



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

Evaluating software academic impact in biomedical research based on large-scale full-text analysis

Yuzhuo Wang, Heng Zhang, Huahao Hu, and Yi Zhao

DOI: <https://doi.org/10.47989/ir31iConf64193>

Abstract

Objective. Analysing patterns of academic software mentions and their academic impact is essential for understanding the scholarly ecosystem and optimising research resources.

Methods. This study focuses on the biomedical domain and investigates the prevalence and impact of software entities in scientific research. Rather than treating impact as a causal effect on research outcomes, we operationalise software impact as its scholarly presence, measured through the frequency of mention. Based on 1,500,334 articles from PubMed Central Open Access (PMC-OA), we collected disambiguated software entities mentioned in the full texts and extracted their features, including mention time, in-text location, and research subfields.

Results. Overall, the impact of software in biomedical research continues to grow, with programs such as SPSS, GraphPad, and Mega demonstrating high academic influence. Within papers, software influence is concentrated in the Methods section and is dominated by general-purpose statistical tools, while other sections display greater diversity. Across fields, domains with high-software mention tend to rely on general-purpose software, whereas more specialised domains adopt software tailored to specific biomedical tasks.

Conclusion. By tackling these questions, this study advances a systematic understanding of software role in biomedical research and provides a basis for further methodological refinement and empirical analysis.

Introduction

In the era of artificial intelligence, with the emergence of the fourth paradigm of scientific research, software has become an indispensable element of modern science. From theoretical derivations to practical explorations, software has been deeply integrated into the entire research process. However, existing studies have shown that the academic value of software has not been fully recognised (Wei et al., 2020). In recent years, scholars have begun to emphasise the importance of software, noting that ‘*virtually all data require some form of software processing*’ (Howison & Bullard, 2015). Some international organizations have also recognised software as a legitimate research output, thereby encouraging scientists to develop and share it (NSF, 2013; REF, 2024). Therefore, exploring software usage behaviors and their academic popularity and impact is of great significance for understanding the dynamics of the scholarly ecosystem, optimizing the allocation of research resources, and enhancing the quality and effectiveness of scientific studies.

Software has become a fundamental component of contemporary scientific research, and the biomedical field is no exception. Given the close connection to human development, biomedical research generates massive volumes of data, rendering software indispensable for effective data management, advanced analysis, and comprehensive visualization (Asada & Fukuda, 2024; Istrate et al, 2022). In this context, disentangling the usage patterns, visibility, and longitudinal evolution of different software is crucial for uncovering their roles in shaping research practices and for understanding how digital infrastructures transform scientific discovery. Previous studies have examined software usage in other fields (Pan et al, 2015, 2019), focusing on quantitative evaluations of software’s academic impact through citation-based or mention-based indicators. However, these studies typically have small sample sizes, which limits their ability to capture the comprehensive picture in different disciplines. In contrast, the biomedical field presents distinct characteristics, including the rapid pace of technological advancements and the interdisciplinary nature of the work. Therefore, it is necessary to explore the dynamic characteristics of software mention behaviors in this field, along with the structural patterns through which software is referenced across articles and sections, using large-scale full-text data.

In this study, we recognise the academic impact of software as its scholarly presence and take biomedical research as a representative case to explore the mention-based impact and developmental trends of software in the era of artificial intelligence. Specifically, we collect the full texts of biomedical articles in PubMed Central OA and extract the software entities mentioned therein to address the following three research questions:

RQ1: How have software mentions in biomedical research articles changed over time?

RQ2: How do software mentions vary across different sections of scholarly articles?

RQ3: How do mention-based influence and representative software differ across biomedical subfields?

Answering these questions will provide a systematic understanding of the role of software in biomedical research and lay a solid foundation for subsequent methodological design and empirical analysis.

Related work

In this study, software entities are conceptualised as a subset of method entities. We review existing research on their recognition and on impact assessment approaches that recognise impact primarily through observable indicators such as mentions, usage descriptions, and textual positioning.

