Information Research - Vol. 30 No. iConf (2025)

Are data papers cited as research data? Preliminary analysis on interdisciplinary data paper citations

Kai Li, Pao-Pei Huang, and Wei Jeng

DOI: https://doi.org/10.47989/ir30iConf46918

Abstract

Introduction. Research data sharing and reuse have become increasingly important in modern science, and data papers represent a new academic publication genre aimed at enhancing the visibility, sharing, and reuse of research data. However, whether citations to data papers reflect actual data reuse remains largely unexplored. This paper presents preliminary findings from a project designed to address this gap.

Method. we conducted a content analysis to manually annotate 437 citation sentences from 309 research articles referencing 50 data papers published in *Data in Brief*, a chief academic journal that only publishes data papers. The data papers were sampled from five knowledge domains based on a paper-level classification system.

Results. Our results show that most citations to all selected data papers (89%) are unrelated to the research data being described in the paper, instead focusing on the research findings or methodologies. This suggests that data papers are being cited similarly to traditional research articles, despite their unique purpose and content.

Conclusion. These findings raise questions about the effectiveness of data papers as representations of research data within the scholarly communication system, as well as their utility in quantitative studies on data reuse.

Introduction

Data papers represent a unique academic genre that primarily serves to document and contextualize research datasets within the broader research landscape (Chavan & Penev, 2011). As Chavan and Penev highlighted, data papers embody the core properties of data publication: availability, documentation, citability, and verification (Chavan & Penev, 2011). They provide comprehensive dataset descriptions, link to the data's location, undergo peer review, and crucially, offer a mechanism for citation credit. Unlike traditional research articles that focus on presenting findings and interpretations, data papers shift the spotlight onto the data itself, offering comprehensive metadata and methodological details. This approach aligns with the growing emphasis on data sharing and open science principles in academia (Kratz & Strasser, 2014). Given its strong relevance to the data-driven research paradigm, this new academic genre has gained traction in recent years, with the number of published data papers surpassing ten thousand and continuing to grow. Some universities and researchers now consider data paper submission a standard practice in data publication (Schöpfel et al., 2020; Thewall, 2020).

The value of data papers extends beyond promoting transparency and reproducibility; they provide an approach for acknowledging the crucial contribution of data to the research process (Gorgolewski et al., 2013). In addition, by offering a citable entity for datasets, data papers also create a pathway for crediting data uses, thereby incentivizing data sharing, research reproducibility, and enhancing the perceived value of research data within the scholarly ecosystem. This addresses a significant challenge identified by Tenopir and colleagues (Tenopir et al., 2011), who noted that the lack of robust reward mechanisms often impedes data sharing activities.

However, despite the increasing prevalence of data journals (such as *Data in Brief* and *Scientific Data*) and data papers, studies indicate that data citation and reuse remain lower than anticipated and that citations to data papers may not equalize data reusing (Jiao & Darch, 2020; Stuart, 2017; Thelwall, 2020). They remain significant challenges to the existing infrastructure to support data sharing and reusing. These discrepancies suggest the presence of factors influencing the impact and integration of data papers in the research system, which are not well understood in the literature. Understanding how data papers are used and cited, particularly the factors behind the citation behavior, is crucial for assessing their effectiveness in promoting data sharing and reuse.

Therefore, our study aims to explore the citation practices surrounding data papers, explicitly focusing on their citation contexts: whether a data paper is cited as research data. This investigation is guided by the overall research interest: What is the role and impact of data papers in scholarly communication, especially from the perspective of the intention of their citations? More specifically, we aim to address:

> RQ1: How are data papers cited in research articles, and for what purposes?
>
> RQ2: How does the context of data paper citations differ between disciplines and citation year gaps?

Through this preliminary analysis, we seek to contribute to the ongoing discussion about the roles of data papers in scholarly communication and their relationship to research data. The findings of this study will have implications for data citation practices, data paper guidelines, and the broader understanding of how data is used and valued across disciplines. As we navigate the evolving landscape of scholarly communication, such insights are crucial for optimizing the impact of data papers and fostering a more data-centric research ecosystem.

