DataCite as a Potential Source for Open Data Indicators

Jonathan Dudek[1], Philippe Mongeon[2], and Josephine Bergmans[1]

[1] *j.dudek@cwts.leidenuniv.nl, j.e.bergmans@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62A, Leiden, 2333AL (The Netherlands)

[2] *philippe.mongeon@ps.au.dk*
The Danish Centre for Studies in Research and Research Policy (CFA), Aarhus University, Bartholins Allé 7, 8000 Aarhus C (Denmark)

Abstract

Evaluating the impact of sharing research data is essential for comprehending the value of such initiatives in the context of Open Science. This study investigates indicators for both the output and the impact of datasets listed in DataCite. Based on metadata available for a single datacenter and research institute from the ocean sciences, the French IFREMER, originators and (re)users of datasets were collected at the levels of publishers, author affiliations, and authors. The results show that for the indicators considered, the metadata obtainable from DataCite is limited in consistency and completeness and does not allow facilitated comparisons of datasets. Consequently, meaningful and comprehensive insights are not easily generated at this point of time. In regard to measuring the (re)use of datasets, we suggest more sophisticated approaches to pursue in the future.

Introduction

Datasets are important scientific records. Making them accessible for broader audiences not only serves the reproduction of scientific findings but allows conducting further research as well. Finally, datasets can be considered a complimentary form of scientific output. In order to know whether such potential is exploited requires insights into how visible, findable, and traceable datasets are. Measuring the production and sharing as well as the (re)use of datasets, metadata plays a crucial role. Metadata are records created for datasets by the entities storing, collecting, or cataloguing them. Accordingly, an investigation of the metadata to be found in current data infrastructures can reveal how consistently and completely this information is provided, and how well datasets thus are comparable. Here, we focus on DataCite as a source of dataset metadata and use a bibliographic database (Scopus) to identify formal citations to these datasets in the scientific literature.

DataCite is an international non-profit consortium established in 2009 and combines the efforts of public research institutions, funding bodies and publishers towards open research data. The central value brought about by DataCite is to provide an infrastructure for data producing entities to assign persistent identifiers (DOIs, digital object identifiers) to datasets. Alongside DOIs, additional information on datasets is being attributed as metadata and retrievable from DataCite. ("Our Mission", n.d.) As of January 2019, DataCite has listed over 16 million data records, with more than 13 million records enhanced by searchable metadata ("DataCite Statistics", n.d.). How valuable is this metadata for a) understanding the origins of datasets, and b) creating links to other forms of scientific output? Approaching this question, we apply a case-study-like procedure, focusing on metadata for datasets from one single data originator. In doing so, we test two different kinds of indicators: *output* indicators and *impact* indicators. The former aim at obtaining an overview of the variety of contributors to datasets covered in DataCite. The latter investigate the frequency of dataset (re)use and the overlap between creators of datasets and (re)users. Following this step, we evaluate DataCite metadata based on how well those indicators reflect the insights sought for.

Data Sources

Each dataset recorded with DataCite originates from a so-called datacenter. Datacenters are not necessarily the entities exclusively dedicated to preserving data. Instead, the term subsumes data repositories as well as libraries, research centers, and publishers. For this study, we selected a datacenter from the ocean sciences, a field in which research data plays an important role. In addition, the datacenter selected should show some indication of data (re)use (i.e. references to the datasets in the scientific literature). A preliminary inquiry had shown that datasets by the *Institut Français de Recherche pour l'Exploitation de la Mer* (IFREMER) received the most citations of all ocean science datacenters. Hence, we selected it. IFREMER is a French research institute that manages oceanographic databases and designs and implements tools for the observation, experimentation and monitoring of the marine environment. It addresses societal challenges around climate change effects, marine biodiversity, pollution prevention, and seafood quality, and allows the scientific community to have access to the development, management and distribution of large research infrastructures, such as fleets, computational resources, testing facilities, and laboratories. ("The Institute", 2018)

We collected all 186 IFREMER datasets included in the CWTS version of DataCite, which dates to April 2018. As a second source, metadata for IFREMER-datasets was retrieved in manual searches from the repositories those datasets can be accessed at online, following their DOIs. This provided additional data on affiliations of authors of datasets, which are not included in metadata directly obtainable from DataCite. For a detailed discussion of metadata provided by DataCite, we refer to Robinson-Garcia et al. (2017). The IFREMER-datasets in our sample were registered with DataCite beginning in 2014; for 134 (72%) of the datasets metadata is provided in English; metadata for the remaining 52 (28%) datasets is in French.