Software entity recognition in academic papers

Early efforts in software identification relied on manual annotation, where researchers labelled software terms from the bibliographic content of academic papers (QasemiZadeh & Schumann,

2016). Compared with metadata, full-text articles offer richer contextual information, allowing researchers to access the software while recording other features (Howison & Bullard, 2015; Ibrahim, 2021). To improve extraction efficiency, rule-based approaches were developed, including dictionary-based matching (Ding et al., 2019) as well as more complex cues such as keywords, linguistic patterns, part-of-speech tags, and lexical positions (Katsurai, 2021; Lam et al., 2016). A notable example is bioNerDS, a rule-based extractor tailored for identifying software entities in bioinformatics (Duck et al., 2013).

With the development of machine learning methods, researchers applied supervised models to extract software and tool entities (Tateisi et al., 2016). Deep learning methods subsequently gained prominence, enhancing performance through transfer learning (Pan & Yang, 2009), active learning (Settles, 2012), and attention mechanisms (Britz, 2016). Building on these advances, pretrained language models have been widely adopted for method entity extraction due to their superior performance. For example, Zheng et al. (2021) obtained an F1 score of 81.4% with SciBERT. Compared with traditional neural network models, pretrained models have consistently demonstrated stronger capability in extracting diverse types of method entities.

Academic impact assessment of method entities

Impact assessment is a classic research topic in quantitative science studies. For method entities such as software, quantitative evaluation is essential to clarify their role and value in scientific research and practice (Wang et al., 2022). The most fundamental metric of impact is frequency, which includes citation counts, mention counts, and usage counts in scholarly literature. For example, Ding et al. (2019) assessed the impact of algorithms in the NLP domain based on the number of citing sentences. However, citation counts tend to overlook informal references. To address this, mention and usage counts have been employed as complementary indicators, capturing entities that contribute to research without being formally cited. Wang and Zhang (2020) analyzed the academic influence and evolutionary patterns of algorithms in NLP by measuring the number of articles. Pan et al. (2015) assessed the impact of software in *PLoS One* by quantifying the number of sentences mentioning, using, or citing software entities. Beyond simple frequency-based measures, prior studies have incorporated additional dimensions such as textual position, rhetorical intent, and co-occurrence structures to obtain a more nuanced understanding of popularity and role of software in scientific discourse (Li & Yan, 2018). These studies highlight the value of frequency-based indicators, suggesting the need for complementary semantic analyses in future research.

Method

Data collection

This study adopts the open dataset released by Istrate et al. (2022) as the primary source of software entities. Their dataset contains automatically extracted and disambiguated software entities from scholarly articles included in commercial category of PMC-OA up to 2020, along with all sentences mentioning software. These 1,452,793 unique software are classified as certain software, not_curated, and unclear. In our initial exploratory analysis, all software entities were included to obtain an overview of software mention patterns. However, to ensure conceptual clarity and analytical robustness, we finally selected certain software and further retrieved the full texts of all articles in PMC-OA that mention software entities. Specifically, we obtained 20,516 software entities, referenced in 8,544,085 sentences across 1,500,334 scholarly articles.

Software feature identification

(1) The time of mentioning software: Based on the full-text data provided by PMC-OA, the year of publication was used as the temporal attribute indicating when a software entity was mentioned in scientific research.

(2) The subfield of papers mentioning software: To identify the disciplinary domains in which software entities were mentioned, we utilised the third level 'Field of Research' classification from the Dimensions database and retrieved the subfield of papers mentioning software entities.

(3) The location of mentioning software: The section in which a software entity was mentioned was treated as its 'location' within the article. Using first-level section headings, we classified software mentions into eight categories: Abstract, Introduction, Related Work, Methods, Results, Discussion, Conclusion, and Appendix. Specifically, we adopted a rule-based approach inspired by Cui et al. (2025), whereby predefined keywords were matched with section titles. If a keyword corresponding to a particular category appeared in the section heading, the section was assigned to that category.

Results

Mention-based impact of software in PMC-OA

This study analyses 1,500,334 papers, where software entities have been disambiguated. On average, each article contains 6.74 sentences referencing 3.42 distinct software tools, indicating their widespread mention in biomedical research. To trace the evolution of software mentions, we examined both the annual number of papers mentioning software and the diversity of tools.

As shown in Figure 1, mention of software remained low and stable between 1989 and 2004 but grew rapidly after 2005, with an accelerated surge post-2015. This trend reflects the dual impact of Internet adoption, which facilitated software dissemination, and the increasing specialization of science, which demanded more sophisticated analytical tools. The expansion of big data and artificial intelligence may further fuel software development and applications, broadening its scope and impact.

