https://doi.org/10.51408/issi2025_100

Harnessing Data Papers: An Analysis of Their Role in Scientific Data Dissemination and Reuse

Liyue Chen[1], Xiaomin Liu[2]

[1] *chenliyue@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, 33 Beisihuan Xilu, Zhongguancun, Haidian District, Beijing (China)

[2] *liuxm@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, 33 Beisihuan Xilu, Zhongguancun, Haidian District, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, 80 East Zhongguancun Road, Haidian District, Beijing (China)

Abstract

This research in progress studies the scholarly role of data papers in scientific data sharing and reuse. Data papers are a crucial publishing form for scientific data, tasked with fully leveraging the value of data science. This study, based on extensive citation context data and large language models, investigates the actual contributions and specific citation purposes of data papers in citing documents. The results indicate that data papers indeed play a role in disseminating scientific data for reuse during scholarly communication, yet their potential has not been fully realized, with certain data papers still serving primarily as methodological support and overviews of data development backgrounds. The impact of data papers also varies across disciplines; fields such as life sciences, natural resources and environmental sciences place greater emphasis on the role of data papers, with richer integration and utilization of scientific data based on them. However, the value of data papers in humanities, social sciences, and some fundamental disciplines remains underexplored.

Introduction

The rapid transformation of scientific paradigms underscores the significance of scientific data. However, as scientific data rapidly accumulates and is widely shared, several issues have become increasingly evident: researchers' reluctance to share data (Gajbe et al., 2021; Mattern et al., 2024), the fragmentation of data resources (Shen et al., 2024), unclear data ownership (Sheng & Yuan, 2021), and challenges in controlling data quality. To address these issues and promote the reproducibility of research findings, the data paper has emerged as an important academic publication format. Data papers are peer-reviewed scholarly publications that provide a standardized description of scientific datasets (Carlson & Oda, 2018; Chavan & Penev, 2011). Typically, data papers detail the methods of data collection and processing, data structure and storage format, methods of data use, and access pathways (Kim, 2020), serving as a 'manual' for the data.

Existing research has analyzed the crucial role of data papers in promoting open data reuse from the perspective of the motivations for data sharing. The reluctance of researchers to share and reuse scientific data can generally be divided into two categories. On one hand, data producers lack the motivation to share, as researchers are uncertain whether their actions of sharing will be properly rewarded (Mattern et al., 2024; Tenopir et al., 2015; Wallis et al., 2013). On the other hand, there is also a lack of sufficient production information to support data reuse (Borgman, 2012; Curty et al., 2017). Compared to other data publication formats, data papers secure the data producers' right to discovery priority and academic reputation through publication and formal citation. They also stimulate data sharing and reuse by controlling the quality of data through a rigorous peer-review process (Thorisson, 2009; Zhao et al., 2018).

A few studies have also explored the actual dissemination impact of data papers on the reuse of scientific data, starting from the citation practices of data papers. Jiao and Darch (2020) analyzed the citation context of 103 data papers in earth sciences and physics through manual interpretation and found that data papers have not fully realized their potential in promoting data reuse. Research on citation behaviors in the biomedical field shows a steady increase in formal citations of data papers for the purpose of data usage (Jiao et al., 2024). Similarly, in the humanities and social sciences, data papers have had a positive impact on the influence of datasets in related research papers (McGillivray et al., 2022). Overall, existing research on data papers is limited in scale, relies primarily on manual annotation for interpreting the scholarly communication role of data papers, and is only conducted within a few disciplines, which does not provide a comprehensive view of the development of data papers.

The purpose of this study is to explore the role of data papers in the open sharing and informed reuse of scientific data during the academic communication. Three research questions are considered: (1) Do data papers make a data-related contribution to the studies that cite them? (2) If they do data-related contribute, for what purposes do the citing works use the scientific data? (3) Does the role of data papers vary across different disciplines? To address these questions, this research conducts a citation context analysis on data papers across all disciplines, which includes identifying actual contributions and analyzing citation purposes, aiming to better understand the facilitative role of data papers in the sharing and reuse of scientific data.

