RESEARCH ARTICLE

JASIST WILEY

The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts

Kai Li[1] | Chenyue Jiao[2]

[1]School of Information Resource Management, Renmin University of China, Haidian District, Beijing, China

[2]School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

Correspondence
Kai Li, School of Information Resource Management, Renmin University of China, 59 Zhongguancun St, Haidian District, Beijing 100872, China.
Email: kai.li@ruc.edu.cn

Abstract

The data paper is an emerging academic genre that focuses on the description of research data objects. However, there is a lack of empirical knowledge about this rising genre in quantitative science studies, particularly from the perspective of its linguistic features. To fill this gap, this research aims to offer a first quantitative examination of which rhetorical moves—rhetorical units performing a coherent narrative function—are used in data paper abstracts, as well as how these moves are used. To this end, we developed a new classification scheme for rhetorical moves in data paper abstracts by expanding a well-received system that focuses on English-language research article abstracts. We used this expanded scheme to classify and analyze rhetorical moves used in two flagship data journals, *Scientific Data* and *Data in Brief*. We found that data papers exhibit a combination of introduction, method, results, and discussion- and data-oriented moves and that the usage differences between the journals can be largely explained by journal policies concerning abstract and paper structure. This research offers a novel examination of how the data paper, a data-oriented knowledge representation, is composed, which greatly contributes to a deeper understanding of research data and its publication in the scholarly communication system.

1 | INTRODUCTION

The growing amount of research data in the 21st century brings significant changes to the position of data in the scientific system, how scientific research is defined, and how it should be conducted (Callaghan et al., 2012; Gitelman, 2013; Hey et al., 2009). In particular, these recent developments call for greater data transparency in the scholarly ecosystem, as summarized into the FAIR Principles of research data stewardship, that is, research data should be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). A major solution to these new needs is the concept of *data publication*, which refers to the pipeline through which data are transformed into discrete, well-documented, publishable, and citable objects (Parsons & Fox, 2013). Under this framework, the practice of publishing data objects as academic papers, or *data papers*, has been embraced by researchers from a broad spectrum of knowledge domains. A data paper is 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 have been proven to be useful in bridging the gaps between the research system and scientific data. First, published articles are highly effective in attracting citations to data objects and making it easier for data works to receive credit (Borgman, 2016; Jiao & Darch, 2020).

Second, academic journals offer a heavily used peer-review pipeline. Despite the known limitations of the peer review system (Lee et al., 2013), publishing data in academic journals is the easiest solution for data to be peer-reviewed (Huang et al., 2013; Mayernik et al., 2015). These reasons support the growing popularity of data papers across various knowledge domains. For example, Candela and colleagues' survey (Candela et al., 2015) identified seven *dedicated, or exclusive, data journals* (journals that only publish data papers) and over a 100 mixed data journals.

As an emerging academic genre, data papers are still far from visible in our research system. A chief reason for the invisibility is a lack of understanding of the textual composition of data papers as an academic genre and how this composition differs from that of regular research articles. On the one hand, data papers are only supposed to describe the collection, processing, and contents of their respective datasets, instead of presenting a study design and results (Callaghan et al., 2012). This definition should give data papers a distinct structure from research articles, for which the introduction, method, results, and discussion (IMRaD) format has been dominant in academic journals during the past several decades (Sollaci & Pereira, 2004). On the other hand, data papers inevitably share many textual similarities with research articles, as both genres are produced under comparable social norms and contexts of scientific knowledge production, such as the peer review system (Li et al., 2020). Besides the dilemma addressed above, the peer review criteria and processes of data journals can also impact the textual features of data publications, which, however, have just recently been studied (Seo & Kim, 2020).

The above issues make the textual features of data papers an important topic in quantitative science studies, because such features bear strong theoretical implications about how knowledge is produced and communicated from a sociotechnical perspective (Bazerman, 1983; Small, 1982; Swales, 1990) and understanding them will facilitate a smoother integration of data papers into the research system. First, sociolinguistic decisions made by researchers could help us better understand the nature of data papers as a representation of scientific knowledge distinct from research articles and to appreciate their functions in the scholarly communication system. Second, data paper composition practices should be more thoroughly considered in future open data policies, because data papers have become one of the most important means of publishing research data. However, no empirical research has yet investigated data papers from this critical perspective.

To fill this gap, the present study offers a quantitative examination of rhetorical moves used in abstracts of data papers. 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). It has been used as a basic research instrument in various research communities focusing on academic writing, ranging from applied linguistics to quantitative science studies (Bazerman, 1988; Cross & Oppenheim, 2006; Swales, 1990; Thelwall, 2019). In this research, we examined the abstracts of data papers because, first of all, the abstract is commonly regarded as an accurate summary of the content of a scientific publication (Hall, 2012). Moreover, the abstract is also the section most frequently accessed by readers and bears a strong news value (Berkenkotter & Huckin, 2016). For these reasons, there is a large body of empirical works examining various linguistic attributes of research article abstracts, especially the move structure of the abstract (Hyland, 2000b; Lorés, 2004; Salager-Meyer, 1992; Stotesbury, 2003).

In this work, we analyzed what rhetorical moves are used, and how they are used, in the abstracts of 360 data papers selected from two flagship data journals, *Scientific Data* and *Data in Brief*. Using this paper sample, the following research questions are pursued in this study:

> **RQ1.** What rhetorical moves are used in abstracts of data papers? This question aims to propose a classification scheme for the rhetorical moves used in data paper abstracts, focusing especially on those moves that are not shared by research articles. In this research, we expanded a highly cited scheme for rhetorical moves in English-language research paper abstracts (Hyland, 2000b) by manually examining 50 randomly selected data paper abstracts. The revised classification system was further validated during the manual classification process. This new system offers a solid methodological foundation for the present work and any future works on similar topics.