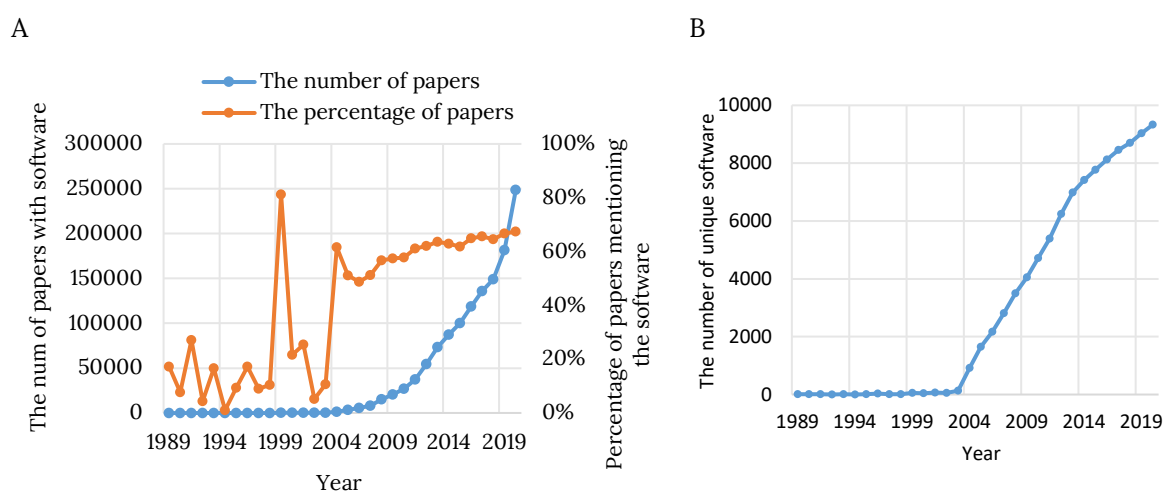


Figure 1. Distribution of the number of papers mentioning software and the diversity of software tools, 1989–2020.

Figure 2 illustrates shifts in the most frequently mentioned software. Early high-frequency software was modestly mentioned with slight variation, but by 2010, SPSS exceeded 4,000, underscoring the rising importance of data-processing tools. In the following decade, highly mentioned software was dominated by two categories: general-purpose statistical software (e.g., SPSS) and specialised biostatistical tools (e.g., GraphPad). Excel, although lower in frequency, also emerged as one of the most mentioned, suggesting a shift toward 'high-frequency general-purpose tools with simple usability.'

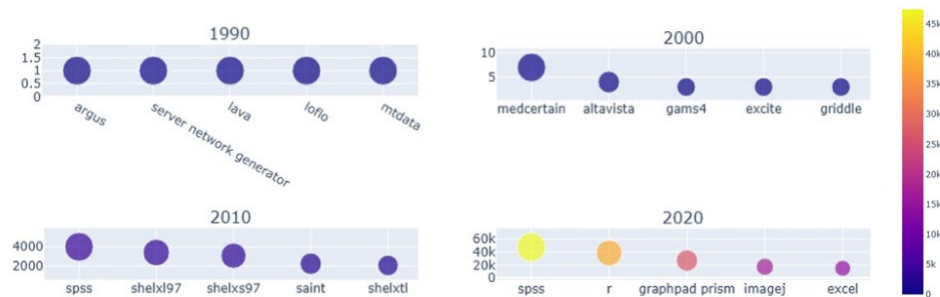


Figure 2. Temporal changes in the usage of representative software tools.

Mention-based impact of software across different sections

We analyzed the frequency of software mentions across different sections of scholarly articles, as shown in Figure 3. The 'Methods' section accounted for the largest share, at 72.84%, followed by 'Results,' 'Discussion,' and 'Introduction.' In contrast, sections such as 'Conclusion' and 'Appendix' contributed relatively little, together accounting for only 1.85%. Overall, the concentration in the Methods section suggests that software is most frequently documented in contexts related to methodological descriptions, such as data processing and analytical procedures. At the same time, software might provide technical support for result presentation and research discussion, reflecting its broad participation across the entire paper presentation.

