Finding the Paper Behind the Data: Automatic Identification of Research Articles related to Data Publications

Barbara McGillivray[1]
King's College London
barbara.mcgillivray@kcl.ac.uk

Kaveh Aryan[1]
King's College London
kaveh.aryan@kcl.ac.uk

Viola Harperath
King's College London
viola.harperath@kcl.ac.uk

Marton Ribary
Royal Holloway, University of London
marton.ribary@rhul.ac.uk

Mandy Wigdorowitz
University of Alabama
University of Cambridge
University of Johannesburg
mwigdorowitz@ua.edu

[1]*Joint first authors*

Abstract

Data papers are scholarly publications that describe datasets in detail, including their structure, collection methods, and potential for reuse, typically without presenting new analyses. As data sharing becomes increasingly central to research workflows, linking data papers to relevant research papers is essential for improving transparency, reproducibility, and scholarly credit. However, these links are rarely made explicit in metadata and are often difficult to identify manually at scale. In this study, we present a comprehensive approach to automating the linking process using natural language processing (NLP) techniques.

We evaluate both set-based and vector-based methods, including Jaccard similarity, TF-IDF, SBERT, and reranking with large language models. Our experiments on a curated benchmark dataset reveal that no single method consistently outperforms others across all metrics, in line with the multifaceted nature of the task. Set-based methods using frequent words (N=50) achieve the highest top-10% accuracy, closely followed by TF-IDF, which also leads in MRR and top-1% and top-5% accuracy. SBERT-based reranking with LLMs yields the best results in top-N accuracy. This dispersion suggests that different approaches capture complementary aspects of similarity (lexical, semantic, and contextual), showing the value of hybrid strategies for robust matching between data papers and research articles. For several methods, we find no statistically significant difference between using abstracts and full texts, suggesting that abstracts may be sufficient for effective matching. Our findings demonstrate the feasibility of scalable, automated linking between data papers and research articles, enabling more accurate bibliometric analyses, improved tracking of data reuse, and fairer credit assignment for data sharing. This contributes to a more transparent, interconnected, and accessible research ecosystem.

1 Introduction

Data sharing and reuse have become increasingly central to research practices, motivating the development of mechanisms to manage, disseminate, and cite datasets effectively. One response has been the emergence of *data papers*, scholarly publications dedicated to describing datasets in detail, including their structure, provenance, and potential applications (Jiao et al., 2023). Unlike traditional research papers, data papers typically do not present new analyses, but instead contain the context of the creation of the dataset being described, the method for its creation, a description of the dataset itself, a measure of its quality and an explanation of its reuse potential (Reymonet, 2017; Kembellec, Gérald and Le Deuff, Olivier, 2022; Li and Jiao, 2022; Liu, 2022). Previous work has found a relatively high amount of variation in the content and structure of data papers (Li et al., 2020; Jihyun, 2020; Li and Chen, 2018). Data papers have been shown to make datasets more discoverable, citable, and reusable across disciplines (Kosmopoulos and Schöpfel, 2024). Although data papers are sometimes perceived as a recent innovation, their development has been gradual: specialized data journals such as the *Journal of Chemical Engineering Data* were established as early as 1956, while more concerted growth in data journals occurred in the past couple of decades (Candela et al., 2015; Walters, 2020). These journals incentivize open sharing according to best practices, offering authors recognition, citation opportunities, and enhanced reuse potential for their resources.

Following a pyramid model of data-driven research projects (McGillivray et al., 2022), the base comprises project repositories containing scripts, notes, and raw files; the next layer is the structured dataset deposited in a public repository; the third layer is the data paper itself, which documents, contextualizes, and links to the dataset; and the

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 34–43
December 23, 2025 ©2025 Association for Computational Linguistics

apex is the research paper, which interprets and analyzes the data. This pyramid makes explicit that the research paper represents only one possible interpretation of the underlying data, while the structured datasets and data papers facilitate transparency, reproducibility, and alternative analyses.

While this model highlights the continuum from raw data to interpretation, the links between its layers are often implicit or missing in metadata. In this study, we focus on automatically reconstructing one of the most critical links: the pairing of a data paper with a related research article. By "related", we mean that the research article is substantively connected to the dataset described in the data paper, for example, because the article is authored by the same team, builds on the same project, or cites the dataset. One such example is a research paper performing semantic profiling of legal language and cluster analysis on Justinian's *Digest*, a historical sourcebook of Roman law compiled under the order of Emperor Justinian in 533 CE (Ribary and McGillivray, 2020). This analysis is based on a relational database of the (mostly) Latin text of the *Digest* created by the same author and reported in a data paper in the same year (Ribary, 2020). By explicitly identifying these pairs of research and data papers, we provide the foundation for more complete reconstructions of dataset–paper–article triangles which is essential for critically assessing the impact of data sharing and enabling reuse.