> **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. Based on the modified classification system, we manually classified all sentences in our paper sample to understand the distribution of rhetorical moves in the space of data paper abstracts. Specifically, we compared how the identified moves were used differently by the journals and sought to explain their differences using the journal policies related to data paper

abstract and paper structure. We also examined how the use of moves evolves over time on the journal level. This question deepens our existing knowledge about data papers as an academic genre by introducing a more dynamic perspective.

RQ3. What is the order in which these moves are structured in the abstract? This question examines the distribution of rhetorical moves within the sequence of abstract texts, with the purpose of understanding how rhetorical moves are used to construct data paper abstracts in a more granular manner. We examined how different rhetorical moves are used in different positions in the abstract, which sheds light on the deeper textual characteristics of different rhetorical moves involved in data papers.

2 | LITERATURE REVIEW

2.1 | Data publication and data papers

With the significant growth of research data, the demand for data sharing has also been increasing, so that more data can be used to support knowledge production and scientific reproducibility (Borgman, 2015). Data publication is an important instrument for making data easier to access, understand, and reuse (Parsons & Fox, 2013), to eventually support more effective reuse of data. Data papers have been generally recognized as a new form of data publication and have been increasingly adopted in many research communities (Candela et al., 2015; Gorgolewski et al., 2013). A major distinction between data papers and research article is their focus: while both types of publications are normally peer-reviewed, data papers are designed to describe data objects (Carlson & Oda, 2018), a major divergence from the IMRaD paper structure (Sollaci & Pereira, 2004).

The unique structure of data papers can greatly contribute to data sharing and reuse. First, effective data sharing requires not only the availability of the data object per se but also sufficient metadata to justify the reusability of the data (Zimmerman, 2007). Data papers thus function as an important incentivizing instrument for researchers to undertake the preparation and sharing of data and its accompanying metadata (Gorgolewski et al., 2013). Second, despite the different paper structures across journals, most require authors to detail methods used to produce the data object and the repository of the dataset (Kim, 2020). Detailed information about how the data object was collected, processed, and shared can support the discovery and evaluation of data papers by other researchers (Faniel et al., 2019; Wang et al., 2021).

During the past few years, data papers have become an increasingly popular type of scientific publication. Candela et al. (2015) identified over 100 journals that accept data papers, and there is evidence that the number of data papers has continued to increase since then (El-Tawil & Agrawal, 2019). Empirical evidence has also shown that data papers are frequently cited, although the citation pattern normally takes on a highly long-tailed distribution (Kotti & Spinellis, 2019), and the papers may be cited for reasons other than reuse of the data (Jiao & Darch, 2020). Much more empirical investigation, using quantitative and bibliographic data, is needed before we can comprehensively understand the data paper's roles and positions in the scholarly communication system, especially as distinct from regular research articles. The present paper, focusing on the rhetorical moves of data papers, is a step toward this goal.

2.2 | Research paper abstract as an academic genre

The abstract is an indispensable component of research articles and is required by most journals as part of an article submission (Martín, 2003). Hyland (2015) observes that "the abstract is generally the readers' first encounter with a text, and is often the point at which they decide whether to continue and give the accompanying article further attention or to ignore it" (p. 63). Underlying this phenomenon is the function of abstracts as a somewhat accurate summary of research articles (Hall, 2012; Salager-Meyer, 1990). This "news value" has been further highlighted in recent years as the pace of academic publication accelerates (Landhuis, 2016).

Given their centrality in the genre of research articles, research article abstracts have been heavily studied in the applied linguistics community, particularly from the perspective of rhetorical moves. For example, dos Santos (1996) and Hyland (2000b) have each proposed a five-step classification system for rhetorical moves used in English-language research article abstracts. The systems are similar in that they cover the most basic functional units in the IMRaD structure, but Hyland's system is regarded as superior because it was built on a much larger sample size from a broader selection of knowledge domains (Suntara & Usaha, 2013).

Apart from these two classification systems, a large number of empirical works have examined how rhetorical moves are used to construct research article abstracts in different knowledge domains and/or countries (Dahl, 2004;

Gillaerts & van de Velde, 2010; Hyland, 2000a; Lorés, 2004; Saeeaw & Tangkiengsirisin, 2014; Samraj, 2002; van Bonn & Swales, 2007). Despite their different scopes, most of these studies (a) have the common aim of supporting pedagogical activities concerned with English-language research article writing and (b) are based on relatively small paper samples and the manual coding method.

Moreover, this line of research is situated in a sociolinguistic approach to scientific documents that can have meaningful conversations with quantitative science studies. Earlier works have shown how textual features, when combined with other quantitative methods (especially the citation analysis method), serve as important indicators of the research community of publications and their roles in scholarly communication (Swales, 1986; White, 2004). We believe this approach will likewise contribute to a deeper understanding of the emerging and heretofore rarely studied genre of data papers. This paper is the first step toward this novel research direction.

3 | METHODS

3.1 | Data collection

In this study, we analyzed data papers included in *Scientific Data* and *Data in Brief*, two domain-independent flagship data journals. The two journals were selected because, first, they are the two leading exclusively data journals, based on journal impact factor, number of publications, and presence in empirical studies (Kim, 2020; Walters, 2020). For this reason, we believe these journals offer a representative sample of data publications. Second, both journals were founded in 2014, which provides a meaningful time frame to conduct temporal analyses. We collected all metadata records in these journals from Scopus on November 15, 2020.

While both journals were categorized as exclusively data journals by Stuart (2017), they contain document types beyond data papers. As an illustration, Scopus classified most publications in *Data in Brief* as *Data Papers*, but almost 75% of *Scientific Data* publications as *Articles*. We validated the Scopus classification with the information on the journal websites. We found that the "data descriptors," as they are classified by *Scientific Data*, were categorized as either *Data Papers* or *Articles* by Scopus, whereas most data papers in *Data in Brief* were classified as such in Scopus. Based on this information, we removed all publications that are NOT classified as data papers by the journal websites, including both research articles and other publication types (most notably, errata and editorial). Our final sample includes 7,712 data papers from both journals, with 6,335 from *Data in Brief* and 1,377 from *Scientific Data*.