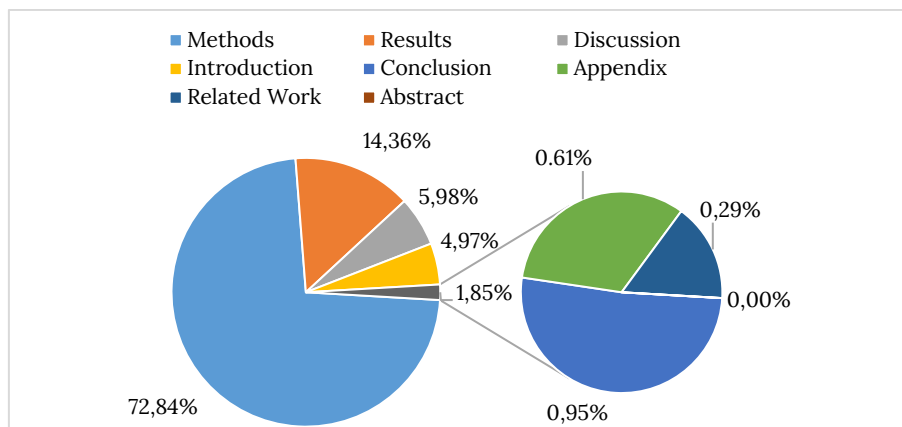


Figure 3. Software presence across different sections of scholarly articles.

We further identified the software with high mention-based impact across different sections (Figure 4). In the 'Methods' section, software mentions were consistently high, dominated by general-purpose statistical tools such as SPSS and GraphPad. In Results and Discussion, the biomedical tool BLAST was most common, reflecting the possible need for specialised computation in interpreting findings. In the 'Conclusion' section, general-purpose programming tools such as R and MATLAB were widely presented, likely due to the potential need for statistical analysis and system simulation in final results. These patterns suggest a tendency for different kinds of software to be documented differently within the structure of scientific articles.

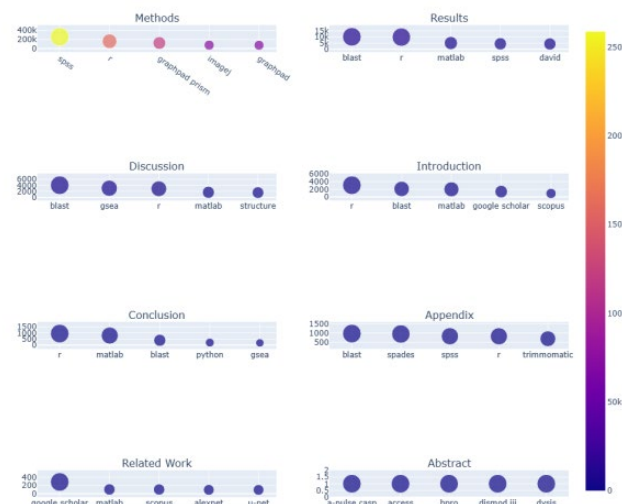


Figure 4. High-impact software by section type.

Mention-based impact across different research subfields

We first calculated and ranked the number of articles mentioning software across 34 subfields. Results show that the three subfields with the highest frequencies of software mentions in PMC-OA are all directly related to biomedicine, while agriculture also demonstrates notable software influence (see Table 1). Building on this, we selected 15 subfields to further examine the distribution of frequently mentioned software. These subfields were categorised into three levels of mention-based impact: high (five subfields with the highest software mentions, ranked 1-5), medium (five subfields with software mention closest to the median, ranked 15-19), and low (five subfields with the lowest software mentions ranking 30-34), as presented in Table 2.

Rank	Sub-field	Frequency
1	A01 Clinical Medicine	354,388
2	A03 Allied Health Professions, Dentistry, Nursing and Pharmacy	353,041
3	A05 Biological Sciences	153,396
4	A06 Agriculture, Veterinary and Food Science	119,325
5	A04 Psychology, Psychiatry and Neuroscience	115,203
6	B12 Engineering	240,047
7	A02 Public Health, Health Services and Primary Care	194,772
8	B11 Computer Science and Informatics	184,585
9	B08 Chemistry	171,645
10	B07 Earth Systems and Environmental Sciences	82,769

Table 1. Ranking of software influence across research domains.