2 Previous work

Despite their importance, links between data papers and the research papers that use the associated datasets are often implicit or missing from bibliographic metadata. Manual curation of these links is labour-intensive and difficult to scale given diverse data-sharing practices and the growing volume of publications. Previous work has addressed aspects of this problem from both bibliometric and computational perspectives. McGillivray et al. (2022) proposed simple heuristic rules for identifying meaningful links between data and research outputs in a manual fashion which were also followed to create the gold standard ground-truth dataset for the present study as reported below in Section 3.1. Ekman et al. (2025) conducted a qualitative analysis of the narrative practices in data papers. Li and Jiao (2021) analysed the rhetorical moves within abstracts of data papers published in the journals *Data in Brief* and *Scientific Data*, including the research article to which the dataset is connected. They found that the related research articles are only mentioned in *Data in Brief* abstracts, but this use has decreased over time, while the description of the data, as well as the introduction and method are among the most frequently used rhetorical moves. Kai et al. (2025) calculate TF-IDF to extract keywords that are distinctive of data papers in relation to their citing research papers using a sample of 10 papers, finding that many of the keywords that are characteristic of data papers did not appear in the abstracts, pointing to the importance of analysing the full texts to gain a better picture of these relations.

No previous study has proposed a method for automatically identifying research articles connected to data papers specifically. Instead, there has been growing research on linking datasets to research papers using a combination of Named Entity detection and disambiguation (Heddes et al., 2021), matching through textual embeddings (Färber and Leisinger, 2021) and large language models with retrieval (Datta et al., 2025).

In this work, we present the results of a series of experiments on fully automatic approaches to link English-language research papers and data papers, and thereby reconstruct the pairs connecting data papers and research articles. We systematically evaluate a spectrum of methods, ranging from simple keyword-based text mining to large language model (LLM)-based approaches. Evaluation on a curated gold standard of research–data paper pairs using metrics such as Mean Reciprocal Rank (MRR) and accuracy shows that our approach robustly identifies links, supporting reproducibility, data reuse, and more comprehensive measurement of scholarly impact.

Our task is related to citation recommendation, dataset discovery, and scientific document linking. While existing research, including models like SciLinkBERT (Yu et al., 2025) and other citation-based approaches (Bouziani et al., 2024), effectively utilizes cross-document relationships to enhance tasks like relation extraction and summarization, their focus remains on general citation networks. However, these approaches typically model broad citation structures rather than the specific, functional connection between a data paper and a related research article. This distinction is important because relatedness here is not captured simply by citation counts or co-occurrence, but by a substantive link that situates the dataset within ongoing research workflows. We contribute to this area by targeting a highly specific and often overlooked functional relationship: the link between a research article and its related data paper. This focus is critical for accurately tracking data reuse and measuring scholarly impact, and requires a dedicated, hybrid approach that goes beyond standard citation analysis. Our work has a number of potential practical applications, including enriching repository and publisher metadata with explicit dataset–article links, enriching dataset discovery tools that help researchers find analyses associated with published data, and enabling open science policy compliance by making data use and credit more transparent.

3 Methods

Figure 1 provides an overview of our data collection and processing pipeline, explained in detail in this section. The code is available on `https://github.com/BarbaraMcG/golden-triangle/tree/main/NLP%20Paper`.

3.1 Curated dataset

To identify a comprehensive list of data journals, we started with those compiled by Candela et al. (2015). From this list, we added data journals that were not included in the original list and selected only data journals that publish primarily data papers. The final list consisted of 11 data journals actively publishing in 2022 (see Table 1 for an overview). These data journals could also be found in OpenAlex by their journal name.

McGillivray et al. (2022) present a manually curated dataset containing 107 pairs of data papers and datasets. The dataset included a subset of 38 triangles where the pair of data paper and dataset could be linked up with an associated research paper. These links were curated from two sources: the *Journal of Open Humanities Data* (JOHD) and the *Research Data Journal for the Humanities and Social Sciences* (RDJ). Each of the pairs were manually validated to ensure that the research paper substantially builds upon the dataset described in the corresponding data paper.