From these journals, we selected only data papers published between 2015 and 2020, as both journals were founded in 2014 and there are insufficient papers to analyze in that year. To enable a more meaningful comparison between the journals, we used a stratified sampling approach in which we randomly selected 30 publications per journal per year. There are 360 data papers in our final sample for manual classification.

For all selected publications, we used the the Natural Language Toolkit (NLTK) Python package (Loper & Bird, 2002), a popular natural language processing library used on scientific corpora, to parse abstracts into sentences. After the automatic processing, we examined the results and manually fixed any mistakenly parsed sentences. A total of 2,182 sentences were parsed and analyzed in the following steps.

3.2 | Manual classification of rhetorical moves

This research aims to identify rhetorical moves used in data paper abstracts on the sentence level. For this purpose, we expanded the classification scheme proposed by Hyland (2000b) for rhetorical moves in English-language research article abstracts, which is composed of *Introduction*, *Purpose*, *Method*, *Product*, and *Conclusion*.

To align this scheme with data papers, two coders manually coded 50 randomly selected abstracts (from the original 7,712 papers) and independently identified new categories not covered by the scheme. In the end, we added four new rhetorical moves found in our paper sample. Our modified scheme is shown in Table 1, where we highlight the four added moves and state how they are defined in this project. Moreover, we changed *Product* to *Results* in our scheme while retaining the original definition proposed by Hyland. This scheme was tested during our manual classification, but we did not find any other category to add.

The majority of the new moves focus on research data, in contrast to the traditional IMRaD-oriented moves. These two orientations support the data publication genre, which follows the basic rhetoric of research articles but is dedicated to the description, discovery, and reuse of data objects. We also identified the move of Related Research Article, which serves to link a data paper to another research article where the dataset was initially used. Examples of these moves from our paper sample are offered in Table 2.

Two coders independently classified all sentences using the modified classification scheme. Our coding also allows the co-existence of multiple moves in one sentence. In the final coding, 77 sentences contain two moves. For moves in these sentences, we used fractional

TABLE 1 Classification and definition of rhetorical moves in data paper abstracts

Move	Definition
Introduction	Context of the paper
Purpose	Purpose or intention of the paper/research
Method	Research design, procedure, assumptions, and approach of the study
Results	Main findings or results
Conclusion	Interpretations of the results beyond the scope of the paper
Data Description	Description of the data object that is the topic of the paper
Data Uses	Proposed uses or implications of the data object
Data Accessibility	How to get access to the data object
Related Research Article	The research article to which the data object is connected

counting to measure the frequencies of moves on the sentence level, counting each move as 0.5 instead of one sentence. The kappa intercoder reliability is 0.706, which indicates a good agreement (Landis & Koch, 1977). All differences between the coders were resolved before data analysis.

3.3 | Abstract-related journal policies

To understand how the rhetorical moves reflect the policies of the two data journals, we also collected policy statements related to data paper abstracts from both journals, focusing on the following two broad areas:

3.3.1 | Abstract content and structure

This area concerns policies that directly regulate the content and structure of data paper abstracts. As shown in Table 3, the journals have very similar policies, except that Scientific Data requires that no reference be given in the abstract

3.3.2 | Paper structure

We also collected paper structure specifications to understand how the paper structure influences the uses of rhetorical moves in paper abstracts. This takes into account

TABLE 2 Examples of rhetorical moves as used in our paper sample

Move	Example
Introduction	"*Aspergillus* sp. plays an important role in lignocellulosic biomass recycling." (Adav et al., 2015)
Purpose	"This data article describes a detailed synthetic strategy and experimental data for the synthesis of brominated fulvene chromophores…" (Budy et al., 2018)
Method	"Firstly, we establish a robust approach to extracting shorelines from vertical aerial photography, validated against LiDAR (Light Detection and Ranging) and coastal topography surveys." (Pollard et al., 2019)
Results	"We found that the depletion of Jumonji domain-containing protein 6 (JMJD6) […] slowed cell proliferation of mouse NIH3T3 fibroblasts." (Hu & Imbalzano, 2016)
Conclusion	"As coastal systems are monitored at greater spatial resolution and temporal frequency there is an unprecedented opportunity to determine how and why coastal systems have changed in the past with a view to informing future forecasting." (Pollard et al., 2019)
Data Description	"The data presented here represents the detailed comparative abundances of diverse groups of biomass hydrolyzing enzymes […] and their post translational modification like deamidation." (Adav et al., 2015)
Data Uses	"The Australian Phytoplankton Database […] allows analysis of ecological indicators of climate change and eutrophication (e.g., changes in distribution; diatom: dinoflagellate ratios)." (Davies et al., 2016)
Data Accessibility	"We have lodged this dataset with the Australian Ocean Data Network (http://portal.aodn.org.au/) allowing public access." (Davies et al., 2016)
Related Research Article	"Here, we provide data related to the research article entitled 'Quantitative proteomics study of *Aspergillus fumigatus* secretome revealed deamidation of secretory enzymes' by Adav et al. (J. Proteomics (2015) [1])." (Adav et al., 2015)

the fact that most traditional rhetorical moves correspond to major IMRaD paper sections. As summarized in Table 4, the two data journals have distinct mandated paper structures. It should be noted that these structures are consistently used in their respective journals based on our manual examination and previous evidence (Li & Chen, 2018).

3.4 | Descriptive analysis: Length and complexity of data paper abstracts

Figure 1 summarizes the frequency of papers with specific numbers of rhetorical moves in the paper abstract.

