Automated Selection and Quality Assessment of Primary Studies: A Systematic Literature Review

YUSRA SHAKEEL, Otto-von-Guericke-University & METOP GmbH, Germany
JACOB KRÜGER, Otto-von-Guericke-University & Harz University of Applied Sciences, Germany
IVONNE VON NOSTITZ-WALLWITZ, Otto-von-Guericke-University & METOP GmbH, Germany
GUNTER SAAKE, Otto-von-Guericke-University, Germany
THOMAS LEICH, Harz University of Applied Sciences & METOP GmbH, Germany

Researchers use systematic literature reviews to synthesize existing evidence regarding a research topic. While being an important means to condense knowledge, conducting a systematic literature review requires a large amount of time and effort. Consequently, researchers have proposed semi-automatic techniques to support different stages of the review process. Two of the most time consuming tasks are (1) to select primary studies and (2) to assess their quality. In this article, we report a systematic literature review in which we identify, discuss, and synthesize existing techniques of the software engineering domain that aim to semi-automate these two tasks. Instead of solely providing statistics, we discuss these techniques in detail and compare them, aiming to improve our understanding of supported and unsupported activities. To this end, we identified eight primary studies that report unique techniques and that have been published between 2007 and 2016. Most of these techniques rely on text mining and can be beneficial for researchers, but an independent validation using real systematic literature reviews is missing for most of them. Moreover, the results indicate the necessity of developing more reliable techniques, providing access to their implementations, and extending their scope to further activities to facilitate the selection and quality assessment of primary studies.

CCS Concepts: • **General and reference → Surveys and overviews**; • **Information systems →** *Digital libraries and archives.*

Additional Key Words and Phrases: Systematic literature review, Primary study assessment, Tertiary study, Quality assessment, Software engineering

ACM Reference Format:
Yusra Shakeel, Jacob Krüger, Ivonne von Nostitz-Wallwitz, Gunter Saake, and Thomas Leich. 2019. Automated Selection and Quality Assessment of Primary Studies: A Systematic Literature Review. 1, 1 (June 2019), 26 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Originating from the medical domain, systematic literature reviews have become an important research method in software engineering to summarize existing evidence [3, 22, 37]. As a result, the number of conducted systematic literature reviews has considerably increased over the last

Authors' addresses: Yusra Shakeel, Otto-von-Guericke-University & METOP GmbH, Magdeburg, Germany, shakeel@ovgu.de; Jacob Krüger, Otto-von-Guericke-University & Harz University of Applied Sciences, Magdeburg & Wernigerode, Germany, jkrueger@ovgu.de; Ivonne von Nostitz-Wallwitz, Otto-von-Guericke-University & METOP GmbH, Magdeburg, Germany, Ivonne.Nostitz@metop.de; Gunter Saake, Otto-von-Guericke-University, Magdeburg, Germany, saake@ovgu.de; Thomas Leich, Harz University of Applied Sciences & METOP GmbH, Weringerode & Magdeburg, Germany, tleich@hs-harz.de.

XXXX-XXXX/2019/6-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

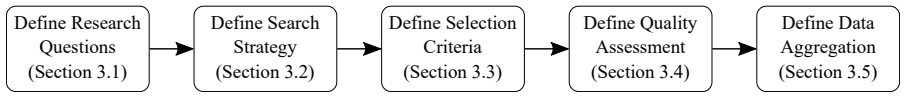

Fig. 1. Overview of our systematic literature review and the sections corresponding to each phase.

10 years [8, 19, 36]. While offering great value for software engineering, conducting systematic literature reviews remains a time consuming and difficult task [13, 15]. In particular, tools to conduct systematic literature reviews in a more efficient and semi-automatic way may support different steps, but often have limited or even missing capabilities. Moreover, tool support is also hampered by technical limitations of digital libraries that threaten their results [21, 32].

While we are focusing on selecting and assessing the quality of papers during a systematic literature review, these two activities are also important in various other contexts. As the number of electronic documents, including scientific papers, regulations, patents, and many more, continuously increases, it becomes more challenging for researchers and organizations to identify and manage those that are relevant for them [31, 39]. For instance, consider a company organizing such documents in a document store or graph database [7] and a connected information system, such as our Hermes [9]. Depending on the database's size, it can be a laborious task to search, select, and assess documents for a specific task, such as conducting a literature review, collecting background for innovations, or simply learning. Consequently, ensuring and assessing the quality of documents is an essential aspect in various domains. As a result, the semi-automated techniques for systematic literature reviews that we review in this article can be helpful to select and assess various documents in different domains—not only during a systematic literature review.

In this article, we report a systematic literature review based on the guidelines of Kitchenham and Charters [21]. Through this systematic literature review, we aimed to identify and synthesize the current state-of-the-art of (semi-)automatic techniques in software engineering for the selection and assessment of papers (called primary studies) for systematic literature reviews. We depict the phases of our systematic literature review in Figure 1 and report details in the displayed sections. This overview is important for both, practitioners and researchers, to support them in their own reviews, to encourage tool development, and to scope further research—not only on systematic literature reviews, but also selection and quality assessment in databases and information systems. The results show that some techniques have been proposed and shown their applicability. However, we found no comparison of these techniques, neither against each other nor to manual systematic literature reviews. Moreover, it seems that only few of the techniques are still accessible.

The remaining article is organized as follows: In Section 2, we discuss works that are related to our study. To ensure transparency and repeatability, we report the details (cf. Figure 1) of our systematic literature review in Section 3. In Section 4, we describe the execution of our systematic literature review. We provide the results, analysis, and answers separated for each of our four research questions in the following Section 5, Section 6, Section 7, and Section 8. Some of the identified techniques have been evaluated with different setups, which we describe in Section 9. In Section 10, we discuss the results and primary findings of our study from a holistic perspective, along with recommendations for future research. We follow with threats to the validity of our study in Section 11 and conclude in Section 12.

2 RELATED WORK

We are aware of few studies that investigate the tool support for systematic literature reviews. The mapping study of Marshall and Brereton [25] provides an overview of tool support to automate systematic literature reviews. They identify that software engineering researchers most commonly aim

to automate the study selection stage. According to the authors' observations, existing approaches mainly represent text mining tools and visualization techniques. Similarly, Hassler et al. [15] conducted a survey in which they identified the tool needs of the software engineering community. The results indicate that existing tools are not appropriate and have to be improved, with the study selection and quality assessment being the most desired features. Both studies indicate the need to improve these two tasks. However, despite providing an overview of existing tools and needs, none of these two works systematically identifies and summarizes techniques that have been proposed (but potentially not implemented) in this regard.

Evidence-based software engineering researchers suggest that quality refers to the methods applied to minimize bias and maximize validity within a study [18, 22]. To rigorously assess the quality of primary studies, Kitchenham and Charters [21] describe a procedure to derive checklists using quality instruments. They suggest that quality assessment checklists must be developed considering problems that may occur at different stages of an empirical study: design, conduct, analysis, and conclusions. Some example questions for assessing quality are:

- Are the research questions and goals of the study clearly stated?
- Does the study method comply with the stated goals?
- Is the purpose of the analysis clear?

Still, such questions are aimed to support researchers during a manual analysis. This may guide the development of automation, but does not consider existing techniques.

To investigate the practices of quality assessment for systematic literature reviews conducted by software engineering researchers, Zhou et al. [40] performed a tertiary study. According to their in-depth analysis, the guidelines for quality assessment defined by Dybå and Dingsøyr [11] and Kitchenham and Charters [21] are mostly used by researchers. Specifically, the most cited checklist is presented by Dybå and Dingsøyr [11] and covers four main categories: rigor, credibility, relevance, reporting. These categories are divided into 11 quality criteria that are again aiming to support manual assessments.

Zhou et al. [40] suggest that software engineering researchers must mainly focus on the credibility of empirical studies for assessing quality, as it directly impacts the validity of the conclusions. In previous works, we analyzed digital libraries to discuss if it is possible to use data from digital libraries to automatically indicate the quality of studies [32, 33]. Still, none of these works provides a comprehensive overview or proposes a new technique to assess the quality of primary studies.