The manual curation process was developed from the heuristic rules established in McGillivray et al. (2022). We linked a research paper to a data paper if:

1. at least one of the following four conditions was satisfied:
   (a) the research paper appeared in the reference list of the data paper;
   (b) the research paper was cited in the dataset repository;
   (c) the research paper cited the data paper;
   (d) the research paper cited the dataset.
2. and the following two conditions were also satisfied:
   (a) at least one person was an author of both the data and the research paper;
   (b) the research paper was a substantial, analytical interpretation of the dataset associated with the data paper.

We recognise that rule 2a and 2b need some justification. Rule 2a expresses the requirement that the data paper and research paper are products of the same research effort as opposed to the reuse of data by others for a new research question. This heuristic rule, therefore, creates a link between a data paper and a research paper where data is interpreted by the person who created that data in the first place. Rule 2b that requires "substantial, analytical interpretation of the dataset" is a matter of subjective judgement, and one which resists to be easily translated to a computer script, but such is the nature of the data and the association of data and research papers we work with. Taking the example of the research and data pair mentioned in section 1, substantial analytical interpretation is understood to be a deep manipulation of the data which generates new insights and goes beyond a simple reference. That is, the research paper goes into significant detail about how the data was used, reorganised and processed to answer a research question that the authors set for themselves. As we continue to expand the ground truth in our future work, we aim to create heuristic guidelines to improve consistency among annotators who are currently enjoying a large degree of discretion appropriate to this early stage of the project. For the purposes of the current study, manual curation involved skim-reading research articles to capture such substantial analytical treatment of data. This resulting curated dataset served as our benchmark to assess how different methods of automatic pairing perform.

We sample 159 data papers from three of the largest data journals (see Table 1). Our curated dataset consists of pairs of data papers published in 2022 and related research papers, 91 from

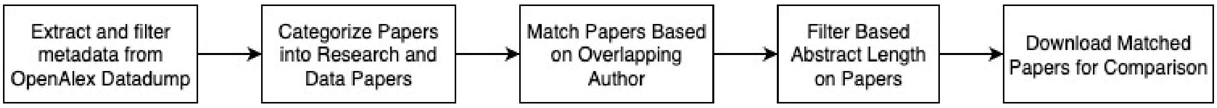
Extract and filter metadata from OpenAlex Datadump
Categorize Papers into Research and Data Papers
Match Papers Based on Overlapping Author
Filter Based Abstract Length on Papers
Download Matched Papers for Comparison

Figure 1: Data Extraction and Processing Steps

Table 1: Data journals used as sources of data papers. The third column contains the number of articles published in the journal in 2022 and the last column contains the number of curated pairs per journal.

Journal	Publisher	2022	Pairs
Biodiversity Data Journal	Pensoft	242	0
Data	MDPI	190	47
Data in Brief	Elsevier	1059	21
Data Science Journal	Ubiquity Press	20	0
Database	Oxford Academic	109	0
Geoscience Data Journal	Wiley	37	0
JOAD	Ubiquity Press	7	0
JOHD	Ubiquity Press	29	0
JOPD	Ubiquity Press	20	0
RDJ	Brill	8	0
Scientific Data	Springer Nature	765	91
Total		**2486**	**159**

Scientific Data, 47 from MDPI's *Data*, and 21 from *Data in Brief*. For our subsequent analysis we filtered out 28 cases where one data paper was matched to multiple research papers, to reduce the complexity of the task for the matching algorithms. The final number of pairs considered was 131. The dataset has been deposited to Figshare (Ribary and Wigdorowitz, 2025).

3.2 Data extraction

To support the experiments, we devised a data extraction pipeline using an openly available repository of research articles. Options for data sources we have considered include OpenAlex, PubMed, Zenodo and Dimensions. After analysis of each source looking at papers they cover, content volume, and metadata we could extract, we chose OpenAlex (Priem et al., 2022) as the source for our dataset. This was due to its comprehensive coverage of multiple disciplines and its extensive volume of more than 200 million works. Initially we queried the OpenAlex API. However, this approach posed challenges due to limitations on query volume, making it difficult to retrieve data at scale, and leading us to download its data dump.

The pipeline consisted of the following steps:

1. Extraction: We used the OpenAlex data dump to retrieve metadata from the papers.

2. Filtering: After extracting all papers, we retained only those that were published open access, were research articles, proceedings papers, or book chapters (excluding review articles and other types).

3. Paper categorization: We automatically categorized the collected papers into data papers (defined as all articles published in our list of data journals described in section 3.1) and research papers (all other articles extracted in the previous step).[1] We focussed on data papers published in 2022, as this year had the greatest coverage in the set of papers used for the ground truth. A five-year range was chosen before 2022 and to the date of analysis (2025), since data papers are published before and after their related research papers (McGillivray et al., 2022).