Data and Methods

In our study, we limited the document type to 'data paper' within the Web of Science Core Collection, covering the period from 1980 to 2024. We retrieved a total of 17,318 data papers, of which 82.24% were cited, with an average citation count of 15.99 per paper. To categorize the disciplines of these data papers, we used the InCites citation topics (macro) schema to map the papers onto 10 disciplines. To analyze the citation context characteristics of these papers, we collected citation context data from citing documents via the *Scites* platform (https://scite.ai). Ultimately, 12,422 data papers were matched with 219,707 citation context entries.

The identification of the actual contributions of data papers employed the automatic recognition method proposed in our previous study (Chen et al., 2024), which classifies the contributions of papers into five categories: theoretical, experimental, methodological, data-based, and other. This classification is performed using a fine-tuned Llama2-13B large language model, which achieved an accuracy rate of 0.94.

To automatically identify the citation purposes of data papers, we utilized large language model techniques in our experiments. Building on existing research (Gregory et al., 2019), we categorized citation purposes into seven types: background, calculation, integration, verification, inspiration, and other. We tested various prompt schemes and different large models (including DeepSeek-R1 and GPT-4o), and ultimately determined that GPT-4o had the best recognition performance, with an F1-score of 0.81.

Results

In-text citation characteristics of data papers

We first analyzed the frequency of in-text mentions of data papers in citing documents. The frequency of in-text mentions for the data papers analyzed was 1.673, which is similar to that of traditional academic papers (Chen et al., 2022; Hsiao & Chen, 2018). In citing documents, instances where data papers were mentioned only once accounted for approximately 66.87%, while mentions two times or more accounted for 33.13%. When we conducted the analysis by different disciplines, we found that data papers in fields such as Physics, Earth Sciences, Agriculture, Environmental and Ecology had an average in-text mention frequency higher than that of all disciplines. In contrast, the Art and Humanities, Social Sciences, and Mathematics had lower in-text mention frequencies. This outcome reflects, to some extent, the varying degrees of emphasis placed on scientific data across different research areas.

We also analyzed the distribution of in-text locations for data papers within the citing studies. As shown in Figure 1, data papers are most frequently mentioned in the 'Materials and Methods' and 'Introduction' sections, accounting for 24.28% and 23.33% respectively. In contrast, they appear less frequently in the 'Results' and 'Discussion' sections. Compared to traditional academic papers, which are primarily mentioned in the 'Introduction' and 'Discussion' sections (Bertin et al., 2016; Voos & Dagaev, 1976), data papers indeed show a distinct characteristic of explicit data support.

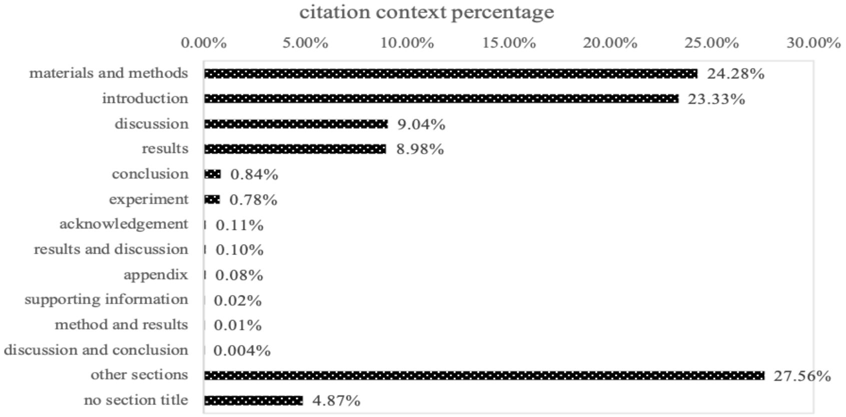

Figure 1. The distribution of in-text locations for data papers within the citing studies.

