DOI: 10.1002/pra2.316

SHORT PAPERS

WILEY

The role of the data paper in scholarly communication

Chenyue Jiao | **Peter T. Darch**

University of Illinois Urbana-Champaign, Champaign, Illinois, USA

Correspondence
Chenyue Jiao, University of Illinois Urbana-Champaign, Champaign, IL.
Email: cjiao4@illinois.edu

Abstract

Data sharing and reuse promise many benefits to science, but many researchers are reluctant to share and reuse data. Data papers, published as peer-reviewed articles that provide descriptive information about specific datasets, are a potential solution as they may incentivize sharing by providing a mechanism for data producers to get citation credit and support reuse by providing contextual information about dataset production. Data papers can receive many citations. However, does citation of a data paper mean reuse of the underlying dataset? This paper presents preliminary findings from a content-based citation analysis of data papers ($n = 103$) published in two specialized data journals, one in earth sciences and one in physical and chemical sciences. We conclude that while the genre of data papers facilitates some data sharing and reuse, they fail to live up to their full potential. Further, practices of reuse of datasets from data papers vary considerably between disciplines. We propose measures for academic publishers to enhance the data paper's role in scholarly communication to attract more attention from researchers and to inform discipline-specific policy and practices related to data publication.

KEYWORDS

data paper, data sharing, data reuse, data citation, scholarly communication

1 | INTRODUCTION

Increased sharing of research data promises many benefits to science by enabling datasets to be reused by other scholars for purposes including reproducibility and production of new scientific knowledge (Borgman, 2015). However, many researchers are reluctant to share and reuse data (Tenopir et al., 2015). Two factors contributing to this reluctance are a lack of incentives for data producers to share data and insufficient contextual information about dataset production to support reuse (Borgman, 2012; Curty et al., 2017). Data papers, published as peer-reviewed articles in journals, are searchable and citable documents that provide descriptive information on "a particular online accessible data set, or a group of data sets" (Chavan & Penev, 2011, p. 3). This genre of publication may help to address these two factors by providing a mechanism for data producers to get citation credit and by provision of contextual information to support reuse (Kim, 2020; Piwowar & Vision, 2013).

Data papers have often been cited in scholarly communication. For example, the Nature publication *Scientific Data* contains more than 700 data papers that have been cited more than 8,000 times (Nature, 2019). However, does citing a data paper really indicate reuse of the underlying dataset? This study addresses the following questions:

RQ1: For what purposes are data papers cited?

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

RQ2: Do the purposes of citing data papers vary according to scientific discipline?

This paper presents preliminary findings from a content-based citation analysis of data papers ($n = 103$) published in two data journals to assess the extent to which citation of data papers indicates reuse of the underlying dataset. We conclude that data papers fail to live up to their full potential in facilitating data sharing and reuse, and that citation practices vary significantly between scientific disciplines. These findings can inform discipline-specific policies and practices related to data publication.

2 | BACKGROUND

2.1 | Data reuse: Opportunities and challenges

Successful reuse of research data requires both effective sharing practices by dataset producers and effective support of prospective dataset reusers to find, access, interpret, and assess reusability of these datasets (Bishop et al., 2019).

Assessing reusability involves evaluating whether the dataset addresses the phenomena in which the prospective dataset user is interested. The second dimension is whether the dataset is trustworthy (Yoon, 2017). Assessing trustworthiness may involve judging the person or people involved in producing the data (are they competent and honest?), and the appropriateness of the methods used to produce the dataset. When making these evaluations, a prospective dataset user requires contextual metadata describing the context where the dataset was produced (Faniel & Yakel, 2011; Faniel et al., 2019).

Effective sharing on the part of a data producer requires the producer to release the dataset and to provide sufficient contextual metadata to support reuse. However, producers may be unwilling to devote time, effort, and resources to effective data sharing practices because they are uncertain about whether they will receive meaningful rewards if other researchers reuse their datasets (Wallis et al., 2013). Variations exist across disciplines regarding data sharing. For instance, an open and collaborative disciplinary culture can encourage researchers to share data (Kim & Yoon 2017).

2.2 | Data papers: A solution?

Data papers are peer-reviewed articles that provide descriptive information about the data (Chavan & Penev, 2011). The structure and publishing process of data papers vary according to the journal. Kim (2020) found information about methods to produce datasets, repositories where the data are stored, and reuse information regarding terms of use and advice on reuse are elements of data papers often required by journals. For instance, papers in the data journal *Scientific Data* typically include details about methods, data records, technical validation, usage notes, and availability. The journal's peer review criteria focus primarily on the rigor and quality of data collection and the completeness of the data description (Nature, 2014).