4. Author overlap: For each research paper, we checked whether any of its authors also appeared among the authors of the data papers. We only retained research papers that shared at least one author with a data document, helping to reduce redundancy and increase the relevance of the comparison set.

5. Length filtering: To ensure a fair comparison and remove anomalies, papers with abstracts shorter than 300 characters or longer than 2,000 characters were excluded.

6. Full-text extraction: We extracted the full text of the papers selected in the above step. This ensured that the storage and computational capacities were used effectively. The full texts were obtained from OpenAlex, which provided open access locations along with landing page URLs that could be used for download.

Table 2 presents the statistics for our dataset construction pipeline, showing the progression from initial OpenAlex metadata extraction through filtering and full-text extraction.

[1] We acknowledge that this heuristic may miss data papers

Table 2: Number of research papers, data papers and total number of papers at the different steps of our data extraction pipeline.

Data extraction Step	**Total**	**Research Papers**	**Data Papers**
Filtering	49,224,956	—	—
Categoration + Author Overlap	457,599	455,105	2,494
Length Filtering	305,308	303,063	2,245
Full-text Extraction	244,571	243,350	1,221

For each paper, we retrieved the following metadata fields: Title, Abstract, DOI, OpenAlex ID, Author IDs, Publication Year and Journal/Source of publication. In what follows, we will use the terms “data papers“ and “research papers” to refer to the final set resulting from applying all 6 pipeline steps described above. For any given data paper d, we will refer to its set of candidate matching research papers r_i as the research papers selected by the 6 pipeline steps and which shared at least one author in common with d and were published up to 5 years before and after d.

3.3 Matching algorithms

We experimented with different methods to calculate the similarity between data papers (d) and research paper (r) and therefore identify research papers that are related to a given data paper. To evaluate both surface-level and semantic similarity in linking data papers to research articles, starting from the set of data papers in the dataset, we implemented and evaluated two complementary approaches: set-based matching and vector-based matching, as summarised in Table 3. Set-based approaches and TF-IDF offer interpretable and computationally efficient ways to measure lexical overlap between documents by identifying shared terminology. SBERT approaches capture deeper semantic relationships by representing documents in high-dimensional spaces.

3.4 Set-Based Matching

For the set-based matching methods we preprocessed the texts with tokenization, lowercasing, and stopword removal to ensure that similarity measures focus on semantically meaningful content rather than common function words. We applied two Jaccard-based metrics to three sets: the top N (N=10, 20, and 50) most frequent words in texts of research papers and data papers, the named entities (NEs) entracted from the texts using the `spaCy` [2] library in Python, and on all tokens in the texts. These methods were applied to both abstracts and full texts.

3.4.1 Jaccard

For each data paper d and research paper r among its candidate matches, let $\text{Tokens}(d)$ and $\text{Tokens}(r)$ denote the sets of unique tokens extracted from d and r, respectively. We computed the Jaccard similarity based on the sets of unique tokens extracted from each document:

$$S_{\text{J}}(d,r) = \frac{|\text{Tokens}(d) \cap \text{Tokens}(r)|}{|\text{Tokens}(d) \cup \text{Tokens}(r)|} \qquad (1)$$

3.4.2 Multi-set Jaccard

To account for term frequency, we also implemented a multiset version of Jaccard similarity, which compares token counts rather than just presence or absence (da Fontoura Costa, 2021). Let T be the total number of distinct tokens in the union of the data paper d and research paper r. The token frequencies in each document are represented as vectors: $[d_1, d_2, \ldots, d_T]$ and $[r_1, r_2, \ldots, r_T]$, where d_i and r_i denote the frequency of token i in d and r, respectively.

The multiset Jaccard similarity is defined as follows and rewards documents that not only share vocabulary, but also use it with similar frequency:

$$S_{\text{MJ}}(d,r) = \frac{\sum_{i=1}^{T} \min(d_i, r_i)}{\sum_{i=1}^{T} \max(d_i, r_i)} \qquad (2)$$

3.5 Vector-based Matching

Vector-based matching methods represent documents as numerical vectors and compute similarity using distance metrics such as cosine similarity. While TF-IDF captures lexical overlap and term salience, SBERT captures deeper semantic relationships through contextualized embeddings.

published in generalist journals or misclassify non-data papers in data journals.

[2] https://spacy.io/ (last accessed 05/10/2025).

Table 3: Overview of algorithms for matching data papers with research papers. We implemented two groups of methods: set-based and vector-based. Each method was applied to different scopes of textual content. We analysed three values for N:10, 20, and 50.

