Toward Reliable Forward Snowballing in Systematic Literature Reviews: A Comparative Study and Framework Proposal

Jailma Januário*, Maria Isabel Nicolau*, Katia Romero Felizardo†, Juliana Alves Pereira*
*Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil
†Federal University of Technology – Paraná (UTFPR), Cornélio Procópio, Brazil

Abstract

Systematic Literature Reviews (SLRs) play a vital role in the software engineering field by synthesizing existing knowledge, identifying research gaps, and guiding future investigations with methodological rigor. Given the rapid growth of published research, employing techniques such as snowballing is essential to complement traditional database searches, ensuring a comprehensive coverage of the relevant literature. Forward snowballing, in particular, helps to discover newer studies that cite key seed papers, enhancing the completeness of SLRs. Despite its value, forward snowballing remains a labor-intensive and error-prone task when performed manually. To address this challenge, we mapped existing tools that support forward snowballing. Our analysis focuses on the capabilities of these tools to automate core tasks, including the identification of relevant articles, the extraction of bibliographic metadata, and deduplication. We critically examine their reliability by comparing their outputs against a manually curated forward snowballing process. Key evaluation criteria include the completeness and relevance of the metadata retrieved, the quality of the retrieved scientific databases, and the ease of integration into SLR workflows. As future research, based on our findings, we will propose a framework that takes advantage of existing tools to automate forward snowballing.

Keywords

Systematic Literature Review, Automation, Forward Snowballing

1 Introduction

Systematic literature review (SLR) is widely adopted in software engineering (SE) to structure and synthesize existing research on specific topics. Unlike traditional or ad hoc literature reviews, SLRs follow a well-defined protocol aimed at ensuring methodological rigor, reproducibility, and comprehensive coverage of relevant evidence. As noted in Wohlin [16], one of the key limitations of database searches in SLRs is their reliance on keyword matching, which may fail to capture important studies due to inconsistent terminology, indexing issues, or publication delays. To mitigate this issue, snowballing has emerged as a valuable search strategy in the context of SLRs, gaining increasing attention for its ability to complement and, in some cases, replace traditional database searches [1, 4, 5, 16]. Previous studies have shown that snowballing can produce results comparable to conventional search approaches when a carefully selected seed set of articles is used [1]. The guidelines proposed by Wohlin [16] have further established snowballing as a cost-effective and reliable alternative, particularly when forward snowballing is applied via platforms like Google Scholar, using a previous SLR and its included studies as the seed set. Subsequent research has reinforced the potential of snowballing, not only to identify new evidence [4, 5], but also to support hybrid search strategies that integrate snowballing and database searches.

In particular, forward snowballing is a valuable strategy for expanding the search space, uncovering recent studies, and tracing the evolution of knowledge through citations of a selected set of studies. It is especially valuable for updating SLRs, as it helps identify new evidence based on existing research [4, 5]. The motivation behind this study stems from a well-known trade-off in SLRs: *Although forward snowballing is critical to extend and update the coverage of a review, manually performing it is time-consuming and error-prone.* As the volume of scientific publications grows, the ability to automate forward snowballing without compromising quality becomes increasingly important.

The semi-automation (or fully automation) of forward snowballing remains a barrier to its broader adoption in SE. Therefore, the goal of this paper is to assess whether current tools are capable of supporting or partially automating forward snowballing.

Our evaluation is grounded in the core principles of SLRs, such as the definition of research questions, inclusion and exclusion criteria, structured data extraction, and quality assessment, which are all used to compare manual and automated snowballing results. This alignment ensures that our analysis does not focus only on technical capabilities, but on whether those capabilities contribute meaningfully to the integrity and completeness of the review process. By following Wohlin's guidelines [16] and adopting the SLR principles as an evaluation framework, we investigated to what extent existing tools can be reliably incorporated into the SLR process, thus bridging the gap between methodological rigor and practical scalability.