TABLE 3 Journal policies concerning content and structure of data paper abstracts

Policy area	Scientific data	Data in brief
Abstract content and structure	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.[a]	Concisely describes the data, its collection process, analysis, and reuse potential. Do not: provide conclusions, results, or mention the word "study."[b]

[a]https://www.nature.com/sdata/publish/submission-guidelines
[b]https://www.elsevier.com/__data/promis_misc/DIB%20Article%20Template%203.2.1.docx

The graph shows a normal distribution, where over 80% of articles have three to five rhetorical moves. This pattern is relatively similar across the two journals: the mean number of moves per abstract is 3.85 in *Scientific Data* and 3.57 in *Data in Brief*. This difference mirrors the difference in abstract length between these journals (mean of 6.38 sentences in *Scientific Data* and 5.74 in *Data in Brief*).

4 | RESULTS

4.1 | How often are the rhetorical moves used in data paper abstracts?

Table 5 illustrates how the rhetorical moves are distributed in our sample. Introduction, Method, Data Description, and Data Uses are seen to be the most frequently used moves on both the sentence and paper levels. This group of moves covers both IMRaD- and data-oriented functions. A notable finding, contrasting with the journal policies, is that the category of Results is frequently used in abstracts: more than 25% of articles have at least one Results sentence. Following our previous work (Li et al., 2020), this finding challenges the boundaries between research articles and data papers in terms of paper content.

TABLE 4 Journal policies concerning the structure of data papers

Journal	Paper section	Definition
Scientific Data[a]	Background and Summary	This section should provide an overview of the study that generated the data, as well as outlining the potential reuse value of the data.
	Methods	The Methods section in Data Descriptors should describe any steps or procedures used in producing the data.
	Data records	This section should be used to explain each data record associated with this work, including the repository where this information is stored, and to provide an overview of the data files and their formats.
	Technical validation	This section should present any experiments or analyses that are needed to support the technical quality of the dataset.
	Usage Notes	"Usage Notes" is an optional section that can be used to provide information that may assist other researchers who reuse your data.
Data in Brief[b]	Value of Data	• Why are your data useful or important to the research community? • Who can use/benefit from the data? • How can the data be (re)used for further insights or analysis?
	Data Description	Briefly, describe and refer to each data file included in the article and supplementary files or in the repository.
	Experimental Design, Materials, and Methods	Provide a detailed description of how the data were acquired and analyzed. Aid data reproduction and reuse by offering a more comprehensive description than any related article or previous articles.

[a]The required paper structure, as well as the specific requirements of *Scientific Data*, are taken from: https://www.nature.com/sdata/publish/submission-guidelines
[b]The required paper structure of *Data in Brief* is taken from: https://www.elsevier.com/__data/promis_misc/DIB%20Article%20Template%203.2.1.docx.
Specifications of the paper sections are taken from: https://www.elsevier.com/__data/promis_misc/Data%20in%20Brief%20Article%20at%20a%20Glance%20illustration.pdf
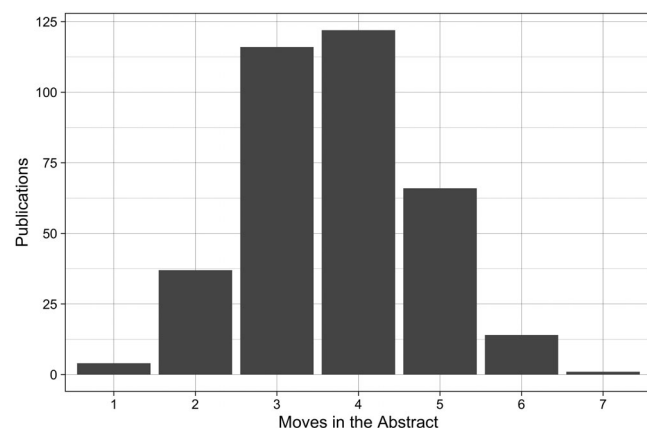
FIGURE 1 Distribution of data papers by number of moves in the abstract ($n = 360$)

TABLE 5 Distribution of rhetorical moves in our sample

Rhetorical moves	Sentences ($n = 2,182$)	Papers ($n = 360$)
Introduction	514	217
Method	527	249
Data Description	357	233
Data Uses	227.5	188
Results	185.5	97
Purpose	149	139
Related Research Article	93.5	93
Conclusion	67.5	52
Data Accessibility	61	67

TABLE 6 Comparison of IMRaD move use in three paper samples

Rhetorical moves	Data papers ($n = 360$)	Amnuai (2019) ($n = 30$)	Farzannia and Farnia (2017) ($n = 30$)
Introduction	60.2%	60%	50%
Purpose	38.6%	90%	86.7%
Method	69.1%	80%	90%
Results	26.9%	90%	93.3%
Conclusion	14.4%	46.7%	63.3%

Abbreviations: IMRaD, introduction, method, results, and discussion.

Based on Table 5, we compared how the IMRaD moves are distributed in data papers versus research articles. As the control group, we selected two recent works, Amnuai (2019) and Farzannia and Farnia (2017), both of which examined rhetorical moves used in English-language journal research articles based on Hyland's classification system. The two works focus on accounting and mining engineering, respectively. We acknowledge the difficulty of finding empirical works that can directly match the scope of the journals analyzed in the present work, given the multidisciplinary nature of both data journals. As such, this comparison is only a preliminary step toward a more comprehensive understanding of the differences between data papers and research articles.

As shown in Table 6, the two distributions based on research articles are highly similar to each other; Purpose, Method, and Results are used in more than 80% of articles in each sample. By comparison, our sample of data papers does not have any dominant categories, partly because of the larger number of categories used in data papers. Purpose, Results, and Conclusion categories are all used much less extensively in data papers than in the two baseline samples; in the latter two cases, this may be explained by the fact that these categories are forbidden by journal policies. On the other side of the story, Introduction is used similarly across corpora, which may indicate that it is one of the most basic building blocks of narratives in both genres.