Method	Top N Frequent Words	Named Entities (NEs)	All Tokens
Set-based Methods			
Jaccard	✓	✓	✓
Multi-set Jaccard	✓	✓	✓
Vector-based Methods			
TF-IDF	–	–	✓
SBERT	–	–	✓
SBERT re-ranked	–	–	✓
SBERT re-ranked with LLM	–	–	✓

3.5.1 TF-IDF

We first applied Term Frequency–Inverse Document Frequency (TF-IDF) (Manning et al., 2008). Each document is represented as a sparse vector where each dimension corresponds to a term in the corpus vocabulary. The TF-IDF score reflects how important a term is to a document relative to its frequency across the corpus.

Let $\text{Tokens}(d)$ and $\text{Tokens}(r)$ denote the sets of unique tokens extracted from data paper d and research paper r, respectively. Let $R(d)$ be the set of candidate research papers for d. The corpus $C(d)$ consists of d and all papers in $R(d)$. For each token t in document $x \in C(d)$, the TF-IDF weight is:

$$\text{TF-IDF}_d(t, x) = \text{tf}(t, x) \times \text{idf}_d(t) \tag{3}$$

where $\text{tf}(t, x)$ is the term frequency of token t in document x, and $\text{idf}_d(t)$ is the inverse document frequency defined as:

$$\text{idf}_d(t) = \log \frac{1 + |C(d)|}{1 + |\{y \in C(d) : t \in \text{Tokens}(y)\}|} + 1 \tag{4}$$

computed over the corpus of the data paper and its candidate research papers, $C(d)$.

For each data paper d and research paper $r \in R(d)$, we compute cosine similarity between their TF-IDF vectors:

$$\begin{aligned} S_{\text{TF-IDF}}(d, r) \\ &= \cos(\text{TF-IDF}(d), \text{TF-IDF}(r)) \\ &= \frac{\text{TF-IDF}(d) \cdot \text{TF-IDF}(r)}{\|\text{TF-IDF}(d)\| \cdot \|\text{TF-IDF}(r)\|} \end{aligned} \tag{5}$$

This method captures term salience and is sensitive to shared terminology, but does not account for synonymy or contextual meaning.

3.5.2 SBERT

To capture deeper semantic relationships, we used Sentence-BERT (SBERT), a transformer-based model that produces dense, contextualized embeddings for sentences and documents (Reimers and Gurevych, 2019). We computed cosine similarity between SBERT embeddings of the titles and abstracts of each data–research paper pair:

$$S_{\text{SBERT}}(d, r) = \cos(\text{emb}(d), \text{emb}(r)) \tag{6}$$

SBERT has been shown to outperform other embedding methods on semantic similarity and transfer learning tasks (Reimers and Gurevych, 2019).

3.5.3 Reranking

To further refine the initial similarity rankings, we implemented two reranking approaches that leverage more sophisticated architectures to capture nuanced relationships between data papers and research papers that may be missed by the initial vector-based methods.

SBERT Cross Encoder Reranking: We employed a cross encoder architecture (Reimers and Gurevych, 2019) that jointly processes pairs of data paper and research paper abstracts. Unlike bi-encoders that generate independent embeddings for each document, cross encoders allow attention mechanisms to operate across both documents simultaneously, capturing fine-grained interactions between their content. The cross encoder takes concatenated representations of the document pair as input and outputs a relevance score. This should lead in principle to superior performance in semantic matching tasks at the cost of increased computational complexity due to longer input sequences.

Listwise LLM-based Reranking: We implemented a listwise approach using Large Language

Models to re-order the top-k research paper candidates from SBERT rankings (Ma et al., 2023). We used GPT-4o-mini (OpenAI, 2024) with a temperature of 0.1 and maximum token limit of 1000. The LLM receives the data paper abstract and a numbered list of candidate research paper abstracts, then outputs a re-ranked ordering based on which papers most likely substantially use or analyze the described dataset. Reranking was performed in a zero-shot setting using prompts that emphasized analytical relevance and dataset usage patterns (see Appendix A for the complete prompt template).

4 Evaluation

In this paper we refer to the curated dataset of manually validated data–research paper pairs as our evaluation set. Since our approach does not involve training a supervised model, we do not distinguish between training, validation, and test subsets. We evaluate all methods using Mean Reciprocal Rank (MRR), a standard metric in information retrieval (Voorhees, 1999). Given a data paper d and a set of candidate research papers $R(d)$, with the correct match $r^* \in R(d)$, the reciprocal rank is defined as:

$$\text{RR}(d) = \frac{1}{\text{rank}(r^*)}$$

The MRR is the average of reciprocal ranks over all data papers in the evaluation set. In case of ties, we use the expected reciprocal rank under uniform random tie-breaking, i.e., the average of the reciprocals of the tied positions. Higher MRR values indicate better performance.

