyses the state of data sharing, reuse and citation for individual datasets. Such work highlights the impermanent and untrackable nature of some citations, such as the widespread use of URLs to reference data (Yoon et al., 2019), or the practice of including data references in the body of articles or in acknowledgement sections, rather than in reference lists (Park et al., 2018). A general laissez-faire approach to data citation has been noted in a number of other studies (Fecher et al., 2015), and persists even in cases where recommended citation formats are provided (Belter, 2014).

Disciplinary differences in data sharing and citation remain a recognized yet unsolved problem. Disciplinary norms play an important factor in the willingness to share datasets (Tenopir et al., 2015). Similarly, early work analyzing the Thomson Reuters (now Clarivate) Data Citation Index (DCI) finds that the hard sciences, specifically biomedical fields, account for the majority (80%) of entries (Torres-Salinas et al., 2014). Lowenberg et al. (2019) note that disciplinary differences in the conception of data themselves demand creating discipline-specific usage statistics (i.e., downloads, views). Peters et al. (2016) also caution that statistics derived from data citations must be interpreted in the context of (disciplinary) data sharing practices and norms, finding that 85% of data remain uncited.

DataCite

The case study described in this research in progress article uses DataCite as the source for our bibliometric investigation. DataCite is an international, non-profit organization that has been assigning persistent identifiers, DOIs, for research data and other artefacts since 2009. DataCite is also actively involved in community outreach and provides data management services and support. Institutions become DataCite members to obtain DOIs for their resources. To receive a DOI, members provide DataCite with metadata describing the given data or artefact. This metadata is provided according to a specialized schema consisting of optional, recommended and mandatory fields (see Table 1).

Table 1. Mandatory Fields by DataCite

Field	*Description*
Identifier	The Identifier is a unique string that identifies a resource.
Creator	The main researchers involved in producing the data, or the authors of the publication, in priority order.
Title	A name or title by which a resource is known.
Publisher	The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource.
Publication Year	The year when the data was or will be made publicly available.
Resource Type	A description of the resource.

DataCite also collects citation data that can be accessed through the GraphQL API. There are two ways in which DataCite collects this data: DataCite members provide the information or DataCite learns about the citation from other academic resources, i.e. CrossRef (Garza, 2020). Recent studies examine DataCite's coverage and potential role in the development of scholarly metrics. Robinson-Garcia et al. (2017), for example, highlight the lack of a standardized vocabulary amongst metadata field entries as a hindrance to its utility as a metrics source. Another recent report examining the role of DataCite in open science practices supports these findings (Dudeck et al., 2019). Using a sample of datasets from an ocean science repository, the authors further conclude that data reuse within this sample is limited to a small number of organizations or to reuse by the original data creators. The study we present here fits into this literature, taking a high-level approach to provide a current analysis of the state of data within DataCite. We consider features that have not been examined before (e.g., citation data) presenting an initial step to more detailed future analyses.

Methodology

To gather data on datasets, we have used DataCite's GraphQL API to collect the metadata. Data was collected between April 15, 2021. Because DataCite is constantly having new submissions and uploads, it was important to collect data in a short time period. With the GraphQL API queries saved, future research collecting data to investigate any rapid changes to DataCite can be done quickly.

One important limitation to mention is that at the time of data collection, the DataCite GraphQL API only returned a maximum of 10 items per query. This is a design feature of the GraphQL API to help optimize performance. Because of this 10-item limitation, accessing the large amounts of data that is offered by DataCite is a challenge, and required work arounds both in terms of methodology and research questions (e.g., running multiple specified queries). There are ways, however, to customize queries in accordance with the DataCite team which can be explored in future work.

Initial Findings

At the time of data collection, there are 8,643,593 total datasets indexed in DataCite (7,440,415 records in 2017; Robinson-Garcia et al., 2017). The most frequent language of datasets is English, with more than 50% of all datasets in DataCite classified as such. Of the ten most common languages of datasets, all but one (Thai) are European languages. Of all the datasets found in DataCite, the majority have been published in the last 10 years (2011-2020). This accounts for 86% of the total number of datasets. It should be noted that these dates are for when the data was published, and not when the data was collected or used. In Table 2, the number of published datasets in DataCite by decade are listed and the general trend of a recent increase is noticeable. One observation is that there is a steep increase 1931-1940 from the previous and following decade. This is a result of a specific repository (University College Dublin) uploading a batch of data during this time period. This repository had a great deal of data connected to this time period as a result of a specific project covering 1937-1938.

