Information and Software Technology 99 (2018) 133–147

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy

Nauman Bin Ali*, Muhammad Usman

Blekinge Institute of Technology, Karlskrona, Sweden

ARTICLE INFO

Keywords:
Secondary studies
Systematic literature reviews
Search strategies
Reliability
Credibility
Guidelines

ABSTRACT

Context: The trust in systematic literature reviews (SLRs) to provide credible recommendations is critical for establishing evidence-based software engineering (EBSE) practice. The reliability of SLR as a method is not a given and largely depends on the rigor of the attempt to identify, appraise and aggregate evidence. Previous research, by comparing SLRs on the same topic, has identified search as one of the reasons for discrepancies in the included primary studies. This affects the reliability of an SLR, as the papers identified and included in it are likely to influence its conclusions.

Objective: We aim to propose a comprehensive evaluation checklist to assess the reliability of an automated-search strategy used in an SLR.

Method: Using a literature review, we identified guidelines for designing and reporting automated-search as a primary search strategy. Using the aggregated design, reporting and evaluation guidelines, we formulated a comprehensive evaluation checklist. The value of this checklist was demonstrated by assessing the reliability of search in 27 recent SLRs.

Results: Using the proposed evaluation checklist, several additional issues (not captured by the current evaluation checklist) related to the reliability of search in recent SLRs were identified. These issues severely limit the coverage of literature by the search and also the possibility to replicate it.

Conclusion: Instead of solely relying on expensive replications to assess the reliability of SLRs, this work provides means to objectively assess the likely reliability of a search-strategy used in an SLR. It highlights the often-assumed aspect of repeatability of search when using automated-search. Furthermore, by explicitly considering repeatability and consistency as sub-characteristics of a reliable search, it provides a more comprehensive evaluation checklist than the ones currently used in EBSE.

1. Introduction

Systematic literature reviews (SLRs) have been influential in other disciplines and are considered to provide the most reliable input for policy decisions. Similarly, evidence-based software engineering (EBSE) envisions to establish software engineering (SE) practice on scientific foundations [1] and relies fundamentally on SLRs to identify, evaluate and synthesize empirical evidence on a topic of interest.

The successful adoption of SLRs in SE is indicated by thousands of citations[1] to the guidelines for conducting SLRs by Kitchenham et al. [2,3], and by a search conducted in Scopus to identify SLRs published in software engineering[2] as shown in Fig. 1. The rapid increase in publications of SLRs is similar to the trend observed in other fields (see e.g. [4]).

However, to sustain and increase the confidence in SLRs as a systematic, objective and robust method that enables evidence-based decision making in SE practice, we must ensure and continuously improve the quality of SLRs [5,6]. The cases where secondary studies on the same topic (see e.g. the pairs [7–10]) come to different conclusions raise questions about the reliability of SLRs. Similar questions about the credibility of SLRs have been raised in other fields as well [4,11,12].

The results of an SLR depend on the papers that will be used in data extraction, analysis, and synthesis. Whether a paper reaches the synthesis phase in an SLR depends mainly on decisions taken during the search, study selection and the quality evaluation of papers [13]. Several studies have evaluated the reliability of SLRs as a method by comparing two SLRs conducted on the same topic [13,14]. Such in-

* Corresponding author.
E-mail addresses: nauman.ali@bth.se (N.B. Ali), muhammad.usman@bth.se (M. Usman).

[1] 2110 citations to the updated guidelines [2], and 1280 to the original guidelines [3] based on results on Google Scholar on May 10, 2017.

[2] (Search string used in Scopus to identify SLRs published in computing TITLE-ABS-KEY("systematic review" OR "systematic literature review") AND PUBYEAR < 2017 AND (LIMIT-TO (SUBJAREA,"COMP ")).

https://doi.org/10.1016/j.infsof.2018.02.002
Received 31 May 2017; Received in revised form 7 February 2018; Accepted 7 February 2018
Available online 09 February 2018
0950-5849/

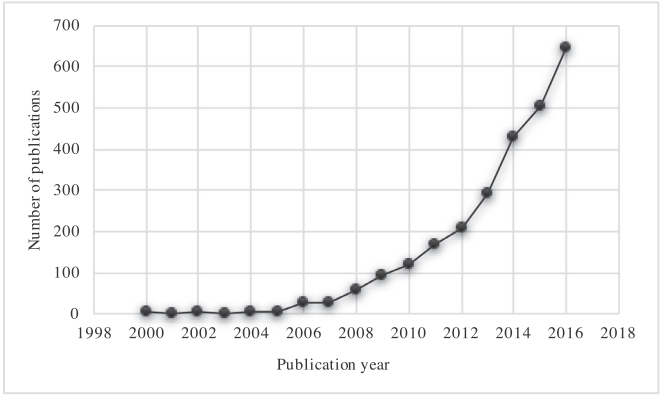

Fig. 1. The increase in number of SLRs in computing literature since 2004.

depth investigations rely on replications and have led to several improvements in the guidelines for conducting and reporting SLRs.