With our goal, the following contributions are made:

- Advancing the understanding about snowballing, complementing Wohlin's guidelines [16] by investigating to what extent existing tools can be reliably incorporated into the SLR process, thus bridging the gap between methodological rigor and practical scalability of the current solutions.
- By conducting a detailed, tool-oriented comparison grounded in empirical SLR practices, this work contributes to the growing body of research focused on open science infrastructure for the automation of forward snowballing in SLRs.
- Analysis of tools supporting forward snowballing, considering aspects such as metadata completeness, citation retrieval accuracy, and literature coverage, providing a comparative foundation for future research and development in the field.
- Identification of practical limitations in the use of automated tools, revealing gaps such as inconsistencies in the retrieval of citing articles and the absence of essential data. These findings guide the scientific community on the necessary precautions when using such solutions in SLRs.

2 Related Works

Systematic Literature Review (SLR) in SE, as in other disciplines, plays a fundamental role in identifying, organizing, and synthesizing the body of evidence available on a specific research topic. However, the accelerated pace of technological advancements in the field presents a major challenge: ensuring that reviews remain up to date over time. To address this, it becomes increasingly necessary to adopt more efficient, automated, and iterative strategies for conducting and updating SLRs. In this context, we carried out an exploratory search of the literature on tools to support the snowballing process. These tools aim to reduce the manual effort involved in identifying relevant studies, improve the scalability of SLRs, and enhance the overall efficiency and reproducibility of SLRs in the SE domain.

Torres-Salinas et al. [13] provide a comprehensive comparative analysis of 44 bibliometric tools, commercial and open source, highlighting the growing diversity of solutions and the importance of interoperability for academic and institutional research. Herrmannova and Knoth [9] present a large-scale evaluation of multidisciplinary bibliographic databases (including Scopus, Web of Science, Crossref, and Microsoft Academic), offering insight into differences in coverage, document types, and disciplinary alignment. In the biomedical domain, Visser et al. [14] compare five open citation sources for PubMed, highlighting variations in citation quality and completeness. Although these studies provide a valuable overview of the capabilities and limitations of bibliographic data sources, most of them focus on general metadata benchmarking rather than on their applicability to advanced search strategies. In contrast, our study addresses this gap by assessing how open tools can be used to support and potentially automate forward snowballing in SLRs.

3 Methodology

To achieve our objective, the study was conducted in two main steps, detailed in 3.2 and 3.3. First, the tools that support forward snowballing were identified through a literature review. A total of five solutions were selected based on the following criteria: (i) they offer free access, at least partially; (ii) they support the execution of forward snowballing; and (iii) they allow exporting the results to an output file. Next, we evaluated the capabilities of these solutions to extract key metadata (*e.g.*, title, authors, DOI, venue), support citation navigation, extract high-quality literature with broader coverage, and manage data deduplication and filtering.

The evaluation was structured around seven dimensions: (i) annual distribution of retrieved articles per tool; (ii) availability and richness of extracted metadata; (iii) overlap of retrieved articles across tools; (iv) duplication rate; (v) scientific database quality of retrieved articles per tool; (vi) support for filtering operations; and (vii) suitability for integration into reproducible review pipelines.

Taking into account the evaluations and limitations identified during Step 1, a framework was proposed to automate forward snowballing in Step 2.

3.1 Research Questions

This study is guided by three research questions (RQs), as follows:

RQ_1 **What types of article's metadata can each tool extract and how complete and accurate is this information?**

Metadata such as title, authors, publication year, venue, and citation links are essential to filter, analyze, and report results in SLRs. This question explores the capacity of different tools to retrieve metadata, as well as the consistency of those data across platforms. The value of a tool depends not just on the quantity of articles recovered but also on the quality of the contextual information it provides for each article. Understanding metadata completeness is crucial for automation to be truly effective.

RQ_2 **Do the articles retrieved by these tools maintain the quality and relevance expected from the relevant SE scientific databases?**

Although automation can reduce effort and human error, it should not compromise the quality of manually conducted SLRs. This question investigates whether automated tools retrieve articles that are comparable in quality from relevant scientific SE databases. The goal is to assess whether the tools support the same level of rigor and trustworthiness that researchers rely on in manual SLRs, which is critical for ensuring that automation enhances rather than weakens the review process.

RQ_3 **How consistent and reliable are the results produced by each tool compared to a manually performed snowballing approach?**

Evaluating the consistency and reliability of automated tools requires a direct comparison with results obtained through manual forward snowballing. This question investigates potential divergences in the set of retrieved articles, examining whether key publications are missed or whether irrelevant results are introduced. By analyzing agreement levels and gaps, we can determine which tools are robust enough to support or even replace parts of this manual process in SLRs.