3 RESEARCH METHOD

An important step for conducting a systematic literature review is to define the review protocol [21]. In the following, we describe each step of the applied search process, including the *research questions*, *search strategy*, *selection criteria*, *quality assessment*, and *data aggregation* (cf. Figure 1).

3.1 Research Questions

The goal of our systematic literature review was to summarize existing techniques that support the selection and quality assessment of primary studies during a systematic literature review. To facilitate and improve the results, automated tools and techniques are necessary [15, 30]. Consequently, it is equally important to know about the current state-of-the-art of automated techniques for systematic literature reviews—not only to use them, but also to determine what has been addressed and to scope future research. As we depict in Figure 2, we aimed to identify and discuss techniques that have been proposed to facilitate the *selection of primary studies* and *assessment of study quality*.

Through our systematic literature review, we sought to analyze what existing (semi-)automatic techniques support the systematic literature review process, what are their underlying concepts

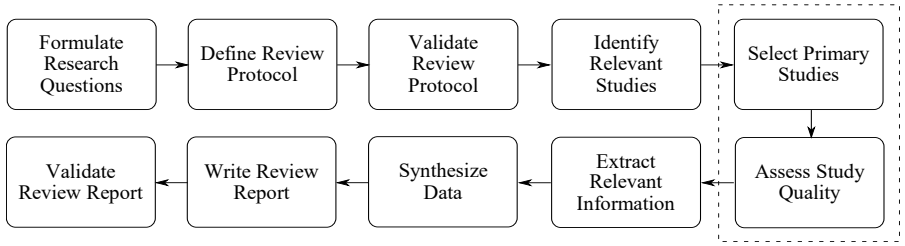

Fig. 2. The systematic literature review process [21]. We framed the steps we focus on in this study.

and limitations, and what areas can be improved in the future. For this purpose, we defined the following research questions:

RQ$_1$ *What techniques have been proposed to select and assess the quality of primary studies?*
First, we aim to provide an overview of techniques for semi-automatically selecting and assessing the quality of primary studies. To this end, we summarize these techniques and describe different aspects, such as, the underlying concepts, conducted evaluations, and results. Our findings help practitioners and researchers who search for a technique to use or who aim to develop a new one.

RQ$_2$ *What underlying concepts do the identified techniques rely on?*
For this research question, we examined the underlying concepts in more detail. We identified the inputs, processing, and outputs of each identified concept to compare them. Thus, it is possible to identify the concepts that may be more suitable for certain use cases or that can be combined with each other.

RQ$_3$ *How are the identified techniques connected?*
Through our third research question, we provide an overview about the relations between the identified techniques. We investigated whether any technique adapts a previously proposed one. These results help to identify whether combinations seem promising and whether there are gaps that have not been addressed, yet.

RQ$_4$ *What are the limitations of existing techniques?*
Finally, we performed a critical analysis of the identified techniques. We discuss the techniques' shortcomings that have been highlighted by the authors themselves or that we identified by comparison. Thus, we determine research directions for future work and limitations that tool developers have to take into account while adopting such techniques.

Overall, we provide a detailed overview of (semi-)automatic techniques to select primary studies and to assess their quality.

3.2 Search Strategy

Our search strategy comprised two phases: First, we conducted an automatic search on four digital libraries and identified relevant papers as primary studies out of the results. To increase the chances of identifying the most relevant works, we selected established scientific databases for software engineering as data sources. Second, to overcome limitations of automatic searches [16, 32], we applied forward and backward snowballing to identify further papers [38].

3.2.1 Automatic Search. To obtain as many relevant results as possible, we extracted broad search terms from the review questions. Through our search string, we aimed to retrieve studies that describe a methodology to assist the conduction phase, specifically, the selection and quality assessment of primary studies for systematic literature reviews. Prior to the actual search for

relevant studies, we performed searches on Google Scholar to inspect whether we would identify relevant literature for our analysis. In addition, we also performed ad hoc queries using databases, namely ACM Digital Library and ScienceDirect, to ensure the inclusion of all key terms. Our key terms included "systematic literature review", as we focus on this research methodology, and "quality assessment" to address this specific stage of the process. We also included the terms "data mining"—as it is a key concept used to extract patterns or models from data, which in our case is a collection of papers—and "relevance categorization"—to identify tools that classify papers based on their relevance for a defined search topic (i.e., selecting them). Then, we constructed a search string by adding synonyms, variations, and related words of these key terms. We connected the terms with logical operators (i.e., `AND`, `OR`) and finally defined the following search string:

```
(approach OR support OR method) AND
(“systematic literature review” OR “systematic review” OR “systematic literature
reviews” OR “systematic reviews” OR SLR) AND
(“quality assessment” OR “literature quality”) AND
(“data mining” OR “recommender system”) AND
(“relevance categorization” OR “relevant study”)
```

All search engines we selected supported this search string at the point we employed our search. We adapted the delimiters for exact searches (i.e., " ") accordingly for ScienceDirect (i.e., { }) [32].

3.2.2 Data Resources. After defining our search string, we selected the scientific databases in which we searched. We targeted those databases that are focusing on software engineering, however we are aware that there may be others with relevant studies not included in our selection. Nevertheless, this can hardly be avoided and we included four libraries that are established in computer science:

- ACM Digital Library (http://portal.acm.org)
- IEEE Xplore (http://ieeexplore.ieee.org)
- ScienceDirect (http://www.sciencedirect.com)
- SpringerLink (http://www.springerlink.com)

We conducted all searches on the full texts of papers between September and November 2016. The search process was performed by the first author and validated by the second and third authors. We remark that we did not include Google Scholar, because it does not allow to download a collection of papers, usually yields a large amount of irrelevant papers, and truncates the results to the 1,000 most cited papers, similar to other search engines. To tackle such problems and complement our results, we applied snowballing.

3.2.3 Snowballing. We first screened the title, abstract, and keywords of each paper obtained during the automatic search to achieve an initial set of potentially relevant papers. Thereafter, we extended our search by performing citation-based analysis using this initial set. Forward and backward snowballing are useful to derive further relevant papers that may be overlooked during the database search [5, 32, 38]. We analyzed the references of each selected primary study (also for those we identified during snowballing) to identify cited papers and used Google Scholar to obtain citing papers [17]. Afterwards, we applied our selection criteria and quality assessment on all identified results.

3.3 Selection Criteria

To identify research that is relevant for our systematic literature review, we defined selection criteria. They are our first way to ensure a certain quality of the identified papers, for example, by including only reviewed papers. We list separated *inclusion* and *exclusion* criteria below:

- Inclusion criteria:
  - The paper must address a technique focusing on the relevance categorization or quality assessment for systematic literature reviews.
  - The paper must evaluate the proposed technique.
  - The paper has been published between 2007, when Kitchenham and Charters [21] proposed their guidelines, and 2016, the year we conducted our review.
  - The paper is peer-reviewed and published in a journal, conference, or workshop.
- Exclusion criteria:
  - The paper is not written in English.
  - The paper is only an abstract or a presentation.
  - The paper is solely published as technical report, bachelor, or master thesis.
  - The paper is published with incomplete or missing information about the publisher or publication type (gray literature).

With these selection criteria, we only included papers that underwent a review process for a scientific venue and are accessible for most researchers.

3.4 Quality Assessment

We assessed the quality of the identified papers to increase the reliability of our results [40]. For this purpose, we followed the guidelines described by Kitchenham and Charters [21] and Kitchenham et al. [20]. For the manual assessment, our checklist comprised the following questions:

QA_1 *Is there a clear statement defining the objective and goal of the work?*
If the goal of the research reported in a paper is not explicitly stated, the understandability can be biased. As scores, we assigned *yes* if the goal is clear, *partly* if some information is missing, and *no* in any other case.

QA_2 *Is there an adequate description justifying the choice of the research method?*
Different methods may be used to design new techniques for assessing papers. We differentiated whether the selected method is described (*yes*) or not (*no*).

QA_3 *Is the research method appropriate to address the defined goal?*
The authors should explain the proposed method completely, including all intermediate steps and their purpose. If this information is provided, we assigned the *yes* score, while unclear or missing descriptions received *partly* and *no* scores, respectively.

