https://doi.org/10.51408/issi2025_083

Annotation and Identification of Scientific Data Sharing Information from *Data Availability* Section

Shuo Xu[1], Jiahao Li[2], Xin An[3], Shengnan Wang[4], Jianhua Liu[5], Yuefu Zhang[6]

[1] *xushuo@bjut.edu.cn*, [2] *lijh0707@emails.bjut.edu.cn*, [6] *yaogeng_z@163.com*
School of Economics and Management, Beijing University of Technology, Beijing (China)

[3] *anxin@bjfu.edu.cn (Corresponding author)*, [4] *18706438326@163.com*
School of Economics and Management, Beijing Forestry University, Beijing (China)

[5] *liujh@wanfangdata.com.cn*
Beijing Wanfang Data Co., LTD, Beijing (China)

Abstract

With the advancement of the open science movement, an increasing number of institutions and journals now require authors to explicitly state data availability in their publications, thus promoting the open sharing and accessibility of scientific data. The aim of this study is to extract scientific data sharing information from data availability statements in scientific papers. In more detail, this study annotates 8,508 data availability statements in research papers from the PLOS corpus over a period of nearly 16 months. In the end, a total of 35,010 entities and 8,524 relations covering 8 types of entities and 2 types of semantic relationships are ultimately annotated. Based on the annotated data, the model on the basis of Universal Information Extraction (UIE) is fine-tuned to automatically identify entity and relation mentions from data availability statements of the remaining scholarly articles. Experimental results show that our model is capable of extracting scientific data sharing information.

Introduction

With the development of open science movement, the open sharing of scientific data has progressively become a significant trend (Xu et al., 2021; Lu et al., 2024). Numerous countries and funding organizations worldwide have actively implemented policies to promote the public availability and standardized management of data (Jiao, Qiu, Ma, & Yang, 2024). In the context of increasing emphasis on the openness and transparency of research data, the emergence of data sharing information within scientific data statements has further laid the groundwork for the standardization and institutionalization of data sharing practices (Yang, Zhang, & Huang, 2023). By data sharing information within scientific data statements, we mean the declarations within scientific publications that outline how scientific data is stored, shared, and accessed.

To enhance the digital ecosystem of scientific data in the process of open sharing, Wilkinson et al. (2016) systematically introduced and defined the FAIR (i.e., Findable, Accessible, Interoperable, and Reusable) principles, which provide an internationally recognized framework for the management and sharing of scientific data. Correspondingly, all journals published by PLOS[1] and Springer Nature[2]

[1] https://journals.plos.org/plosone/s/data-availability

[2] https://www.springernature.com/gp/authors/research-data-policy

issued new open data policies. The submission guidelines explicitly state that all scientific data supporting conclusions must be stored in public data repositories that comply with FAIR principles and provide corresponding DOIs or access numbers. Additionally, the data availability statement must clearly outline any access restrictions or special conditions, such as limitations due to legal or ethical constraints or the requirement of an application for access. These statements are usually located in *Data Availability* section. This enables the accessibility and evaluation of scientific data sharing information at large scale.

Federer et al. (2018) collected data availability statements from the articles published in PLOS ONE journal between March 2014 and May 2016, and found that only approximately 20% of the statements indicated the data were stored in a repository. After then, the long-term availability of URLs and DOIs mentioned in the data availability statements of PLOS ONE articles were further examined. Federer (2022) observed that approximately 80% of the resources could be successfully retrieved, whereas the retrieval rate relying on author contact to locate data was substantially lower, ranging from 10% to 40%. Subsequently, Jiao et al. (2024) took into consideration the articles published in PLOS ONE journal from 2014 to 2020, and employed the rules on the basis of regular expressions to extract data sharing mechanisms and repositories from the data availability statements. Jiao et al. (2024) argued that although data continued to be primarily shared through the main article or its supplementary materials, the use of data repositories exhibited a steady growth trend.

It is not difficult to see that previous studies are just limited to the articles published in PLOS ONE journal. In addition, since sharing information often appears in the form of diverse and irregular expressions, this results in unsatisfactory performance in sharing information extraction with rule-based approaches. Hence, this study considers all the articles published in the journals by PLOS publisher, and annotates a large-scale and high-quality dataset for data sharing information, encompassing eight types of entities and two types of semantic relationships. What's more, an automated identification model is constructed with the help of Universal Information Extraction (UIE) (Lu et al., 2022).