3.2 Step 1: Mapping Existing Solutions to Support Forward Snowballing

We map the landscape of available solutions to support forward snowballing, documenting their core features and accessibility. We adopted this overview as the basis for the subsequent analysis of our research questions (see Section 3.1) and to help SE researchers make informed decisions when selecting tools for evidence retrieval in SE and related fields.

To identify existing tools in the literature that support the application of forward snowballing, a search was carried out for online and freely accessible solutions. Based on the literature, including [3], a set of five solutions was selected for this study. Table 1 summarizes their main features.

ResearchRabbit [3]. It is a scientific literature exploration platform supported by Artificial Intelligence (AI). The tool enables users to locate academic works related to one or more initial publications, using visual maps and lists of prior, subsequent, or similar studies. ResearchRabbit was designed to support exploratory and unstructured search processes while maintaining a visual trace from the original publication to selected authors or articles.

Litmaps [12]. It is an AI-based tool developed to facilitate the mapping of academic citations. It offers an intuitive interface that

Table 1: Comparison of solutions for forward snowballing automation.

Features	ResearchRabbit [3]	Litmaps [12]	Semantic Scholar [6]	OpenAlex [11]	Google Scholar [8]
Considered Databases	PubMed, Semantic Scholar	Semantic Scholar, OpenAlex, Crossref	MAG, Crossref, PubMed, PMC, arXiv, Springer, Elsevier, IEEE, ACL, bioRxiv, medRxiv	MAG, Crossref, ORCID, DOAJ, Unpaywall, arXiv, Zenodo	Various academic sources (unspecified)
Open Source	No	No	No	Yes	No
Input Types	Article title	Article title, DOI	Keyword search, filters, article ID	Keyword search, specific IDs, DOI, ORCID	Title, DOI
Output Information	Visualization maps, publication lists, full metadata*	Interactive citation maps, publication lists, full metadata*	Full metadata*	Full metadata*	Articles, theses, books, abstracts
Tool Integration	Yes, with Zotero	Yes, with Zotero	API available	API available	Not available
Update Frequency	Not specified	Not specified	Not specified	Monthly	Continuous
Documentation Availability	Yes	Not available	Yes	Yes	Not available

*title, authors, citations, ISSN, DOI, abstract.

allows users to visually explore the complex landscape of scientific production, efficiently identifying connections between different publications. The platform provides free access for up to two projects per user.

Semantic Scholar [6]. It is a tool maintained by the Allen Institute for AI and uses Natural Language Processing (NLP) techniques to enrich extracted metadata. Semantic Scholar provides a public API with access to detailed information on articles, authors, citations, and references. It uses machine learning models to identify influential citations, infer semantic connections, and automatically organize knowledge through thematic areas and venues.

OpenAlex [11]. It is a public API maintained by the non-profit organization OurResearch. Its architecture is based on an interconnected data model inspired by OpenCitations and the concept of academic knowledge graphs. Each entity (authors, institutions, venues, articles) has a unique identifier and is represented relationally, enabling complex queries through structured filters. The API supports pagination, sorting, term searches in titles and abstracts, publication year filters, and filters by field of knowledge. OpenAlex is particularly robust for studies that require institutional, geographic, or interdisciplinary analysis, due to the richness of its metadata and its integration with CrossRef[1] and ORCID[2].

GoogleScholar [8]. It is a free academic search and reference tool that indexes content from various open access sources, such as SciELO, IEEE, and ACM. The platform includes scientific articles, theses, abstracts, monographs, dissertations, and books.

3.3 Step 2: Solutions Evaluation

To evaluate the solutions selected in *Step 1* (see Section 3.2) and address our three research questions, we selected the SLR titled: *"Learning Software Configuration Spaces: A Systematic Literature Review"* [10]. This SLR was chosen for three main reasons. First, it addresses a topic of high relevance within the SE community. The paper was published in 2021 and has more than 100 citations[3]. Second, the authors bring deep domain expertise in this field, ensuring methodological rigor and an insightful interpretation of the results. Third, the original review is currently being manually updated by forward snowballing by members of the same research group, which guarantees that both manual and automated approaches operate under identical inclusion and exclusion criteria. This facilitates a direct comparison of the results retrieved and also allows us to perform a more nuanced evaluation of each tool's ability to capture relevant high-quality studies.

