Aslib Journal of Information Management

Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science

Philippe Mongeon, Nicolas Robinson-Garcia, Wei Jeng, Rodrigo Costas,

Article information:

To cite this document:
Philippe Mongeon, Nicolas Robinson-Garcia, Wei Jeng, Rodrigo Costas, (2017) "Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science", Aslib Journal of Information Management, Vol. 69 Issue: 5, pp.545-556, https://doi.org/10.1108/AJIM-01-2017-0024
Permanent link to this document:
https://doi.org/10.1108/AJIM-01-2017-0024

Incorporating data sharing to the reward system of science

Linking DataCite records to authors in the Web of Science

Philippe Mongeon
École de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal, Montréal, Canada
Nicolas Robinson-Garcia
INGENIO (CSIC-UPV), Polytechnic University of Valencia, València, Spain
Wei Jeng
Department of Library and Information Science, National Taiwan University, Taipei, Taiwan and University of Pittsburgh, Pittsburgh, Pennsylvania, USA, and
Rodrigo Costas
CWTS – Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands and Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Stellenbosch, South Africa

Received 13 January 2017
Revised 4 May 2017
11 May 2017
16 May 2017
Accepted 26 May 2017

Abstract

Purpose – It is widely recognized that sharing data is beneficial not only for science but also for the common good, and researchers are increasingly expected to share their data. However, many researchers are still not making their data available, one of the reasons being that this activity is not adequately recognized in the current reward system of science. Since the attribution of data sets to individual researchers is necessary if we are to include them in research evaluation processes, the purpose of this paper is to explore the feasibility of linking data set records from DataCite to the authors of articles indexed in the Web of Science.
Design/methodology/approach – DataCite and WoS records are linked together based on the similarity between the names of the data sets' creators and the articles' authors, as well as the similarity between the noun phrases in the titles of the data sets and the titles and abstract of the articles.
Findings – The authors report that a large number of DataCite records can be attributed to specific authors in WoS, and the authors demonstrate that the prevalence of data sharing varies greatly depending on the research discipline.
Originality/value – It is yet unclear how data sharing can provide adequate recognition for individual researchers. Bibliometric indicators are commonly used for research evaluation, but to date no large-scale assessment of individual researchers' data sharing activities has been carried out.

Keywords Web of Science, Bibliometrics, Data sharing, Research evaluation, DataCite, Reward system of science

Paper type Research paper

Introduction

The idea that the raw data used in scientific research should be made available to other scholars is more than a century old. When the journal *Biometrika* was founded in 1901, it aimed to publish raw biometric data, as well as the results of future analyses based on these data (The spirit of *Biometrika*, 1901). Raw data available to the public and other researchers were perhaps not easily done in the print era, hence, research data have traditionally been left out of the social contract of scientific publishing (Vision, 2010) as well as from the reward system of science (Costas *et al.*, 2013). Calls for the development of a data sharing culture and

Aslib Journal of Information Management
Vol. 69 No. 5, 2017
pp. 545-556

2050-3806
DOI 10.1108/AJIM-01-2017-0024

its necessary infrastructures have been heard since the 1970s. With the digital era came new possibilities for data sharing and a growing belief that "data should be openly available to the maximum extent possible" (Arzberger *et al.*, 2004, p. 136) and that a reward system that stimulates data sharing should be put in place in the near future (Costas *et al.*, 2013).

Prior works has widely recognized the importance of data sharing (Corti *et al.*, 2014). It reinforces open science (Fienberg *et al.*, 1985), enables the reuse of research data for different purposes (Arzberger *et al.*, 2004), allows a more efficient use of scientific resources (Piwowar, 2011), and provides possibilities for training new students and researchers (Tenopir *et al.*, 2011). Data sharing also promotes transparency (Lyon, 2016), and makes research more valid and rigorous by facilitating reproducibility and promoting a replication culture (Ioannidis, 2014).

Costello (2009) argues that researchers may be compromising scientific development by not giving access to their data to the public. Making data available to the public also has numerous social benefits, such as encouraging citizen science (Kowalczyk and Shankar, 2010), improving public health, and stimulating economic growth (Renolls, 1997). Langat *et al.* (2011) argue that researchers have a moral duty to share their data, especially in the context of public health emergencies: "Data sharing is the morally sound default position" (p. 6). Another aforementioned argument is that, since public funds are often used to cover the cost of data collections, "data ownership can be broadened to include the public that funds it" according to Langat *et al.* (2011, p. 6).