Data papers are hailed as potential solutions to the challenges of facilitating dataset reuse (Piwowar & Vision, 2013). First, they may help prospective data reusers find, and assess the reusability of, datasets because by providing contextual metadata (Kim, 2020). Their narrative structure may help the reader to understand how the dataset was produced (Chavan & Penev, 2011). Second, they provide opportunities for dataset producers to receive scholarly credit through peer-reviewed publications and citations (Zhao et al., 2018). These opportunities may incentivize producers to carry out the work necessary to share datasets. However, few studies have been conducted to assess the extent to which data papers in practice live up to their potential. Some have argued that these papers can induce false expectations and fail to provide consistent and complete data-related information (Li et al., 2020; Parsons & Fox, 2013).

3 | METHODS

Content-based citation analysis is the analysis of citation content to explain the "how" and "why" of citation behavior (Ding et al., 2014). We employ content-based citation analysis of articles from two journals that specialize in data papers, *Earth System Science Data (ESSD)*, in the discipline of earth sciences, and the *Journal of Physical and Chemical Reference Data (JPCRD)*, in the disciplines of physics and chemistry. We selected these disciplines because they have well-established data use and citation practices (Silvello, 2018). These two journals were chosen because they have the highest impact factors in their respective disciplines (Clarivate Analytics, 2018). *ESSD* was first published in 2009, while *JPCRD* was first published in 1972. We collected all (n = 103) data papers published in these two journals during 2017 and 2018 (this time period was chosen to strike a balance between enabling the study of recent citation practices and allowing sufficient time for papers to receive substantial numbers of citations) and all

English-language research papers retrieved from Web of Science that cited the data papers up to February 2020 (n = 433). Each citation was analyzed based on the coding scheme shown in Table 1.

All coding was performed by the first author. If researchers integrated the datasets described in a data paper for new analysis, the citation was coded as reuse (3.1) for integration (4.4). If they used the dataset for background information required for the new research, the citation was coded as reuse (3.1) for background (4.1). However, if researchers cite data papers just for providing relevant information that is not required for the new research, the citation was coded as non-reuse (3.2). Most citations in the introduction part are to mention previous studies (5.1) or to provide contextual information (5.2).

TABLE 1 Coding scheme of data paper citation

Code	Description
1. Type of data 1.1. Observational 1.2. Computational 1.3. Experimental 1.4. Records	Refers to the primary type of data in the data paper. This category was developed based on a report of the U.S. National Science Board (2005).
2. Location 2.1. Introduction 2.2. Method 2.3. Result 2.4. Discussion	Refers to in which section the data paper was cited. IMRaD is the most prominent structure of scientific articles (Sollaci & Pereira 2004).
3. Data reuse 3.1. Yes 3.2. No	Refers to whether the data paper was cited for data reuse. If the data is reused for a new purpose other than the original one, code it with 3.1 and continue to 4; if not, code it with 3.2 and continue to 5.
4. Reuse purpose 4.1. Background 4.2. Calculation 4.3. Comparison 4.4. Integration 4.5. Verification 4.6. Other	Refers to the primary reuse purpose of citing data paper. The category of reuse purposes was developed based on Gregory et al. (2019).
5. Non-reuse purpose 5.1. Giving credit for previous studies 5.2. Background 5.3. Same approach 5.4. Implication 5.5. Other	Refers to the primary non-reuse purpose of citing data paper. The difference between 4.1 and 5.2 is for what purpose the background information is used. Background reuse means data usage as supporting materials that is required and can affect the new research. By contrast, background non-reuse just provides trivial information whose deletion would not impact on the new research.

4 | RESULTS

We present the citation practices of data papers in each data journal and the disciplinary distinctions between these two disciplines. Tables 2, 3, and 4 give an overview of data citation and reuse in the two journals.

4.1 | Earth system science data

Sixty-three data papers (90.0%) in *ESSD* were cited at least once for reuse purposes while seven data papers (10.0%) were cited for non-reuse purposes only. Three hundred and ten papers cited these data papers, of which 210 (67.7%) cited for reuse purposes while 100 (32.3%) cited for non-reuse purposes only (Figure 1).

Of a total of 584 citations, 328 (56.2%) indicated reuse of the underlying dataset. The most common form of reuse was integration (31.4%) with other datasets. The percentages of citations for calculation, comparison, background information, and verification are quite similar at 19.2, 16.2, 13.7, and 12.8% respectively (Table 3).

Two hundred and fifty-six citations did not indicate dataset reuse; instead, the most common purposes for citing in these cases were citing previous studies (45.3%) and providing background information (35.2%) (Table 4).

4.2 | Journal of physical and chemical reference data

Of the 33 data papers, 26 (78.8%) data papers were cited at least once for reuse purposes. Of the 56 citing papers, 27 (45.5%) included citations indicating dataset reuse. Of 273 citations, 103 (37.8%) indicated reuse, while 170 did not (62.2%) (Figure 1). Calculation (43.7%), verification (28.2%), and comparison (14.6%) were the most common types of reuse (Table 3). The least common types of reuse were integration (3.9%) and background (1.9%).