The study of Pereira et al. [10] is an SLR that focuses on the use of learning techniques applied to configurable software systems. Conducted up to 2019, the study systematically examined the existing literature with the objective of understanding how learning models have been used to predict performance, detect invalid configurations, automate testing and optimize configuration decisions. Following the rigorous application of inclusion and exclusion criteria, 59 relevant articles were selected for detailed analysis. The review offers a comprehensive categorization of the types of approach adopted, the problems addressed, and the characteristics of the datasets utilized, thus providing a consolidated overview of the current state of the art in this emerging research field.

To collect data for the evaluation of the five selected solutions, we applied rigorous strategies. For both *ResearchRabbit* and *Litmaps*, one researcher was required to create a user account and manually enter the title of the seed article in the search interface. This action enabled retrieval of the *cited by* count provided by the tool.

For *Semantic Scholar* and *OpenAlex*, data collection was performed using their publicly available APIs. This approach allowed for automated retrieval of cited articles and associated metadata, such as titles, authors, publication venues, DOIs, and abstracts.

For *Google Scholar*, the search was performed by entering the title of the article in the search bar. The list of citation articles accessible through the *Cited by* link was then used as the basis for the analysis. To download the metadata of the cited articles, *Google Scholar* offers an option that allows all records to be downloaded at once, allowing bulk download of all listed records, which facilitated the consolidation of citation data in a single step.

This data collection approach ensured consistent handling of the seed article in all state-of-the-art tools, enabling a fair and reliable comparison of metadata richness, the quality of retrieved scientific databases, and the accuracy in retrieving relevant articles.

The number of articles retrieved by the different tools varies significantly, as shown in Figure 1. Google Scholar stands out by returning a substantially higher number of articles compared to ResearchRabbit, while Semantic Scholar yields a volume of results more comparable to that of Google Scholar.

[1]https://www.crossref.org/
[2]https://orcid.org/
[3]Indicated by Google Scholar in 21st of June.

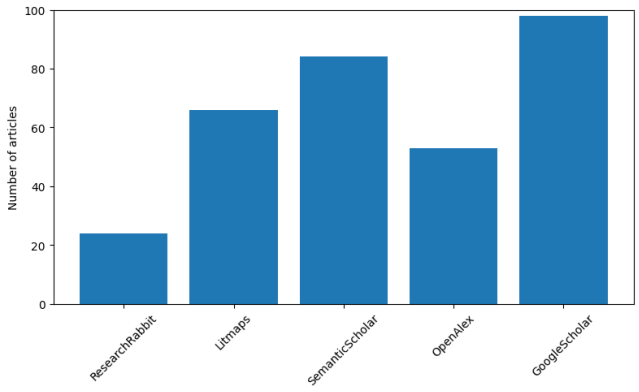

Figure 1: Number of articles retrieved by each tool supporting forward snowballing.

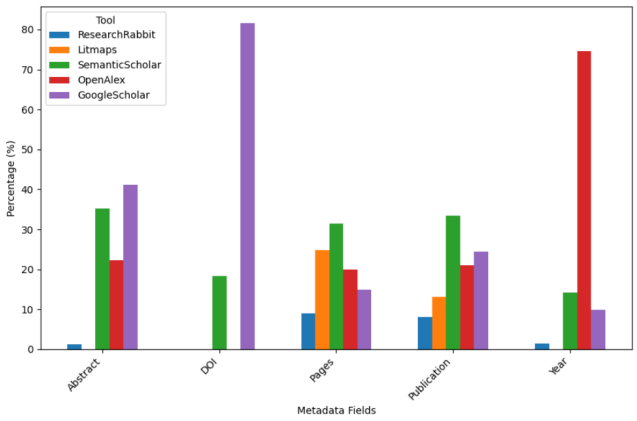

Figure 2: Distribution of missing metadata fields per tool.

4 Results

The following sections present the results for our research questions detailed in Section 3.1. Based on the results, we highlight the most important findings.