Despite the strengthening of the open science movement, apparent benefits of data sharing, and the increasing mandates by funding agencies and journals, there is still a considerable portion of researchers who withhold their raw data (Andreoli-Versbach and Mueller-Langer, 2014). For instance, Savage and Vickers (2009) issued requests to obtain data from ten articles published in *Public Library of Science* (*PLoS*) *Medicine* and PLoS Medical Trials and were only able to obtain only one in return. Note that Savage and Vickers' work was conducted prior to changes on the journals' data sharing policy (Bloom *et al.*, 2014). It is important to remark also that there are still multiple technical issues that are yet to be resolved (e.g. incompatibilities in machine and software systems, data file structures, data storage, compatibility, access, et cetera (cf. Groves, 2010), and that there are also complex social processes at stake (Arzberger *et al.*, 2004). The relatively low prevalence of data sharing indicates that making data publicly available is more than a simple technicality, and having the technological means to easily share data at minimal costs does not suffice (Borgman, 2012).

Prior studies have investigated barriers, incentives, researchers' perceptions, and mandates for sharing data. Aside from the researchers' sense of responsibility, journals' policy that requires researchers to share raw data when publishing their research is one of the main reasons why researchers do so. Piwowar and Chapman (2008) investigated the data sharing policies of 70 journals and found that researchers more frequently share data when journals have such a policy, and that the probability of sharing data correlates positively with the strength of the policy. Since 2003, the National Institutes of Health (NIH) requires that grant applicants requesting $500,000 or more, must include a data sharing plan in their submission (National Institutes of Health, 2002). Preliminary results of an investigation by Piwowar and Chapman (2010) suggest that receiving funding from the NIH is associated with a higher prevalence of data sharing, especially when the policy applies. In another survey of 1,317 researchers in science, technology, engineering, mathematics, Kim and Stanton (2016) found that normative pressure from the scientific field, the existence of disciplinary data repository and journals' data sharing policies had significant positive effects on the data sharing behavior of researchers.

Barriers to data sharing reported by previous studies include the fear of hindering one's own professional career (Langat *et al.*, 2011), the lack of awareness, the amount of efforts

required (Kim and Zhang, 2015), and the fear that data might be misused. Campbell *et al.* (2002) surveyed 1,240 geneticists and found that the lack of resources and concerns about scientific priority are the main barriers to data sharing. More recently, Tenopir *et al.* (2011) report that many researchers mentioned a lack of institutional support for short and long-term data management. Kim and Stanton (2016) also found that perceived effort was a deterrent to data sharing. Several researchers who refused to share their data in the study by Savage and Vickers (2009) said it was too much effort. Enke *et al.* (2012) found that many researchers had concerns relating to the potential loss of control over the way data are used, the lack of data sharing standards, the amount of time required to make data available, and the lack of acknowledgment received. In fact, one general complaint of scholars is that data sharing is not an element considered for promotion and evaluation (Schäfer *et al.*, 2011) and therefore not part of the reward system. In line with this last observation, many have argued (e.g. Arzberger *et al.*, 2004; Piwowar *et al.*, 2008; Ioannidis, 2014) that data creators, curators, and managers need to be appropriately rewarded if we are to overcome these barriers and make science effectively more open. This lack of reward creates a paradox that can be termed as the "data sharing vicious circle:" scholars do not share their data because they feel they are not rewarded for this, thus limiting the development of "data metrics" (i.e. indicators and reward schemes based on data sharing activities) that would encourage the inclusion of data metrics in the reward system, and the lack of this reward system discourages even further scholars to share their data (cf. Costas *et al.*, 2013).

Sharing data can be time consuming and costly, outweighing the perceived benefits. The current academic reward system focuses mainly on published work and citations to these publications, and does not consider other research activities. The function of the reward system is to maximize the production of knowledge by identifying and rewarding those who best fulfill their role as researchers (Merton, 1957; Cole and Cole, 1973). Therefore, we argue that since sharing research data has become an integral part of the role of researchers, the reward system must evolve to reflect this new reality and bring proper recognition to those who contribute to the advancement of knowledge, as suggested by recent work (Tenopir *et al.*, 2015; Borgman, 2015).

Data papers and data journals can be considered as means to incorporate data sets in the traditional reward system (Ball, 2013). Also, some studies suggest that providing access to data when publishing a paper might have a citation advantage (Piwowar *et al.*, 2007; Piwowar and Vision, 2013). This can be seen as an indirect way of including data sharing in the current reward system. In such cases, the central element remains a published paper, while the raw data are presented as a complementary asset. However, data themselves have a value (for the originally intended purposes or other purposes) that does not depend on their ties to published research, and that value is not recognized in the current reward system. We thus argue that it is necessary to expand it by including data sets (and potential data citations) as complements to current evaluation systems (cf. Costas *et al.*, 2013).