An analysis of the in-text locations of data papers across various disciplines reveals that Mathematics data papers are most frequently mentioned in the 'Materials and Methods' section, accounting for 42.54% of mentions (Figure 2). This is followed by Earth Sciences and Clinical & Life Sciences. In contrast, in fields such as Humanities and Social Sciences, Engineering and Materials Science, data papers are more often cited in the 'Introduction' section, seemingly serving more as a background overview. Moreover, compared to other fields, data papers in Chemistry and Clinical & Life Sciences are relatively more frequently cited in the 'Results' section. This trend may stem from the reliance of these disciplines on scientific experimental processes and experimental data, using cited scientific data from other studies for comparative analysis and validation of results in their current research.

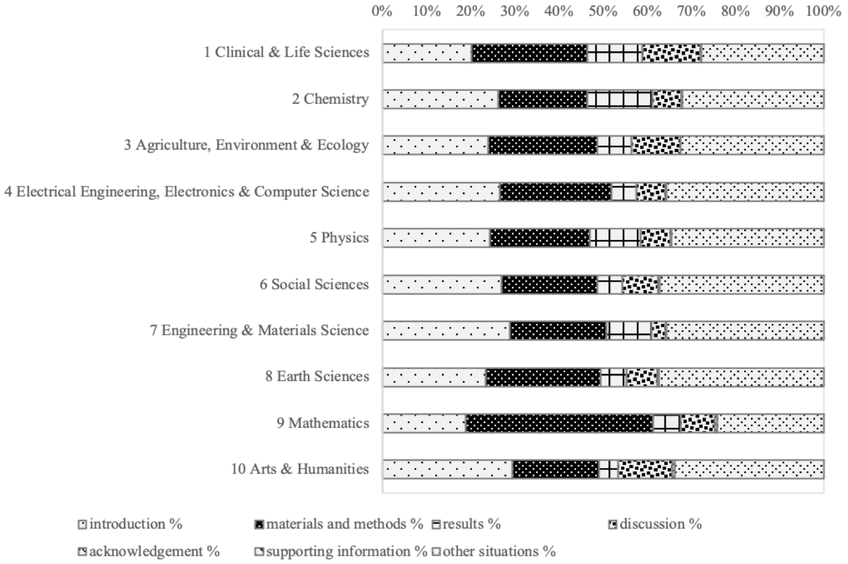

Figure 2. The distribution of in-text locations for data papers across 10 disciplines.

The actual contribution roles of data papers

By identifying the actual contribution types of data papers, it was found that data papers indeed make a data-based contribution to the citing research in 42.7% of cases. However, in more than half of the citation context instances, data papers play roles in other aspects, including providing methodological support for the citing research, being used for experimental comparison and result validation, and offering background information for the cited studies.

The actual contributions of data papers in various research areas also show significant differences (Figure 3). In fields such as Electrical Engineering, Earth Sciences, and Agronomy, the proportion of data-based contributions is relatively high, whereas in Engineering and Materials Science, Mathematics, and Chemistry, the proportion of data contributions is relatively lower. In Engineering and Materials Science, data papers are more focused on corroborating and supporting experimental results, while in Mathematics and Chemistry, data papers play a role in methodologies such as mathematical formulas and experimental schemes.