4.1 Metadata Information (RQ_1)

The quality of automated solutions depends on the completeness of the extracted metadata. In this section, we evaluate which metadata elements are retrieved by each of the analyzed tools and how this affects their suitability for supporting forward snowballing in SLRs. We examine the presence or absence of key metadata elements, and also the depth and structure of the information retrieved from our selected case study. This analysis is crucial for understanding the solution's ability to support automated screening, enable reproducibility, and ensure the integrity of bibliographic datasets.

The solutions used for semi- or fully automated forward snowballing tasks provide different types of information, as illustrated in Table 2. Metadata such as DOIs, authors, title, publication year, abstracts, citation count (cited by), and publication venue are commonly mapped in SLRs [10]. These elements not only enable more precise bibliometric analyses, but also support the application of inclusion and exclusion criteria, assisting researchers in the initial assessment of the relevant articles.

We observe substantial heterogeneity in metadata support across the evaluated tools. For example, Google Scholar does not return the DOI, a fundamental identifier to ensure and enable seamless traceability integration with external databases, when we follow the most automatic procedure available on this platform. Consequently, researchers must manually locate and record DOIs, introducing additional effort and the potential for errors. Such discrepancies undermine the efficacy of automated screening, compromise the completeness of the extracted dataset, and impede reproducibility.

Furthermore, the absence of specific metadata elements (for example, date, language, open access status, or keywords, which are provided by OpenAlex) further fragments the acquisition of information. By combining metadata from various sources, one can assemble a more consistent and robust dataset, thus enhancing the quality of automated forward snowballing and supporting more reliable SLR outcomes.

> **Finding 1.** OpenAlex provides the most complete metadata elements among the evaluated tools. However, in general, the evaluated solutions retrieve only part of the key metadata elements. Therefore, the combination of different solutions can enrich the metadata and improve the consistency of the dataset for reproducibility.

Although Table 2 catalogs the metadata elements that each tool can provide, the utility lies in whether those elements are actually populated. Figure 2 therefore reports the percentage of missing values for every element supported by at least three of the five evaluated solutions. Please, refer to our supplementary material for the complete list (see Section 6).

Several key observations emerge:

(1) Google Scholar's gaps. Except for title and authors, which all tools retrieved completely, Google Scholar shows null entries in nearly all metadata elements. For example, nearly 82% of the DOI entries are null and more than 40% of the abstract, journal, pages, publication, and publisher elements are missing. This sparsity forces manual procedures and undermines any attempt to automate the process.

(2) Inconsistent support in Litmaps and ResearchRabbit Both platforms exhibit high null rates (50–85%) for pages, publication, and publisher, indicating an inability to capture these citation details reliably. Their abstract coverage is better rated (missing in 36–40% of records).

(3) Superior completeness in OpenAlex and Semantic Scholar These two tools generally maintain missing value rates below 25% for most elements. OpenAlex in particular populates DOI, abstract, and open-access, while Semantic Scholar shows robust coverage of journal, year, and cited by counts. However, they struggle most with the pages and the publisher (closer to 20–30%).

This pattern indicates that none of these solutions is able to extract these metadata elements accurately or consistently. Such

Table 2: Metadata fields extracted by each analyzed tool.

Tool	id	DOI	title	year	date	language	open_access	authors	keywords	abstract	venue	references	cited by	URL	publisher
ResearchRabbit	–	✓	✓	✓	–	–	–	✓	–	✓	✓	–	–	–	–
Litmaps	✓	✓	✓	✓	✓	–	–	✓	–	✓	✓	✓	✓	–	–
Semantic Scholar	✓	✓	✓	✓	–	–	✓	✓	–	–	✓	–	✓	✓	–
OpenAlex	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Google Scholar	–	–	✓	✓	–	–	–	✓	–	–	✓	–	✓	✓	–

deficiencies may undermine the completeness and reliability of bibliographic datasets, posing challenges for bibliometric analyzes and SLRs that require comprehensive and precise source information. Consequently, the lack of these metadata elements limits the effectiveness of these tools in supporting rigorous academic research.

> **Finding 2.** Tools such as OpenAlex and Semantic Scholar, which deliver more richly populated records, are better positioned to support end-to-end SLR workflows. In contrast, the data sparsity in Google Scholar, Litmaps, and ResearchRabbit requires combining multiple tools to ensure that bibliographic datasets are comprehensive and reliable.