As shown in Table 2, SPSS, Excel, and R are widely used across domains, reflecting their broad presence in the scholarly literature. However, disciplinary differences are apparent. The top five domains, mainly in biomedicine, feature frequent mentions of general-purpose tools such as SPSS, GraphPad, and ImageJ. Medium domains and low-prevalence subfields exhibit mixed mention patterns, combining general-purpose and specialised tools such as ArcGIS, Oxcals and synbio, which tend to appear primarily within narrowly defined disciplinary contexts. Therefore, general-purpose, and specialised software exhibit distinct patterns of mention across domains.

Subfield Software	high					medium					low				
	A01	A03	A05	A06	A04	C20	B10	B09	C13	C15	D30	D27	D25	D31	D29
spss	87842	107723	9126	21241	31542	1018	140	31	671	123	5	12	35	4	
r	77283	42792	46635	38706	21755	440	4438	85	678	755	8	52	8		
imagej	60485	46442	38031	7437	12295	3	139	313	26	64			1		2
graphpad prism	66399	49932	18109	10033	7979	8	13	29	16	6		1	1		
blast	12405	13900	41132	27154	511	4	143	16	21	85		2	1		
graphpad	35022	26028	8540	5570	4524	6	23	21	6	3					
excel	20033	28911	11485	11175	4478	496	228	57	351	96	4	21	20		2
matlab	10808	9718	13247	2541	31215	14	2385	776	260	57	6	7	3	3	1
stata	21912	26165	394	2576	6200	1182	278	4	274	8	1	2	11		1
mega	5015	5185	15149	14403	128	4	37	2	1	81			10		
blastp	2729	4464	17661	8951	84	1	31	6		13		2			
covid	10181	6445	627	421	5754	1132	432	3	401	1	40	31	18	13	1
edger	5240	2129	4060	2861	516		506	2	4	1					
lasso	6353	2811	442	741	976	7	671	4	58						
freesurfer	264	229		1	11139		10	33			1				
nvivo	196	8184	3	181	753	943	1		133		2	4	3	2	
arcgis	921	1009	1983	2250	134	94	76	12	908	272		1	13		1
admixture	802	91	2762	1820	58		116			261			9		
twitter	188	1758	45	48	768	141	32	20	23			70	4	2	
mathematica	404	257	826	117	179	56	553	105	9	2	1	3			
medline	483	1534	18	49	203	63	5	2	14		2		2	25	
pythia	49	6						1482							
sage	277	296	283	120	81	126	18		22		45				
vosviewer	223	595	33	41	97	24	3		60				10	8	

Note: B09: Physics; B10: Mathematical Sciences; C13: Architecture, Built Environment, and Planning; C15: Archaeology; C20: Social Work and Social Policy; D25: Area Studies; D27: English Language and Literature; D29: Classics; D30: Philosophy; D31: Theology and Religious Studies

Table 2. Representative software in domains with varying levels of influence.

Conclusion

Based on 1,500,334 full-text articles from PubMed Central OA, this study extracted software entities and analyzed their mention-based impact. Our results show that software presence in biomedical research has grown markedly, especially after 2005, and has been accelerating post-2010. Additionally, clear differences were observed across subfields with varying levels of software mentions: subfields with high mention were more frequently associated with general-purpose statistical software, whereas subfields with low mention were characterised by the prominence of more specialised tools. Within articles, software mentions were most concentrated in the 'Methods' sections. General-purpose statistical tools, such as SPSS and GraphPad, consistently maintained a high impact across multiple sections due to their broad applicability.

Generally, this study highlights the trajectory and patterns of software use in biomedical research and provides empirical evidence of its role as a knowledge tool. As an initial step toward assessing software impact, this study adopts mention-based indicators, which capture the explicit visibility and relative impact of software in scholarly articles rather than its causal influence on research methods or outcomes. However, longitudinal trends in software mentions may be affected by changes in reporting practices over time, particularly in earlier periods when software use was not consistently recorded. Therefore, future research could consider the combination of sentiment features and mention frequency, further explore the relationship between software use and

research outputs to establish a more comprehensive and dynamic framework for evaluating software impact.

