How are data paper abstracts constructed? Preliminary analysis of rhetorical moves in data paper abstracts from *Scientific Data* and *Data in Brief*

Kai Li[1] and Chenyue Jiao[2]

[1] *kai.li@ruc.edu.cn*
Renmin University of China, School of Information Resource Management, 15 Zhongguancun St., Beijing, China, 100080

[2] *cjiao4@illinois.edu*
University of Illinois Urbana-Champaign, School of Information Sciences, 501 E. Daniel St., Champaign, IL, United States, 61820

Abstract

The data paper is an emerging academic genre that responds to the rising importance of data in the scientific enterprise. It is a type of academic publication that focuses on the description of data objects. With a large number of data papers published in recent years, it is critical to understand their presence in the scholarly communication system. Our project aims to achieve this goal by investigating the rhetorical functions played by data papers, as a means to understand their scientific-rhetorical characteristics. This research-in-progress paper reports preliminary results from this project. In this work, we expanded an established classification system of rhetorical moves in research article abstracts to make it applicable to data paper abstracts. We further applied this system to classify all sentences in 360 abstracts of data papers in two leading data journals, *Scientific Data* and *Data in Brief*. We identified four new rhetorical moves specific to data papers and examined how all rhetorical moves are distributed across the two journals. This work illustrates some important characteristics of data papers as a rhetorical device and informs future research directions towards a more comprehensive appreciation of data papers in the scientific system.

Introduction

Data have risen to be one of the most prominent epistemic objects in the scientific enterprise (Silvello, 2018; Wynholds, 2011). The growing amount of data paves ways for new research questions and methods as well as large-scale scientific collaborations (Hey et al., 2009). All these changes, in return, require higher transparency of data in the scholarly ecosystem, as summarized into the FAIR Principles of research data stewardship, i.e., research data should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016).

One approach to addressing the new needs for research data in the scholarly system is through the concept of *data publication*. Data publication refers to the pipeline through which data are transformed into discrete, well-documented, publishable, and citable objects (Parsons & Fox, 2013). An implementation of this concept that is gaining momentum is to publish data objects as academic papers, as shown in the emerging academic genre of *data papers*. A data paper is defined as a "scholarly publication of a searchable metadata document describing a particular online accessible dataset, or a group of datasets, published in accordance with the standard academic practices" (Chavan & Penev, 2011, p. 3).

Data papers are becoming more popular from the mid-2010s, partly supported by the establishment of journals dedicated to this genre, or *data journals*. Candela and colleagues' survey (2015) identified seven dedicated data journals and over a hundred academic journals accepting data papers along with research articles. One of the earliest exclusively data journals is *Earth System Science Data* that was founded in 2009 (Pfeiffenberger & Carlson, 2011). In 2014, the journal *Scientific Data* was developed by Nature Publishing Group (Hrynaszkiewicz & Shintani, 2014) and *Data in Brief* by Elsevier (Thelwall, 2020), both later grew into the most important exclusively data journals in the market (Walters, 2020).

Given the short history of data papers, we are still far from a clear understanding of their presence in the scholarly communication system. One critical aspect of this knowledge is how these publications are constructed differently from research articles from a rhetorical perspective, as textual features of scientific publications bear strong theoretical implications towards the publication and communication of scientific knowledge (Small, 1982). In this regard, data papers offer a valuable site to observe the diversity of scientific rhetoric. On the one hand, data papers have distinct purposes from research articles: they are only supposed to describe the datasets per se, instead of offering information about the research design and results (Callaghan et al., 2012), which is reflected in the distinct content requirements by data journals (Kim, 2020). On the other hand, data papers inevitably share many similarities with research articles, as both genres are produced in the same scholarly system and highly similar research contexts (Li et al., 2020). However, how these two genres are compared with each other has never been examined by empirical studies.