Table 7 summarizes how the moves are used differently by the two journals. There are two notable differences: first, the Related Research Article is only used in *Data in Brief*. This can be explained by the fact that *Data in Brief* asks authors to include a citation to the research article when the data paper is submitted via another Elsevier journal (co-submission).[1] However, references are specifically forbidden in *Scientific Data*.

Second, both Introduction and Data Uses are adopted much more heavily in *Scientific Data*, which echoes the

inclusion of corresponding paper sections in *Scientific Data*'s paper structure. Introduction directly corresponds to the Background and Summary section; Data Uses, to the Usage Notes section. Moves covered by both structures, especially Methods and Data Description, seem to be used similarly across journals. Even though it is difficult to find a direct match between every rhetorical move and paper section, there seems to be a general relationship between the availability of a paper section and how its corresponding move is used in the abstract.

Figure 2 shows the temporal trend of how moves are used on the journal level. We found that most of the moves are relatively consistently used over time by the two journals, despite yearly fluctuations, indicating a relatively stable landscape in the paper sample we examined.

4.2 | Order of rhetorical moves in abstracts

To understand the order of rhetorical moves in data paper abstracts, we measured which moves are used as the first and last moves in the abstract. Table 8 shows the top five moves used as the first move in the abstract in

TABLE 7 Summary of rhetorical move use on sentence- and paper-levels in the two data journals

Moves	Sentence-level		Paper-level	
	Count and percentage in Data in Brief ($n = 1,033$)	Count and percentage in Scientific Data ($n = 1,149$)	Count and percentage in Data in Brief ($n = 180$)	Count and percentage in Scientific Data ($n = 180$)
Introduction	169 (16.4%)	345 (30%)	76 (42.2%)	141 (78.3%)
Method	303.5 (29.4%)	223.5 (19.5%)	130 (72.2%)	119 (66.1%)
Data Description	142.5 (13.8%)	214.5 (18.6%)	101 (56.1%)	132 (72.3%)
Data Uses	61 (5.9%)	166.5 (14.5%)	53 (29.4%)	135 (72.5%)
Results	115.5 (11.2%)	70 (6.1%)	52 (28.9%)	45 (25%)
Purpose	79.5 (7.7%)	69.5 (6.1%)	77 (42.8%)	62 (34.4%)
Related Research Article	93.5 (9.1%)	0	93 (51.7%)	0
Conclusion	39.5 (3.8%)	28 (2.4%)	29 (16.1%)	23 (12.6%)
Data Accessibility	29 (2.8%)	32 (2.8%)	31 (16.3%)	36 (20%)

FIGURE 2 Use of rhetorical moves in the two journals over time
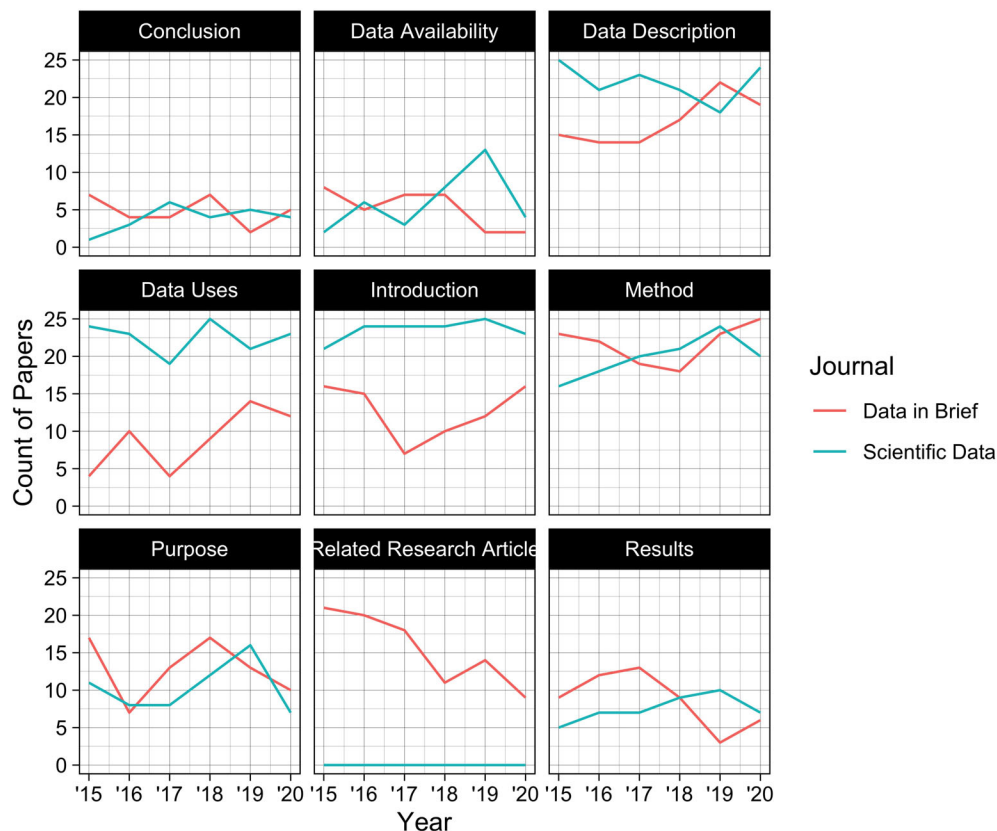
our overall sample (left panel), *Data in Brief* (middle panel), and *Scientific Data* (right panel). The results show that Introduction is the opening move in more than 50% of abstracts, followed by Data Description and Related Research Article. However, this pattern is vastly different across the two journals: in *Scientific Data*, nearly 75% of articles use Introduction as the first move, while this number is lower than 40% in *Data in Brief*. Aligned with results in the previous section, this difference can be explained by the fact that *Scientific Data* has an Introduction section at the beginning of its paper structure. In *Data in Brief*, however, Related Research Article, Data Description, and Purpose can frequently be used to start the abstract despite their quite different rhetorical functions. This result suggests a higher level of standardization in data paper abstracts in *Scientific Data*.