Data Annotation

Data sources

Since 1 March 2014, PLOS has implemented a data availability policy, requiring all submitted manuscripts to provide a detailed description of data sharing compliance within the data availability statement (Bloom, Ganley, & Winker, 2014). Therefore, the PLOS corpus [3] is selected as the data source in this study. This corpus was downloaded on August 21, 2023, comprising a total of 338,810 papers (excluding *correction* and *expression of concern* articles), in which 189,369 papers are attached with a section of data availability statements. On preliminary

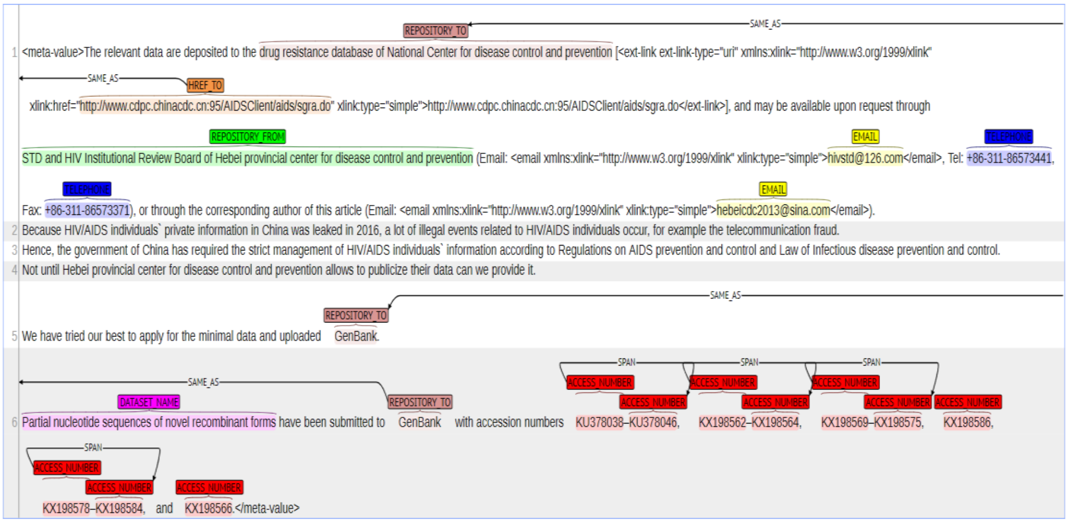
[3] https://plos.org/text-and-data-mining/

analysis, we observe that many data availability statements are very simple, such as "All relevant data are within the paper and its Supporting Information files.", and "All relevant data are within the paper." As for these cases, several rules based on regular expressions are manually curated to match 107,747 scientific publications. In this way, 81,622 articles remain, from which 8,508 ones are randomly drawn for annotating entities and semantic relationships.

Definition of Entities and Relations

This study defines 8 types of entities: DATASET_NAME, ACCESS_NUMBER, REPOSITORY_FROM, REPOSITORY_TO, HREF_FROM, HREF_TO, TELEPHONE, and EMAIL, along with 2 types of relationships: SPAN and SAME_AS.

An example of data availability statements with annotated entity and relation mentions is illustrated in Figure 1. From Figure 1, it is easy to understand the DATASET_NAME, ACCESS_NUMBER, TELEPHONE, EMAIL, and SPAN. As for REPOSITORY_FROM, REPOSITORY_TO, HREF_FROM, and HREF_TO, the suffix "FROM/TO" can distinguish between data source repositories/hyper-references and data storage repositories/hyper-references. The semantic relation SAME_AS is mainly used to establish clear and standardized connections between different repository or URL mentions. Note that the SAME_AS holds between the entities with the following types: REPOSITORY_FROM vs. REPOSITORY_FROM, HREF_FROM vs. REPOSITORY_FROM, REPOSITORY_TO vs. REPOSITORY_TO, and HREF_TO vs. REPOSITORY_TO.



1 <meta-value>The relevant data are deposited to the drug resistance database of National Center for disease control and prevention [<ext-link ext-link-type="uri" xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://www.cdpc.chinacdc.cn:95/AIDSClient/aids/sgra.do" xlink:type="simple">http://www.cdpc.chinacdc.cn:95/AIDSClient/aids/sgra.do</ext-link>], and may be available upon request through STD and HIV Institutional Review Board of Hebei provincial center for disease control and prevention (Email: <email xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple">hivstd@126.com</email>, Tel: +86-311-86573441, Fax: +86-311-86573371), or through the corresponding author of this article (Email: <email xmlns:xlink="http://www.w3.org/1999/xlink" xlink:type="simple">hebeicdc2013@sina.com</email>).