The present project aims to fill this gap by evaluating various aspects of rhetoric in data papers and investigate how they are connected to the roles played by data publications in scholarly communication. In the present work, we report our preliminary findings by identifying and classifying rhetorical moves in abstracts of a selected sample of data papers in two flagship data journals, *Scientific Data* and *Data in Brief*. The rhetorical move is commonly defined as a "discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse" (Swales, 2004, p. 268). Thus, by identifying rhetorical moves in data papers, we strive to understand what rhetorical units and functions are included in this academic genre. These goals will ultimately produce deeper knowledge of the new, data-driven mode of knowledge production represented in data publications and more effective extraction of data-related information from these publications. Specifically, the following two questions are pursued in this research-in-progress paper:

RQ1: What rhetorical moves are used in abstracts of data papers? This question aims to examine all rhetorical moves used in data paper abstracts and identify new moves that are specific to data-related contexts. We used the manual coding method and applied a modified classification system to identify all moves from abstracts of a selected sample of data papers.

RQ2: How are these moves distributed in the two data journals? This question strives to understand the cross-journal differences in the use of rhetorical moves. In this preliminary study, we only connected the differences to the journal policies concerning abstracts to draw preliminary explanations.

The rest of the paper is organized as follows. The Method section discusses our data sample. This is followed by the section ("Classification of rhetorical moves in abstracts") modified the classification system of rhetorical moves and used it in manual coding. The Results and discussion section illustrates preliminary results based on the classification, which is followed by Conclusions addressing implications of our results and the next steps of this project.

Method

In this research-in-progress paper, we only used data papers included in *Scientific Data* and *Data in Brief*, two domain-independent flagship data journals, from Scopus on November 15, 2020. While both journals were categorized as *exclusively data journals* by Stuart (2017), they may contain document types beyond data papers, especially comments and reviews. We removed all other document types and retrieved 7,712 data papers from these journals, with 6,335 from *Data in Brief* and 1,377 from *Scientific Data*.

These two journals were selected due to the following reasons. First, they are the two leading, domain-independent exclusively data journals based on the journal impact factor, the number of publications, and the presence in empirical studies (Kim, 2020; Stuart, 2017). Second, both journals were founded in 2014, which facilitates meaningful comparisons over time. Third,

these journals have a few differences in the author guideline, which can shed light on the diversity of data papers. Based on these reasons, we believe our sample is able to reflect how data papers in a broad array of research domains are composed.

This preliminary study focuses on the journal policies of the paper abstract, which are summarized in Table 1. While rules on both journals are similarly worded, especially what should and should not be included in the paper abstract, one major difference is that *Scientific Data* directly states that no reference should be cited in the abstract.

Table 1. Abstracted-related policies from the two journals

Scientific Data[1]	*Data in Brief*[2]
- They should succinctly describe the study, the assay(s) performed, the resulting data and their reuse potential. - It should not make any claims regarding new scientific findings. - No references are allowed in this section.	- Concisely describes the data, its collection process, analysis and reuse potential. - Do not: provide conclusions, results, or mention the word 'study'.

We selected all data papers published between 2015 and 2020, as both journals were founded in 2014 and there are not enough papers to be analyzed in that year. To enable a more meaningful comparison between the journals, we used a stratified sampling approach where we take 30 publications for each journal in each year. There are finally 360 data papers in our sample for manual classification, to be described in the next section.

We used the NLTK package of the Python language (Loper & Bird, 2002) to parse paper abstracts into sentences. A total of 2,182 sentences were parsed from all selected data papers.

Classification of rhetorical moves in abstracts

We classified all parsed sentences based on its rhetorical function(s) in the texts. For this purpose, we modified the classification scheme of rhetorical moves in research article abstracts proposed by Hyland (2000) that is composed of *Introduction*, *Purpose*, *Method*, *Product*, and *Conclusion*.

Based on existing empirical evidence, these moves are not sufficient for data papers, due to the distinct functions of the latter genre (Kim, 2020; Li & Chen, 2018). As a result, we expanded this system using 50 randomly-selected abstracts (from the original 7,712 papers), where two coders independently reviewed them to identify any additions to this system. In the end, we added four new rhetorical moves that are specific to data papers to Hyland's scheme. Our modified scheme is illustrated in Table 2, with the four added categories highlighted. Moreover, we changed *Product* into *Results* in our scheme, even though we retained its original definition.