Regardless of how data sharing activities can effectively be incorporated into the reward system of science, we argue that to better understand the different data sharing cultures and practices in academia, an important first step is to link the population of data creators with that of published scientific authors. This is the main objective of this paper: to present a preliminary method to link the data creators included in DataCite to authors of articles published in the Web of Science (WoS). We choose DataCite as our source to identify data creators as this is the most comprehensive source of open data currently available (Peters *et al.*, 2016). By attributing data sets to the authors, we can quantitatively assess the contribution of individual researchers, institutions, and countries to data repositories across disciplines. This is a crucial first step toward the development of data sharing and citation

metrics, and thus toward the inclusion of data sets and data citations in large-scale quantitative assessments of scholars, institutions, and countries' contribution to the production of knowledge.

Data collection

Dataset and cleansing

In April 2016, we downloaded all data sets records[1] published in 2015 from DataCite through their public API (https://api.datacite.org/) and parsed the retrieved records into a relational database. The data comprises 1,059,890 records of data sets and 1,429,298 creator entries. The format of the creator names is inconsistent: some records have multiple authors listed in the same field, using inconsistent delimiters, but most often delimited with commas or semicolons. Other creator fields use the "last name, first name" format, while other use a "first name last name" format. Additionally, some creator names include additional information such as job titles, affiliation, prefixes (e.g. Dr) or suffixes (e.g. PhD). Finally, in some cases the listed creators are not individuals but institutions. Consequently, a first necessary step was to extensively clean and process data to a standardized format. Out of 121,148 distinct creator entries, we were able to identify a total of 111,873 distinct creator names (92.3 percent) on DataCite (see Table I).

For each distinct creator name found, we extracted the last name(s), first name(s), and initial(s) put them into distinct columns (see Table II). For those cases where the field contains more than two names and no comma to distinguish the last name(s) and first name(s), we created a new entry for the same name for each possibility (the case of Ashley Simpson Baird in Table II is an example). In addition, in cases where the full first name is not available, we leave the first name column blank and use only the initials for the matching.

Dataset of researchers. Our scientific publications and authors data set was created by retrieving all publications from WoS over the 2013-2015 period. Researchers have been disambiguated using the algorithm developed by Caron and van Eck (2014). The resulting data set includes 4,520,672 publications and 8,026,780 distinct authors.

Table III shows the format of the WoS author data used to match the authors and creators. It should be noted that first names of authors were included in the WoS author records since 2008 and that even in 2015 still many records do not include the author first name. This is considered in the matching procedure described below.

Table I. Distribution of creator entries by format

Format	Example	Entries *n*	Entries %
First name, last name	John S. Smith | J.S. Smith	73,244	60.5
Last name, first name	Smith, John or Smith, J.S.	38,629	31.9
Other/unidentified format		9,275	7.7
Total		121,148	100.0

Table II. Format of DataCite creator name for matching with WoS authors

ID	Original name	Last name	First name	Initials
1	Herzog, Max Carl	Herzog	Max Carl	MC
2	Rufenach, C.L.	Rufenach	CL	CL
3	Ruf, W.	Ruf	W	W
4	Ashley Simpson Baird	Simpson Baird	Ashley	A
4	Ashley Simpson Baird	Baird	Ashley Simpson	AS

Matching procedure

First, we identified all the potential matches between creators and authors, using three sequential steps:

(1) exact match between creators and authors using their full name;

(2) partial match between creators and authors using their last name and full initials; and

(3) partial match between creators and authors using last name and first initial only.

We identified 96,343 creator-author name matches (see Table IV) accounting for 160,772 data sets (i.e. 46.5 percent of the data sets for which we identified at least a name, and 15.0 percent of the total 2015 data sets). Table IV shows the number of creators with a potential match.

The matching of authors with creators resulted in more than 72 million data set-author links, which, of course, include multiple false positives. Therefore, in the second step of the process, we attempted to eliminate as many of these false positives as possible. To do so, we used multiple elements of the dataset and publication records. First, we extracted noun phrases from all data sets titles and WoS publications (title and abstracts), which we used to calculate the cosine similarity of data sets and publications. The higher the cosine similarity between the paper and the data set, the higher the probability that the creator-author match is a true positive. Second, we counted the number of matched co-creators' and co-authors' names for each data set and publication, and calculated the cosine similarity between the list of creators and authors. We assume that the higher the number of shared co-author and co-creator matches between a data set and a paper, the higher the probability that the creator-author match is a true positive. Third, as described above, some author-creator matches are more precise than others. For instance, a match based on the full first name is more precise than a match based on two initials, which is in turn more precise than a match based on a single initial. Thus, we also assume that the higher the precision of the name match, the higher the probability that it is a true positive.