Methods

We took our sample from the journal of Data in Brief, an exclusively data journal founded by Elsevier in 2014. As one of the few journals that only publishes data papers, it has been frequently analysed to understand how research data is published and the relationship between data publication and scientific research (Chen et al., 2022; Fu et al., 2023; Li & Jiao, 2022; Thelwall, 2020).

We retrieved the metadata and citations of all data papers in the journal from the Dimensions database (https://www.dimensions.ai/). Dimensions is a well-established scholarly database that contains more than 146 million publications and have been heavily used in various scientometrics and science of science studies (Herzog et al., 2020; Martín-Martín et al., 2021). We only considered all data papers that have at least 10 citations indexed in the Dimensions database, so that we will have enough citations to analyse for each paper.

We specifically traced the domain information of all data papers from Dimensions that is based on the Australian and New Zealand Standard Research Classification (ANZSRC). It contains 22 broad divisions (FoR2) and 159 detailed groups of these divisions (FoR4). This classification was implemented in Dimensions on the paper-level using a machine learning algorithm (https://plus.dimensions.ai/support/solutions/articles/23000018820-which-research-categories-and-classification-schemes-are-available-in-dimensions), which can more accurately categorize the discipline of data papers published in the same journal than journal-level classification.

We mapped the FoR2 categories into five major knowledge domains for the next step of analysis. These five categories include: biomedical (*Bio*), environmental science (*Env*), physical science (*Phy*), social science and humanities (*Soc*), and Technologies and engineering (*Tech*). The mapping table, as well as the size of publications in each FoR2 category, can be accessed in our complementary materials from Zenodo (Li et al., 2024). For each of these five categories, we randomly selected 10 data papers into our final sample. And for each data paper, we randomly selected 10 publications that cite the paper to analyse in the next step.

We manually examined and filtered out the following research papers from the last step: (1) any publication that is not a research paper, such as review paper, data paper, and corrections, (2) any publication that is not published in English, and (3) any publication that is not citing the target data paper, despite the citation information supplied by the Dimensions database. Some of these issues are derived from the incomplete or inaccurate metadata information from the Dimensions database, which is a major issue in the research infrastructure to support research data. After this step, 309 research papers remain in our final sample.

For every publication, we manually collected sentences where an original data paper was cited. There are 437 sentences in total, given that a paper maybe cited multiple times in the citing publication. Two coders independently annotated every sentence using the following scheme we developed using another 50 randomly selected sentences taken prior to the final sample. The scheme focuses on distinguishing sentences describing or mentioning the data described in the data paper (i.e., data description, data use, and data background) or those that are not related to the data at all (i.e., research concept or background, research method, and research finding). Beyond the definitions presented in Table 1, examples of these categories can be accessed in our complementary materials shared on Zenodo (Li et al., 2024).

Category	Definition
Data-related	
Data Description (DP)	The sentence describes the dataset(s) per se, without indicating actual usage or how it was collected.
Data Use (DU)	The sentence describes the dataset is used in the research, including being used as a baseline.
Data Background (DB)	The sentence provides contextual information about the dataset, without describing its specific contents or usage.
Not data-related	
Research Concept or Background (NDC)	The sentence references the research topic or theoretical concepts of the cited data papers.
Research Method (NDM)	The sentence references the research methods, including but not limited to process, tools, techniques, used in the cited data paper.
Research Finding (NDF)	The sentence references the research findings, including results and their implications, derived from data in the cited data paper.

Table 1. Scheme for manual annotation of all citation sentences

We measured the inter-coder reliability of the annotation using Cohen's kappa implemented in the '*psych*' package of R (Revelle, 2024). The unweighted Cohen's kappa value is 0.73, which suggests substantial agreement (McHugh, 2012). We revisited sentences that are disagreed by the two coders.

Results

How are data papers cited in research articles, and for what purposes?