Table 2. DataCite Datasets Published by Decade

Decade	*Number of Dataset*	*Decade*	*Number of Datasets*
2011-2020	7,129,806	1961-1970	8,766
2001-2010	815,399	1951-1960	544
1991-2000	87,782	1941-1950	212
1981-1990	30,249	1931-1940	48,073
1971-1980	10,694	1921-1930	53

There are currently 535,449 datasets in DataCite that have a discipline specified, ca. 6% of all datasets. This is not enough data to do a thorough investigation of discipline behaviors when it comes to data sharing and citation practices. However, based on personal communication with technical experts at DataCite, there will be over four million datasets that will gain discipline classification in 2021. This will allow for a more robust study of this area. Of the datasets that have a discipline identified, the ten disciplines with the most published datasets are listed in Table 3. For each discipline, the number of datasets with a citation are also listed. It should be noted that there are very few datasets in DataCite that currently have citations. In total, 97,734 of the datasets have at least one citation, ca. 1% of all datasets.

Table 3. Ten Most Common Disciplines Listed for DataCite Datasets

Discipline	*Number of Datasets*	*Number of Datasets with ≥1 Citation*
Biological sciences	289,137	286
Earth + related environmental sciences	89,931	414
Health sciences	77,601	29
Chemical sciences	63,285	9
Computer + information sciences	61,483	52
Clinical medicine	58,702	33
Sociology	39,144	166
Mathematics	32,901	12
Physical sciences	17,660	24
Psychology	15,450	25

The size of the dataset is listed for 2,485,517 datasets, or 28% of the total number of datasets. The reporting of the data for this variable is not standardized which makes its analysis challenging. Authors of datasets input information on size free of formatting. For example, some datasets express size by how many bytes a dataset takes up (e.g. 2GB) while others record this in other ways (e.g., number of rows, number of items). With the lack of standardization for size, there is a limited amount of data that can be provided by DataCite and, therefore, the depth of analysis we can conduct.

A mandatory field when inputting a dataset into DataCite is publisher. This variable, like size, does not have standardized data input, which makes evaluation of the metadata challenging and not something that is available via the GraphQL API. However, there are some general observations that can be made based on internal DataCite data. Global Biodiversity Information Facility is currently the most frequently identified publisher (941,335 datasets) and The Cambridge Structural Database is the second most frequent (889,586). There is a discrepancy between our findings and those reported in Robinson-Garcia et al. (2017) that requires further investigation in future work. Again, with the lack of standardization, there is a limitation to the data provided by DataCite and the analysis that we can conduct.

The license that is assigned to a dataset is important because it has a direct effect on the options to reuse a dataset. In Table 4, we have listed the top 10 most common licenses for datasets in DataCite. These licenses have been listed in order of the year in which the license was most used, and the number of datasets for that year and total datasets are listed. Because there has been such a recent uptick in use of DataCite, it is not that surprising that so many of the licenses have been published 2020. It is interesting to note how CC-BY-3.0 was so frequently used in 2006. This license was released in 2007 which would explain why it would have been so heavily adopted by these datasets.

Table 4. DataCite Datasets Published by License

License	*Year of Most Published*	*Number of Datasets for Most Published Year*	*Number of Total Datasets*
CC-BY-4.0	2020	171,807	625,685
CC-BY-NC-4.0	2020	102,918	214,311
CC-BY-NC-SA-4.0	2020	9,487	15,630
CC-BY-SA-4.0	2020	2,776	5,586
CC-BY-NC-3.0	2020	1,314	4,123
CC-BY-NC-ND-4.0	2020	634	2,378
CC-BY-1.0	2020	188	399
MIT	2018	1,036	4,281
CC0-1.0	2016	32,098	140,342
CC-BY-3.0	2006	55,848	283,489

Discussion

The goal of this research in progress article is to give initial insights into dataset reuse, availability, and sharing behaviors. To examine variables including datasets' age, size, license, publisher, language, citations, and discipline, we have used the DataCite GraphQL API. With that we contribute to the bibliometric meta research on datasets, hoping to encourage other bibliometricians to explore this type of scholarly output in more detail.