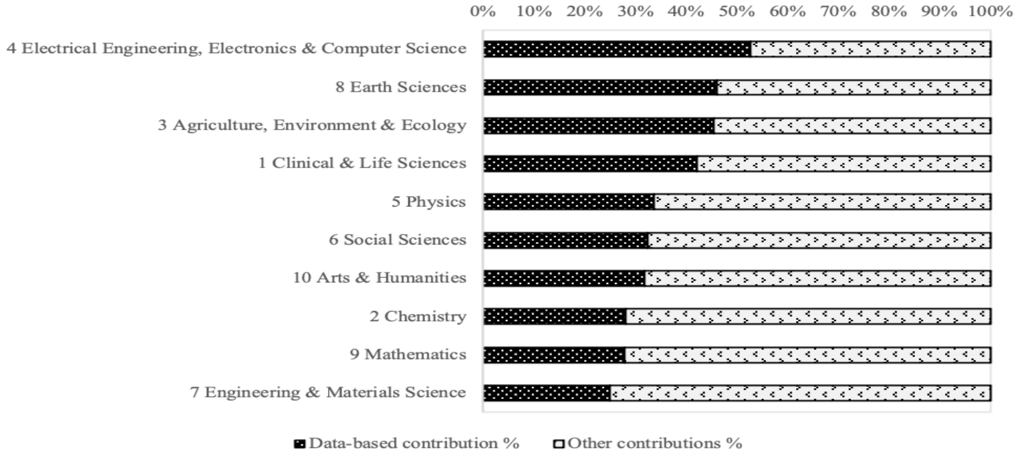

Figure 3. Distribution of actual contribution types made by data papers in 10 research areas.

The purpose of citing scientific data in data papers

The type of actual contribution reflects whether data papers have made a contribution related to citing works, focusing specifically on the data core of the cited papers. To further understand the purposes behind the citation of data papers in citing studies, we conducted a citation purpose analysis on the contexts where data papers clearly have a data-based contribution. As Figure 4 shows, in nearly 50% of cases, although data papers provide a data-based contribution, they are merely mentioned by the citing studies to elaborate on the background knowledge of the research. Secondly, a certain proportion of citations in citing studies are due to their use in experimental calculations, accounting for 15.19%. Additionally, citing papers use the cited literature for data integration, comparative data analysis, and data benchmarking and validation, with these three citation purposes having similar distributions. Very few data papers serve as a source of research inspiration for the citing literature.

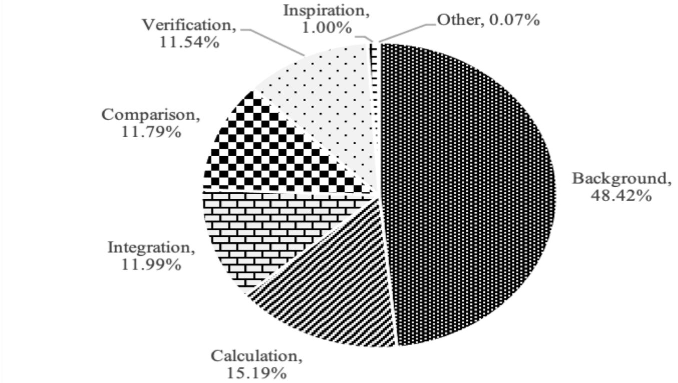

Figure 4. Citation purposes distribution of data papers with data contributions.

Figure 5 shows that the distribution of citation purposes in cited data papers is relatively similar across various disciplines. Data papers in Life Sciences and Earth Sciences have a higher proportion of citations for usage purposes (including calculation, comparison, integration and verification) compared to the average across all disciplines. In contrast, these usage purposes are relatively lower in Art and Humanities, as well as Mathematics.

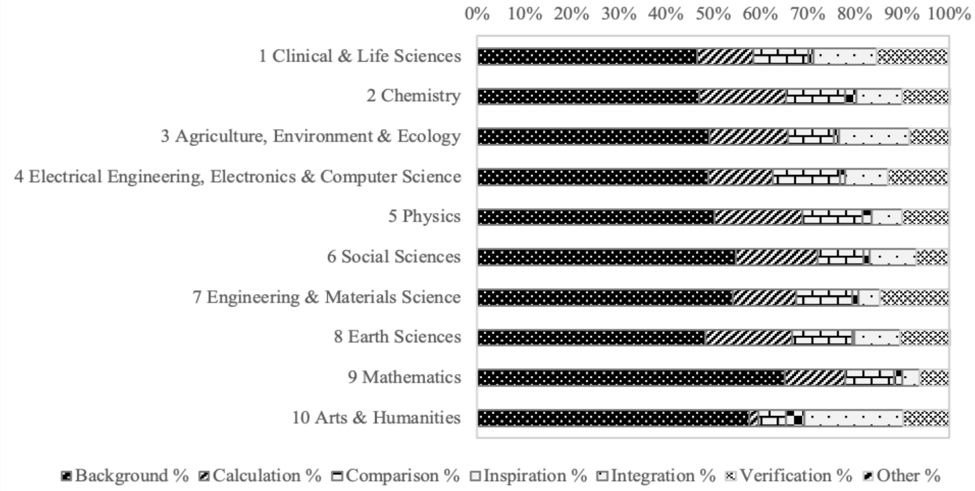

Figure 5. Citation purposes distribution of data papers with data contributions (10 disciplines).