Indicators

Measuring output

The datasets observed originate from several different entities, which varied depending on the source the datasets were extracted from, i.e. the publishing organisations. Among the points of origin, there are *publishers, authors, principal investigators, custodians, originators, resource providers*, and *affiliations*. However, not all datasets have all those entities assigned. Metadata in French returns even more terms. We focused on three points of origin: *authors*, *affiliations* (of authors), and *publisher*.

Not all datasets originate from IFREMER directly. Instead, various publishers and data repositories act as intermediaries. One of the most pronounced institutions is SEANOE, a publisher of scientific data in the field of marine sciences. ("About SEANOE", n.d.) Altogether, 103 (55%) datasets originate from this publisher (See Figure 1.)

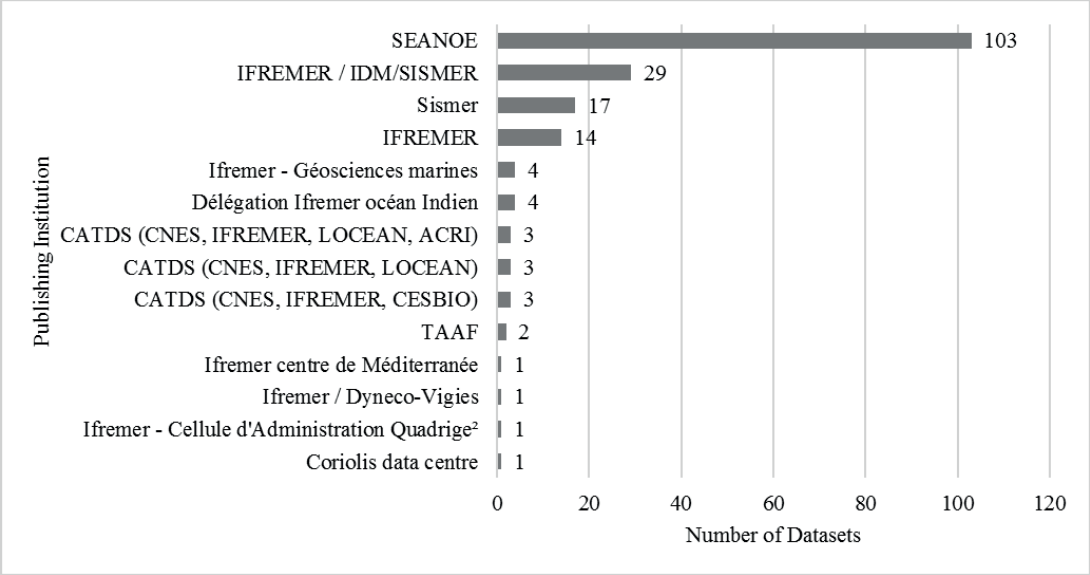

Figure 1: Number of datasets per publisher.

Authors are not necessarily affiliated with the institution serving as the publisher of a dataset. Since many datasets are results of team efforts, author teams with very mixed affiliation backgrounds can be observed. Unsurprisingly, IFREMER is the most prominent affiliation, with 133 authors affiliated to it or to a subsidiary organisation of IFREMER (see Figure 2 for the top ten affiliations of dataset authors).