We also evaluated based on the retrieval accuracy on top-N and top-$N\%$ selection. Ties in ranking scores were handled in a tie-aware fashion. If the correct data paper was part of a tie block fully above the cutoff (e.g., top-10), the prediction was counted as correct. If the tie block straddled the cutoff, we assigned fractional credit proportional to the number of tied items within the cutoff (e.g., if three papers tied for ranks 9–11 and two were within the top-10, the correct paper received a score of $2/3$). If the entire tie block fell outside the cutoff, the prediction was considered incorrect. This approach prevents arbitrary tie-breaking and ensures consistency across metrics.

4.1 Comparing abstracts and full texts

One of the key practical considerations in designing systems to match data papers with research papers is the availability and granularity of textual content. While full texts may offer richer information, abstracts are more readily accessible and computationally efficient to process. To determine whether this trade-off affects matching performance, we conducted statistical significance tests on a subset of our set-based and vector-based method, comparing results obtained from abstracts and full texts.

Table 4: Significance test results comparing abstracts vs. full texts for a subset of methods. Each method was applied to different scopes of textual content: top N frequent words (Freq) for $N = 10$, and all tokens (All). $\alpha = 0.05$.

Method	Scope	p-value	Stat
Jaccard	Freq	0.59	54.55%
	All	0.058	63.64%
Multi-Jaccard	Freq	0.59	54.55%
	All	0.237	58.62%
TF-IDF	All	0.77	47.06%

We applied the paired sign test (Gibbons, 1993) to compare the performance of methods when using abstracts versus full texts. This non-parametric test was chosen because it makes minimal assumptions about the data distribution under the null hypothesis. It assesses whether there is a statistically significant difference in performance (i.e., ranking) between the two conditions across all data papers in our curated dataset.

Table 4 reports the results of statistical significance tests (with significance threshold $\alpha = 0.05$) comparing the performance of set-based methods applied to abstracts vs. full texts. The *Stat* column shows the percentage of times the abstract method achieved a better ranking than the full-text method. Across all scopes and methods, the p-values indicate that the differences are not statistically significant, suggesting that using abstracts yields comparable performance to using full texts.

5 Results

Table 5 presents the performance of all matching methods when applied to abstracts. It shows that the best-performing methods vary depending on the evaluation metric and reflects the diverse strengths of each method. This dispersion reflects the complementary strengths of different approaches: lexical overlap, semantic similarity, and contextual reasoning each contribute uniquely to match qual-

Table 5: Matching performance using abstracts. Metrics include Mean Reciprocal Rank (MRR), top-N accuracy, and top percentile accuracy. Each method was applied to different scopes of content: top N frequent words (Freq), named entities (NE), and all tokens (All). For SBERT re-ranked methods, only the top 50 candidates were re-ranked.

Method	Scope	MRR	Top-5	Top-10	Top-50	Top-1%	Top-5%	Top-10%
Jaccard	Freq=10	0.38	44.40%	52.61%	77.14%	39.68%	60.76%	72.39%
	Freq=20	0.31	45.99%	55.73%	81.92%	35.37%	64.33%	77.49%
	Freq=50	0.44	54.01%	63.54%	87.37%	47.66%	72.71%	**82.90%**
	NE	0.10	14.06%	14.84%	22.53%	36.19%	54.76%	60.49%
	All	0.40	53.91%	62.50%	89.06%	46.88%	67.19%	79.69%
Multi-Jaccard	Freq=10	0.38	44.40%	52.61%	77.14%	39.68%	60.76%	72.39%
	Freq=20	0.31	45.99%	55.73%	81.92%	35.37%	64.33%	77.49%
	Freq=50	0.44	54.01%	63.54%	87.37%	47.66%	72.71%	**82.90%**
	NE	0.07	8.60%	12.06%	21.94%	25.35%	31.20%	34.82%
	All	0.41	50.78%	64.06%	89.06%	45.31%	71.88%	79.69%
TF-IDF	All	**0.45**	41.18%	70.31%	84.38%	**51.56%**	**75%**	82.81%
SBERT	All	0.40	53.91%	68.75%	92.19%	35.16%	53.91%	73.44%
SBERT re-ranked	All	0.39	53.91%	65.62%	92.19%	31.25%	56.25%	70.31%
SBERT re-ranked + LLM	All	0.44	**62.50%**	**71.88%**	**92.19%**	34.38%	62.50%	75.00%

ity. TF-IDF achieved the highest MRR (0.45) and top-1% and top-5% accuracy (51.56% and 75%, respectively), outperforming other vector-based methods in those metrics. SBERT re-ranked with LLMs showed the best overall performance in top-N accuracy, with top-10 and top-50 scores reaching 71.88% and 92.19%, respectively.