Discussion and Conclusion

Data papers not only promote transparency and reproducibility in scientific research but also confer academic credit to data producers. It is very meaningful to explore how data papers can help scientific data fully realize its innovative value.

Our study finds that data papers indeed play a clear role in the dissemination and reuse of scientific data, yet there is substantial space for improvement. A significant number of citations to data papers still stem from methodological support, experimental comparisons, and result validation, or describing the current state of data development relevant to the research questions. By comparing citation context characteristics of data papers across various research areas, we observe differences in attention and usage levels towards data papers among disciplines. Areas such as Clinical & Life Sciences, Agriculture, Environment & Ecology are experiencing rapid development in scientific data and data papers, with a more pronounced trend in data usage based on citations to data papers. However, fields like the Arts & Humanities, and Mathematics have smaller volumes of data papers, with citations primarily focusing on confirming research viewpoints and describing relevant backgrounds. Our subsequent research questions include: (1) By conducting annual statistics, we will explore whether the role or function of data papers has changed. (2) From the perspective of academic publishing standards, we will further investigate the publishing attributes of cited data papers and the citing literature.

References

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology, 67*(1), 164-177.

Borgman, C. L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078.

Carlson, D., & Oda, T. (2018). Data publication - goals, practices and recommendations. *Earth System Science Data, 10*(4), 2275-2278.

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *Bmc Bioinformatics, 12,* S2.

Chen, L., Ding, J., Song, D., & Qu, Z. (2024). Exploring Scientific Contributions through Citation Context and Division of Labor. *arXiv* preprint arXiv:2410.13133.

Chen, L. Y., Ding, J. L., & Lariviere, V. (2022). Measuring the citation context of national self-references. *Journal of the Association for Information Science and Technology, 73*(5), 671-686.

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.

Gajbe, S. B., Tiwari, A., Gopalji, & Singh, R. K. (2021). Evaluation and analysis of Data Management Plan tools: A parametric approach. *Information processing & management, 58*(3),102480.

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A. and 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*: 419-432.

Hsiao, T. M., & Chen, K. H. (2018). How authors cite references? A study of characteristics of in-text citations. *Proceedings of the Association for Information Science and Technology, 55*(1), 179-187.

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), e316.

Jiao, H., Qiu, Y. H., Ma, X. W., & Yang, B. (2024). Dissemination effect of data papers on scientific datasets. *Journal of the Association for Information Science and Technology, 75*(2), 115-131.

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.

Mattern, J. B., Kohlburn, J., & Moulaison-Sandy, H. (2024). Why academics under-share research data: A social relational theory. *Journal of the Association for Information Science and Technology, 75*(9), 988-1001.

McGillivray, B., Marongiu, P., Pedrazzini, N., Ribary, M., Wigdorowitz, M., & Zordan, E. (2022). Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences. *Publications, 10*(4), 39.

Shen, Z.H., Zhu, X.J., Wang, H.J., et al. (2024). Research data network: concept, systems and applications. *Frontiers of Data & Computing, 6*(4), 3-21.

Sheng, X.P., & Yuan, Y. (2021). Data rights governance in open sharing of scientific data. *Journal of Library Science in China,* (5), 80-96.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *Plos One, 10*(8), e0134826.

Thorisson, G. A. (2009). Accreditation and attribution in data sharing. *Nature Biotechnology, 27*(11), 984-985.

Voos, H., & Dagaev, K. S. (1976). Are All Citations Equal? Or, Did We Op. Cit. Your Idem? *Journal of Academic Librarianship, 1*(6), 19-21.

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.

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.