Of the 170 citations indicating non-reuse, 103 (60.6%) were for giving credit for previous studies, while 39 (22.9%) were to give credit for methods (Table 4).

4.3 | Disciplinary distinctions

Clear contrasts exist between our earth sciences and our physics and chemistry samples. Data papers in *ESSD* have more reuse than non-reuse citations while those in *JPCRD* have more non-reuse than reuse citations. A particular contrast is integration (31.4% in earth sciences vs. 3.9% in physics and chemistry). The percentage of calculation and verification in physics and chemistry (43.7%

		ESSD			*JPCRD*		
		2017	**2018**	**Total**	**2017**	**2018**	**Total**
Data papers		14	56	**70**	18	15	**33**
Citing papers		94	216	**310**	79	44	**123**
Citation	Reuse	92	236	**328**	63	40	**103**
	Non-reuse	59	197	**256**	95	75	**170**
	Total	**151**	**433**	**584**	**158**	**115**	**273**

TABLE 2 Overview of data citation

TABLE 3 Overview of purposes in reuse citations

Reuse	***ESSD***	***JPCRD***
Background	45 (13.7%)	2 (1.9%)
Calculation	63 (19.2%)	45 (43.7%)
Comparison	53 (16.2%)	15 (14.6%)
Integration	103 (31.4%)	4 (3.9%)
Verification	42 (12.8%)	29 (28.2%)
Other	22 (6.7%)	8 (7.8%)
Total	328	103

TABLE 4 Overview of purposes in non-reuse citations

Non-reuse	***ESSD***	***JPCRD***
Background	90 (35.2%)	16 (9.4%)
Giving credit for previous studies	116 (45.3%)	103 (60.6%)
Same approach	28 (10.9%)	39 (22.9%)
Implication	12 (4.7%)	7 (4.1%)
Other	10 (3.9%)	5 (2.9%)
Total	**256**	**170**

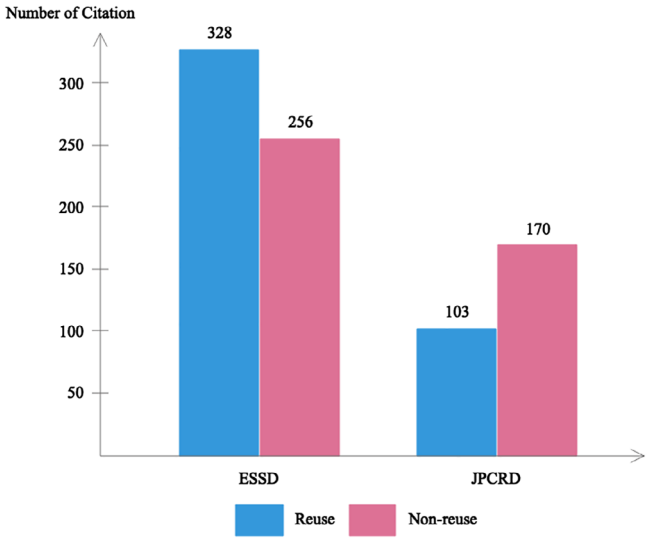

FIGURE 1 Citations for reuse and non-reuse purposes

and 28.2% respectively), is more than twice that in earth sciences (19.2% and 12.8%). However, crediting previous studies is the primary non-reuse purpose of citing data papers in both earth sciences and physics and chemistry (45.3% and 60.6% respectively).

5 | DISCUSSION AND CONCLUSIONS

Our study shows only about half of the citations in our sample indicate data reuse. Thus, while the genre of data papers facilitates some data sharing and reuse, they appear to fail to live up to their full potential. One possible reason is that data papers have not received as much attention as research papers from scientific communities. Data papers are often not counted in academic evaluation systems, such as those for tenure and promotion, which limits incentives for scholars to produce these papers. Besides, the contextual information commonly requested by data publishers is not sufficient to develop trust in data. Often, data journals do not impose strict, standardized policies on authors regarding structure and content.

Another possible reason is that not all data papers contain high-value datasets that can be reused for other purposes. Currently, review criteria for data papers focus more on the quality of datasets than on their value and reuse potential. For example, the main goal of *ESSD* is to "provide quality assessment for datasets", and whether the data or data publication is of high quality is the most important criterion (Carlson & Oda, 2018). The contextual information required for authors may also be insufficient to support reuse. Data journals should consider what contextual information is needed to support data reuse, what data should be published as data papers, and how to evaluate data reusability and how to track its impact. For example, information about curatorial actions taken by data repositories should be included in data papers to help develop prospective reusers' trust in data, and papers about data that require considerable investment of resources to collect should be prioritized for inclusion in data journals. Another issue to be

considered is to balance the emphasis on data quality, data value, and data reusability.