First, for each data set, we ranked the candidate matches based on each of the following criteria:

- number of creator-author matches;
- cosine similarity between the list of creators of the dataset and the list of authors of the publication;

WoS_name	Last name	First name	Initials
Herzog-MC	Herzog	Max Carl	MC
Rufenach-CL	Rufenach	Clifford L.	CL
Ruf-W	Ruf	Walter	W
SimpsonBaird-A	Simpson Baird	Ashley	A

Table III. Format of WoS author names for matching with DataCite creators

Method	Creators with at least one match	
	n	%
Last name + first name	60,180	53.8
Last name + 3 initials	516	0.5
Last name + 2 initials	14,798	13.2
Last name + 1 initials	16,958	15.2
Last name + first initial	3,891	3.5
Not matched	15,530	13.9
Total	111,873	100.0

Table IV. Results of the author-creator matching procedure

- cosine similarity between the title of the dataset and the title and abstract of the publication; and
- precision of the author-creator match (i.e. full name, three initials, two initials, one initials, first initial only).

We then assign each of the 160,772 data sets to the publication/authors for which the sum of the ranks is the lowest. For example, if the same publication ranks first with for each criterion, then the dataset is attributed to the authors of this publication. While this allows to retain only the publications that are the most similar to each data set, some pairs may still have a very small similarity, especially when the match is based on a single author, and the cosine similarity is null. We thus considered as true positive only pairs with a minimum of two author-creator matches, or with a non-null cosine similarity.

The "relation" field of DataCite records contains information (typically DOIs) of other records that have some relationship with the dataset (e.g. it describes the data set, related data sets, etc.). The "relation" may point to records that are available in DataCite, but also to documents that are not (Robinson-Garcia *et al.*, forthcoming). We used information in this field to assign some of the remaining (non-assigned) data sets to authors. Data sets were thus assigned to the same author as the related documents, provided that this author also appears as a creator in the data set. For example, if data set A has been assigned to author Z, and data set B is related to data set A (as indicated in the "related document" field), then data set B is assigned to author Z, as long as the name of author Z matches the name of a creator of data set B.

Results

We linked a total of 70,701 DataCite records to 60,784 distinct authors from WoS. This represents a relatively small proportion (5.7 percent) of the total records included in DataCite for the year 2015. However, this is not surprising given that not only researchers, but also governments, non-research organization, companies, and other professionals may also be creators of data sets recorded in DataCite. The ten most prolific data producers collectively account for about 5 percent of the attributed data sets. At the top of the list is Gordon Dooley, the Director of the company that developed the Cochrane Register of Studies and a Member of the Cochrane Linked Data Project[2]. We also find a data scientist at Berkeley and two of his collaborators, three botanists from the Canadian Museum of Nature, and three researchers in development biology, public health and environmental science. Furthermore, a manual search on the DataCite website using these researchers' names retrieved a large number of data sets. This provides a certain assurance of the validity of our matching algorithm and of the resulting data set.

Figure 1 displays the distribution of data sets by discipline. The discipline is based on the journal in which the paper that was linked to the dataset was published, and the journal's discipline is based on the National Science Foundation classification. We observe that data sharing is very frequent in biomedical research, biology, clinical medicine and chemistry (25.4, 18.5, 17.4 and 12.2 percent of data sets, respectively), but less so in disciplines of the social sciences (i.e. social sciences, health, psychology and professional fields), which, combined, account for 7 percent of the data sets. A very small number of data sets (0.3 percent) were linked to researchers in arts and humanities.

When focusing on the distribution of the attributed data sets by data centers (Figure 2), we observe that 44,366 (62.7 percent) of the attributed data sets are from Digital Science (figshare), followed by Dryad with 14,142 data sets (20.0 percent). Overall, there were 18 repositories for which we attributed more than 100 data sets to WoS authors. These account for 98.5 percent of the data sets, while the remaining 1.5 percent are dispersed in 101 data centers.