In bibliometric studies, discipline is an important scholarly variable which we have found to also need more attention in dataset studies. However, with only 6% of datasets in DataCite currently having a discipline classification, there is not enough data to do a robust analysis of datasets by discipline. More (approximately 50% of total) needed discipline classification will be added to DataCite in 2021. Other possibilities for expanding the amount of discipline data that is available will have to be considered, which could include integrating metadata from other sources to infer disciplines. As well, getting publishers and repositories to deliver data on disciplines in the future by asking them to select from controlled vocabulary (such as those provided by Organisation for Economic Co-operation and Development) could be necessary. We believe that this aspect is a rich field for future bibliometric research.

There are other variables, in addition to the ones presented in this research in progress article, that would be useful for this type of analysis. For example, additional variables on dataset reuse (e.g., downloads), the number of authors/contributors or the countries of origin could reveal insights into the nature of dataset sharing. There are challenges to doing this right now, mostly surrounding not having enough data to study it. With the increasing adoption of DataCite as well as DataCite adding more variables to their API we project that this will become an even more important area of research imminently.

Acknowledgments

The Alfred P. Sloan Foundation (Sloan Grant G-2020-12670) supported this work. We would like to thank Daniella Lowenberg (DL), Kristian Garza (KG) and Martin Fenner (MF). Stefanie Haustein, Isabella Peters, MF, DL, and KG collaborate in the 'Make Data Count initiative'.

References

Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. PLoS ONE, 9(3), e92590. https://doi.org/10.1371/journal.pone.0092590

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world.

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In C. R. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 93–116). De Gruyter. https://doi.org/10.1515/9783110308464-008

Burton, A., Aryani, A., Koers, H., Manghi, P., La Bruzzo, S., Stocker, M., Diepenbroek, M., Schindler, U., & Fenner, M. (2017). The Scholix Framework for Interoperability in Data-Literature Information Exchange. D-Lib Magazine, 23(1/2). https://doi.org/10.1045/january2017-burton

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, *18*(1), 9. DOI: http://doi.org/10.5334/dsj-2019-009

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? PLOS ONE, 10(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. Scientific Data, 6(1), 28. https://doi.org/10.1038/s41597-019-0031-8

Garza, K. (2020). Datacite Citation Display: Unlocking Data Citations. DataCite. https://doi.org/10.5438/1843-K679

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). Open Data Metrics: Lighting the Fire. Zenodo. https://doi.org/10.5281/zenodo.3525349

Morissette, Erica, Peters, Isabella, & Haustein, Stefanie. (2020). Research data and the academic reward system. Zenodo. http://doi.org/10.5281/zenodo.4034585

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. Journal of the Association for Information Science and Technology, 69(11), 1346-1354.

Parsons, M. A., Duerr, R. E., & Jones, M. B. (2019). The History and Future of Data Citation in Practice. *Data Science Journal*, *18*, 52. https://doi.org/10.5334/dsj-2019-052

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

Rauber, A., Asmi, A., van Uytvanck, D., & Proell, S. (2015). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). http://dx.doi.org/10.15497/RDA00016

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. Journal of Informetrics, 11(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

Silvello, G. (2018). Theory and practice of data citation. Journal of the Association for Information Science and Technology, 69(1), 6-20.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE, 10(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Torres-Salinas, D., Martín-Martín, A., & Fuente-Gutiérrez, E. (2014). Analysis of the coverage of the Data Citation Index – Thomson Reuters: Disciplines, document types and repositories. Revista Española de Documentación Científica, 37(1), e036. https://doi.org/10.3989/redc.2014.1.1114

Yoon, J., Chung, E., Lee, J. Y., & Kim, J. (2019). How research data is cited in scholarly literature: A case study of HINTS: How HINTS data is cited in scholarly literature. Learned Publishing, 32(3), 199–206. https://doi.org/10.1002/leap.1213