Our results show that the majority of data papers are NOT cited as research data in research articles. Table 2 shows the distribution of all citation sentences across the six categories. Among all 437 citations, only 48 of them (11.0%) belong to the data-related categories. And the largest category is *Research finding* (*NDF*), despite the fact that no data paper is supposed to present research design and findings based on its original definition (Chaven & Penev, 2011). However, there is also a decent share of citations to data papers that concerns research methods (i.e., *NDM*). Even though data papers are not cited as research data in these cases, this citation context is still relatively close to research data. Overall, we find a diverse spectrum of citation practice around data papers that are strongly deviant from the original purposes of publishing data papers.

Category	# Sentences	Share of Sentences
Data-related	**48**	**11.0%**
Data Use (DU)	27	6.2%
Data Description (DP)	14	3.2%
Data Background (DB)	7	1.6%
Not data-related	**425**	**89.0%**
Research Finding (NDF)	207	47.4%
Research Method (NDM)	130	29.7%
Research Concept or Background (NDC)	52	11.9%
All categories	**473**	**100%**

Table 2. Distribution of all citation sentences across the six categories

The above numbers on the level of sentences are translated into 16 data papers with at least one data-related citation. Among the 16 papers, only four of them are cited primarily data as research data (i.e., more than 50% of all citation sentences a paper received are data-related). Even though the reasons for such differences between data papers are not covered by this preliminary research, this would be an interesting question for the next step of our investigation.

We specifically examined how citation contexts are related to the paper sections in which a data paper is cited, given the strong relationship between these variables in citation analysis (Tahamtan & Bornmann, 2019). We classified all paper sections into the following four categories: Introduction (including literature review), Methods, RDC (i.e., Results, Discussion, and Conclusion), and Others (i.e., appendix and acknowledgement). The key statistics are presented in Table 3. For both data and method citations, they are more prevalent in the Methods section. In addition, even though only a few data papers are cited in appendix and acknowledgement, this section is also quite strongly connected to the data-related citation contexts.

Section	# Sentences	% Data citation	% Method citation
Introduction	149	3.4%	21.5%
Methods	106	23.6%	92.6%
RDC	158	4.4%	15.2%
Others	24	45.8%	7.7%

Table 3. Data citation index by paper section (% Data citation is the share of all data-related sentences among all sentences; % method citation is the share of all research meth-od sentences among all non-data sentences.)

Key factors behind the citation behavior: discipline and the citing-cited year gap

We examined two key factors (i.e., disciplines, citation year gaps) behind the citation contexts of data papers, which is reported in this section.

Discipline of data papers

We analysed the relationship between citation contexts and the discipline of data papers based on the ANZSRC classification system. Table 4 presents the results, which shows largely similar share of data citation to data papers from all domains. The higher share of data citations in the category of Soc (social sciences and humanities) is largely attributed to the single paper '*Residential electric vehicle charging datasets from apartment buildings*' (Sørensen et al., 2021), which has all citations related to data. And we believe this categorization is highly debatable, which is further highlighted in our discussion.

Discipline	# Sentences	% Data citation
Soc	75	14.7%
Phy	76	11.8%
Bio	72	11.1%
Tech	101	9.9%
Env	113	8.8%

Table 4. Data citation context by the discipline of data papers

Year gap of citing paper

We further investigated how the year gap between data papers and citing papers could be related to the citation contexts. Results, as shown in Table 5, illustrates that when the year gap widens, there is a growing possibility that the data paper is cited as research data. However, even after five years of the data paper, the share of data citation is still only at 20%. Our findings show similar patterns with citation contexts of method papers and method objects (Li, 2021; Small, 2018), where older research instruments are more likely to be cited as instruments, possibly because of the demonstrated validity.

Section	# Sentences	% Data citation
Year 0-1	119	5.0%
Year 2-3	145	15.2%
Year 4-5	121	9.1%
Year > 5	45	20.0%

Table 5. Data citation context by citation year gap between citing and cited documents

To further understand how the two variables are statistically correlated to the outcome, we constructed a logistic regression model to understand the correlation between the above variables as well as the publication year of the data papers as ***independent variables*** and whether a sentence is data-related or not as the ***categorical dependent variable***. The summary of the regression model is shown in Figure 1, with the y-axis showing how much each category contributes to the possibility of a sentence being a data-related citation. It shows that comparing to the domain of biomedical science (which is the baseline not shown in the graph), all other domains are at least equally likely having data-related citations, even though none of the other domains are significantly more likely so. Similarly, comparing to the year of 2016, data papers published in the years of 2017 and 2021 are significantly less likely to be cited as data, which is likely due to the smaller data points in these two years. We also did not find the year gap between citing and cited papers to be a significant factor for data citation.