4.2 Quality of Scientific Databases (RQ_2)

In addition to analyzing metadata coverage and technical capabilities, recent work [2, 15] has highlighted the importance of considering the scientific relevance of the literature retrieved by automated tools. In this context, indicators such as the SCImago Journal Rank (SJR) [7] have been widely used to assess the prestige of the journal by accounting for the citation counts and the influence of the sources of the citation. Including such indicators in tool-based evaluations helps to understand whether automation supports meet the standards of quality typically expected in SLRs.

Both the JCR and the Journal Impact Factor (JIF) provide information about journals, publishers, total citations, quartiles, and H-index, among other metrics. The JCR dataset covers information from more than one thousand articles, while the JIF dataset includes more than 31 thousand articles. Thus, we used both datasets to ensure a wider coverage of the analyzed articles.

In this study, we analyze the quartile indicator, which classifies journals and conferences into four groups based on impact metrics, where Q1 represents the top 25% most prestigious in the field, and Q4 groups the bottom 25% with the lowest impact. Therefore, an article published in a Q1 journal is generally recognized as more relevant within the scientific community compared to those published in journals of the lower quartiles [2, 15].

Figure 3 illustrates the distribution of retrieved articles across quartiles (Q1 to Q4) for the five tools: ResearchRabbit, Litmaps, Semantic Scholar, OpenAlex, and Google Scholar. The distribution of citations by quartile indicates that most of the articles retrieved by the tools are published in high impact conferences or journals (Q1), except ResearchRabbit. In particular, the Google Scholar solution stands out by presenting a higher correspondence rate with the SJR/JCR rankings and a larger number of citations in Q1 journals. Specifically, the proportion of Q1 articles retrieved was 72% for Google Scholar (36 out of 50), 68% for Litmaps (17 out of 25), 57% for Semantic Scholar (17 out of 37), 47% for OpenAlex (7 out of 15),

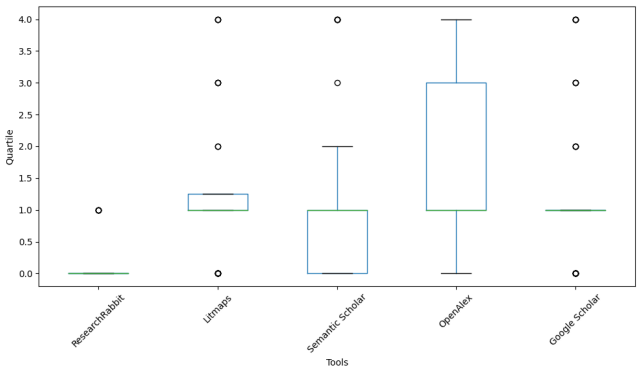

Figure 3: Distribution of citations per quartile (Q1–Q4) for articles retrieved by each tool.

and only 17% for ResearchRabbit (4 out of 24), further confirming the relevance of incorporating such indicators when assessing the scientific quality of automated literature retrieval in SLRs.

> **Finding 3.** The results indicate that most of the articles retrieved by the evaluated tools are published in high-impact venues (Q1), with Google Scholar standing out for its stronger alignment with established rankings and a higher proportion of citations from top-quartile journals.

4.3 Results Consistency and Reliability (RQ_3)

When forward snowballing is performed manually, the researcher uses predefined selection criteria in the protocol to select relevant studies [16]. To answer this question, two researchers (experts in SLR and configurable systems) manually classified the articles following the process defined in Pereira et al. [10] and using the inclusion and exclusion criteria defined in the original paper. Thus, we assessed which of the tools used in this study presents the greatest consistency in its results.

The evaluation was carried out based on the data retrieved from Google Scholar, which totaled 98 articles initially collected. After applying the inclusion and exclusion criteria, 24 articles were considered relevant for further analysis. Of these 24 selected articles, only three were identified in the Semantic Scholar database, while no matches were found in the other tools (OpenAlex, ResearchRabbit, and Litmaps). The low retrieval rate of these tools may be related to differences in indexing mechanisms or the frequency with which their repositories are updated. These results suggest that, depending on the criteria defined by the expert, the use of certain tools may not provide additional benefits to the study selection process.