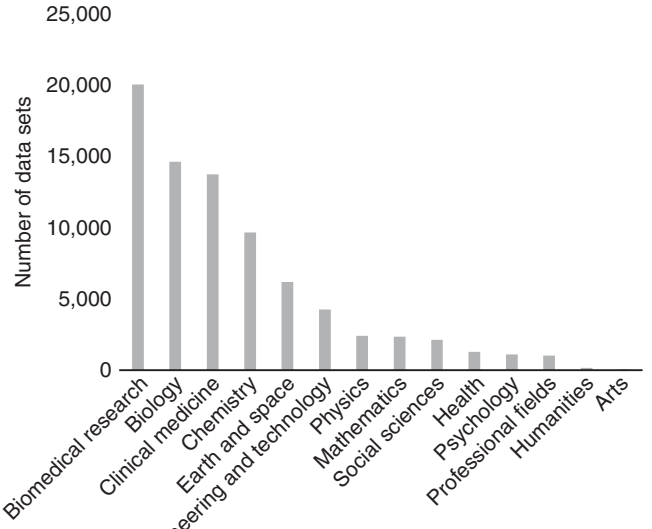

Figure 1.
Number of data sets attributed to WoS authors by discipline

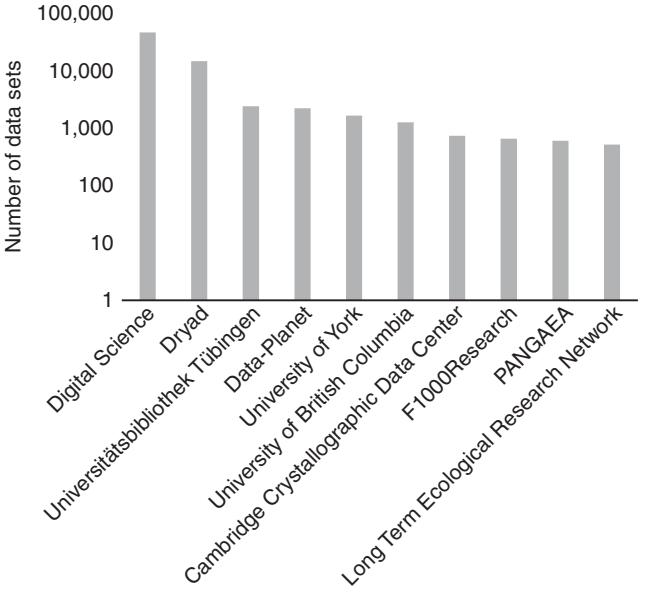

Figure 2.
Number of data sets attributed to WoS authors by data center

Discussion

The results of this study provide useful insights on the possibilities DataCite offers for large-scale automatic attribution of data sets to authors of scientific publications. This has several practical implications.

One-step further toward effectively assessing data sharing practices

First, our method enables us to link data sets to scientific authors without relying on personal identifiers like ORCID. These identifiers would greatly facilitate the attribution of data sets to authors if everyone used them, but this is not currently the case. Another practical application is that our methods are a step forward in the assessment of data sharing practices, the attribution of due recognition to researchers who make their data available, and thus in the provision of additional incentives for data sharing in academia. The methods used in this study could also be used to link other types of data and records (e.g. patents, social media profiles, GitHub records), and help provide a more holistic portrait of scholarly activities and outputs of researchers.

We would like to note that for many DataCite records, there is a lack of information that could be used to link the two sets of records. Indeed, while some records have long titles or many authors listed with their full names, other records have very short title and a single creator with only the last name and initials available. In addition, the lack of affiliation data limits the certainty of the matches. Something also observed in the Data Citation Index (Robinson-Garcia and Torres-Salinas, 2015). Finally, the fact that records registered in DataCite originate from multiple sources leads to a high heterogeneity in the format and quality of the data.

In this study, we make no distinction between record types. We did, however, exclude records with the data type "text" because many of these records are journal articles. However, some other data types may in fact include textual data, and sometimes also full published papers. This is the case, for example, of records from the Universitätsbibliothek Tübingen Data Center, many of which are actually papers, but appear with "collection" data type, which may have influenced the results presented above. This further highlights the potential issues resulting from the inconsistencies of DataCite records.

Discipline unevenness of data sharing: the discipline culture, nature, or infrastructure?

Our study also provides a broad picture of the prevalence of data sharing in the different scientific disciplines. The results show that data sharing seems more prevalent in the life sciences and natural sciences than in the arts, humanities, and social sciences. This is in line with the previous studies that reported that some research fields like genomics, for instance, have long ago developed a data sharing culture as well as the necessary infrastructure (e.g. Anagnostou *et al.*, 2015; Choudhury *et al.*, 2014; Kaye *et al.*, 2009) While the results presented here may indeed reflect the influence of disciplinary cultures and available infrastructures on the propensity of researchers to share their data, we should keep in mind that disciplines may also differ in the way and the extent to which they produce and or use data. There may also exist barriers to data sharing that are out of the researchers' control. For instance, they might be using data that are proprietary (as is it often the case in the field of bibliometrics, for instance) or of a sensitive nature. Thus, one should be careful when interpreting disciplinary differences in the results, as a lower output in data sets does not necessarily mean that there is less sharing occurring, but may also point to such differences in the use of data and in the type of data used. Thus, any assessment of the level of data sharing must take into account what could (or should) have been shared, rather than the raw output.