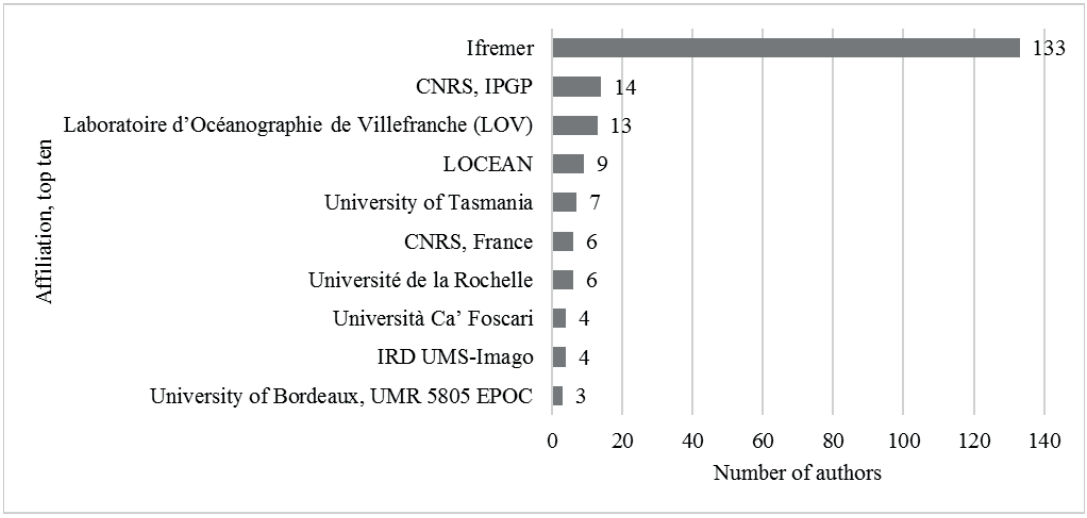

Figure 2. Top ten affiliations of authors.

The *author* field in DataCite usually contains individuals. However, there are 24 cases where organisations are listed as authors. In some of these cases, principal investigators are then provided additionally. As this is not consistently done, for the analysis of authors we focused on any entity called authors, i.e. both individuals and organisations, and did not replace institutional authors with principal investigators.

Accordingly, a total of 280 distinct authors can be identified for the datasets observed. Datasets usually are the result of several contributing investigators, with four authors per dataset on average. 71 datasets share at least one author with another dataset. At the same time, a few authors are highly prevalent, with three of them (co-)authoring more than 50 datasets.

Measuring impact

Regarding impact indicators, we sought empirical evidence of usage of IFREMER datasets by looking at the cited references of all documents indexed in the Scopus database. Overall, we identified 43 such references for a total of 12 distinct datasets. This shows that references to IFREMER datasets are quite rare. Furthermore, those few references are highly concentrated, with one single dataset out of the 12 cited datasets attracting 30 (70%) of all references. Previous work (Park, You, & Wolfram, 2018) has found that (re)used datasets are often not listed in the references, but rather mentioned in the articles' text or acknowledgements. A search for mentions of IFREMER datasets in abstracts of Scopus articles with the two keywords "dataset" and "IFREMER" returned 21 entries. The same keyword search in acknowledgements documented in the Web of Science returned 1,000 entries. This shows that there is a potential for discovering mentions of datasets in abstracts or acknowledgement texts of publications beyond formal citations in publications.

The second part of our investigation of impact aimed to provide an overview of dataset (re)users and their relationship with data producers/creators. In total, 208 different authors were found citing IFREMER datasets (our analysis is limited to formal citations), affiliated to 77 different research organizations. Figure 3 shows the top ten of those organizations.