As expected, among set-based methods performance generally improves with larger token scopes (e.g., Freq=50 and All), which suggests that richer lexical context enhances matching accuracy. Named Entity-based matching underperforms across all metrics, showing that entity-level overlap alone is insufficient to capture the nuanced relationships between data and research papers. Vector-based methods show strong performance, with SBERT re-ranked using LLMs achieving the highest scores in top-N accuracy. This highlights the value of semantic understanding and contextual reasoning in identifying meaningful links. Interestingly, the performance gap between SBERT and SBERT re-ranked is modest, which suggests that initial semantic similarity captures much of the relevant signal, and reranking offers incremental gains. For the subset of methods tested (set-based approaches and TF-IDF), the lack of statistically significant differences between abstracts and full texts supports the feasibility of using abstracts for these approaches, especially when full texts are unavailable or costly to process. Further evaluation is needed for semantic embedding methods.

6 Limitations and Conclusion

Our findings demonstrate that it is feasible to automatically identify research papers related to data publications using NLP-based methods and that the best-performing method varies depending on the metric used. They also underscore the importance of evaluating methods across multiple metrics to avoid over-reliance on a single performance indicator. Hence, hybrid systems combining multiple matching strategies may offer the most robust performance. Set-based approaches, particularly Multi-set Jaccard and Jaccard with frequent words or all tokens, offer interpretable and computationally efficient solutions. TF-IDF achieves the highest MRR, indicating strong precision in ranking the correct match highly. Vector-based methods, especially SBERT with LLM-based reranking, provide superior performance in terms of top-N accuracy, though at higher computational cost. For set-based and TF-IDF methods, abstracts appear sufficient for effective matching, which has practical implications for scalability, since abstracts are more readily available and less resource-intensive to process than full texts.

As to this study's limitations, we rely on metadata availability, which may not generalize to all disciplines or publication venues. Future research could explore avenues to generalize our approach. First, expanding the curated dataset to include more data journals and multilingual content would help assess cross-domain applicability. Second, incorporating citation networks, dataset repository metadata, and author affiliations could enrich the matching process and reduce reliance on textual similarity alone. Third, exploring the analysis of matches between data papers and research papers where more than one research paper corresponds to the same data paper. Finally, developing hybrid models that combine lexical, semantic, and structural features may improve performance, especially in cases where abstracts are sparse or ambiguous.

7 Authors' contributions

BMcG designed the study, supervised the project, wrote sections 1, 2, 4, 5, and 6, reviewed and edited the manuscript. KA designed the taxonomy of methods and statistical tests; conducted experimentation on article abstracts; processed the OpenAlex data dump; wrote sections 3.5.2, 3.5.3, and Appendix A; and parts of sections 3.3 – 3.5. VH downloaded article full texts and conducted experiments for sections 3.4 and 3.5.1; wrote Figure 1, Table 2 and part of sections 3.2 – 3.5. MR conceptualized the study, the manual linking process, compiled the curated pairs, wrote parts of section 1 and 3, reviewed and edited the manuscript. MW conceptualized the study, compiled the curated pairs, wrote Table 1, and parts of section 3, reviewed and edited the manuscript.

References

Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann, and Andrea Pierleoni. 2024. REXEL: An end-to-end model for document-level relation extraction and entity linking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 119–130, Mexico City, Mexico. Association for Computational Linguistics.

Leonardo Candela, Donatella Castelli, Paolo Manghi, and Alice Tani. 2015. Data journals: A survey. *Journal of the Association for Information Science & Technology*, 66(9):1747–1762.

Luciano da Fontoura Costa. 2021. Further generalizations of the jaccard index. *ArXiv*, abs/2110.09619.

Priyangshu Datta, Suchana Datta, and Dwaipayan Roy. 2025. Raging against the literature: Llm-powered dataset mention extraction. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '24, New York, NY, USA. Association for Computing Machinery.

Stefan Ekman, Olle Sköld, and Isto Huvila. 2025. Functions of paradata in data papers. *Journal of Documentation*, 81(7):253–272.

Michael Färber and Ann-Kathrin Leisinger. 2021. Recommending datasets for scientific problem descriptions. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3014–3018, New York, NY, USA. Association for Computing Machinery.