2 Because HIV/AIDS individuals` private information in China was leaked in 2016, a lot of illegal events related to HIV/AIDS individuals occur, for example the telecommunication fraud.

3 Hence, the government of China has required the strict management of HIV/AIDS individuals` information according to Regulations on AIDS prevention and control and Law of Infectious disease prevention and control.

4 Not until Hebei provincial center for disease control and prevention allows to publicize their data can we provide it.

5 We have tried our best to apply for the minimal data and uploaded GenBank.

6 Partial nucleotide sequences of novel recombinant forms have been submitted to GenBank with accession numbers KU378038–KU378046, KX198562–KX198564, KX198569–KX198575, KX198586, KX198578–KX198584, and KX198566.</meta-value>

Figure 1. An example of data availability statements with annotated entity and relation mentions (DOI = "10.1371/JOURNAL.PONE.0171481").

Data labeling

The data annotation process is empowered by the web-based annotation tool BRAT (Wang et al., 2023). Our team consists of 3 members with one team leader (the first author of this work), and spends approximately 16 months. To ensure consistency, the team strictly manages the online+offine annotation process. First, before annotating, all annotators are trained. Second, the team leader regularly conducts sample audits of the annotation results and provides corrections and guidance for typical errors. Finally, the workload of each annotator is adjusted periodically based on their annotation results. The annotators with lower accuracy experience a corresponding reduction in their workload.

Throughout the annotation process, three to four rounds of refinement are involved. After each round is completed, all annotated mentions are reviewed by our team leader, the resulting feedbacks are incorporated to optimize and adjust the annotation guidelines. Taking REPOSITORY as an example, the annotation guidelines underwent the following changes: In the first iteration, we focus on annotating repositories in the papers that explicitly mention data storage locations, with popular repositories such as Figshare, the NCBI database, and the Genbank database. In the second iteration, the rules for ethics committees are added. If a paper references an ethics committee providing dataset access, such as "Requests for access to the data should be made to the Medical Ethics Committee of the Second Affiliated Hospital of Nantong University," this entity mention should be annotated. In the third iteration, the annotation guidelines for organizations are introduced. In this case, the organizations related to data requests are considered. For instance, in the statement "The data are not publicly available owing to privacy or ethical restrictions, as they contain sensitive information. The data are held by the Anhui Provincial Tuberculosis Institute. Requests to access the data can be sent to Xiao-Hong Kan, Chief of Science and Education at the Anhui Provincial Tuberculosis Institute." where it is explicitly stated that data requests should be directed to Anhui Provincial Tuberculosis Institute, this entity should be annotated. Finally, in the fourth iteration, the annotation rules are established for supplemental files. In the end, a total of 35,010 entity mentions and 8,524 relation mentions are annotated. The distribution is illustrated in Figure 2.

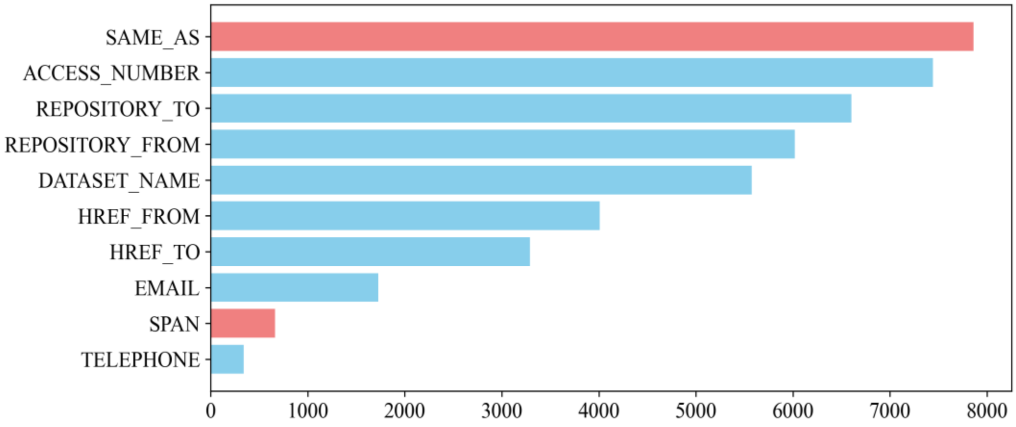


Figure 2. Distribution of number of entity and relation mentions in the annotated dataset.