QA_4 *Is the applied evaluation feasible to achieve the desired results?*
The results of an evaluation or feasibility study are necessary to show how promising a technique is. Also, limitations must be identified and discussed. In this regard, we assigned *yes* if the evaluation is fully described using tables and graphs, *partly* if some information is missing, or *no* if no evaluation is reported.

QA_5 *Are the findings precisely and coherently reported?*
For any technique, the authors should report valid results as an individual section of the paper, based on the proposed method. If these findings are reported, we assigned *yes*; and *no* in any other case.

QA_6 *Is the value of the work evident?*
To answer this question, we identified whether the practical applicability and value are described. We assigned *yes* if this is the case and *no* otherwise.

In the remaining article, we use symbols to represent the values for *no* (○), *partly* (◐), and *yes* (●).

The quality assessment criteria we list above cover the four main areas of empirical research as explained by Dybå and Dingsøyr [10]:

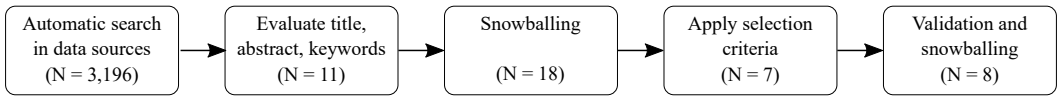

Fig. 3. Conducted steps in our systematic literature review and number of identified papers.

- *Reporting:* Criteria QA_1 and QA_2 relate to the quality of reporting an empirical study based on the goal and its description.
- *Rigor:* Criteria QA_3 and QA_4 address how appropriate and thorough the applied research is for achieving valid results.
- *Credibility:* Criterion QA_5 focuses on the reliability of findings concluded through a study.
- *Relevance:* Criterion QA_6 is concerned with the scientific relevance of papers.

To assess the quality, we analyzed each selected paper. We assigned an overall score based on the aforementioned quality criteria. For this purpose, we assigned a 1 for the *yes* score, 0.5 for *partly*, and 0 for *no*. This procedure is recommended by Kitchenham et al. [19] and the highest possible score is six.

3.5 Data Aggregation

After selecting and rating the quality of identified papers, we extracted the relevant information to answer our research questions. Further, we extracted the following standard data for each paper:

- Primary study ID
- Authors
- Title
- Publisher
- Publication year
- Publication venue and details
- Number of pages

We synthesized the following specific information into a table to answer our research questions:

- Summary of the goal of the paper
- Name and descriptions of the proposed technique
- Explanation of the applied methodology
- Summary of the findings and limitations
- Results of the evaluation

We studied the selected papers carefully to extract this information and report and discuss our results in the following sections.

4 CONDUCT

In this section, we describe how we *executed* our review. We explain the procedure for identifying and assessing papers based on the criteria we defined in the previous section.

4.1 Identification of Primary Studies

We illustrate the complete execution process in Figure 3. First, we applied our search string on the selected databases to identify an initial set of papers. In total, we received 3,196 results of which most are from the ACM Digital Library, as we show in Table 1.

In the next step, we initially screened all papers by reading their titles, abstracts, keywords, and the full text if necessary. We focused on whether a paper is related to our research topic and identified 11 papers to be potentially relevant to answer our research questions. To retrieve further

Data Source	Results	Selected
ACM Digital Library	2,076	2
IEEE Xplore	31	3
ScienceDirect	474	1
SpringerLink	615	1

Table 1. Number of studies identified in each data source.

ID	Title	Reference	Year	Publisher	Venue
1	A Visual Text Mining Approach for Systematic Reviews	Malheiros et al. [24]	2007	IEEE	ESEM
2	Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews	Felizardo et al. [13]	2011	IEEE	ESEM
3	Linked Data Approach for Selection Process Automation in Systematic Reviews	Tomassetti et al. [34]	2011	IET	EASE
4	Applying Information Retrieval Techniques to Detect Duplicates and to Rank References in the Preliminary Phases of Systematic Literature Reviews	Abilio et al. [1]	2015	SciELO	CLEIEJ
5	Semantic Enrichment for Recommendation of Primary Studies in a Systematic Literature Review	Rizzo et al. [29]	2015	Oxford	JDSH
6	Semi-Automatic Selection of Primary Studies in Systematic Literature Reviews: Is it Reasonable?	Octaviano et al. [27]	2015	Springer	EMSE
7	Improvements in the StArt Tool to Better Support the Systematic Review Process	Fabbri et al. [12]	2016	ACM	EASE
8	PaperQuest: A Visualization Tool to Support Literature Review	Ponsard et al. [28]	2016	ACM	CHI

Table 2. Selected primary studies.

papers, we performed forward and backward snowballing on these papers, resulting in seven additional ones. Afterwards, we applied our selection criteria and selected seven papers as primary studies. In the next step, all authors validated the process and we performed snowballing on the seven included papers. Based on the snowballing, we identified one more paper, resulting in the eight primary studies we present in Table 2.

4.2 Quality Assessment

Before investigating our defined research questions, we assessed the quality of the identified primary studies. For each study, we assigned a value for each quality criterion and accumulated them to an overall score. We display the assigned values in Table 3.

As we can see, most of the identified studies are from more recent years and all of them are of rather high quality—almost fulfilling all quality criteria. The criterion we found to be only partly answered in every case is QA_4 (feasibility of the applied evaluation). Two other criteria are only partly fulfilled by other studies: QA_3 (appropriate research method) and QA_5 (reporting of the results). However, in many cases the authors themselves reported the limitations we identified and for which we reduced the scores. For example, Felizardo et al. [13] and Malheiros et al. [24] perform case studies with limited paper samples. Thus, the results cannot be completely transferred to extensive systematic literature reviews.

Similarly, the assessment strategies recommended by Abilio et al. [1] and Octaviano et al. [27] are only evaluated on parts of the papers (i.e., title, abstract, and keywords), but not the full texts. While this is also a problem of accessibility [32], the results are still biased, wherefore we reduced the scores for these studies. Nonetheless, considering also the publishers and venues, the papers we included are of high quality.

Study ID	QA_1	QA_2	QA_3	QA_4	QA_5	QA_6	Score
1	●	●	●	◐	●	●	5.5
2	●	●	●	◐	●	●	5.5
3	●	●	◐	◐	●	●	5.0
4	●	●	●	◐	●	●	5.5
5	●	●	●	◐	●	●	5.5
6	●	●	◐	◐	●	●	5.0
7	●	●	●	◐	◐	●	5.0
8	●	●	●	◐	◐	●	5.0

○ no (0) - ◐ partly (0.5) - ● yes (1)

Table 3. Assessed quality of the identified primary studies.

5 RQ_1: PROPOSED TECHNIQUES FOR STUDY SELECTION AND ASSESSMENT

In this section we present the *identified techniques* proposed to support the selection and assessment of primary studies for systematic literature reviews. Thus, we address our first research question.

5.1 Results

We identified eight different techniques proposed by software engineering researchers to support the selection and quality assessment of primary studies. A description of the research method, the performed evaluation, and results is provided in each of the identified papers. In Table 4, we provide the name, evaluation method, and a summary of the results for each technique. Observing the results of experimental studies, it is evident that the proposed techniques improve the primary study selection activity. However, we further investigated the evaluation methods to determine whether these would also prove to be useful for real systematic literature reviews. To this end, we summarize the identified techniques in the following.

(1) Malheiros et al. [24]. report a feasibility study to evaluate how visual text mining can support systematic literature reviews. For this purpose, three researchers performed systematic literature reviews on the same topic, including 100 papers from the IEEE Digital Library. Two of them (one specialist in visual text mining) were using a visual text mining technique, while one performed a completely manual review. The used visual text mining technique was capable to organize and select papers from the found set. Both researchers who used the technique, had to mark regions of interest in the presented view, which were used to identify further papers. The researcher performing the manual search evaluated papers by reading the abstracts.

Overall, 40 different papers were selected by all four researchers. Of these, 24 were considered relevant as they were included in at least two reviews. The manual review comprised 26 of the 40 papers and required three hours. In contrast, the other two reviews included 22 and 23 papers, requiring 49 and 51 minutes, respectively. Thus, the researchers using visual text mining required considerably less time to perform the same systematic literature review with a comparable quality.