Table 2. The new classification of rhetorical moves in data paper abstracts

Move	*Definition*
Introduction	Context of the papers
Purpose	Purpose or intention of the paper/research
Method	Research design, procedure, assumptions, approach of the study
Results	Main findings or results
Conclusion	Interpretations of the results beyond the scope of paper
Data description	**Description of the data object that is the topic of the paper**
Data uses	**How the data object is supposed to be used or its implications**

[1] https://www.nature.com/sdata/publish/submission-guidelines

[2] https://www.elsevier.com/journals/data-in-brief/2352-3409/guide-for-authors

Data accessibility	**How to get access to the data object**
Related research article	**The research article to which the data object is connected.**

In this classification system, we specifically separated moves that are focused on the data object per se and those on the research behind the data object, as the data-research dichotomy is an important distinction between data papers and research articles. In the former group, four added categories specifically focus on any information about the data object being described in the data paper. Examples of these four categories are given below:

- **Data description:** "The data presented here represents the detailed comparative abundances of diverse groups of biomass hydrolyzing enzymes including cellulases, hemicellulases, lignin degrading enzymes, and peptidases and proteases; and their post translational modification like deamidation." (Adav et al., 2015)
- **Data uses:** "These unique data sets can be used by the wider community to implement analog approaches for characterizing reservoir and aquifer formations." (Bayer et al., 2015)
- **Data availability:** "For public access, mass spectrometry raw data are available via ProteomeXchange with identifier PXD002153." (Sikulu et al., 2015)
- **Related research article:** "This paper provides data in support of the research article entitled 'DPF2 regulates OCT4 protein level and nuclear distribution'." (Liu et al., 2015)

On the other hand, while some traditional rhetorical moves are supposed to be used in the data paper abstracts, especially *Introduction* and *Method*, moves like *Results* and *Conclusion* are clearly discouraged to be used based on the journal policies in Table 1.

Using the modified classification scheme, two coders independently classified all sentences. The intercoder reliability between the two coders is 0.706, indicating a good agreement (Landis & Koch, 1977). All differences between the coders were resolved before data were analyzed.

In our coding, we allowed the co-existence of multiple moves in the same sentence, given the complexity of human language. In our final result, we identified 77 sentences with two moves. We used a fractional counting method for these sentences in the next section, with a sentence being counted as 0.5 for each move its covers.

Results and discussion

Table 3 summarizes the counts of sentences and papers with the nine moves. The table shows that *Introduction*, *Method*, *Data description*, and *Data uses* are the most frequently used moves on both the sentence- and paper-levels. This group of moves is composed of both research- and data-oriented functions. A notable finding, contrasting to the journal policies, is that *Results* and *Conclusion* are often used in abstracts, especially that more than 50% of articles have at least one *Results* sentence. Following our previous work (Li et al., 2020), this finding, again, sheds doubt on the boundaries between research articles and data papers.

Table 3. Counts of sentences and papers with moves

Move	*# Sentences (fractional count; n = 2,182)*	*# Papers (n = 360)*	*Sentences per paper*
Introduction	514	217	2.37
Method	527	249	2.12
Data description	357	233	1.53
Data uses	227.5	188	1.21
Results	185.5	97	1.91
Purpose	149	139	1.07
Related research article	89.5	90	0.99

Conclusion	67.5	52	1.30
Data availability	61	67	0.91

It is also worth noting that when these moves are used in a paper, there is a large variance in terms of the number of sentences with the specific move. For example, *Introduction* and *Method* have over two sentences per abstract as compared to *Related research article* and *Data availability* with lower than one. The latter moves have fewer than one sentence per abstract because they can be co-used with other moves in the same sentence. One obvious explanation for such differences is the different amounts of details related to all rhetorical moves: research-oriented moves tend to be richer in details than their data-oriented counterparts.