TABLE 8 Most frequent first moves in data paper abstracts from the full sample and individual journals

Full sample		Data in Brief		Scientific Data	
Move	**Percentage of articles**	**Move**	**Percentage of articles**	**Move**	**Percentage of articles**
Introduction	55.6%	Introduction	36.7%	Introduction	74.4%
Data Description	16.9%	Related Research Article	21.1%	Data Description	13.9%
Related Research Article	10.6%	Data Description	20%	Method	4.4%
Purpose	8.3%	Purpose	13.3%	Purpose	3.3%
Method	4.7%	Method	5%	Data Uses	1.1%

TABLE 9 Most frequently last moves in data paper abstracts from the full sample and individual journals

| Full sample		Data in Brief		Scientific Data	
Move	**Percentage of articles**	**Move**	**Percentage of articles**	**Move**	**Percentage of articles**
Data Uses	37.5%	Data Uses	19.4%	Data Uses	55.6%
Data Description	9.4%	Related Research Article	17.8%	Conclusion	9.4%
Conclusion	9.2%	Results	12.2%	Data Description	8.3%
Data Accessibility	9.2%	Method	11.7%	Data Accessibility	7.8%
Related Research Article	8.9%	Data Description	10.6%	Method	5%
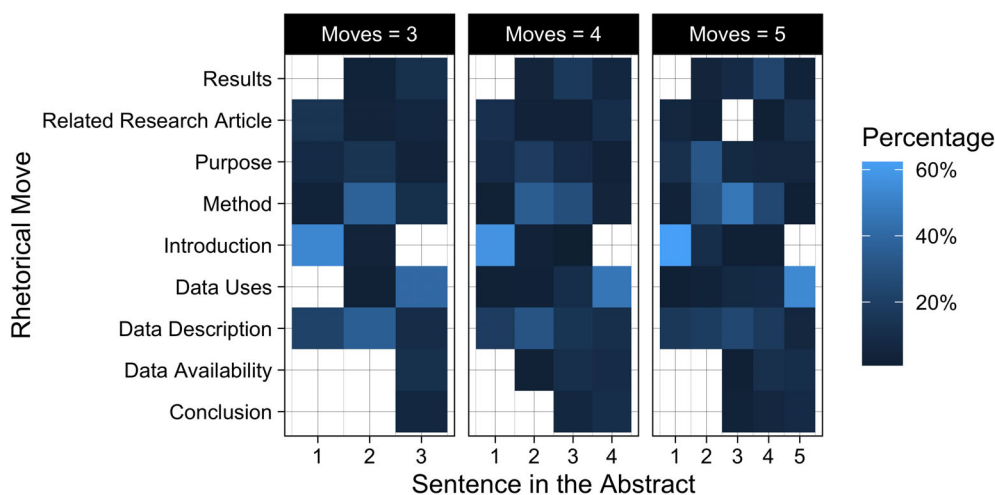
FIGURE 3 Likelihood of rhetorical move use by position in abstracts with three to five moves

Moreover, it is worth noting that the categories of Results, Data Accessibility, and Conclusion are never used as the first move in the abstract.

Table 9 shows the distribution of the last moves, following the structure of Table 8. The results show that the last move is much less consistent than the first; however, as in Table 8, *Scientific Data* is more standardized than *Data in Brief*, as illustrated by the fact that Data Uses appears as the last move in more than 50% of its publications, whereas no single move closes the abstract of more than 20% of *Data in Brief* publications. The fact that Data

Uses is the most frequently used last move in *Scientific Data* may be partly explained by the fact that Usage Notes is the last section in its paper structure.

Figure 3 shows the distribution of moves in each move sequence when move complexity is between three and five, the three most dominant levels based on Figure 1. The color of each box represents the percentage of all moves in the position that belongs to a specific category. For example, in all three scenarios, Introduction is the most likely to be used as the first move in the abstract. Overall, the graph shows a strong similarity across the three scenarios, which suggests

that even though new moves may be added when the abstract becomes longer, the basic relationship between rhetorical moves and data paper abstracts remains consistent.

5 | DISCUSSION

5.1 | Rhetorical moves used in data papers

The present research strives to understand what rhetorical moves are used in the emerging genre of data paper abstracts. By using the manual coding method, we found that data paper abstracts involve both traditional IMRaD rhetorical moves and new data-focused moves. These two general categories support the two-sided identity of the data paper, that is, it is both a scientific publication following traditional writing norms and a descriptor of data objects. These two aspects of identity embodied in the genre of data papers are not easily distinguishable from each other based on our results as well as earlier evidence (Li et al., 2020).

Our results show that traditional rhetorical moves, as summarized by Hyland (2000b), still have a strong presence in data paper abstracts. For example, Introduction and Methods are the two most frequently used moves in our sample, each appearing in over 60% of examined articles. Moreover, even though both journals specifically required that results not be discussed in data papers, we find that the Results move is used in about 25% of data papers in the sample. One potential explanation of this ostensible difference is that results may have different meanings in these genres, which warrants a future study. Overall, these results indicate that the traditional building blocks of narratives in research articles remain useful in constructing the new genre of data papers, even though they are no longer the sole focus of the new genre of data papers.