While independent replications are highly desirable and are perhaps a true test of reliability, they are an expensive undertaking. Therefore, we need evaluation criteria that can help assess the likely reliability of an SLR. Among the three main strategies for conducting search in an SLR (i.e. manual-search [15], reference-based search [16–18], and automated-search [2,19]), we have focused on automated-search. In this search strategy, a reviewer primarily relies on a keyword-based search in electronic databases. In SE, a majority of SLRs use automated-search as their primary search strategy, and may only complement the search using manual and reference-based search.

In this study, by critically analyzing the existing evaluation criteria and by applying these on a set of SLRs, we highlight the need for comprehensive criteria to evaluate the search strategy (i.e. the decisions and actions taken to search for relevant literature) used in an SLR. To develop a comprehensive evaluation checklist, we utilized existing design and reporting guidelines for SLRs in SE. We pursued this approach because the guidelines for conducting and reporting SLRs are intended to achieve reliable results by design. Thus, an assessment of conformance to those design and reporting guidelines should allow us to reason about the reliability of search in an SLR. Fig. 2 positions the contribution of this study and the role of existing guidelines.

To illustrate the usefulness of the proposed checklist, we applied it to a set of recent papers reporting SLRs. The assessment revealed that the proposed checklist highlights both significantly more and more serious threats to the reliability of automated-search than current checklists.

The remainder of the paper is structured as the following: Section 2 presents related work on the quality assessment of SLRs with a particular focus on the search-related aspects of SLRs. Section 3 further motivates the need for this research and defines the basic constructs used in the paper. Section 4 presents the research approach. Sections 5 and 6 present the results of the study. Section 7 discusses the results, the broader implications of the findings of the study, and presents some future directions for this work. Section 8 concludes the paper.

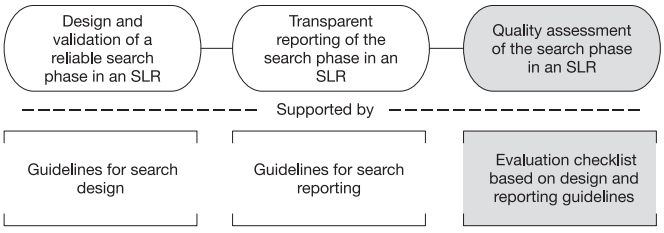

Fig. 2. This study contributes to the quality assessment of an SLR's search phase.

2. Related work

It is often assumed that conducting search systematically is fairly straightforward and more judgment is involved in the later phases of an SLR [17] e.g. selection and quality assessment of papers (e.g. [20,21]). Furthermore, several secondary studies that have investigated the quality of existing SLRs using the following four questions [22] have found relatively good performance on the question related to search quality (i.e. Q-2)):

Q-1: Are the reviewâs inclusion and exclusion criteria described and appropriate?
Q-2: Is the literature search likely to have covered all relevant studies?
Q-3: Did the reviewers assess the quality/validity of the included studies?
Q-4: Were the basic data/studies adequately described?

Kitchenham et al. [23] evaluated 33 studies and assessed only 3 of these (less than 10%) as not "likely to have covered all relevant studies". Similarly, Silva et al. [24] classified only 8 out of the 67 studies (less than 12%) as not "likely to have covered all relevant studies". The score on quality of search was relatively better than the score for Q-3 and Q-4 (Table 1). Furthermore, the score for "search quality" also improved between the two time periods investigated.

Also, some discussions in the community are focused on the overall search approach whether to use manual-search [15], reference-based search [16–18], or mainly automated-search strategy [2,19].

In other studies, about the reliability of SLRs, researchers have compared the results of two or more secondary studies that have investigated the same or similar topics [13,14]. MacDonell et al. [14] compared the results of two independently conducted SLRs. They identified several differences between the two studies in their search approach, data extraction, and analysis. However, both studies had a high overlap in the included primary studies and came to the same conclusions. Wohlin et al. [13] performed a detailed analysis of two independently conducted systematic mapping studies. They conclude that the study cohorts or samples differ between SLRs mainly due to decisions in search, inclusion and exclusion and quality evaluation. Similar results have been found in medicine, Rosan and Suhami [26] performed a comparative analysis of two SLRs on the same topic. They also found search methods among others as a reason for differences in study cohorts.

Tubío et al. [27] in particular analyzed different strategies to search for studies reporting experiments exhaustively, optimally or in an acceptable way. Tubío et al. [27], Zhang et al. [28] and Kitchenham et al. [29] all use precision and recall to make an informed decision about the completeness of a search strategy.

Zhang et al. [28] analyzed 38 published SLRs to identify the search strategy (automatic, manual or a combination), digital libraries and publication venues frequently used in the studies. They further propose a systematic search process that emphasizes the evaluation of automated-search results using a set of known papers. In a similar proposal, Kitchenham et al. [29] recommend splitting the known papers into two sets: for developing a search strategy and for evaluating the probable completeness of the search.

In the studies discussed above the focus has not been on whether the

Table 1
Quality assessment results for SLRs.

	# of SLRs	Quality assessment results			
		Q1	Q2	Q3	Q4
Kitchenham et al. [25]	33	80%	75%	21%	57%
Silva et al. [24]	67	85%	80%	32%	61%