Integrating data sharing to the current reward system

Thinking about the potential integration of data sharing in the reward system of science requires also a reflection on what actually constitutes a data unit. Are there types of output that should be excluded? For example, records from the Digital Science data center (figshare) are often figures from published papers, and each figure from a paper will be in a separate record. Since figures often are visual representations of a dataset, should they be

considered as data themselves? Should each figures from an academic article be considered as a single data record, or should they be grouped into a single dataset? We will not attempt to provide answers to these questions here, but simply highlight the necessity of taking such factors into account in the interpretation of the results presented above, as well as in future studies.

One limitation comes from the use of publication data from WoS to populate our list of authors and papers for the matching. As Mongeon and Paul-Hus (2016) reported, some the journals in Arts and Humanities and Social Sciences are underrepresented in WoS, and so are journals in languages other than English or published in countries other than the UK, the USA, the Netherlands, Germany, Switzerland, and France. Moreover, WoS does not include books, which are still one of the main means of knowledge dissemination in arts and humanities and, to a lesser extent, in social sciences (Larivière *et al.*, 2006). As a result, it is likely that the results presented in this paper amplify the disciplinary differences in data sharing practices.

Conclusion

This study is a first attempt to link research data creators to scientific authors. The method developed allowed us to attribute 70,701 data sets to 60,784 distinct authors of WoS publications, using mostly basic information such as the creators and authors' names, the titles of the data sets and publications, and the abstract of the publications. Using these linkages, we are able to obtain a broad picture of data sharing activities across all scientific disciplines, showing that data sharing is much more prevalent in some disciplines, like biology and biomedicine, than in others. If we are to broaden out the scope of outputs used in the scientific reward system by including data sets, a first step is to be able to attribute different products to researchers. While the use of author IDs such as the ORCID or ResearcherID intends to establish such linkages, as self-reported tools they still are more of a promise than a reality. In this sense, methodologies such as the one presented here may contribute to respond to such demand.

However, if we are to expand the reward system to recognize data sharing, the attribution of data sets to individual researchers, though it may be a crucial step, may not be sufficient. Indeed, while it may provide indications on the data sharing practices of individuals or groups as well as measures of their data set output, it does not provide information the impact of this output (the extent to which these data sets are reused by others inside and outside of academia). Thus, the next step will be to analyze this impact using, for instance, citations and acknowledgments data. While some efforts have been made in this regard by looking at data citations (Robinson-García *et al.*, 2016) or altmetrics (Peters *et al.*, 2016), our study is expected to contribute to further studies by making it possible to link the reused or cited data sets to the researchers who made them available, and consequently to measure the impact of these individual or groups achieve through data sharing. It will also be necessary, in future work, to reflect on ways to deal with the diversity and heterogeneity of data types. We need to better understand how these different types of data are produced, shared and reused. As we mentioned in the discussion, there may also be certain types of records in DataCite that we may want to exclude from future studies depending of their purpose. In any case, thinking about data production and sharing as monoliths is certainly not the way forward as it would fail to account for the complexity and diversity of practices and outputs.

Being able to, at least, quantify the data sharing activities of individual scholars as recorded in DataCite introduces an important step toward large-scale empirical analyses of data sharing in academia and the development of data sharing metrics which can better recognize responsible practices and open science, ensuring greater transparency and data reuse.

Notes

1. The different data types in DataCite are data set, image, collection, software, audiovisual, film, physical object, event, model, interactive resources, sound, workflow, service, and other (an open-ended entry that allows a free-text input value). Despite the fact that textual documents can often constitute the raw data in some fields of study, in this paper, we broadly define data sets as all DataCite records excluding the data type "text", because these are typically journal articles, conference papers and reports.
2. http://community.cochrane.org/tools/project-coordination-and-support/transform/project-transform-team

References

Anagnostou, P., Capocasa, M., Milia, N., Sanna, E., Battaggia, C., Luzi, D. and Destro Bisol, G. (2015), "When data sharing gets close to 100%: what human paleogenetics can teach the open science movement", *PLoS ONE*, Vol. 10 No. 2, p. e0121409, doi: 10.1371/journal.pone.0121409.

Andreoli-Versbach, P. and Mueller-Langer, F. (2014), "Open access to data: an ideal professed but not practised", *Research Policy*, Vol. 43 No. 9, pp. 1621-1633, doi: 10.1016/j.respol.2014.04.008.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. and Wouters, P. (2004), "Promoting access to public research data for scientific, economic, and social development", *Data Science Journal*, Vol. 3, pp. 135-152, doi: 10.2481/dsj.3.135.

Ball, A. (2013), "Making data count", Digital Curation Centre, available at: www.dcc.ac.uk/blog/making-data-count (accessed April 22, 2017).

Bloom, T., Ganley, E. and Winker, M. (2014), "Data access for the open access literature: PLOS's data policy", *PLoS Biology*, Vol. 12 No. 2, p. e1001797, doi: 10.1371/journal.pbio.1001797.