Finding 4. The absence of results from OpenAlex, ResearchRabbit, and Litmaps, combined with the limited coverage of Semantic Scholar, highlights significant inconsistencies across platforms. These findings suggest that relying on a single tool may not be sufficient for comprehensive and consistent literature identification, reinforcing the need to combine multiple sources to support robust and reproducible SLR processes.

4.4 Threads to Validity

The use of Google Scholar as a data source presents challenges to reproducibility, as its results may vary over time, across geographic regions, and due to frequent updates in citation counts and article visibility. These dynamics impact the consistency of the collected data, particularly in studies that require stability for comparison or replication. However, this is also the case for manually performed SLRs. To mitigate this issue, we recorded search logs and locally stored the datasets used in our analyzes, ensuring traceability and auditability. Furthermore, the availability and quality of the metadata provided by the other tools can fluctuate as new content is indexed or the systems are updated; most of the tools did not specify the update frequency, as seen in Table 1.

Another threat to validity is that the analyzes were based on a single SLR [10], restricting the generalizability of the results. Moreover, although our study focused on the Software Product Line domain, we recognize that metadata richness can vary between fields such as between medicine and computer science, which can influence both the retrieval of studies and the effectiveness of automated screening techniques in broader applications.

5 Proposed Tool

Given the limitations observed in current tools, we propose the development of an integrated forward snowballing solution that combines the advantages of different tools, scientific databases applied to SLRs, and state-of-the-art techniques.

The proposed tool offers automated mechanisms for metadata collection, enrichment, and deduplication, integrating the descriptive richness of sources such as OpenAlex and Semantic Scholar with the broad coverage and scientific relevance of Google Scholar. In addition to prioritizing the completeness of essential metadata (such as title, authors, references, year, publication venue, and impact metrics), the solution will allow for cross-referencing data across multiple sources, increasing the consistency and reliability of the generated datasets.

Unlike previous approaches, our proposal goes beyond merely automating the search and screening phases. It places a strong emphasis on ensuring the scientific quality of the retrieved articles by integrating impact-based filters derived from established ranking systems such as SJR (SCImago Journal Rank) and JCR (Journal Citation Reports). In addition, it addresses key concerns in the evidence-based SE community by enhancing the traceability and reproducibility of the review process.

The proposed tool is designed to support not only the early stages of an SLR but also subsequent SLR stages, such as filtering by impact indicators (*e.g.*, journal quartile), automatic classification of studies based on customizable inclusion and exclusion criteria, and the ability to export data in standardized formats. These features aim to simplify the workflow and reduce manual effort while ensuring methodological rigor. Ultimately, our goal is to deliver a robust, extensible, and scalable solution that meets the practical and methodological needs of the SE research community.

A functional prototype of the tool is available as part of our supplementary material (see Section 6). The current version automates the retrieval of articles that cite a given set of seed papers, and provides interactive visualizations for preliminary analyses, such as the distribution of publications by year and venue. Moreover, it includes an automated screening module that supports the application of inclusion and exclusion criteria defined by researchers, facilitating a more efficient and reproducible selection process.

6 Conclusion

This study aimed to analyze five existing tools that support the forward snowballing process in SLRs, with the goal of identifying limitations and proposing solutions that can better support the scientific community. The analysis revealed significant discrepancies among the tools in terms of the number of articles identified through the "cited by" metadata field, as well as the absence of key metadata elements essential to conducting high-quality reviews.

Furthermore, the assessment of the quality of the retrieved results and their consistency with studies previously reviewed by experts showed a low level of correspondence between the evaluated tools. Based on our findings with this comparative study, we developed a functional web-based prototype tool designed to support researchers throughout the SLR process, specifically the forward snowballing and screening phases. For the forward snowballing phase, it combines different tools to maximizes the retrieval of relevant studies, enables metadata enrichment and deduplication. For the screening of the retrieved articles based on inclusion and exclusion criteria, we applied an automatic approach using large language models (LLMs). Thus, contributing to more complete, reliable, and reproducible SLRs in SE.

