



Information Research - Vol. 31 No. iConf (2026)

# Hybrid weak supervision model with manual and pseudo labels: a novel approach for scientific text mining

Shengli Deng, Rongrong Xiang, Qiuyu Zhu, and Fang Chen

DOI: <https://doi.org/10.47989/ir31iConf64192>

## Abstract

**Introduction.** Although deep learning has advanced scientific text mining, the high cost of manual annotation remains a significant bottleneck. To address this, we propose a hybrid weak supervision model (HWSM) that balances manual annotation costs, computing resources, and model performance.

**Method.** HWSM integrates a manually annotated subset with LLM-generated pseudo-labels and trains a deep learning model on this hybrid corpus, leveraging both human expertise and large-scale automated labeling in entity recognition tasks.

**Analysis.** The model was evaluated against supervised deep learning and few-shot LLM baselines on 5,000 LIS papers, measuring precision, recall, F1, and cost-performance trade-offs.

**Results.** HWSM achieves a competitive F1-score (within 0.01 of GPT-4.1) while significantly reducing labeling and inference costs. Optimal performance is observed when incorporating 80% pseudo-labeled data.

**Conclusion.** HWSM provides a robust, cost-effective solution for large-scale scientific entity recognition. It is particularly suited for scenarios with limited labeling budgets, offering a practical alternative to resource-intensive deep learning or expensive LLM-only approaches.

## Introduction

As a crucial component of Named Entity Recognition (NER), Method Entity Recognition serves as an important means of domain knowledge extraction and forms the foundational task for constructing domain knowledge bases (Al-Moslmi et al., 2020). In existing research, method entity recognition is commonly performed by training deep learning models on manually annotated datasets to achieve automatic identification of method entities (Wang et al., 2022). In recent years, some scholars have also explored the use of large language models for this task. Furthermore, while retrieval-augmented generation (RAG) and model distillation are emerging alternatives, they often entail complex infrastructure or high computational overhead for continuous large-scale inference.

However, traditional supervised learning requires high manual annotation costs when applied to large-scale data, while relying solely on large language models often results in issues such as unclear entity boundaries and high economic costs. To address these challenges, this paper proposes a hybrid weak supervision model (HWSM) for domain-specific Method Entity Recognition. This method not only ensure the performance of model, but also significantly reduces manual annotation costs and computing when dealing with large-scale data.

## Related work

Existing approaches to method entity recognition in academic papers can be categorised into four types: rule-based methods, traditional machine learning methods, deep learning methods, and inference-based methods using large language models. Rule-based methods rely on manually defined patterns to extract entities (Vlachidis & Tudhope, 2016), while traditional machine learning methods treat the task as a classification or sequence labeling problem (Heffernan & Teufel, 2018; Tarasova et al., 2022). The BiLSTM+CRF architecture is widely used (Luo et al., 2018; Ma & Yuan, 2019; Qiu et al., 2019). With the advent of pre-trained language models such as BERT and SciBERT (Beltagy et al., 2019; Shen et al., 2023), deep learning approaches have achieved notable success in this domain (Puccetti et al., 2023), such as SciBERT+CRF and SciBERT+BiLSTM+CRF proposed to further enhance performance (Zhang et al., 2023). Recently, large language models (LLMs) such as GPT have been applied to entity recognition tasks via few-shot prompting, demonstrating promising yet limited performance compared to deep learning model (Hao et al., 2025). While deep learning models offer superior accuracy, they require large amounts of high-quality annotated data. In contrast, LLMs reduce annotation costs but needs more computing resources.

To address these questions, we propose the HWSM that combines a set of manually annotated data with pseudo- annotated data generated by LLMs to train supervision models. The method aims to balance annotation efficiency, computing resources, and recognition performance to achieve better recognition performance. The Processing flow of our proposed model is shown in figure 1.