Borgman, C.L. (2012), "The conundrum of sharing research data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 6, pp. 1059-1078, doi: 10.1002/asi.22634.

Borgman, C.L. (2015), *Big Data, Little Data, No Data: Scholarship in the Networked World*, MIT Press, Cambridge, MA.

Campbell, E.G., Clarridge, B.R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N.A. and Blumenthal, D. (2002), "Data withholding in academic genetics: evidence from a national survey", *JAMA*, Vol. 287 No. 4, pp. 473-480, available at: www.ncbi.nlm.nih.gov/pubmed/11798369

Caron, E. and van Eck, N.J. (2014), "Large scale author name disambiguation using rule-based scoring and clustering", in Ed, N. (Ed.), *Proceedings of the 19th International Conference on Science and Technology Indicators*, CWTS-Leiden University, Leiden, pp. 79-86.

Choudhury, S., Fishman, J.R., McGowan, M.L. and Juengst, E.T. (2014), "Big data, open science and the brain: lessons learned from genomics", *Frontiers in Human Neuroscience*, Vol. 8 No. 239, pp. 1-10, doi: 10.3389/fnhum.2014.00239.

Cole, J.R. and Cole, S. (1973), *Social Stratification in Science*, University of Chicago Press, Chicago, IL.

Corti, L., Van den Eynden, V., Bishop, L. and Woollard, M. (2014), *Managing and Sharing Research Data: A Guide to Good Practice*, Sage, Los Angeles, CA.

Costas, R., Meijer, I., Zahedi, Z. and Wouters, P.F. (2013), "The value of research data metrics for datasets from a cultural and technical point of view. A knowledge exchange report", available at: www.knowledge-exchange.info/event/value-reseach-data-metrics (accessed April 22, 2017).

Costello, M.J. (2009), "Motivating online publication of data", *BioScience*, Vol. 59 No. 5, pp. 418-427, doi: 10.1525/bio.2009.59.5.9.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B. and Gemeinholzer, B. (2012), "The user's view on biodiversity data sharing: investigating facts of acceptance and requirements to realize a sustainable use of research data", *Ecological Informatics*, Vol. 11, pp. 25-33, doi: 10.1016/j.ecoinf.2012.03.004.

Fienberg, S.E., Martin, M.E. and Straf, M.L. (1985), *Sharing Research Data*, National Academy Press, Washington, DC.

Groves, T. (2010), "The wider concept of data sharing: view from the BMJ", *Biostatistics*, Vol. 11 No. 3, pp. 391-392, doi: 10.1136/bmj.b3928.G.

Ioannidis, J.P.A. (2014), "How to make more published research true", *PLoS Medicine*, Vol. 11 No. 10, p. e1001747, doi: 10.1371/journal.pmed.1001747.

Kaye, J., Heeney, C., Hawkins, N., de Vries, J. and Boddington, P. (2009), "Data sharing in genomics: re-shaping scientific practice", *Nature Reviews Genetic*, Vol. 10 No. 5, pp. 331-335, doi: 10.1038/nrg2573.

Kim, Y. and Stanton, J.M. (2016), "Institutional and individual factors affecting scientists' data-sharing behaviors: a multilevel analysis", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 4, pp. 776-799, doi: 10.1002/asi.23424.

Kim, Y. and Zhang, P. (2015), "Understanding data sharing behaviors of STEM researchers: the roles of attitudes, norms, and data repositories", *Library & Information Science Research*, Vol. 37 No. 3, pp. 189-200.

Kowalczyk, S. and Shankar, K. (2010), "Data sharing in the sciences", *Annual Review of Information Science and Technology*, Vol. 45, pp. 247-294.

Langat, P., Pisartchik, D., Silva, D., Bernard, C., Olsen, K., Smith, M., Sahni, S. and Upshur, R. (2011), "Is there a duty to share? Ethics of sharing research data in the context of public health emergencies", *Public Health Ethics*, Vol. 4 No. 1, pp. 4-11, doi: 10.1093/phe/phr005.

Larivière, V., Archambault, É., Gingras, Y. and Vignola-Gagné, É. (2006), "The place of serials in referencing practices: comparing natural sciences and engineering with social sciences and humanities", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 8, pp. 997-1004, doi: 10.1002/asi.20349.

Lyon, L. (2016), "Transparency: the emerging third dimension of open science and open data", *LIBER Quarterly*, Vol. 25 No. 4, pp. 153-171.

Merton, R.K. (1957), "Priorities in scientific discovery: a chapter in the sociology of science", *American Sociological Review*, Vol. 22 No. 6, pp. 635-659.

Mongeon, P. and Paul-Hus, A. (2016), "The journal coverage of Web of Science and Scopus: a comparative analysis", *Scientometrics*, Vol. 106 No. 1, pp. 213-228, doi: 10.1007/s11192-015-1765-5.