As observed in Figure 2, several key characteristics can be observed as follows. (1) ACCESS_NUMBER (7,442) has the highest annotation count. This indicates that data access numbers are most frequently referenced in the data availability statements of research papers, highlighting their central role in data sharing. (2) The relatively high annotation frequency of REPOSITORY_FROM and REPOSITORY_TO suggests that the formal storage and traceability of scientific data are of significant concern in research. Notably, REPOSITORY (6,604) is annotated far more frequently than HREF (3,289), reflecting a tendency among researchers to directly reference data storage platforms or databases rather than individual web links. (3) The relatively low annotation frequencies of EMAIL (1,727) and TELEPHONE (340) suggest that instances of restricted data access still exist, albeit to a limited extent.

Entity and Relation Mentions Recognization Framework

The UIE framework

UIE (Universal Information Extraction) (Lu et al., 2022) represents a comprehensive framework for information extraction. Based on this framework, the PaddleNLP has developed and open-sourced the inaugural UIE model, with the ERNIE 3.0 as knowledge-enhanced pre-trained architecture. This model exhibits significant advantages in cross-domain adaptability, few-shot fine-tuning and efficient task transfer. More notably, UIE provides strong support for customizable model fine-tuning, allowing one to further refine the model using domain-specific data to optimize its performance in specialized fields or tasks.

In this study, the extraction of scientific data sharing information primarily involves two tasks: entity recognition and relation extraction. Traditional approaches (Chen et al., 2020) typically necessitate the independent training of two separate models, which significantly increases training complexity and may lead to a loss in predictive accuracy. In contrast, the UIE, by sharing network parameters, enables both tasks to be handled simultaneously within a unified framework, reducing computational redundancy and resource waste. Moreover, UIE offers enhanced flexibility and scalability, making it more suitable for addressing complex application scenarios like our case.

UIE Model Fine-tuning

When training and inference are based on the BERT model, the maximum length of each input text is typically limited to 512 tokens (Xu et al., 2024), a constraint inherent to its architectural design. Since the UIE utilizes BERT as the underlying pre-trained model, it is similarly constrained by input length during the fine-tuning process (defaulting to 512 tokens). While the length can be extended, doing so significantly increases the consumption of computational resources.

This study further analyzes the textual characteristics of PLOS corpus. It is found that most samples adhere to the 512-token limit, although a subset of texts exceeds this length. To enhance the capacity to handle long texts, we select 786 tokens as the maximum input length for fine-tuning, taking into account both the input limitations of the pre-trained model and computational costs. This length accommodates the majority of samples, minimizes the loss of information due to excessive truncation, and improves the model's understanding of long texts. In more detail, 7,441 samples have a text length not exceeding 786 tokens in our annotated dataset.

To ensure consistency, we exclude the samples with text length more than this limitation in the training phrase. During model training, the dataset is further split into training, validation, and test sets in a 7:2:1 ratio, with 5,209 samples used for training, 1,488 for validation, and 744 for testing. Samples with a text length exceeding 786 tokens total 1,067. To handle the samples with text length more than this limitation, we employ a sliding window approach for segmentation and evaluation. Specifically, a fixed-size window (Xu et al., 2024) is applied to the original text, with a certain overlap maintained during each segmentation. This method ensures that more coherent semantic information is captured when processing long texts.

During the fine-tuning process, the UIE model demonstrates strong entity recognition capabilities on both the validation and test sets. As shown in Table 1, our model performs well on most entity types in terms of Precision, Recall, and F1-score. EMAIL and TELEPHONE nearly achieved perfect recognition performance, while entity types such as ACCESS_NUMBER, REPOSITORY_TO, HREF_FROM, and HREF_TO also maintained evaluation scores above 0.95, reflecting excellent recognition performance. However, the UIE model demonstrated relatively weaker performance on DATASET_NAME and REPOSITORY_FROM, particularly in terms of Recall. In our opinion, this issue is partly related to the nature of the entities in the data availability statements themselves. For instance, it is usually difficult to determine the connotation and denotation of a dataset name. This enables many annotated entity mentions with the DATASET_NAME category not to always point to a publicly available dataset name, such as "raw metagenomic sequencing data".

A similar issue is observed in long texts. As shown in Table 1, although the overall prediction accuracy remains at a commendable level, certain entity types, such as DATASET_NAME and ACCESS_NUMBER, exhibit a noticeable decline in terms of Precision and F1-score. This indicates that while the UIE demonstrates the capacity to some extent for handling long texts, its generalization ability may be limited in cases involving complex information.