(2) Felizardo et al. [13]. adapted and extended the concept proposed by Malheiros et al. [24] to develop the SLR-VTM technique. To support their findings, the authors conducted an experiment using one existing systematic literature review of 37 papers as baseline. The experiment involved four doctoral students having prior experience in conducting systematic literature reviews. After randomly splitting them into two groups, one group performed the paper selection activity manually by reading the abstracts (Group 1) while the other used the visual text mining tool (Group 2). The participants of Group 2 analyzed the collection of papers—organized by the visual text mining

Study ID	Name	Study Method	Evaluation Results
1	VTM-Based SR	Experiment	Participants were faster with similar to higher precision.
2	SLR-VTM	Experiment	Participants performed better with more reliable outcomes.
3	-	Case study	Reduction of required manual review by 20%.
4	-	Experiment	One strategy achieved 50% precision and 80% recall.
5	-	Case study	Reduction of human workload by 18% with 95% of recall.
6	SCAS	Case study	Strategy reduces average effort by 58.2% with an average error rate of 12.98%.
7	StArt	Tool	Implementation and semi-automation of SCAS.
8	PaperQuest	Tool	Visualization and suggestion of papers based on citations.

Table 4. Overview of the identified techniques.

tool—based on the topics of clusters, the frequency of occurrences of user defined expressions, and neighborhood connections of a relevant paper. In addition to the content analysis, papers are also categorized using their citation relationships.

The manual review of the two students in Group 1 resulted in 25 and 22 papers being correctly assessed, respectively. It took them 85 and 54 minutes to perform the selection activity. In contrast, the two students in Group 2 correctly assessed 27 and 28 papers based on the visual text mining technique. The time to perform the same activity were 30 and 58 minutes, respectively. Additionally, considering the studies incorrectly judged, Group 1 had higher false-negative than false-positive decisions, which can effect the accuracy of the systematic literature review. Thus, the overall results suggest that the use of a visual text mining technique can help to improve the performance of the study selection activity.

(3) Tomassetti et al. [34]. have used semantic web and text mining techniques in the context of a linked data technique to semi-automate the selection of primary studies. The authors describe a supervised, iterative process to assess and categorize papers. A text classifier is used to filter the potentially relevant papers from a search space, producing a reduced set of papers. For the final selection, researchers examine the reduced set that is considered to contain more relevant studies.

A prototype has been implemented to perform a case study, for which an existing systematic literature review with 111 papers was used as baseline. The results show a reduction of 20% in the workload compared to a manual selection with a recall of 100%. Thus, this technique can enable researchers to assess papers with reduced workload and less subjectivity bias.

(4) Abilio et al. [1]. propose two strategies based on information retrieval techniques to rank papers in decreasing order of importance for a systematic literature review. The ranking is based on the relevancy of papers considering specific parts of their content and the search string. In both strategies, a vector model is used and the weight of terms is defined by their frequency of occurrences in the text. Strategy 1 uses the traditional form of the vector model that considers the partial contribution of each query term. In contrast, Strategy 2 defines a ranking function that simulates the Boolean expressions of the search string.

The authors performed an experimental evaluation using real data sets obtained from two existing systematic literature reviews. In these two systematic literature reviews, 13 and 44 papers had been selected, respectively. For the first systematic literature review, Strategy 2 resulted in 80% recall with 50% precision, outperforming Strategy 1 that resulted in a precision of 17.2%. Similarly, for the second systematic literature review, precision values of 50.6% and 52% were achieved by Strategy 1 and Strategy 2, respectively—both with 90% recall. Consequently, the overall results show that the proposed strategies minimized the effort for assessing papers to some extent, but quite differently.

(5) Rizzo et al. [29]. have extended the previous work of Tomassetti et al. [34] and report a semi-automated technique based on text mining and semantic enrichment techniques to reduce human workload for the selection of papers. Unlike the previous supervised, iterative process, this extended work relies on a semi-supervised technique to reduce the set of interesting papers for researchers to evaluate. Thus, researchers can directly discard papers selected by automatic classifiers by looking at their title and abstract without necessarily reading the full text.

For evaluating their technique, the authors used an existing, manually performed systematic literature review with 2,215 studies as a benchmark. This extends the previously used data size of 111 papers to an approximately 20 times larger baseline. The implemented tool was trained with different configurations of relevant papers and tested as an empirical study. Overall, the results of this case study show a reduction of manual workload by 18% with a recall of 95%.

(6) Octaviano et al. [27]. have proposed the Score Citation Automatic Selection (SCAS) technique that adapts the concept proposed by Malheiros et al. [24]. The authors extend that concept with features of an existing visual text mining tool. Papers are automatically categorized, using the SCAS technique, based on their content relevance scores and number of citation links.

To evaluate the feasibility of SCAS, an exploratory case study has been conducted using three manually performed systematic literature reviews. The goal of the evaluation was to compare the accuracy of selecting papers manually and by using the SCAS technique. Using SCAS to replicate the manually conducted systematic literature reviews reduced the effort by 61.85%, 54.04%, and 58.71% with error rates of 4.12%, 18.91%, and 15.90%, respectively. These results indicate that using the SCAS technique minimizes the required manual effort without necessarily affecting the overall results of a systematic literature review.

(7) Fabbri et al. [12]. have implemented the SCAS technique, presented by Octaviano et al. [27], as a new feature of an existing tool. State of the Art through Systematic Review (StArt) is one of the relevant tools that exist to support the entire systematic literature review process and that has been preliminary evaluated. The added feature allows users to automatically set papers as accepted or rejected based on the SCAS recommendations. This implementation makes StArt more robust, improving the support provided to researchers for the conduct of systematic literature reviews.

(8) Ponsard et al. [28]. present a visualization tool to support efficient reading decisions by only displaying information useful at a given step of the review. PaperQuest finds and sorts papers that are likely to be relevant to users based on papers they have already expressed interest in and the number of citations. A preliminary feedback has been collected from four doctoral students, one postdoctoral researcher, and three researchers working in the field of information visualization. The results indicate that PaperQuest may be helpful for conducting systematic literature reviews.

5.2 Summary

Considering the experiments performed to evaluate the visual text mining based techniques, VTM-Based SR (1), and SLR-VTM (2), we see that these techniques seem to considerably facilitate the assessment and selection of papers. For instance, comparing the precision of selected papers, the manual review had a rate of 83.87% and the visual text mining reviews achieved comparable or higher rates with 81.28% and 92%, respectively. As a result, the identified techniques relying on visual text mining seem to accelerate systematic literature reviews, while not reducing the precision. However, we must emphasize that the experiments are performed on relatively small datasets compared to real systematic literature reviews, miss replications, and were not evaluated with experts in the domains. Thus, to verify the indicated suitability of these techniques, further evaluations are necessary.

Similarly, the overall results of the case study performed to evaluate the SCAS technique (6) also show a reduction in average effort by 58.2% with an average error rate of 12.98% [27]. The manual workload can be further reduced by increasing the automatic assessment of papers. However, this results in an increase of error, due to false-negative decisions, meaning that relevant evidence is lost. This result indicates a common problem of all the identified techniques: As not all information that are needed to assess a paper's quality can be identified without reading, there is a trade-off between automation and precision. The question arises, to what extent this trade-off can be accepted or is small enough to not negatively impact the outcome?

Other techniques that rely on semantic web or text mining also seem to be useful. For instance, some results (3) show that more than 20% of the manual work with respect to the original manual workload can be reduced, without missing relevant papers [34]. Furthermore, the authors extended the technique to measure the effect of enrichment on the precision of the classifier, observing a gain of up to 5%. Still, the analyses are based only on the abstract and introduction of the papers. Finally, we remark that the identified techniques seem more concerned with the relevance of the analyzed papers and their selection. Assessing the actual quality of papers is only scratched by most techniques.

5.3 Concluding Remarks for RQ_1

We identified eight papers that present different techniques, published between 2007 and 2016 in the software engineering domain. These papers provide experimental or case study evaluations of the proposed techniques. Most results show that the techniques are effective and facilitate the selection activity up to a certain extent. However, a complete formal evaluation is missing for most of them. In particular, comparisons between the techniques and to experts are missing. Thus, to meet this deficiency of reliable techniques to accelerate the execution phase of systematic literature reviews, further research is necessary. Nonetheless, the existing techniques do not only provide appropriate starting points, but they are already promising to be further integrated into tools and work flows.