National Institutes of Health (2002), "Final NIH statement on sharing research data", available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html (accessed January 4, 2017).

Peters, I., Kraker, P., Lex, E., Gumpenberger, C. and Gorraiz, J. (2016), "Research data explored: an extended analysis of citations and altmetrics", *Scientometrics*, Vol. 107 No. 2, pp. 723-744, doi: 10.1007/s11192-016-1887-4.

Piwowar, H.A. (2011), "Who shares? Who doesn't? Factors associated with openly archiving raw research data", *PLoS ONE*, Vol. 6 No. 7, p. e18657, doi: 10.1371/journal.pone.0018657.

Piwowar, H.A. and Chapman, W.W. (2008), "Identifying data sharing in biomedical literature", *AMIA Annual Symposium Proceedings*, pp. 596-600.

Piwowar, H.A. and Chapman, W.W. (2010), "Public sharing of research datasets: a pilot study of associations", *Journal of Informetrics*, Vol. 4 No. 2, pp. 148-156, doi: 10.1016/j.joi.2009.11.010.

Piwowar, H.A. and Vision, T.J. (2013), "Data reuse and the open data citation advantage", *PeerJ*, Vol. 1, p. e175, doi: 10.7717/peerj.175.

Piwowar, H.A., Day, R.S. and Fridsma, D.B. (2007), "Sharing detailed research data is associated with increased citation rate", *PLoS ONE*, Vol. 2 No. 3, p. e308, doi: 10.1371/journal.pone.0000308.

Piwowar, H.A., Becich, M.J., Bilofsky, H. and Crowley, R.S. (2008), "Towards a data sharing culture: recommendations for leadership from academic health centers", *PLoS Medicine*, Vol. 5 No. 9, p. e183, doi: 10.1371/journal.pmed.0050183.

Renolls, K. (1997), "Science demands data sharing", *British Medical Journal*, Vol. 315 No. 7106, pp. 486-487.

Robinson-Garcia, N. and Torres-Salinas, D. (2015), "Countries and universities rankings of their research output according to Thomson Reuters' citation indexes, 2010-2014", figshare, available at: https://dx.doi.org/10.6084/m9.figshare.1287652.v3

Robinson-García, N., Jiménez-Contreras, E. and Torres-Salinas, D. (2016), "Analyzing data citation practices using the data citation index", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 12, pp. 2964-2975, doi: 10.1002/asi.23529.

Robinson-Garcia, N., Mongeon, P., Jeng, W. and Costas, R. (forthcoming), "DataCite as a novel bibliometric source: coverage, strengths and limitations", *Journal of Informetrics*.

Savage, C.J. and Vickers, A.J. (2009), "Empirical study of data sharing by authors publishing in PLoS journals", *PLoS ONE*, Vol. 4 No. 9, p. e7078, doi: 10.1371/journal.pone.0007078.

Schäfer, A., Pampel, H., Pfeiffenberger, H., Dallmeier-Tiessen, S., Tissari, S., Darby, R., Giaretta, K., Giaretta, D., Gitmans, K., Helin, H., Lambert, S., Mele, S., Reilly, S., Ruiz, S., Sandberg, M., Schallier, W., Schrimpf, S., Smit, E., Wilkinson, M. and Wilson, M. (2011), "Baseline report on drivers and barriers in data Sharing", available at: www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1_0_public_final.pdf (accessed April 22, 2017).

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. and Frame, M. (2011), "Data sharing by scientists: practices and perceptions", *PLoS ONE*, Vol. 6 No. 6, p. e21101, doi: 10.1371/journal.pone.0021101.

Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. and Dorsett, K. (2015), "Changes in data sharing and data reuse practices and perceptions among scientists worldwide", *PLoS ONE*, Vol. 10 No. 8, p. e0134826, doi: 10.1371/journal.pone.0134826.

The spirit of *Biometrika* (1901), *Biometrika*, Vol. 1 No. 1, pp. 3-6.

Vision, T.J. (2010), "Open data and the social contract of scientific publishing", *BioScience*, Vol. 60 No. 5, pp. 330-331, doi: 10.1525/bio.2010.60.5.2.

Corresponding author
Philippe Mongeon can be contacted at: philippe.mongeon@umontreal.ca

This article has been cited by:

1. Paul-HusAdèle, Adèle Paul-Hus, DesrochersNadine, Nadine Desrochers, de RijckeSarah, Sarah de Rijcke, RushforthAlexander D., Alexander D. Rushforth. 2017. The reward system of science. *Aslib Journal of Information Management* **69**:5, 478-485. [Citation] [Full Text] [PDF]