In the relation extraction task, SPAN benefits from its clear structural characteristics, maintaining strong recognition performance in both short and long texts. In contrast, SAME_AS involves more complex structure and a wider range of entity types, which increases the difficulty of relation extraction. Specifically, in long texts, where more intricate contextual information and potential ambiguities arise, SAME_AS faces greater challenges.

Table 1. Evaluation Performance of UIE Model on the Validation Set / Test Set / Long Texts.

	Precision	Recall	F1-score
DATASET_NAME	0.8133 / 0.8449 / 0.5947	0.6657 / 0.6765 / 0.6708	0.7321 / 0.7321 / 0.6304
ACCESS_NUMBER	0.9852 / 0.9879 / 0.6698	0.9926 / 0.9712 / 0.9721	0.9889 / 0.9795 / 0.7931
REPOSITORY_FROM	0.8725 / 0.8720 / 0.7583	0.7802 / 0.7967 / 0.7555	0.8238 / 0.8326 / 0.7569
REPOSITORY_TO	0.9602 / 0.9526 / 0.8812	0.9468 / 0.9393 / 0.8892	0.9534 / 0.9459 / 0.8852
HREF_FROM	0.9939 / 0.9867 / 0.7842	0.9290 / 0.8810 / 0.9462	0.9604 / 0.9308 / 0.8576
HREF_TO	0.9871 / 0.9955 / 0.7610	0.9147 / 0.8975 / 0.8118	0.9495 / 0.9440 / 0.7856
TELEPHONE	1.0000 / 1.0000 / 0.8818	0.9756 / 1.0000 / 0.9418	0.9877 / 1.0000 / 0.9108
EMAIL	1.0000 / 1.0000 / 0.8682	1.0000 / 1.0000 / 1.0000	1.0000 / 1.0000 / 0.9295
SPAN	0.9806 / 0.9831 / 1.0000	0.9712 / 0.9831 / 0.9730	0.9759 / 0.9831 / 0.9863
SAME_AS	0.9250 / 0.8889 / 0.8443	0.8216 / 0.8085 / 0.7030	0.8703 / 0.8468 / 0.7672

Model Prediction and Analysis

During the model prediction phase, we primarily focus on the data availability sections of the remaining 73,114 papers. As shown in Figure 3，the identification results exhibit a pronounced long-tail distribution. Among the entities, REPOSITORY_TO (139,774) exhibits the highest frequency, emphasizing the central role of repository storage in data sharing. REPOSITORY_FROM (48,973) follows, slightly surpassing ACCESS_NUMBER (47,532). Entities with moderate frequencies include DATASET_NAME (37,129), HREF_TO (31,129), and HREF_FROM (26,944). Among the relation types, SAME_AS (51,223) dominates.

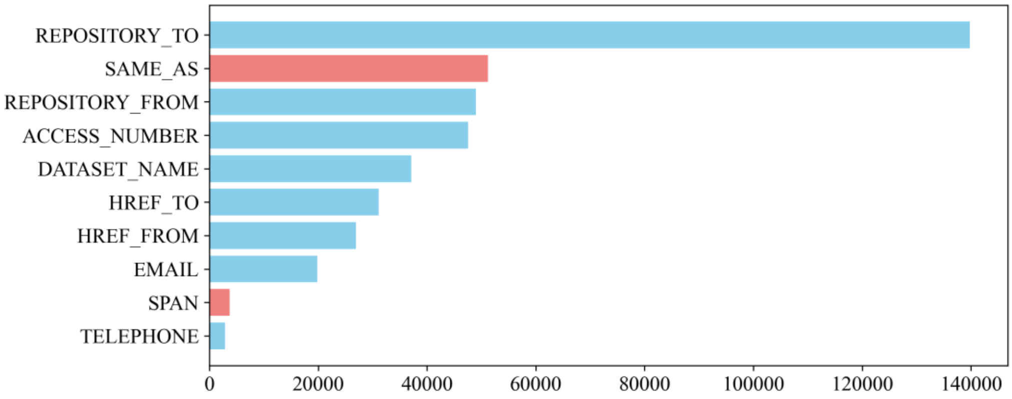
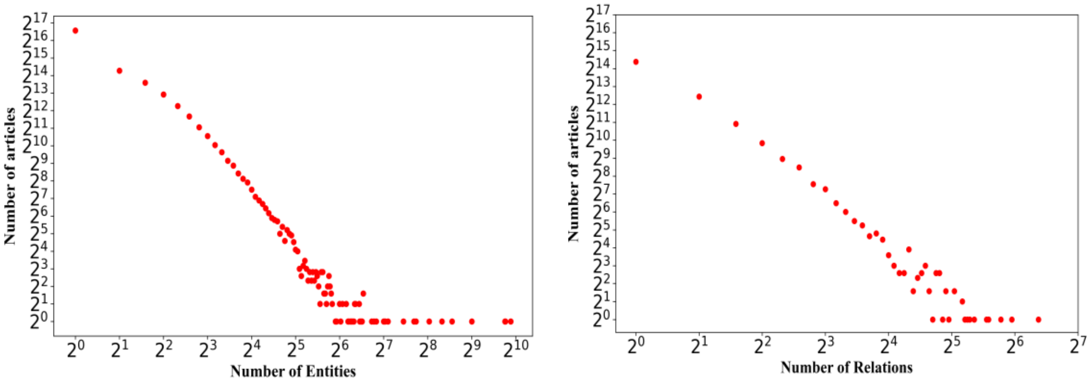