6 RQ_2: UNDERLYING CONCEPTS

In this section, we consolidate and explain the *underlying concepts of the identified techniques* to answer our second research question.

6.1 Results

After reviewing the selected papers, we found that although each study describes a different technique to semi-automate paper selection and quality assessment, some rely on the same concepts. For example, the techniques described in most papers (1, 2, 6, and 7) are based on visual text mining techniques. We provide a comparative overview of the underlying concepts in Table 5. We identified four concepts that we describe in more detail in the following, focusing on the information required as input, their processing, and the resulting output.

Visual Text Mining. The process of extracting useful information from textual documents is referred to as text mining [35]. To improve and simplify the discovery of information, interactive visualization techniques are combined with text mining algorithms. Visual text mining puts large textual sources in a visual hierarchy or map and provides browsing capabilities to support efficient exploration of documents [24]. For a systematic literature review, visualization tools (i.e., PEx and Revis) are used for applying visual text mining techniques to help users select relevant papers without actually reading the complete texts [24, 25].

Concept	VTM				SW & TM		IR	CL
Study ID	**1**	**2**	**6**	**7**	**3**	**5**	**4**	**8**
Input								
Abstract	●	●	●	●	●	●	●	●
Citations	○	●	●	●	○	○	○	●
Conclusion	●	○	○	○	●	○	○	○
Full Text	●	○	○	○	○	○	○	○
Introduction	●	○	○	○	●	●	○	○
Keywords	●	●	●	●	○	○	●	○
References	○	●	○	○	○	○	●	●
Search query	○	○	●	●	○	○	●	○
Seed papers	○	○	○	○	●	●	○	●
Title	●	●	●	●	●	○	●	●
Output								
Citation network	○	●	○	○	○	○	○	○
Document map	●	●	○	○	○	○	○	○
Edge bundle	○	●	○	○	○	○	○	○
List	○	○	○	○	○	○	●	●
Quadrants	○	○	●	●	○	○	○	○
Set	○	○	○	○	●	●	○	○

VTM: Visual text mining; TM: Text mining; SW: Semantic web; IR: Information retrieval; CL: Citation links

Table 5. Summarized concepts of the identified techniques.

Input. Considering the technique proposed by Malheiros et al. [24] (1), the visual text mining tool expects the full texts of papers in raw ASCII format as input. Whereas, the three other proposed techniques (2, 6, 7) require title, abstract, keywords, and citations of papers—partly enriched with meta-data or references. All these techniques derive the initial set of papers by applying the defined search query on the selected venues.

Processing. The visual text mining techniques process the input to compute the similarity between papers. This process involves the conversion of all papers into multi-dimensional vectors based on the extracted terms. Thus, the term-matrix for a document comprises the *term frequency, inverse document frequency (tf-idf)* measure. Finally, the techniques compute paper similarity based on the cosine similarity.

Output. All techniques can output the relationships of papers in a two-dimensional paper map [24]. Other visual text mining techniques support additional representations, such as, an edge bundle and a citation network [13] (2). To create this output, the user must provide references of papers along with their title, abstract, and keywords. As the final paper selection is done manually using specified exploration strategies, a representation of citation relationships can be beneficial. Furthermore, the visual text mining tool by Octaviano et al. [27] (6) can also automatically classify studies into three categories and four quadrants, as we display in Figure 4. This supports users in identifying papers that require manual reviewing based on the relevancy score and citation links.

Semantic Web and Text Mining. Researchers use automated text classification techniques to reduce the workload while selecting and assessing papers. We identified two papers (3, 5) that adopt combinations of semantic web and text mining techniques to enrich feature selections based on linked data. To this end, a resource description framework repository, such as DBpedia [2], is

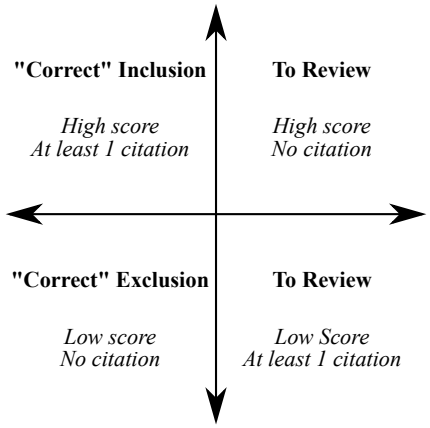

Fig. 4. Classification of papers using the SCAS technique of Octaviano et al. [27] (6).

useful to link information between papers. Thus, the data space of papers is enriched with further information to identify concepts for the classification.

Input. An initial set of relevant papers is required to build a model based on the occurrences of terms. The supervised iterative technique proposed by Tomassetti et al. [34] (3) considers the title, abstract, introduction, and conclusion for the model building. In contrast, only abstracts and introductions are used for the semi-supervised technique proposed by Rizzo et al. [29] (5).

Processing. The techniques build an initial model by comparing the papers that the user provides. A text classifier filters potentially relevant papers within this search space and produces a reduced set containing papers with higher similarity—based on the ratios of identical terms used—to the initial set. To this end, different text classifiers, such as, Naive Bayes, decision trees, neural networks, support vector machines, and hybrid approaches can be used. Every re-classification of papers as relevant makes the used model obsolete and requires rebuilding. Each iteration progressively tailors the model according to the domain of interest of the user, allowing refinements of the search process. Thus, these techniques depend on the dimensions of the search space: The lager it is, the more effective the technique becomes.

Output. The two techniques we identified return a set of potentially relevant papers. This set is a reduced set of the overall search space provided by the user, which may comprise any number of papers. For the final selection, reviewers must manually review the obtained papers.

Information Retrieval. Retrieving information from resources based on user's requirements is referred to as information retrieval [1]. Different techniques have been proposed to analyze text documents, web pages, online catalogs, multimedia objects, as well as structured and semi-structured records. To effectively identify relevant papers with information retrieval techniques, a classic algebraic model can be used, for instance, a vector model. We identified one technique (4) that represents papers and queries as vectors in a t-dimensional space, with t denoting the number of distinct terms.

Input. To facilitate systematic reviews, the user has to provide a search query that can be executed on the search venues. Furthermore, the references of a paper must be provided. The technique itself analyzes the abstract, keywords, and title of each paper.

Processing. First, the provided input is preprocessed, including the removal of stop words, punctuation, and symbols as well as the conversion of letters to lowercase. In the following steps, the technique inserts distinct tokens—obtained from the analyzed papers—into an inverted index

structure composed of key-value pairs. Each pair consists of a key as a distinct token and the value as a list of occurrences. The frequency of occurrences for each token in the provided query and the inverted index is then verified. Lastly, the weight of each token in the query (Strategy 1) or in the group (Strategy 2) for a new paper is calculated to determine its relevance.
Output. For the identified information retrieval technique, the output comprises a list of papers. This list also comprises relevance scores that are used to order papers in descending order. A paper that is higher in the list should have a better chance of being relevant, and thus selected.

Citation Links. The citation relationships of papers can be used to identify and assess further related papers, which is the idea of snowballing searches. This can be helpful, as authors build their research upon previous works, by themselves and others. Consequently, interpreting citation links in a network helps to indicate how papers are related.
Input. The user performs an initial search to retrieve relevant papers in their field of interest. For the technique we identified (8), a dataset with meta-information of the retrieved papers is provided as input, comprising: Title, Digital Object Identifier (DOI), publication year, venue, authors, abstract, and references to other papers. Furthermore, the technique builds on the citation counts of an external library, such as Google Scholar, to extend the dataset. Using citation links of these seed papers, users can discover and assess other relevant papers.
Processing. Browsing citation links of the seed papers can retrieve large numbers of papers, as these may refer dozens and may be cited by hundreds. Consequently, filtering the retrieved papers and identifying those with the most relevant information to the current goal is crucial. The identified technique applies a multi-level decision process. To this end, a minimal amount of information is gathered from each paper to decide whether to keep the paper for further processing. The process is flexible and iterative: After reading some papers, the users can get a better understanding of the domain of interest and can gather new papers to filter and read. As the assessment of papers is an exploration of the search space of published papers, this space can be divided into [28]:

- *Core*: Papers the users have read and obtained their understanding from.
- *Fringe*: Papers the users have access to because they are linked to core papers.
- *To Read*: Papers the users assessed as relevant in the fringe but did not read, yet.
- *Unknown*: All other papers that are not related to the potentially relevant ones.