More interestingly, data papers introduce new rhetorical moves that are strongly related to the goals of promoting data object sharing and reuse (Jiao & Darch, 2020), including Data Description, Data Uses, Data Accessibility, and to a lesser extent, Related Research Article. The first three moves serve to offer more granular descriptions of the data objects that are the focus of data papers. Data Description and Data Uses are each used in more than 50% of articles, showing the popularity of these moves. Moreover, both Data Description and Data Uses have been increasingly used in our sample. Generally, these data-related moves have become important components in the narratives of data papers, where they are combined with traditional blocks in the IMRaD structure.

5.2 | How these moves are used across journals and over time?

Based on our modified classification scheme, this paper quantitatively examines how the rhetorical moves are used by the two selected journals to gain a dynamic understanding of the genre of data papers. Our results testify that there is general consistency in how rhetorical moves are used in data publications. Most of the move categories are used by both journals and many of them are used similarly across journals, in terms of both frequency and order. This indicates that the data paper can be regarded as a distinct and somewhat stabilized academic genre in terms of its linguistic features.

Despite this general consistency, however, there are some important differences between the journals, which sheds important light on the diversity of the genre of data papers. Moreover, we found that many of these differences can be at least partly explained by journal policies about how the abstract should be constructed or the overall paper structure.

One of the starkest differences lies in the fact that the Related Research Article move is only used in *Data in Brief*, which can be explained by the fact that this journal allows a reference to another research article that is co-submitted with the data paper, whereas *Scientific Data* specifically forbids the use of references in paper abstracts. Moreover, the much higher usage of Introduction and Data Uses moves in *Scientific Data* mirrors the use of paper sections in this journal with similar rhetorical functions, which supports the abstract's role as a somewhat accurate summary of the paper content as reported in applied linguistics (Hall, 2012).

On a higher level, these differences indicate the diversity of practices in using the emerging genre of data papers to describe data objects in the current sociotechnical environment. One consequence of this diversity is that different journal norms or policies may not have the same capacity to support the discovery and reuse of data objects, which is an important topic to study in future works. Moreover, it can be expected that more conversations will be needed in the future to develop more fine-grained policies about what data papers should describe, as this genre becomes more popular and undertakes greater responsibility in scholarly communication.

6 | CONCLUSION

This paper offers a quantitative study of the rhetorical moves used in data paper abstracts, by analyzing 360 data papers acquired from *Scientific Data* and *Data in Brief*, two flagship data journals. We developed a classification

scheme for rhetorical moves used in data papers by expanding a similar scheme focusing on research paper abstracts. Using the scheme as the basis of manual classification, we analyzed how these moves are used between the two journals. We found that IMRaD-oriented moves and data-focused moves are mixed in data paper abstracts, supporting a two-sided and intertwined identity of this emerging genre. Moreover, in comparing the journals, we found that journal policies are a major factor underlying the different usage of these moves, which may have long-term implications for the communication of data-focused knowledge and outputs in the research system.

This research addresses the urgent need to understand data papers as an emerging academic genre in the scholarly communication system, particularly from a sociolinguistic perspective, because of the increasing popularity and unique linguistic features of this genre. This work illustrates the diversity and uncertainty of practices in the early years of a new data-centric academic genre and shows how sociotechnical factors, especially journal policies, influence these practices. As a result, this work sheds light on how research data are involved in new modes of knowledge production and highlights important issues in the identity of the data publication. Our results will also serve as an important baseline dataset for future works on the rhetorical structure of data papers, especially those focusing on the disciplinarity of data papers and the differences between data papers and research articles.

Despite its timely contributions, this paper is only the first step toward the goal of understanding data publication as rising academic practices and outputs, and it has a few limitations that should be addressed in future works. First, we only analyzed two domain-independent data journals in this work and did not consider how the disciplinarity of data papers influences their textual features. We acknowledge that academic writing practices are deeply situated in the disciplinary contexts, as shown in earlier empirical evidence (Saeeaw & Tangkiengsirisin, 2014; Suntara & Usaha, 2013). Given the limitations of journal-based classification systems implemented in the Web of Science or Scopus (Shu et al., 2019), we believe that this issue will need to be addressed on the infrastructure level, by developing a more effective approach to classifying data papers by research domain. Second, this study offers only a very preliminary comparison between data papers and research articles; future work in this vein will greatly contribute to a more accurate distinction between data papers and research articles in scientific databases, especially in mixed data journals (i.e., journals accepting both genres discussed above). Third, our conclusions may also be influenced by the different peer-review policies and practices implemented by data journals, even though we believe this is not a central factor

that determines how the paper abstracts were constructed. We plan to address all these limitations in the next step of our project to better situate data papers in our existing scholarly communication system.

ORCID

Kai Li https://orcid.org/0000-0002-7264-365X

ENDNOTE

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.

Amnuai, W. (2019). Analyses of rhetorical moves and linguistic realizations in accounting research article abstracts published in international and Thai-based journals. *SAGE Open*, *9*(1), 2158244018822384.

Bazerman, C. (1983). Scientific writing as a social act: A review of the literature of the sociology of science. In *New essays in technical and scientific communication: Research, theory, practice* (pp. 156–184). Baywood.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science* (Vol. 356). University of Wisconsin Press Madison.

Berkenkotter, C., & Huckin, T. N. (2016). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Routledge.

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

Borgman, C. L. (2016). Data citation as a bibliometric oxymoron. In *Theories of informetrics and scholarly communication* (pp. 93–115). De Gruyter.

Budy, S. M., Son, D. Y., Balaich, G. J., & Iacono, S. T. (2018). Synthesis of reactive 1,3-diphenyl-6-aryl-substituted fulvene chromophores. *Data in Brief*, *19*, 1638–1642. https://doi.org/10.1016/j.dib.2018.06.026

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

Carlson, D., & Oda, T. (2018). Editorial: Data publication–ESSD goals, practices and recommendations. *Earth System Science Data*, *10*, 2275–2278.

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(15), 1.