Figure 3. Distribution of entities and relations in the predicted dataset.

The log-log curves in Figures 4(a) and 4(b) illustrate the relations between the number of articles and the number of entity mentions, and the number of articles and the number of relation mentions, respectively. As the number of entity/relation mentions increases, the number of articles follows a typical power-law trend. To say it in another way, most articles contain fewer entities or relations, while articles containing a large number of entities or relations are relatively rare.



Figur4. Log-log curve of the number of articles and the number of entities (a), and the number of articles and the number of relations (b).

(a) Log-log curve between the number of articles and the number of entities

(b) Log-log curve between the number of articles and the number of relations

Conclusions and Limitations

In the context of the growing openness and transparency of scientific data, data availability statements, as one of the primary means of data sharing, have been widely implemented and received significant attention across various academic journals. Previous studies primarily focused on the articles in PLOS ONE journal, rule-based approaches were usually resorted for extracting shared information, resulting in an unsatisfactory performance.

Therefore, this study randomly selects 8,508 articles published in the journals by PLOS publisher for the annotation of entities and semantic relationships. Through rigorous multiple rounds of manual annotation and quality review, this study ultimately constructs a high-quality corpus containing 8 types of entities and 2 types of semantic relationships, with a total of 35,010 entity mentions and 8,524 relation ones. Building on this, the study fine-tunes a model based on the UIE information extraction framework to achieve automated identification of entities and relations.

Though, there is still some room to improve our study as follows. The UIE framework under-performs when handling low-frequency entity types and relationships with ambiguous boundaries. Moreover, the performance in processing long texts needs to be further improved.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 72474016, 72004012 and 72074014).

References

Bloom, T., Ganley, E., & Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLoS Medicine*, 11(2), e1001607.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., & Yang, G. (2020). A deep learning based method for extracting semantic information from patent documents. *Scientometrics*, 125(1), 289-312.

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y. L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of data availability statements. *PloS ONE*, 13(5), e0194768.

Federer, L. M. (2022). Long-term availability of data associated with articles in PLOS ONE. *PloS ONE*, 17(8), e0272845.

Jiao, C., Li, K., & Fang, Z. (2024). Data sharing practices across knowledge domains: A dynamic examination of data availability statements in PLOS ONE publications. *Journal of Information Science*, 50(3), 673-689.

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

Lu, L., Zhong, Y., Luo, S., Liu, S., Xiao, Z., Ding, J., ... & Xu, J. (2024). Dilemmas and prospects of artificial intelligence technology in the data management of medical informatization in China: A new perspective on SPRAY-type AI applications. *Health Informatics Journal*, 30(2), 14604582241262961.

Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., ... & Wu, H. (2022). Unified structure generation for universal information extraction. *arXiv preprint arXiv*:2203.12277.

Wang, Z., Xu, S., Wang, Y., Chai, X., & Chen, L. (2023). Bureau for rapid annotation tool: Collaboration can do more among variance annotations. *Aslib Journal of Information Management*, 75(3): 523-534.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9.

Xu, S., Zhang, Y., Chen, L., & An, X. (2024). Is metadata of articles about COVID-19 enough for multi-label topic classification task? *Database*, 2024, baae106

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., ... & Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4), 100179.

Yang, N., Zhang, Z., & Huang, F. (2023). A study of BERT-based methods for formal citation identification of scientific data. *Scientometrics*, 128(11), 5865-5881.