A major interest of this paper is to investigate how the rhetorical moves are used differently across the two journals. Table 3 illustrates the number of papers with specific moves in the two journals. There are two stark differences between the journals. First and foremost, *Related research article* is only used in *Data in Brief*. This can be explained by the fact that nearly all such sentences are accompanied by an in-text citation, which is disallowed by *Scientific Data*. Second, both *Introduction* and *Data uses* are adopted much more heavily in *Scientific Data* than *Data in Brief*. For both moves, their different usages in these journals cannot be explained by policies concerning paper abstracts, but these may be connected to the journals' different paper structures and peer review criteria, which warrants a future research.

Table 3. Counts of papers with moves in the data journals (for each journal, n = 180)

Move	*Scientific Data*	*Data in Brief*
Introduction	141	76
Method	119	130
Data description	132	101
Data uses	135	53
Results	45	52
Purpose	62	77
Related research article	0	90
Conclusion	23	29
Data availability	36	31

Moreover, we also examined how these moves are used in these journals over time. A few key findings emerge from Figure 1. First, most of the moves are used similarly in both journals and consistently used over time. Second, for the three moves that are used differently across the two journals, there seems to be a generally converging trend between the journals. For example, the use of *Related research article* has been decreasing in *Data in Brief* over time and the opposite trend can be observed for *Data uses*.

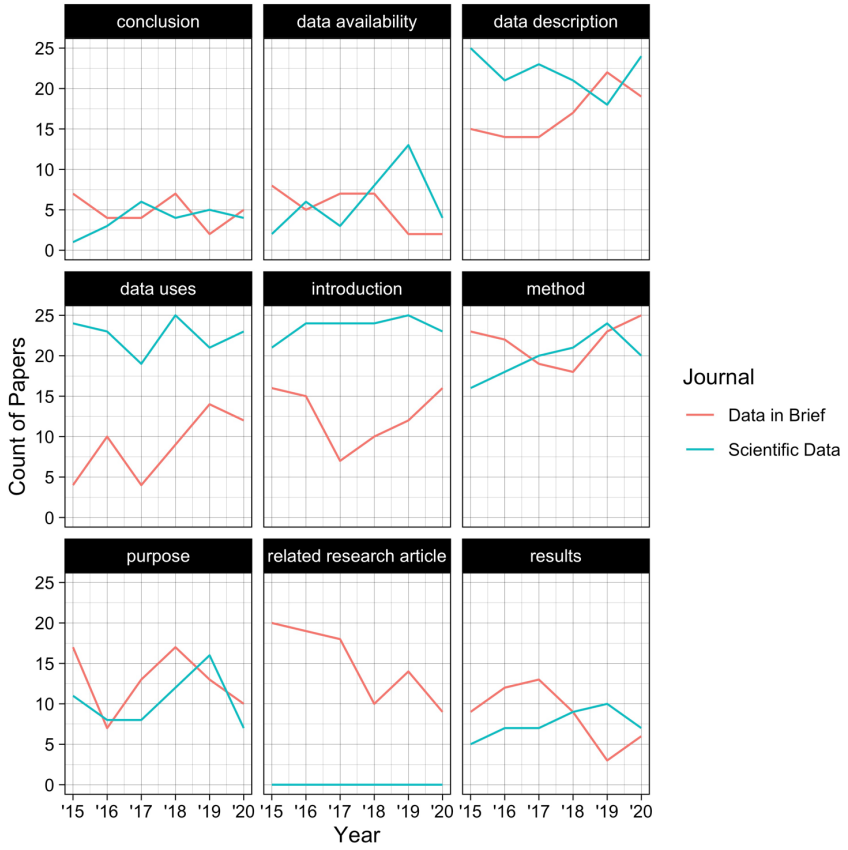

Figure 1. Number of papers with specific moves over time by journal

Conclusions

This research-in-progress paper is part of our larger research project aiming to investigate the rhetoric in data papers, so as to better locate this new academic genre in the scholarly communication system. One step towards this broad goal is to evaluate the differences between data papers and research articles in terms of used rhetorical moves. In this work, we present some preliminary findings concerning how rhetorical moves are used in data paper abstracts in two prominent data journals, *Scientific Data* and *Data in Brief*. We expanded an existing classification system of rhetorical moves in research article abstracts and identified four extra moves that are commonly used in the data-oriented contexts: *Data description*, *Data uses*, *Data accessibility*, and *Related research article*. We found that *Data description* is among the most frequently used rhetorical moves in our sample, along with *Introduction* and *Method*. Moreover, we also identified some notable differences in the use of rhetorical moves between the two journals, some of which can be explained by the differences in their journal policies about paper abstracts.