The technique itself assesses the relevance of papers in each of these subspaces to reassign them. For this purpose, it compares and weights the links of papers, using three metrics: Internal citation count of the provided papers, external citation count based on a digital library (i.e., Google Scholar), and a connectedness measure.
Output. While there are four subspaces of papers, the output is a single list. This list displays the papers ranked from most relevant to least relevant according to the metrics. Thus, this list can support users in focusing on important papers and may enable them to automatically discard papers below a certain threshold.

6.2 Summary

We identified four concepts that are used to select and assess papers for systematic literature reviews. The most commonly applied concept is text mining, partly enriched with additional concepts or refinements. In particular, visual text mining has been used to assess the relevance of papers. Two techniques additionally incorporate semantic web concepts to enable further automation. Other concepts, namely information retrieval and citation links, each have been used once.

Despite these efforts, all concepts we have identified heavily rely on the user to provide suitable data, thus requiring manual effort. Some of these concepts solely require input documents without further information. Consequently, the automation could be increased for those concepts, if they

are integrated with tools for automated searches. However, we see some issues with the identified concepts, for instance, all concepts rely on rather simplistic metrics that may not be appropriate to assess papers [33]. This seems even more problematic, as some techniques are only applied on parts of the papers, potentially for accessibility reasons. While analyzing the concepts' processes, we faced mainly four open questions:

- Are the used metrics meaningful to assess the relevance and quality of a paper?
- How can we further automate the identified concepts for systematic literature reviews?
- Can we incorporate actual quality criteria (in natural language) that are defined by users?
- Are the results' representations useful for users conducting a systematic literature review?

These questions may guide further research and also help to identify other concepts to support the assessment of papers. For example, machine learning and other data mining techniques may enable users to define fine-grained quality criteria to assess papers.

6.3 Concluding Remarks for RQ_2

To summarize our findings, we identified four concepts that are used to identify and assess papers. Apart from one technique that entirely relies on citation relationships to retrieve relevant papers, all others use a content-based analysis. However, we observe that only specific parts of the content are considered, such as title, abstract, introduction, and conclusion. The support for full-text analysis of papers seems to be missing for these techniques, while the concepts could use the full texts. Thus, further research in this regard can facilitate and improve the reliability of paper assessments.

7 RQ_3: CONNECTING TECHNIQUES

To address our third research question, we investigated whether the identified techniques have been combined and derived from each other. As they all share the goal of facilitating the selection and assessment of papers, it seems reasonable to combine techniques to utilize their strengths.

7.1 Results

We show the temporal development of all identified techniques in Figure 5. Besides the fact that most techniques have been proposed in recent years, we can also see that the different concepts are still strictly separated. In the area of visual text mining, essentially the same technique has been extended with additional refinements or implementations, for example, further visualizations (2). Similarly, the technique proposed by Tomassetti et al. [34] (3) has been extended within the same concept area. While some of the concepts are related, as they rely on text mining and citation links, we did not identify a direct combination of the corresponding techniques. According to our results, visual text mining represents the concept for most of the identified techniques and has been used continuously throughout the analyzed period. In contrast, other concepts have only been adopted later on and may be further advanced.

7.2 Summary

All identified techniques are used to select and assess papers for a systematic literature review. Surprisingly, most techniques rely on a single underlying concept (cf. Section 6) and are not combined, yet. Considering the temporal evolution, we further argue that there are other techniques in the same concept areas (e.g., other visual text mining techniques) that can be adopted and are not only based on previously proposed techniques. Thus, it seems surprising that not more concepts are used together to improve the quality based on combinations. Instead, we see a rather fixed separation between all concepts with currently no steps taken to integrate them. This could help to

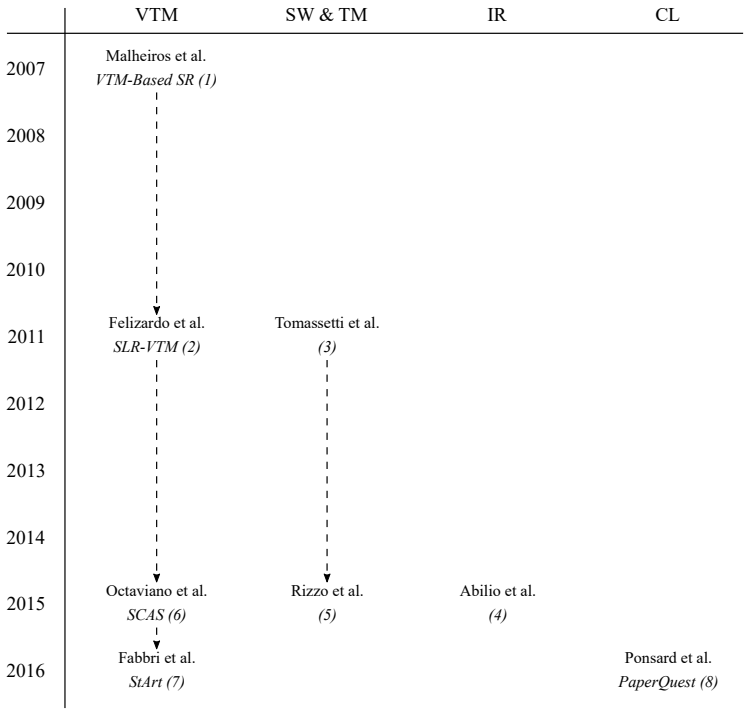

Fig. 5. Temporal development and connections between the identified techniques and concepts.

overcome limitations of some techniques to improve the quality of their results and, finally, achieve the common goal of facilitating the conduct of systematic literature reviews.

7.3 Concluding Remarks for RQ_3

Based on our results, we determine that text mining, particularly visual text mining, is the most established underlying concept for techniques that support selection of primary studies for a systematic literature review. While some techniques utilize other concepts, we found no combination of these established concepts or techniques. Consequently, we argue that it should be a goal for researchers to not only verify and extend existing or propose new techniques, but to also investigate combinations. In particular, comparative evaluations of the techniques' performances and establishing reliable tools are important concerns. To this end, we discuss limitations of the techniques we identified in Section 8 and comparative experiments in Section 9.

8 RQ_4: LIMITATIONS OF IDENTIFIED TECHNIQUES

We examined the identified papers to analyze the applied techniques and underlying concepts in the previous sections. In this section, we discuss *limitations* and the required *manual effort* of the identified techniques.

8.1 Results

Although the proposed techniques proved to be useful for supporting the systematic literature review process, there are some limitations in their usage or their evaluations. For example, some shortcomings of the evaluation procedure are highlighted as threats to validity by Malheiros et al.

[24] (1) and Felizardo et al. [13] (2). These mainly include the use of small samples and limited numbers of participants. Considering the experiment described by Malheiros et al. [24], the authors specify that the paper evaluation is subjective and the influence of researchers' judgment on the analysis. We further have to remark that the experiment involved one participant who manually reviewed papers based only on abstracts. However, due to the absence of a unique ontology between abstracts and full text elaborations, some abstracts may poorly reflect a paper's actual content. Thus, such interpretations can be misleading and may show worse or better results. Additionally, users who use visual text mining tools require training and have to prepare the input data. Subsequently, this results in manual effort and can also bias the results that have been obtained.

Tomassetti et al. [34] (3) also report several threats to their technique and its evaluation. Again, one threat is the subjective bias of users composing the initial set of papers. As it represents just a portion of the entire literature regarding a field of interest, automated classifications may discard all resources that are not part of the described niche. Proceeding to the subsequent steps of their technique, there is a construct threat in the linked data enrichment: The technique may not be applicable to a paper, because some relevant terms are not selected. Furthermore, an imprecise classification condition can bias the derived conclusions. Finally, the evaluation has been performed only on some papers of the systematic literature review that is used as baseline, weakening the derived conclusions.