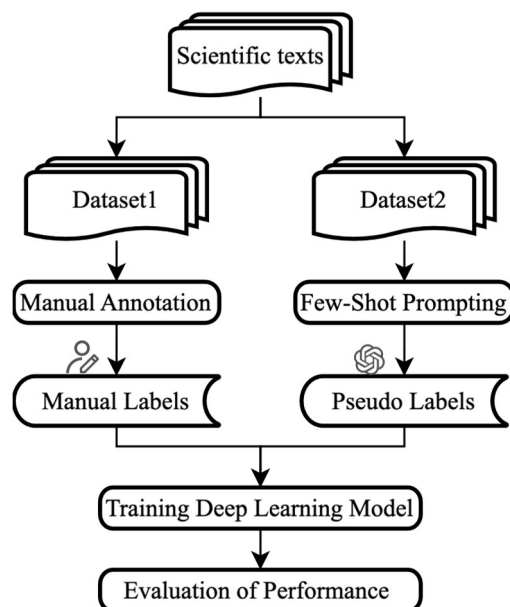


**Figure 1.** Progress of hybrid weak supervision model.

In traditional supervised learning, the training set consists of manually annotated data; in large language model inference learning, the training set is automatically generated by a large language model using few-shot prompting; and in hybrid learning, the training set comprises a small amount of manually annotated data combined large amounts of pseudo labels generated by the large language model. Additionally, the test set for all three experiments is the same manually annotated dataset.

## Experiment

### Datasets

Based on 21 core journals in the field of library and Information science (Zhang et al., 2023), this study collected bibliographic data of papers from each journal from 1995 to 2024 via the Web of Science Core Collection. Data sampling was performed according to the publication size of each journal (Tuomaala et al., 2014), resulting in a total of 5,000 records with non-empty abstracts (Table 1).

| Journal name                                                      | Sample/Number of papers |
|-------------------------------------------------------------------|-------------------------|
| ASLIB JOURNAL OF INFORMATION MANAGEMENT                           | 200                     |
| COLLEGE & RESEARCH LIBRARIES                                      | 200                     |
| ELECTRONIC LIBRARY                                                | 200                     |
| INFORMATION & CULTURE                                             | 200                     |
| INFORMATION PROCESSING & MANAGEMENT                               | 400                     |
| INFORMATION RESEARCH-AN INTERNATIONAL ELECTRONIC JOURNAL          | 200                     |
| INFORMATION TECHNOLOGY AND LIBRARIES                              | 200                     |
| INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT                   | 200                     |
| JOURNAL OF DOCUMENTATION                                          | 200                     |
| JOURNAL OF INFORMATION SCIENCE                                    | 200                     |
| JOURNAL OF LIBRARIANSHIP AND INFORMATION SCIENCE                  | 200                     |
| JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY | 400                     |
| LIBRARY & INFORMATION SCIENCE RESEARCH                            | 200                     |
| LIBRARY COLLECTIONS ACQUISITIONS & TECHNICAL SERVICES             | 200                     |
| LIBRARY QUARTERLY                                                 | 200                     |
| LIBRARY RESOURCES & TECHNICAL SERVICES                            | 200                     |
| LIBRARY TRENDS                                                    | 200                     |
| LIBRI-INTERNATIONAL JOURNAL OF LIBRARIES AND INFORMATION STUDIES  | 200                     |
| ONLINE INFORMATION REVIEW                                         | 200                     |
| REFERENCE & USER SERVICES QUARTERLY                               | 200                     |
| SCIENTOMETRICS                                                    | 400                     |

**Table 1.** Description of data sources.

The selection of 5,000 papers for manual annotation represents a balanced trade-off between experimental rigor and annotation costs. While we acknowledge the benefits of larger datasets, it is important to note that all 5,000 records in this study were meticulously annotated by human experts to serve as a high-quality standard (Hong et al., 2020). For a manually labeled corpus, this scale provides a statistically significant foundation for evaluating method entity recognition. Furthermore, the limitation in manual labeling capacity underpins the primary Motivation of this research: developing a HWSM that can maintain high performance while reducing the reliance on massive human-annotated datasets. We intend to progressively expand the data scale in future studies as more resources become available.