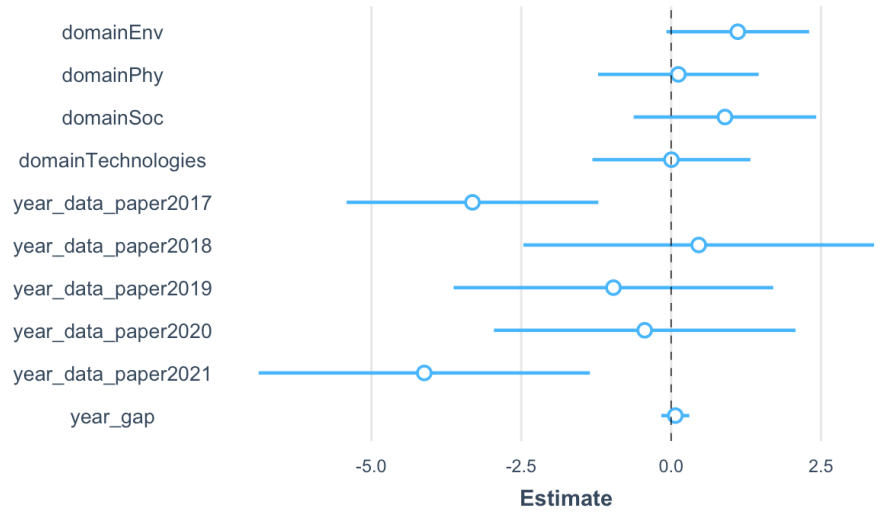

Figure 1. Summary of the statistical model

Discussion and conclusion

This short paper presents preliminary findings from our project to understand the citation contexts of data papers in English-language research articles. We examined the general patterns of whether data papers are cited as research data, an important argument to support the usefulness of this new genre, with special focus on some key factors behind the citation behavior, such as the discipline of the data paper and citation time gap. Our results reconfirm existing findings that data papers can be cited for various reasons beyond just those related to research data (Jiao & Darch, 2020; Thelwall, 2020). However, the fact that the majority of data papers are not cited as research data raises a critical concern about whether data papers can help to give the proper credit (i.e., credit of creating and publishing research data) to the authors.

Additionally, building upon existing efforts, our paper strives to offer a more comprehensive empirical investigation on factors behind the data paper citation behavior. We find that the general pattern stays valid across knowledge domains and does not change significantly as the year gap between the citing and cited documents widens.

Beyond our major findings, our research design and process also reveal some key issues in the infrastructure to support data publication. One critical issue is the identification of the discipline of data papers. While many data papers are published in multidisciplinary data journals, such as *Data in Brief* and *Scientific Data*, the journal-level classification system in many scholarly databases (such as the Web of Science) is inadequate for determining the discipline of such publications. Even by using a novel paper-level classification, we find cases where the classification may not be accurate, which could have strong impact on future quantitative studies on this topic. We argue that establishing a more robust and accurate system to evaluate the discipline of data papers and research data is a critical next step; and this system should consider not just paper metadata, but also the authors' affiliations and other attributes of the research datasets to achieve better performance.

In the next step of the project, we will expand the presented investigation by using a larger sample size as well as considering more factors to explain the citation behavior of data papers, such as the discipline of the citing paper and whether the citation is from the data paper authors themselves (i.e., self-citation). We expect that these extra factors will have strong impacts on how a data paper is understood and discussed in citing publications. In addition, we will also use advanced statistical

models to understand more factors behind the citation behavior of data papers, which will contribute to the literature of citation context analysis.