Cross, C., & Oppenheim, C. (2006). A genre analysis of scientific abstracts. *Journal of Documentation*, *62*(4), 428–446.

Dahl, T. (2004). Some characteristics of argumentative abstracts. *Akademisk Prosa*, *2*(1), 49–67.

Davies, C. H., Coughlan, A., Hallegraeff, G., Ajani, P., Armbrecht, L., Atkins, N., Bonham, P., Brett, S., Brinkman, R., & Burford, M. (2016). A database of marine phytoplankton abundance, biomass and species composition in Australian waters. *Scientific Data*, *3*(1), 1–12.

Dos Santos, M. B. (1996). The textual organization of research paper abstracts in applied linguistics. *Text & Talk*, *16*(4), 481–500.

El-Tawil, S., & Agrawal, A. K. (2019). Data papers: A new submission category. *Journal of Structural Engineering*, *145*(12), 1619001.

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.

Farzannia, S., & Farnia, M. (2017). Genre-based analysis of English and Persian research article abstracts in mining engineering journals. *Beyond Words*, *5*(1), 1–13.

Gillaerts, P., & van de Velde, F. (2010). Interactional metadiscourse in research article abstracts. *Journal of English for Academic Purposes*, *9*(2), 128–139. https://doi.org/10.1016/j.jeap.2010.02.004

Gitelman, L. (2013). *Raw data is an oxymoron*. MIT Press.

Gorgolewski, K., Margulies, D. S., & Milham, M. P. (2013). Making data sharing count: A publication-based solution. *Frontiers in Neuroscience*, *7*, 9.

Hall, G. M. (2012). *How to write a paper*. John Wiley & Sons.

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: Data-intensive scientific discovery* (Vol. 1). Microsoft Research.

Hu, Y.-J., & Imbalzano, A. N. (2016). Global gene expression profiling of JMJD6-and JMJD4-depleted mouse NIH3T3 fibroblasts. *Scientific Data*, *3*(1), 1–6.

Huang, X., Hawkins, B. A., & Qiao, G. (2013). Biodiversity data sharing: Will peer-reviewed data papers work? *Bioscience*, *63*(1), 5–6.

Hyland, K. (2000a). *Disciplinary discourses. Social interactions in academic writing*. Routledge.

Hyland, K. (2000b). Speaking as an insider: Promotion and credibility in abstracts. In *Disciplinary discourses: Social interactions in academic writing* (pp. 63–84). Routledge.

Hyland, K. (2015). *Teaching and researching writing*. Routledge.

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.

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.

Kotti, Z., & Spinellis, D. (2019). Standing on shoulders or feet? The usage of the MSR data papers. 2019 IEEE/ACM 16th international conference on mining software repositories (MSR), pp. 565–576.

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, *535*(7612), 457–458.

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

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17.

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.

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

Lorés, R. (2004). On RA abstracts: From rhetorical structure to thematic organisation. *English for Specific Purposes*, *23*(3), 280–302.

Martín, P. M. (2003). A genre analysis of English and Spanish research paper abstracts in experimental social sciences. *English for Specific Purposes*, *22*(1), 25–43. https://doi.org/10.1016/S0889-4906(01)00033-3

Mayernik, M. S., Callaghan, S., Leigh, R., Tedds, J., & Worley, S. (2015). Peer review of datasets: When, why, and how. *Bulletin of the American Meteorological Society*, *96*(2), 191–201. https://doi.org/10.1175/BAMS-D-13-00083.1

Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*, WDS32–WDS46.

Pollard, J. A., Brooks, S. M., & Spencer, T. (2019). Harmonising topographic & remotely sensed datasets, a reference dataset for shoreline and beach change analysis. *Scientific Data*, *6*(1), 42. https://doi.org/10.1038/s41597-019-0044-3

Saeeaw, S., & Tangkiengsirisin, S. (2014). Rhetorical variation across research article abstracts in environmental science and applied linguistics. *English Language Teaching*, *7*(8), 81–93. https://doi.org/10.5539/elt.v7n8p81

Salager-Meyer, F. (1990). Discoursal flaws in medical English abstracts: A genre analysis per research-and text-type. *Text-Interdisciplinary Journal for the Study of Discourse*, *10*(4), 365–384.

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes*, *11*(2), 93–113.

Samraj, B. (2002). Disciplinary variation in abstracts: The case of wildlife behavior and conservation biology. *Academic Discourse*, *40*, 56.

Seo, S., & Kim, J. (2020). Data journals: Types of peer review, review criteria, and editorial committee members' positions. *Science Editing*, *7*(2), 130–135.

Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225.

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

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.

Stotesbury, H. (2003). Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes*, *2*(4), 327–341.

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

Suntara, W., & Usaha, S. (2013). Research article abstracts in two related disciplines: Rhetorical variation between linguistics and applied linguistics. *English Language Teaching*, *6*(2), 84–99.

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics*, *7*(1), 39–56. http://applij.oxfordjournals.org/content/7/1/39.short

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

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

Thelwall, M. (2019). The rhetorical structure of science? A multidisciplinary analysis of article headings. *Journal of Informetrics*, *13*(2), 555–563.

van Bonn, S., & Swales, J. (2007). English and French journal abstracts in the language sciences: Three exploratory studies. *Journal of English for Academic Purposes*, *6*(2), 93–108.

Walters, W. H. (2020). Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insight*, *33*(1), 18.

Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, *72*(9), 1161–1182. https://doi.org/10.1002/asi.24483

White, H. D. (2004). Citation analysis and discourse analysis revisited. *Applied linguistics*, *25*(1), 89–116+132. https://doi.org/10.1093/applin/25.1.89

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. 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*, 160018. https://doi.org/10.1038/sdata.2016.18

Zimmerman, A. S. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1–2), 5–16 http://link.springer.com/article/10.1007/s00799-007-0015-8