Jean Gibbons. 1993. Location tests for single and paired samples (sign test and wilcoxon signed rank test). In *Nonparametric Statistics*, pages 5–25. SAGE Publications, Inc., Thousand Oaks, California. Accessed 2025-10-02.

Jenny Heddes, Pim Meerdink, Miguel Pieters, and Maarten Marx. 2021. The automatic detection of dataset names in scientific articles. *Data*, 6(8).

C. Jiao, K. Li, and Z. Fang. 2023. How are exclusively data journals indexed in major scholarly databases? an examination of four databases. *Scientific Data*, (737).

Kim Jihyun. 2020. An analysis of data paper templates and guidelines: types of contextual information described by data journals. *Science Editing*, (7):16–23.

Naoto Kai, Hayato Tomisu, Toshiki Shimbaru, and Tomoki Yoshihisa. 2025. Study on extracting keywords that reveal the value of research data through comparisons between academic and data papers. In *Advances in Internet, Data and Web Technologies*, pages 1–8, Cham. Springer Nature Switzerland.

Kembellec, Gérald and Le Deuff, Olivier. 2022. Poétique et ingénierie des data papers.

Christine Kosmopoulos and Joachim Schöpfel, editors. 2024. *Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*. Collection Humanités numériques et science ouverte. Septentrion Presses Universitaires.

Kai Li and Pei-Ying Chen. 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.

Kai Li, Jane Greenberg, and Jillian Dunic. 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.

Kai Li and Chenyue Jiao. 2021. How are data paper abstracts constructed? Preliminary analysis of rhetorical moves in data paper abstracts from Scientific

Data and Data. In *Proceedings of 18th International Conference on Scientometrics & Informetrics*.

Kai Li and Chenyue Jiao. 2022. The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts. *Journal of the Association for Information Science and Technology*, 73(6):834–846.

Xiaozheng Liu. 2022. Discussion on the structural model and constituent elements of data papers. *Resources Data Journal*, 1:2–9.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*, page 100–123. Cambridge University Press.

Barbara McGillivray, Paola Marongiu, Nilo Pedrazzini, Marton Ribary, Mandy Wigdorowitz, and Eleonora Zordan. 2022. Deep impact: A study on the impact of data papers and datasets in the humanities and social sciences. *Publications*, 10(4):39.

OpenAI. 2024. Gpt-4o technical report. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-10-02.

J. Priem, H. Piwowar, and R. Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Nathalie Reymonet. 2017. Améliorer l'exposition des données de la recherche : la publication de data papers. Ce texte présente la structure et le contenu d'un “ data paper ” ainsi que des exemples de revues qui publient de tels articles.

Marton Ribary. 2020. A relational database of roman law based on justinian's digest. *Journal of Open Humanities Data*.

Marton Ribary and Barbara McGillivray. 2020. A corpus approach to roman law based on justinian's digest. *Informatics*, 7(4).

Marton Ribary and Mandy Wigdorowitz. 2025. Manually curated links between data papers and research papers. Figshare. DOI: `https://doi.org/10.6084/m9.figshare.3058945`.

E. M. Voorhees. 1999. TREC-8 Question Answering Track Report. In *Proceedings of the 8th Text Retrieval Conference*, pages 77–82.

William H. Walters. 2020. Data journals: incentivizing data access and documentation within the scholarly communication system. *Insights: the UKSG journal*, 33(18).

Ju-Yeon Yu, Donghun Yang, and Kyong-Ha Lee. 2025. Scilinkbert: A bert-style language model for understanding scientific texts with citations. *Journal of Supercomputing*, 81(1356).

A LLM Reranking Prompt Template

The following prompt template was used for GPT-4o-mini reranking with temperature=0.1 and max_tokens=1000:

LLM Reranking Prompt Template

```
You are helping to find research papers
that are related to a given data paper.

DATA PAPER:
Title: {query_title}
Abstract: {query_text}

Below are {len(candidate_list)} research
papers ranked by semantic similarity.
Your task is to rerank them based on how
likely each research paper is to be
related to the data paper above (i.e.,
the research paper likely uses or
references the dataset described in the
data paper).

CANDIDATES TO RERANK:
{candidates_text}

Please provide your reranking as a
comma-separated list of numbers, with
the most relevant paper first.
For example: 3,1,7,2,5,4,6

Your reranking:
```

Where `{query_title}`, `{query_text}`, `{len(candidate_list)}`, and `{candidates_text}` are dynamically filled with the data paper title, abstract, number of candidates, and formatted list of candidate papers respectively.