Among 5,000 papers, using the doccano platform, 5,000 records were manually annotated for method entity recognition to serve as learning labels and model evaluation standards. Annotated entity types included 7 kinds of research methods, and details are shown in Table 2.

| Entity type     | Definition                                                                             | Sample                                                                                                                       |
|-----------------|----------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| Theory          | Theoretical framework, paradigm or school of thought that characterises the research   | DIKW model, technology acceptance model, constructivism, social cognitive theory                                             |
| Data Collection | Means used to collect data for the study                                               | questionnaires, interview, web collecting, observations, Delphi, focus group                                                 |
| Data Analysis   | Methods used to process and analyse the data                                           | regression analysis, structural equation model (SEM), content analysis, co-word analysis, machine learning, network analysis |
| Dataset         | Referring to the specific data sources used (data objects, platforms, databases, etc.) | Weibo, WOS, Scopus, Facebook, Twitter                                                                                        |
| Indicator       | Nomenclature of specific indicators to measure the research object or variable         | user satisfaction, information utilisation, click-through rate, response time, citation frequency                            |
| Tool            | Tools or software used to perform analysis, coding, modeling                           | SPSS, Gephi, Python, R, Stata                                                                                                |
| Other           | Terms related to the research methodology but difficult to categorise                  | labeling system, indicator system                                                                                            |

**Table 2.** Description of method entities.

### Experiment design

To demonstrate the cost-effectiveness of our method, This paper sets up three types of comparison groups: (1) supervised learning models based on manually annotated data, including SciBERT+CRF, BiLSTM+CRF, and SciBERT+BiLSTM+CRF; (2) large language model inference-based learning, including GPT-4o, GPT-4.1, and DeepSeek-V3, using API Call and few shots to achieve tasks; (3) our proposed hybrid weak supervision model, which integrates manually annotated data, pseudo-labels generated by large language models, and traditional supervision models. The study further analyses the recognition performance of the hybrid model under different scales of pseudo-labeled data. The experimental design is shown in Figure 2.

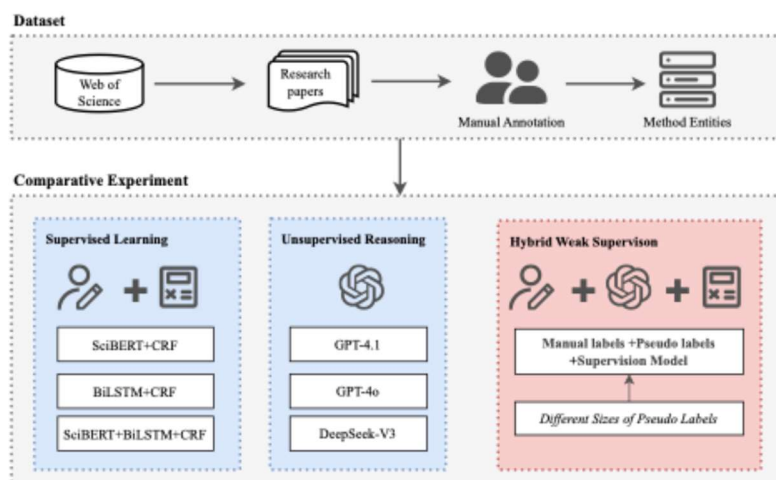


**Figure 2.** Experiment design of method entity recognition model.

## Results

### Performance of baseline model and HWSM

Table 3 shows the evaluation results of commonly used three typical deep learning models, two large language models, and our proposed HWSM. In HWSM, manually annotated data and LLM-annotated data each account for half, and SciBERT+BiLSTM+CRF for train and test.

| Model              | Precision | Recall | F1-score |
|--------------------|-----------|--------|----------|
| SciBERT+CRF        | 0.6249    | 0.5714 | 0.5970   |
| BiLSTM+CRF         | 0.5667    | 0.5914 | 0.5788   |
| SciBERT+BiLSTM+CRF | 0.6521    | 0.6273 | 0.6395   |
| GPT-4.1            | 0.6601    | 0.6321 | 0.6458   |
| GPT-4o             | 0.6354    | 0.6026 | 0.6186   |
| DeepSeek-V3        | 0.6151    | 0.5528 | 0.5823   |
| HWSM (proposed)    | 0.6498    | 0.6255 | 0.6374   |

**Table 3.** Performance of baseline models and HWSM

First, GPT-4.1 achieved the best entity extraction performance, followed by SciBERT+BiLSTM+CRF and HWSM. As shown in Table 3, the F1 scores of GPT-4.1, SciBERT+BiLSTM+CRF, and HWSM are 0.6458, 0.6395, and 0.6374, respectively—clearly outperforming the other models, with less than 0.01 difference among these three. Second, although our proposed HWSM is not the top performer, its F1 score is within 0.01 of the best model.