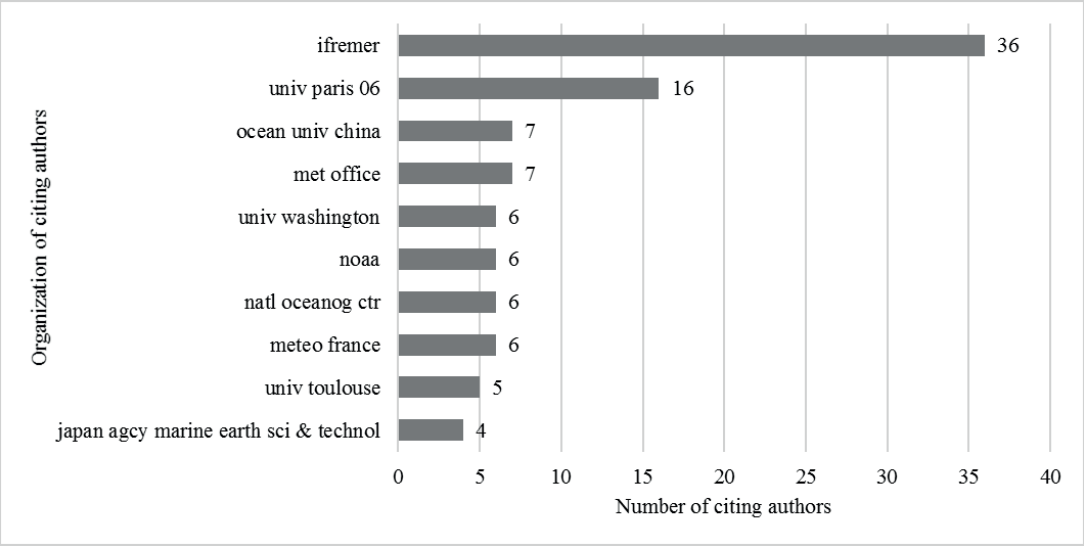

Figure 3. Top ten affiliations of authors citing datasets.

We found that, just like the data producers, the (re)use of datasets is highly concentrated: of all organizations serving as affiliations of citing authors, a small number is responsible for most of the identified instances of data (re)use. In this case, it is IFREMER leading the list, with a total of 36 affiliated authors (17% of all citing authors).

A further analysis investigated the overlap of authors of datasets and citing authors. Nine out of the twelve datasets cited share at least one author with the publication it is cited by; of the 208 unique citing authors, 31 (15%) are also authors of datasets.

From a copyright perspective, (re)use of datasets requires the permission to do so. Most datasets (67%) included in our sample are labelled with a Creative Commons (CC) license, establishing an indicator of potential (re)use. CC-licenses specify in which contexts and how intellectual work can legally be (re)used. ("About The Licenses", n.d.) For the remaining 33% of datasets, licenses are not explicitly stated; however, verbal statements on (re)use possibilities of datasets are provided in almost all cases. Figure 4 shows the share of datasets by license type and language; license types are ordered from the least restrictive (CC0) to the most restrictive (BY-NC-ND). Apparently, the extent to which datasets show a CC-license may partly depend on the language of origin, with 52% of datasets with French metadata having no license at all (compared to only 25% of datasets with English metadata).

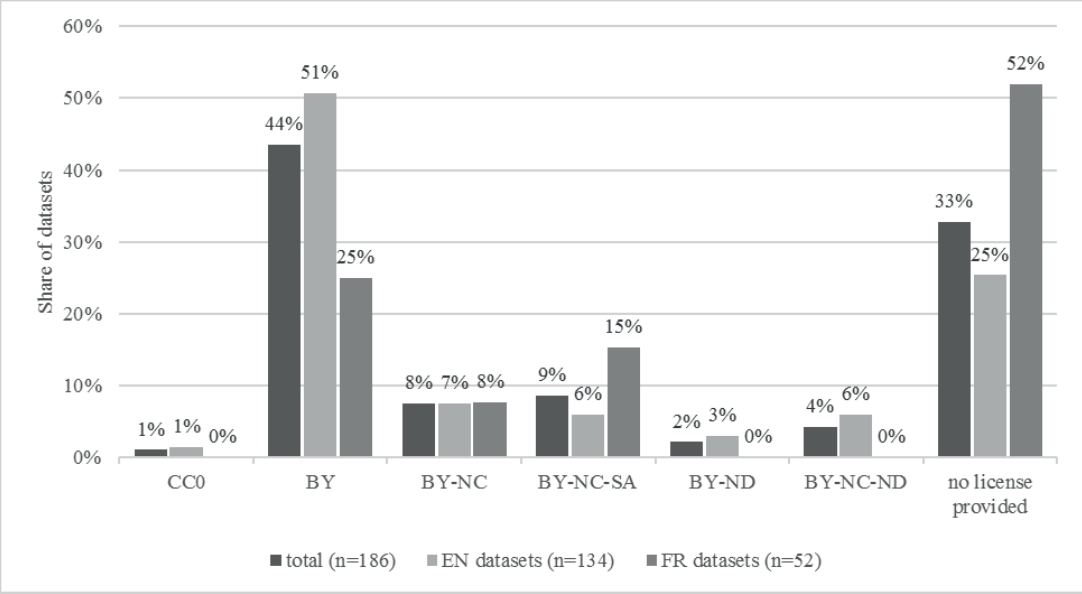

Figure 4. Shares of datasets per CC-licensing type and language of datasets.