This work serves as a pointer to some important research directions to be pursued during the next steps of this project. First, it is important to connect the use of rhetorical moves to broader journal policy contexts and the domains in which data papers are produced. For example, some differences in our results may be explained by the different paper structures in these journals. Second, the combination and order of rhetorical moves in the abstract are better indicators of the story being told in academic publications. As a result, they will be studied in our future work to better understand key characteristics of data papers as a rhetorical device. Moreover, we believe a broader selection of data journals will help to reveal the diversity of data publications in the different venues, which will also be considered in our future work.

References

Adav, S. S., Ravindran, A., & Sze, S. K. (2015). Data for iTRAQ secretomic analysis of Aspergillus fumigatus in response to different carbon sources. *Data in Brief*, *3*, 175–179.

Bayer, P., Comunian, A., Höyng, D., & Mariethoz, G. (2015). High resolution multi-facies realizations of sedimentary reservoir and aquifer analogs. *Scientific Data*, *2*(1), 1–10.

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, *7*(1), 107–113. https://doi.org/10.2218/ijdc.v7i1.218

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747–1762. https://doi.org/10.1002/asi.23358

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(15), 1. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S15-S2

Hey, T., Tansley, S., Tolle, K. M., & others. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA. https://www.fh-potsdam.de/fileadmin/user_upload/fb-informationswissenschaften/bilder/forschung/tagung/isi_2010/isi_programm/TonyHey_-__eScience_Potsdam__Mar2010____complete_.pdf

Hrynaszkiewicz, I., & Shintani, Y. (2014). Scientific data: An open access and open data publication to facilitate reproducible research. *J Inf Process Manag*, *57*, 629–640.

Hyland, K. (2000). Speaking as an insider: promotion and credibility in abstracts. *Disciplinary Discourses: Social Interactions in Academic Writing*, 63–84.

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.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. http://www.ncbi.nlm.nih.gov/pubmed/843571

Li, K., & Chen, P. (2018). The narrative structure as a citation context in data papers: A preliminary analysis of Scientific Data. *Proceedings of the Association for Information Science and Technology*, *55*(1), 856–858.

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.

Liu, C., Zhang, D., Shen, Y., Tao, X., Liu, L., Zhong, Y., & Fang, S. (2015). Data in support of DPF2 regulates OCT4 protein level and nuclear distribution. *Data in Brief*, *5*, 599–604.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. *ArXiv Preprint Cs/0205028*.

Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*(0), WDS32--WDS46. http://jlc.jst.go.jp/DN/JST.JSTAGE/dsj/WDS-042?from=Google

Pfeiffenberger, H., & Carlson, D. (2011). " Earth System Science Data"(ESSD)-A Peer Reviewed Journal for Publication of Data. *D-Lib Magazine*, *17*(1/2).

Sikulu, M. T., Monkman, J., Dave, K. A., Hastie, M. L., Dale, P. E., Kitching, R. L., Killeen, G. F., Kay, B. H., Gorman, J. J., & Hugo, L. E. (2015). Mass spectrometry identification of age-associated proteins from the malaria mosquitoes Anopheles gambiae ss and Anopheles stephensi. *Data in Brief*, *4*, 461–467.

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

Small, H. (1982). Citation context analysis. *Progress in Communication Sciences*, *3*, 287–310.

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

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Thelwall, M. (2020). Data in Brief: Can a mega-journal for data be useful? *Scientometrics*, *124*(1), 697–709.

Walters, W. H. (2020). Data journals: incentivizing data access and documentation within the scholarly communication system. *Insights*, *33*(1).

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

Wynholds, L. (2011). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. *International Journal of Digital Curation*, *6*(1), 214–225. https://doi.org/10.2218/ijdc.v6i1.183