Although GPT-4.1 achieves the highest F1-score (0.6458), the performance of HWSM (0.6374) remains highly competitive with a marginal difference of less than 0.01. This minuscule performance gap justifies the use of HWSM in large-scale scenarios. For instance, in an environment involving 10,000 papers or more, the local deployment of a SciBERT-based HWSM reduces the economic costs, latency and privacy risks associated with external LLM APIs. Therefore, HWSM is not merely a compromise on performance but a deliberate architectural choice to optimise the cost-performance trade-off for scientific text mining. Furthermore, for large-scale entity extraction tasks, HWSM can maintain competitive performance while reducing the manual annotation cost required by deep learning models and the computational resources required by large language models.

### Performance of HWSM under different proportions of pseudo labels

To observe whether the proportion of pseudo labels and manual labeling will affect the performance of the HWSM method, we conducted a comparative experiment on the HWSM F1 Score under different proportions of pseudo labels, as shown in Figure 3.

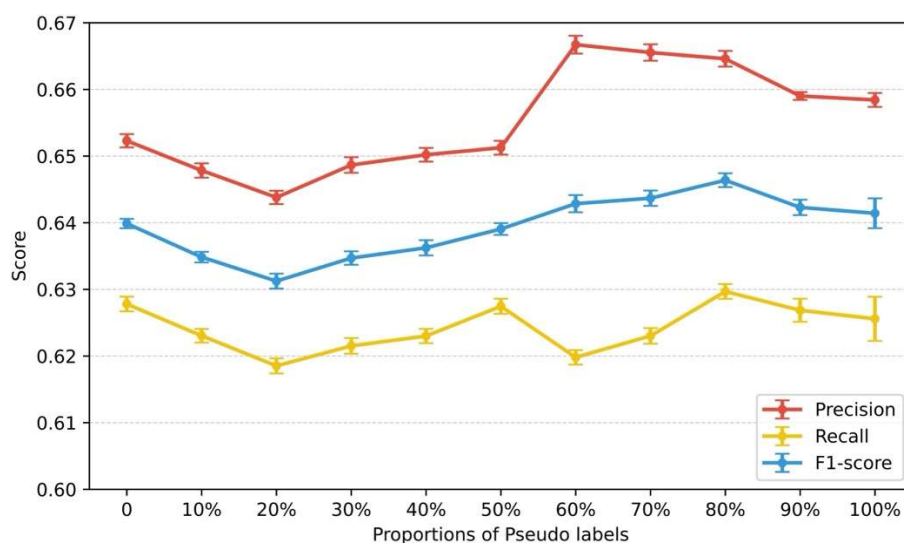


**Figure 3.** Performance of HWSM under different sizes of pseudo labels (mean  $\pm$  SD of five experiments).

As is shown in Figure 3, the HWSM method achieves the best performance, exhibiting the highest precision, recall, and F1 score under the 80% pseudo-labeling configuration. The inclusion of error bars representing the standard deviation across five independent experiments indicates that the model's performance is statistically stable. Our experiments demonstrate that HWSM consistently maintains its optimal F1 score within a narrow variance range across different pseudo-label proportions. Notably, the small standard deviation at the 80% mark confirms that the synergy between 20% manual anchors and 80% pseudo-labels provides a robust and repeatable performance peak.

Overall, regardless of the exact proportion of pseudo labels, the entity extraction performance remains within an acceptable range. Researchers can therefore choose different configurations based on manual annotation cost, computational resources, and budget to determine the ratio between pseudo labels and human-annotated data.

## Discussion

**Trade-off between performance and cost:** Experimental results show that while GPT-4.1 achieves the highest F1-score, our proposed HWSM maintains a highly competitive performance with a marginal gap of less than 0.01. This minuscule difference justifies the architectural choice of HWSM for large-scale scientific text mining. In real-world scenarios involving millions of academic papers, direct inference via LLM APIs is often hindered by prohibitive economic costs, API rate limits, and latency. By contrast, HWSM utilises a lightweight SciBERT-based backbone that can be deployed on local GPUs, offering a significantly higher throughput and lower operational cost.