Discussion

The study at hand reveals some of the intricacies of generating insights into the origins and the (re)uses of research data based on the metadata available from data infrastructures. Focusing on a subset of datasets originating from a selected datacenter listed in DataCite, we collected the publishers, the authors of datasets, and their affiliations. Further on, we investigated the impact of datasets by measuring counts of citations per dataset, the distributions of citing authors and their respective affiliations, and the overlap of authoring entities and citing entities. A final indicator of (potential) impact were CC-licenses assigned to datasets.

In the course of testing those indicators, the biggest challenge encountered is what we call a lack of metadata control. Herein, the necessity to extract metadata from different sources is a first hurdle: Metadata for the indicators devised is not entirely available from DataCite alone but requires querying publishers' repositories as well (next to a database like Scopus). Secondly, the metadata observed differed in how entities of origin are named and how they are listed, as well as how CC-licenses are assigned. This shows that with DataCite as a single point of access, information cannot be assembled sufficiently – even if only for the same datacenter. Instead, it appears necessary to consider metadata characteristics at the level of publishers.

In the light of the FAIR-principles of data sharing (Wilkinson et al., 2016), a dataset fulfills the requirement of *findability* by being listed in DataCite. However, in order to cover the full range of FAIR-principles for a given dataset (e.g., *reusability*), additional sources need to be included as well. For gathering and comparing datasets, this might constitute a considerable barrier: Depending on the scope of analysis and the point of entry – either starting e.g., an exploration of datasets in DataCite records, or in publisher's repositories; and either comparing datasets across publishers, or only those by a certain publisher – adaptations to different metadata can be necessary. When, as in our case, such dataset origins and usages are to be measured, this barrier becomes even more relevant.

We have shortly mentioned references to datasets beyond what can be found in reference sections (e.g., in abstracts or the acknowledgements) as a further means for estimating the (re)use of datasets. Providing respective metadata would be a worthwhile next step to pursue in addition to reporting citation metrics and serve a better understanding of (re)use, and hence, a dataset's potential for open use. However, both at the levels of DataCite, and of the publishers a consistent framework for reporting such information would need to be set in place. The urgency of this depends on the desirability of indicators of (re)use. Enabling a thorough evaluation of the opening and sharing of research data, though, does require such action.

Our investigation shows that output as well as impact indicators based on DataCite metadata alone do not represent a complete picture, necessitating caution in research evaluation. It should be noted, though, that this conclusion is limited as far as we have focused on one particular datacenter only, from one field of research only. Further research is needed into the data sharing practices of the whole of a scientific field, and then, also, regarding the comparison of different fields. With measures in place to track (re)use of datasets, broader and more general conclusions should become possible. Still, our work shows how the different sources of metadata (can) interact and currently need to be considered when evaluating the state of open data. With DataCite as a major infrastructure provider, fortunately, a central point for enhancing the visibility, comparability and traceability of research data exists. Thus, the necessary foundations for understanding better the origins and (re)uses of datasets may eventually be provided.

Acknowledgements

This study is part of the Open Science Monitor which is funded by the European Commission under grant number PP-05622-2017.

References

About SEANOE. (n.d.). Retrieved January 24, 2019, from https://www.seanoe.org/html/about.htm

About The Licenses (n.d.). Retrieved May 28, 2019, from https://creativecommons.org/licenses

DataCite Statistics. (n.d.). Retrieved January 24, 2019, from https://stats.datacite.org

Our Mission. (n.d.). Retrieved January 28, 2019, from https://www.datacite.org/mission.html

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. DOI: 10.1002/asi.24049

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. DOI: 10.1016/j.joi.2017.07.003

The Institute. (2018). Retrieved January 24, 2019, from https://wwz.ifremer.fr/en/The-Institute

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*. DOI: 10.1038/sdata.2016.18