About the authors

Kai Li is Assistant Professor in School of Information Sciences, University of Tennessee, Knoxville, USA. He received his Ph.D. from the Department of Information Science at Drexel University, and his research interest are scholarly communication, quantitative science studies. He can be contacted at kli16@utk.edu.

Pao-Pei Huang is Ph.D. student in School of Information and Library Science, University of North Carolina, Chapel Hill, USA. Her research interest are open science practices, scholarly communication, and social informatics. She can be contacted at paopei@unc.edu.

Wei Jeng is Associate Professor at Department of Library and Information Science, National Taiwan University, Taiwan and Director of Talent Empowerment Center at National Institute of Cyber Security, Taiwan. She received her Ph.D. from the school of computing and information at University of Pittsburgh, and her research interest are open science practices and research data infrastructure. She can be contacted at wjeng@ntu.edu.tw.

References

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, 12, 1-12. https://doi.org/10.1186/1471-2105-12-S15-S2

Chen, P. Y., Li, K., & Jiao, C. (2022). A preliminary analysis of geography of collaboration in data papers by S&T capacity index. *Proceedings of the Association for Information Science and Technology*, 59(1), 642-644. https://doi.org/10.1002/pra2.676

Fu, J., Tian, L., Zhang, C., & Li, J. (2023). Opening research data contributes to the citations of related research articles: Evidence from Data in Brief. *Learned Publishing*, 36(3), 426-438. https://doi.org/10.1002/leap.1551

Gorgolewski, K.J., Margulies, D.S., & Milham, M.P. (2013). Making data sharing count: A publication-based solution. *Frontiers in neuroscience*, 7, 9. https://doi.org/10.3389/fnins.2013.00009

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative science studies*, 1(1), 387-395. https://doi.org/10.1162/qss_a_00020

Jiao, C., & Darch, P.T. (2020). The role of the data paper in scholarly communication. *Proceedings of the Association for Information Science and Technology*, 57(1), p.e316. https://doi.org/10.1002/pra2.316

Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. *F1000Research*, 3. https://doi.org/10.12688/f1000research.3979.3

Li, K. (2021). The re-instrumentalization of the Diagnostic and Statistical Manual of Mental Disorders (DSM) in psychological publications: A citation context analysis. *Quantitative Science Studies*, 2(2), 678-697. https://doi.org/10.1162/qss_a_00124

Li, K., & Jiao, C. (2022). The data paper as a sociolinguistic epistemic object: A con-tent analysis on the rhetorical moves used in data paper abstracts. *Journal of the As-sociation for Information Science and Technology*, 73(6), 834-846. https://doi.org/10.1002/asi.24585

Li, K., Huang, P. P., & Jeng, W. (2024) Dataset for "Are data papers cited as research data? Preliminary analysis on interdisciplinary data paper citations" [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13763303

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906. https://doi.org/10.1007/s11192-020-03690-4

McHugh, M.L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Revelle, W. (2024). Package 'psych' version 2.4.3. The comprehensive R archive net-work. https://CRAN.R-project.org/package=psych

Schöpfel, J., Farace, D., Prost, H., & Zane, A. (2020). Data papers as a new form of knowledge organization in the field of research data. *Knowledge Organization*, 46(8), 622-638. https://doi.org/10.5771/0943-7444-2019-8-622

Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, 12(2), 461-480. https://doi.org/10.1016/j.joi.2018.03.007

Sørensen, Å.L., Lindberg, K.B., Sartori, I., & Andresen, I. (2021). Residential electric vehicle charging datasets from apartment buildings. *Data in Brief*, *36*, 107-105. https://doi.org/10.1016/j.dib.2021.107105

Stuart, D. (2017). Data bibliometrics: Metrics before norms. *Online Information Review*, *41*(3), 428-435. https://doi.org/10.1108/OIR-01-2017-0008

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635-1684. https://doi.org/10.1007/s11192-019-03243-4

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PloS one*, 6(6), p.e21101. https://doi.org/10.1371/journal.pone.0021101

Thelwall, M. (2020). Data in Brief: Can a mega-journal for data be useful?. *Scientometrics*, *124*(1), 697-709. https://doi.org/10.1007/s11192-020-03437-1