**Rationality of the hybrid weak supervision approach:** The superiority of HWSM over simple LLM-based labeling or traditional supervised learning lies in the synergy between manual and pseudo-labels. LLMs possess vast general knowledge but may suffer from hallucinations or inconsistent boundary recognition in niche scientific domains. In HWSM, the small set of manually annotated data serves as learning objects. These anchors help the model calibrate the noise introduced by LLM-generated pseudo-labels, ensuring that the model learns domain-specific features without being entirely misled by LLM errors.

## Conclusion

This study proposes a novel HWSM approach, which is designed to balance manual annotation costs, computing resources, and model performance in the task of entity recognition in academic

papers. An empirical analysis was conducted on 5,000 scientific papers in the field of library and Information science, comparing traditional supervised learning models, large language model-based learning, and the proposed HESM. Experimental results demonstrate that the performance of HEWM method is like the best model GPT 4.1, its F1 score is within 0.01. Furthermore, our findings highlight several key insights and offer suggestions for future research and methodological improvements.

**Directions for model improvement:** The observation that both deep learning models and large language models exhibit higher precision than recall indicates a conservative prediction behavior, where models tend to avoid false positives at the cost of missing less salient or ambiguous entities. This suggests the necessity of incorporating recall-enhancing strategies, such as span-based detection or prompt calibration (Balakrishnan et al., 2016), to ensure comprehensive coverage of method-related entities.

**Control of pseudo labels utilisation:** Experimental results indicate that the optimal performance in method entity recognition is achieved when the scale of pseudo-labeled data generated by large language models is approximately 80%. However, as our training set consists of only 5,000 manually annotated samples, whether this ratio holds and continues to yield superior performance in larger-scale scenarios remains an open question for future investigation (Chen et al., 2024).

**Conditional limitations of HESM:** The proposed model, HESM, is designed for method entity recognition tasks in large-scale data scenarios and its aim is to balance costs and performance. It may not be suitable for small-scale datasets or situations with extremely limited manual annotations. Specifically, when sufficient manually labeled data is available, deep learning models should be prioritised due to their superior performance (Li et al., 2025). In contrast, in low-resource settings where efficiency is critical and labeled data is scarce, large language models offer a more practical alternative (Santoso et al., 2024).

## Acknowledgements

This study was supported by the National Social Science Foundation of China under the project ‘Research on the Construction and Application of Knowledge Base for Research Methods in Information Resource Management Discipline’ (Grant No. 23&ZD229) and the Special Scholarship for Graduate Students Going Abroad for Exchange of Wuhan University.

## About the authors

**Shengli Deng** is a Professor in the School of Information Management, Wuhan University, China. He received his PhD from Wuhan University. His research interests include information resource management and information behavior. He can be contacted at [victorydc@sina.com](mailto:victorydc@sina.com)

**Rongrong Xiang** is a PhD candidate in the School of Information Management, Wuhan University, China. She received her master’s degree from Central China Normal University. Her research interests include scientific text mining and knowledge discovery. She can be contacted at [xiang\\_rongrong@whu.edu.cn](mailto:xiang_rongrong@whu.edu.cn)

**Qiuyu Zhu** is a PhD candidate in the School of Information Management, Wuhan University, China. She received her master’s degree from Wuhan Institute of Technology. Her research interests include natural language processing and information behavior. She can be contacted at [zqy18371988085@163.com](mailto:zqy18371988085@163.com)

**Fang Chen** is a PhD candidate in the School of Information Management, Wuhan University, China. He received his master’s degree from Nanjing University. His research interests include text mining and information resource management. He can be contacted at [2025101040022@whu.edu.cn](mailto:2025101040022@whu.edu.cn)