Acknowledgements

This work is supported by the youth project of the National Social Science Fund of China (Grant No. 24CTQ027).

About the authors

Yuzhuo Wang is an Associate Professor in the School of Management at Anhui University, China. Her main research focuses on text mining and scientometrics.

Heng Zhang is a Lecturer in the School of Information Management at Central China Normal University, Wuhan, China. His research interests include academic text mining and natural language processing.

Huahao Hu is a graduate student in the College of Computer Science at Sichuan University, Chengdu, China. His research interests include software engineering.

Yi Zhao is a Lecturer in the School of Management at Anhui University, Hefei, China. His research interests include scientific collaboration and scientific text mining. He is the corresponding author and can be contacted at yizhao93@ahu.edu.cn

References

- Asada, M., & Fukuda, K. (2024). Enhancing Relation Extraction from Biomedical Texts by Large Language Models. In Proceedings of H. Degen & S. Ntoa (Eds), ARTIFICIAL INTELLIGENCE IN HCI, PT III, AI-HCI 2024 (14736, 3-14). Springer International Publishing Ag.
- Britz, D. (2016). Attention and memory in deep learning and NLP. Retrieved from <http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/>.
- Cui, H., Wang, Y., & Li, K. (2025). Beyond Citations: Tracing and Validating the Rapid Adoption of AlphaFold in Biomedical Research Through Full-Text Analysis. 2002–2009.
- Ding, R., Wang, Y., & Zhang, C. (2019). Investigating Citation of Algorithm in Full-text of Academic Articles in NLP domain: A Preliminary Study. Proceedings of the 17th International Conference on Scientometrics and Informetrics (ISSI 2019), 2726–2727.
- Duck, G., Nenadic, G., Brass, A., Robertson, D. L., & Stevens, R. (2013). BioNerDS: Exploring bioinformatics' database and software use through literature mining. BMC Bioinformatics, 14(1), 194.
- Howison, J., & Bullard, J. (2015). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. Journal of the Association for Information Science and Technology, 67(9), 2137–2155.
- Ibrahim, B. (2021). Statistical methods used in Arabic journals of library and information science. Scientometrics, 126(5), 4383–4416.
- Istrate, A.-M., Li, D., Taraborelli, D., Torkar, M., Veytsman, B., & Williams, I. (2022). A large dataset of software mentions in the biomedical literature (arXiv:2209.00693). arXiv. <https://doi.org/10.48550/arXiv.2209.00693>

- Katsurai, M. (2021). Adoption of Data Mining Methods in the Discipline of Library and Information Science. *Journal of Library and Information Studies*, 19(1), 1–17.
- Lam, C., Lai, F.-C., Wang, C.-H., Lai, M.-H., Hsu, N., & Chung, M.-H. (2016). Text Mining of Journal Articles for Sleep Disorder Terminologies. *PLOS ONE*, 11(5), e0156031.
- Li, K., & Yan, E. (2018). Co-mention network of R packages: Scientific impact and clustering structure. *Journal of Informetrics*, 12(1), 87–100.
- Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
- Pan, X., Yan, E., Cui, M., & Hua, W. (2019). How important is software to library and information science research? A content analysis of full-text publications. *Journal of Informetrics*, 13(1), 397–406.
- Pan, X., Yan, E., Wang, Q., & Hua, W. (2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics*, 9(4), 860–871.
- QasemiZadeh B & Schumann A K. (2016). The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1862–1868.
- Settles, B. (2012). *Active learning*. Springer International Publishing.
- Tateisi, Y., Ohta, T., Pyysalo, S., Miyao, Y., & Aizawa, A. (2016). Typed Entity and Relation Annotation on Computer Science Papers. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3836–3843.
- Wang, Y., & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics*, 14(4), 101091.
- 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.
- Wei, Q., Zhang, Y., Amith, M., Lin, R., & Xu, H. (2020). Recognizing software names in biomedical literature using machine learning. *Health Informatics Journal*, 26(1), 21–33.
- Zheng, A., Zhao, H., Luo, Z., Feng, C., Liu, X., & Ye, Y. (2021). Improving On-line Scientific Resource Profiling by Exploiting Resource Citation Information in the Literature. *Information Processing & Management*, 58(5), 102638.

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