Artifact Availability

We have made all artifacts from our study publicly available at https://github.com/aisepucrio/forward-snowballing

Acknowledgements

This research was partially funded by the Brazilian funding agencies CAPES (Grant 88881.879016/2023-01), CNPq (Grant 302339/2022 – 1), FAPESP (Grant 2023/00811-0), and the Binational Cooperation Program CAPES/COFECUB (Ma1036/24). We also acknowledge the Brazilian companies Stone[4] and Flopo[5] for their financial support.

References

[1] D. Badampudi, C. Wohlin, and K. Petersen. 2015. Experiences from using snowballing and database searches in systematic literature studies. In *International Conference on Evaluation and Assessment in Software Engineering* (Nanjing, China) *(EASE' 15)*. ACM, New York, NY, USA, Article 17, 10 pages. https://doi.org/10.1145/2745802.2745818

[2] Lutz Bornmann and Loet Leydesdorff. 2014. Scientometrics in a changing research landscape. In *Springer Handbook of Science and Technology Indicators*. Springer, 347–360. https://doi.org/10.1007/978-3-030-02511-3_15

[4] https://www.stone.com.br/
[5] https://flopo.com.br/

[3] Victoria Cole and Mish Boutet. 2023. ResearchRabbit. *The Journal of the Canadian Health Libraries Association* 44, 2 (2023), 43.

[4] K. R. Felizardo, A. Y. I. da Silva, E. F. de Souza, N. L. Vijaykumar, and E. Y. Nakagawa. 2018. Evaluating Strategies for Forward Snowballing Application to Support Secondary Studies Updates: Emergent Results. In *Brazilian Symposium on Software Engineering* (São Carlos, Brazil) *(SBES' 18)*. Sociedade Brasileira de Computação, Brazil, 184–189.

[5] K. R. Felizardo, E. Mendes, M. Kalinowski, E. F. Souza, and N. L. Vijaykumar. 2016. Using Forward Snowballing to Update Systematic Reviews in Software Engineering. In *International Symposium on Empirical Software Engineering and Measurement* (Ciudad Real, Spain) *(ESEM' 16)*. ACM, New York, United States, Article 53, 6 pages.

[6] Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA* 106, 1 (2018), 145.

[7] Borja Gonzalez-Pereira, Vicente P. Guerrero-Bote, and Félix Moya-Anegón. 2010. The SJR indicator: A new indicator of journals' scientific prestige. *Journal of Informetrics* 4, 3 (2010), 379–391. https://doi.org/10.1016/j.joi.2010.03.002

[8] Michael Gusenbauer. 2024. Beyond Google scholar, Scopus, and web of science: an evaluation of the backward and forward citation coverage of 59 databases' citation indices. *Research Synthesis Methods* 15, 5 (2024), 802–817.

[9] Drahomira Herrmannova and Petr Knoth. 2021. Large-Scale Comparison of Bibliographic Data Sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies* 2, 1 (2021), 20–41. https://doi.org/10.1162/qss_a_00114

[10] Juliana Alves Pereira, Mathieu Acher, Hugo Martin, Jean-Marc Jézéquel, Goetz Botterweck, and Anthony Ventresque. 2021. Learning software configuration spaces: A systematic literature review. *Journal of Systems and Software* 182 (2021), 111044.

[11] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).

[12] Dwi Sulisworo. 2023. Exploring research idea growth with litmap: visualizing literature review graphically. *Bincang Sains dan Teknologi* 2, 02 (2023), 48–54.

[13] Daniel Torres-Salinas, Nicolas Robinson-Garcia, and Enrique Herrera-Viedma. 2023. New Trends in Bibliometric APIs: A Comparative Analysis. *Journal of Informetrics* 17, 3 (2023), 101389. https://doi.org/10.1016/j.joi.2023.101389

[14] Martijn Visser, Nees Jan van Eck, and Ludo Waltman. 2021. Finding Citations for PubMed: A Large-Scale Comparison Between Five Freely Available Bibliographic Data Sources. *Journal of the Association for Information Science and Technology* 72, 9 (2021), 1077–1090. https://doi.org/10.1002/asi.24433

[15] Ludo Waltman and Nees Jan van Eck. 2013. Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison. *Scientometrics* 96, 3 (2013), 699–716. https://doi.org/10.1007/s11192-012-0913-4

[16] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *EASE '14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 1–10. https://doi.org/10.1145/2601248.2601268