The extended technique proposed by Rizzo et al. [29] (5) reduces subjectivity in the overall process, due to automatically building the used model based on the seed papers. The proposed technique also allows searches in a larger search space, and thus can better capture similar papers and those with conceptual relations to the relevant papers. Still, the semi-supervised, iterative paper classification process is not fully automatic. The users need to manually review the inclusion and exclusion of potentially relevant papers that the classifier selected. To evaluate their technique, the authors only rely on positive examples during the classification, arguing that this yields higher probability that a paper is actually relevant. We argue that further research necessary to investigate whether this is true.

Abilio et al. [1] (4) report that variations in the used search engines [32], for example, IEEE Xplore, Scopus, and ACM Digital Library, limit the applicability of techniques. This is arguably a threat to all techniques and not limited to this one, independent of whether these search by themselves or are provided a set of papers. The authors emphasize the importance of selecting the right search terms. Besides the subjectivity of users applying a technique, there may also be subjectivity in the papers that can negatively impact the results. In particular, synonyms can pose a serious threat, potentially resulting in some relevant papers being not selected or poorly assessed because the defined terms are not used.

Concerning SCAS (6), Octaviano et al. [27] report several threats. As we described in Figure 4, SCAS categorizes papers into quadrants based on their citation links. However, it is unclear how the metrics may negatively influence the selection of recent papers with few cross-citations. Consequently, reviewers have to carefully analyze papers belonging to quadrant four and cannot automatically exclude them based on the recommendation. Additionally, the absence of visualization support for the citation feature is currently a limitation of the tool.

StArt (7) and PaperQuest (8) are still in their development stage. Both lack a formal evaluation using real systematic literature reviews. The developers aim to improve user interactions by using further visualization techniques to represent papers and metrics. Thus, while both tools are promising and seem to aim at addressing some issues we mentioned in previous sections, they are still in a preliminary stage.

8.2 Summary

The evaluations performed for the identified techniques show that most of them facilitate the selection and assessment of systematic literature reviews. However, most evaluations face significant threats, challenging a definitive assessment. Mostly, these threats are concerned with small paper or subject samples. Thus, the obtained findings may not withstand if applied in the real world. The techniques themselves also face restrictions that can lead to poor results. For example, visual text mining techniques have reasonable precision, but require training and careful selection of the initial paper seeds. For information retrieval techniques, the users have to be careful with selecting appropriate search terms.

Considering the reported threats, we see the need to empirically evaluate the proposed techniques. Experiments and case studies can help in this regard. We further argue that it is necessary to conduct surveys and user studies within the software engineering community to understand their needs and assessment criteria. Thus, techniques and their used metrics can be scoped properly to the actual needs and yield better results. Currently, it also seems not possible to reliably use any of the identified techniques. We see a strong need to extend the available tooling and integrate it into an overall work-flow.

8.3 Concluding Remarks for RQ_4

Through our analysis, we determined the current state of tool support, highlighting the fact that all of them partially address the selection and assessment of papers for the systematic literature review process. There are several shortcomings of the techniques and their evaluations. To summarize our findings, we recommend researchers to:

- Use a systematic literature review with a large number of primary studies as baseline.
- Include multiple participants (with domain knowledge) in evaluations who perform comparable systematic literature reviews with different techniques.
- Report what has been measured to what extent (e.g., the time from getting the seed papers to the final selection) and include at least correctness and time.
- Clearly state technical limitations of the proposed technique.
- Discuss the applicability of the technique with experts and experienced reviewers.
- Investigate the manual effort the technique requires.

Moreover, while the manual workload is reduced through these techniques, users still have to manually assess the identified papers. Overall, we see considerable advancements and efforts to improve this part of a systematic literature review, but it still requires further evaluations, scoping, and usable tools.

9 EXTENDED EXPERIMENTAL STUDIES

During the snowballing phase of our systematic literature review, we found two experimental papers that further evaluated two of the identified techniques. In this section, we briefly discuss these two papers.

9.1 Replication Study of Felizardo et al. [13] (2)

Felizardo et al. [14] report a replication study that they performed to evaluate their technique. In their original paper, the authors investigated whether the proposed technique improves the productivity of four PhD students. Within this replication, they involved six PhD and 15 graduate students that were randomly sampled into two groups: Group 1 comprising seven subjects who manually read abstracts and Group 2 comprising eight subjects who were trained on and used the visual text mining tool. The authors ensured that the differences in prior experience with systematic

literature reviews in each group were not significant. No other factor besides the number of participants was changed compared to the previous evaluation (cf. Section 5.1). Thus, the baseline were 37 papers, using the same selection criteria and measuring the required time.
Results. The average time recorded were 70.14 (spanning from 60 to 95) and 54.5 (spanning from 38 to 66) minutes for Group 1 and Group 2, respectively. Both groups had similar standard deviations of 11.56 and 11.60 minutes. The participants of Group 1 correctly selected 21.7 papers, while those of Group 2 correctly selected 24.5 papers on average. Considering these correctly selected papers, the standard deviations were 3.54 and 2.32, respectively.
Conclusion. The results substantiate that the use of the developed visual text mining tool is promising in terms of performance. It accelerates the process of exploring studies by reducing the effort and time spent for the selection activity. However, in terms of selection effectiveness, the results suggest that the decision of PhD students were more consistent compared to the graduate students. This led to the finding that the level of experience in research impacts the primary study selection for a systematic literature review. Still, the PhD students also performed more effectively using the visual text mining tool compared to the manual reading process. The time and slight accuracy improvements indicate that the technique of Felizardo et al. [13] is suitable to support systematic literature reviews. Nonetheless, more extensive evaluations and advanced techniques are necessary. For example, we are still concerned about the small baseline of papers that is used and the opinions of researchers who regularly conduct systematic literature reviews.

9.2 Experimental Study of Octaviano et al. [27] (6)

Octaviano et al. [26] extended their previous evaluation of SCAS with an experimental study. Consequently, the authors' main objective was to determine whether SCAS is more efficient than manual paper selection and assessment. This experiment also investigated the effectiveness of recommendations on resolving decision conflicts. The experiment was based on the Goal-Question-Metric (GQM) model [4] and comprised 21 PhD students, 12 from computing, five from production engineering, and four from education. All participants were divided into five groups according to their research areas, with computer science being divided into three subgroups. The participants had to do both, use SCAS and perform a manual selection of papers, and had to compare their decisions with the SCAS recommendations.
Results. The fully manual selections and assessments required 95 minutes on average. In contrast, SCAS reduced the overall time significantly and needed only 4 minutes to process the considered papers. As evaluation metrics, the authors use *effort reduction*—referring to the number of papers that were not completely read compared to the manual process—and the ratio of papers that SCAS classified incorrectly. All groups faced effort reductions between 13.00% and 27.34% with error rates ranging from 1.95% to 6.35%. To evaluate the support of SCAS for decision conflicts, the authors analyzed the ratio of papers for which the participants and SCAS agreed. Overall, the authors found that SCAS was correct in 58.98% of the cases. Finally, the overall results obtained by Octaviano et al. [26] showed a precision of 65.49% and a recall of 90.24%.
Conclusion. The findings suggest that SCAS improves the speed of selecting and assessing papers, while resulting in only small loss of evidence (false-negative decisions). However, SCAS is not significantly useful to resolve conflicting decisions. Thus, the authors verified their previous evaluation and substantiated the usefulness of the SCAS techniques. Still, being unbiased and ideally complete are the expected characteristics of a systematic literature review and a loss of evidence can be a substantial threat. Moreover, a comparative evaluation of different techniques and critical investigations with practitioners are missing, again.

10 DISCUSSION

With the research questions of our systematic literature review, we aimed to provide a comprehensive overview of existing techniques supporting the selection and quality assessment of papers. Although the reported evaluations show that the identified techniques can be useful, the evaluations also seem to be rather incomplete. Thus, further assessments are necessary to confirm the usability of these techniques. To provide more insights into our research questions, we intended to compare the identified techniques ourselves, based on hands-on experiences as well as metric-based analyses. Unfortunately, when we tried to find and set up the implemented techniques, we realized that most of them are currently not available anymore. We contacted the developers and authors, asking for help, but got few responses. Moreover, in most cases it was still not possible to set up and use the tools. To the best of our knowledge, only two tools are available: PaperQuest (8) and StArt (7).