## References

- Al-Moslmi, T., Gallofre Ocana, M., Opdahl, A. L., & Veres, C. (2020). Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, 8, 32862–32881. <https://doi.org/10.1109/ACCESS.2020.2973928>
- Balakrishnan, V., Ahmadi, K., & Ravana, S. D. (2016). Improving retrieval relevance using users' explicit feedback. *Aslib Journal of Information Management*, 68(1), 76–98. <https://doi.org/10.1108/AJIM-07-2015-0106>
- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3615–3620. <https://doi.org/10.18653/v1/d19-1371>
- Chen, R., Qin, C., Jiang, W., & Choi, D. (2024). Is a Large Language Model a Good Annotator for Event Extraction? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17772–17780. <https://doi.org/10.1609/AAAI.V38I16.29730>
- Hao, J., Chen, Z., Peng, Q., Zhao, L., Zhao, W., Cong, S., Li, J., Li, J., Qian, Q., & Sun, H. (2025). Prompt Framework for Extracting Scale-Related Knowledge Entities from Chinese Medical Literature: Development and Evaluation Study. *Journal of Medical Internet Research*, 27(1), e67033. <https://doi.org/10.2196/67033>
- Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116(2), 1367–1382. <https://doi.org/10.1007/S11192-018-2718-6>
- Hong, Z., Tchoua, R., Chard, K., & Foster, I. (2020). SciNER: Extracting named entities from scientific literature. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12138 LNCS, 308–321. [https://doi.org/10.1007/978-3-030-50417-5\\_23](https://doi.org/10.1007/978-3-030-50417-5_23)
- Li, Q., Wang, P., Liu, C., Li, X., & Hou, J. (2025). Integration patterns in the use of metadata for data sense-making during relevance evaluation: An interpretable deep learning-based prediction. *Journal of the Association for Information Science and Technology*, 76(3), 621–641. <https://doi.org/10.1002/ASI.24961>
- Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8), 1381–1388. <https://doi.org/10.1093/BIOINFORMATICS/BTX761>
- Ma, J., & Yuan, H. (2019). Bi-LSTM+CRF-based named entity recognition in scientific papers in the field of ecological restoration technology. *Proceedings of the Association for Information Science and Technology*, 56(1), 186–195. <https://doi.org/10.1002/PRA2.16>
- Puccetti, G., Giordano, V., Spada, I., Chiarello, F., & Fantoni, G. (2023). Technology identification from patent texts: A novel named entity recognition method. *Technological Forecasting and Social Change*, 186, 122160. <https://doi.org/10.1016/J.TECHFORE.2022.122160>
- Qiu, Q., Xie, Z., Wu, L., Tao, L., & Li, W. (2019). BiLSTM-CRF for geological named entity recognition from the geoscience literature. *Earth Science Informatics*, 12(4), 565–579. <https://doi.org/10.1007/S12145-019-00390-3/METRCS>
- Santoso, J., Sutanto, P., Cahyadi, B. K., & Setiawan, E. I. (2024). Pushing the Limits of Low-Resource NER Using LLM Artificial Data Generation. *Findings of the Association for Computational Linguistics ACL 2024*, 9652–9667. <https://doi.org/10.18653/V1/2024.FINDINGS-ACL.575>

- Shen, S., Liu, J., Lin, L., Huang, Y., Zhang, L., Liu, C., Feng, Y., & Wang, D. (2023). SsciBERT: a pre-trained language model for social science texts. *Scientometrics*, 128(2), 1241–1263. <https://doi.org/10.1007/S11192-022-04602-4>
- Tarasova, O. A., Rudik, A. V., Biziukova, N. Y., Filimonov, D. A., & Poroikov, V. V. (2022). Chemical named entity recognition in the texts of scientific publications using the naïve Bayes classifier approach. *Journal of Cheminformatics*, 14(1), 1–12. <https://doi.org/10.1186/S13321-022-00633-4>
- Tuomaala, O., Järvelin, K., & Vakkari, P. (2014). Evolution of library and information science, 1965–2005: Content analysis of journal articles. *Journal of the Association for Information Science and Technology*, 65(7), 1446–1462. <https://doi.org/10.1002/asi.23034>
- Vlachidis, A., & Tudhope, D. (2016). A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. *Journal of the Association for Information Science and Technology*, 67(5), 1138–1152. <https://doi.org/10.1002/ASI.23485>
- Wang, Y., Zhang, C., & Li, K. (2022). A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics*, 127(5), 2479–2520. <https://doi.org/10.1007/S11192-022-04332-7>
- Zhang, C., Tian, L., & Chu, H. (2023). Usage frequency and application variety of research methods in library and information science: Continuous investigation from 1991 to 2021. *Information Processing & Management*, 60(6), 103507. <https://doi.org/10.1016/J.IPM.2023.103507>

© [CC-BY-NC 4.0](#) The Author(s). For more information, see our [Open Access Policy](#).