Disciplinary differences in practices of citing data papers can be clearly observed from our analysis. *ESSD* has more reuse citations than non-reuse citations while *JPCRD* has more non-reuse citations than reuse citations. This tendency may be due to differences between disciplines in terms of the type of datasets they use. Observational data, requiring huge investment and effort to collect, is the most common data type in earth sciences (Wynholds et al., 2012). This type of data can be hard, even impossible, to reproduce and it often has value beyond its original purpose. By contrast, physics and chemistry often draw on experimental data (Womack, 2015) that may be readily reproduced in laboratories. Moreover, the purposes of citing data papers also vary significantly according to scientific discipline. The difference in the proportion of integration reuse within two disciplines indicates earth scientists tend to reuse the datasets for new analysis.

This study is limited by the selection of disciplines and data journals and the assumption that data would be cited properly if they are reused. The issues of data citation still exist in scholarly communication, so the citation of data papers might not give accurate information about data reuse. In future work, we will interview researchers who cite data papers to discover their reasons for doing so. We also intend to explore further the disciplinary factors that shape citation practices. We recommend that academic publishers should be more aware of the potential of data papers in scholarly communication and develop discipline-specific policies and practices to satisfy various needs of scientific communities. For example, data papers in earth sciences could emphasize the various prospects of data reuse while data papers in physics and chemistry could provide precise documentation of data production. Also, they should enhance the value and reusability of data papers by improving the review process and developing concrete review criteria.

REFERENCES

Bishop, B. W., Hank, C., Webster, J., & Howard, R. (2019). Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. *Proceedings of the Association for Information Science and Technology*, *56*(1), 21–31. https://doi.org/10.1002/pra2.4

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Carlson, D., & Oda, T. (2018). Editorial: Data publication – *ESSD* goals, practices and recommendations. *Earth System Science Data*, *10*(4), 2275–2278. https://doi.org/10.5194/essd-10-2275-2018

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(S15), S2. https://doi.org/10.1186/1471-2105-12-S15-S2

Clarivate Analytics. (2018). *Journal citation reports*. Retrieved from https://jcr.clarivate.com/JCRLandingPageAction.action

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PLOS ONE*, *12*(12), e0189288. https://doi.org/10.1371/journal.pone.0189288

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, *65*(9), 1820–1833. https://doi.org/10.1002/asi.23256

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation*, *75*(6), 1274–1297. https://doi.org/10.1108/JD-08-2018-0133

Faniel, I. M., & Yakel, E. (2011). Significant properties as contextual metadata. *Journal of Library Metadata*, *11*(3–4), 155–165. https://doi.org/10.1080/19386389.2011.629959

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: A review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, *70*(5), 419–432. https://doi.org/10.1002/asi.24165

Kim, J. (2020). An analysis of data paper templates and guidelines: Types of contextual information described by data journals. *Science Editing*, *7*(1), 16–23. https://doi.org/10.6087/kcse.185

Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *68*(12), 2709–2719. https://doi.org/10.1002/asi.23892

Li, K., Greenberg, J., & Dunic, J. (2020). Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. *Journal of the Association for Information Science and Technology*, *71*(2), 172–182. https://doi.org/10.1002/asi.24226

National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century*. Retrieved from https://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

Nature. (2014). *For authors*. Retrieved from https://www.nature.com/sdata/publish/for-authors

Nature. (2019). *Scientific data 5th anniversary*. Retrieved from https://www.nature.com/content/5th-anniversary/index.html.

Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*(0), WDS32–WDS46. https://doi.org/10.2481/dsj.WDS-042

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

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1002/asi.23917

Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association*, *92*(3), 364–367.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... 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

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Womack, R. P. (2015). Research data in core journals in biology, chemistry, mathematics, and physics. *PLoS ONE*, *10*(12), e0143460. https://doi.org/10.1371/journal.pone.0143460

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, data use, and scientific inquiry: Two case studies of data practices. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vol. 12, pp. 19–22). New York, NY: ACM. https://doi.org/10.1145/2232817.2232822

Yoon, A. (2017). Data reusers' trust development. *Journal of the Association for Information Science and Technology*, *68*(4), 946–956. https://doi.org/10.1002/asi.23730

Zhao, M., Yan, E., & Li, K. (2018). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, *69*(1), 32–46. https://doi.org/10.1002/asi.23919

How to cite this article: Jiao C, Darch PT. The role of the data paper in scholarly communication. *Proc Assoc Inf Sci Technol*. 2020;57:e316. https://doi.org/10.1002/pra2.316