10.1 PaperQuest (8)

Although PaperQuest is available, it is limited in its usability, as we would have to rely on the predefined datasets that contain only some conferences. Namely, the available papers include the Conference on Human Factors in Computing Systems (CHI) and User Interface Software and Technology Symposium (UIST) from 1982 until 2010 as well as the Visualization (IEEE VIS) Publication Dataset[1] from 1995 until 2014. The developers claim that a positive feedback was received from researchers using the tool, but the specific details about the experiment performed are missing. This implies that the tool is specific for certain topics of research and is limited in terms of their content. Due to the domain limitations, and our missing expertise in this domain, we could not evaluate this tool to a more detailed extent.

10.2 StArt (7)

StArt is the only tool that is available, supports the entire systematic literature review process, and that we were able to use and analyze. To this end, we conducted the following study: We used two different systematic literature reviews that were performed manually by other researchers, namely by Mahdavi-Hezavehi et al. [23] (SLR_1) and Britto et al. [6] (SLR_2). Our goal was to replicate these by using StArt and compare the results of the tool with the original systematic literature reviews.

At the beginning, StArt asks the user to input the review protocol, for which we used the steps reported in each systematic literature review. For the data sources, StArt supports a number of digital libraries, but also lacks support for some others, such as, CiteSeerX. However, it provides the option to select libraries that are not supported and the results can be manually added to the tool. Regarding the data formats, StArt supports BibTex, Medline, RIS and Cochrane, but lacks support for other formats, such as, comma-separated values (CSV) files, that is specifically important for exporting results from various databases, such as Springer Link [32]. For our experiment, we used Zotero[2] to convert CSV files to BibTex format and imported those into StArt. Once the results from the libraries are inserted into the tool, StArt scores each paper based on a keyword analysis—automatically assigning a reading priority. In the next step, the user has to classify the papers as accepted or rejected based on the study selection criteria followed by completing the quality forms. To compare the results, we selected the option to assign all papers automatically to a specific quadrant (SCAS-based recommendations, cf. Figure 4) and to classify them accordingly.

We noticed that most of the papers originally included as primary studies in the systematic literature reviews were either rejected by StArt or had low scores, meaning that the papers were not considered as relevant. For SLR_1, out of 49 primary studies selected by the authors, 34 were

[1] https://sites.google.com/site/vispubdata/home
[2] https://www.zotero.org/

rejected by StArt and two remained unclassified. We remark that, although we followed the steps defined in the study, our database search did not retrieve 10 of the original primary studies, which could be due to technical problems of digital libraries that we cannot overcome [32]. Similarly, SLR_2 included five primary studies out of which StArt rejected two automatically and assigned the remaining three as unclassified with relatively low scores.

Thus, due to this experience with StArt, we have to highlight a low precision of the automatic classification and relevancy rating of papers when compared to the original systematic literature reviews. Although the developers claim that their users provided positive feedback for the tool, details of the experiments are missing. Consequently, more experiments with real systematic literature reviews must be performed in the future to verify the reliability of results. However, we still believe that StArt is a useful tool provided that users perform the classification and quality assessment of studies mainly manually to improve the precision of results.

10.3 Summary and Outcomes

Considering all results we presented in this article, we argue that promising techniques have been developed to provide assistance with the selection and assessment of scientific papers. Unfortunately, we found that almost none of the identified techniques is available and that a complete evaluation of an existing systematic literature review is barely possible. However, the most crucial step that still needs to be addressed is the quality assessment of papers. Even with StArt, the user initially provides the quality criteria in the protocol, but has to manually fill in the quality forms for the classification of papers. We consider this a major lack in current research, as the quality of papers is still subject to manual analysis to ensure reliable results. Information and data quality is an important concern for software engineering researchers during a systematic literature review to validate the findings. It is equally important in other scenarios and domains, such as the evaluation of books for teaching or reports in an organization's information system. We believe that further research must be performed to overcome current limitations—with particular importance on the availability of the proposed techniques.

A second limitation in current research is the missing comparative analysis of the underlying concepts and the techniques themselves. As we illustrated in Figure 5, visual text mining is mostly adopted for the identified techniques and has been extended by researchers from time to time. To provide concrete evidence on what underlying concept could be preferable for developing such techniques, a detailed comparison must be performed. Due to the current unavailability, we could neither evaluate and compare the techniques nor their underlying concepts. This must be considered as a serious limitation that hampers the implementation and application of such techniques. We urge developers to consider the possibility of making their tools publicly available in the future. To summarize, we see our study as a starting point for further research for tools that support literature analysis, in particular, the selection and quality assessment of primary studies in a systematic literature review.

11 THREATS TO VALIDITY

Construct Validity. A threat to the construct validity of our systematic literature review may occur due to the formulation of our research questions. They mainly focus on papers addressing techniques to support the conduction phase of a systematic literature review. In order to reduce the single reviewer bias, results obtained through the search string were reviewed individually by the first three authors. Nonetheless, we realize that our study can only present the current state of research in automating a certain phase of the systematic literature review process. Thus, our results are valid to a limited extent, especially as the properties of search engines can change rapidly. Due to such factors, other researchers may derive different decisions during their analysis.

Internal Validity. Generally, results of a systematic literature review can be biased, due to the selection of included papers and the data extraction strategy. To minimize the subjectivity involved in our study, we strictly adhered to the search and selection process described in Section 3. We emphasize that the systematic literature review and the results presented in this work are limited to the domain of software engineering. Our investigation is restricted to literature available in the selected digital libraries that has been published during a specific period of time. These factors may cause bias and threaten the internal validity of this systematic literature review, which we cannot completely eliminate. However, the focus on the software engineering domain was intended and following the described protocol should ensure that we identified most relevant papers.
Conclusion Validity. Concerning the conclusions we derived, we believe that our findings provide important insights for the software engineering community to plan future activities on the quality assessment of not only scientific papers, but various sorts of documents. We carefully analyzed each paper to draw meaningful and valid conclusions. Moreover, we documented the individual steps we conducted and report how we derived our conclusions. Thus, other researchers can understand and repeat our procedure to build on and extend our findings.

12 CONCLUSION

In this article, we reported a systematic literature review to summarize and discuss the state-of-the-art in automating systematic literature reviews. The results reported in this study focus on semi-automated techniques that support the selection and quality assessment of primary studies. Overall, we identified eight techniques and our results show that:

- The fundamental concepts of most techniques are based on text mining, showing a missing usage of other advanced concepts and measures.
- A complete validation of the proposed techniques using real systematic literature reviews is missing, threatening their practical applicability.
- Comparisons and integration of different techniques have not been performed, wherefore it is unclear which techniques may perform better or can overcome limitations of others.
- Current techniques face several limitations and, while potentially facilitating the conduct of systematic literature reviews, may cause different biases that have to be analyzed.
- Most techniques focus on the study selection phase and solely scratch the actual quality assessment, which is often still a purely manual task.
- Unfortunately, most techniques were not available during this study, which contradicts all efforts taken to implement them.

Some of the identified limitations are considerably challenging the proposed techniques. Moreover, our observations indicate a deficiency of tool support for the quality assessment of primary studies. Based on the results, we contributed and discussed research opportunities and challenges to support researchers to scope future work.

As study selection and quality assessment are important steps in the systematic literature review process, the identified deficiencies of existing techniques must be addressed by researchers. Furthermore, research in other forms of information visualization may be useful, as most of the techniques present the results only as some sort of list. Our further work will include the development of more advanced techniques to automate the quality assessment, building on the results presented in this article. To this end, it is also interesting to compare our findings with techniques that are used in other domains, for example, to ensure the quality of data in information systems and databases.

ACKNOWLEDGMENTS

This research has been supported the German Research Foundation (DFG) grants LE 3382/2-1, LE 3382/2-3, SA 465/9-1, SA 465/49-3, and the German Academic Exchange Service (DAAD) STIBET Matching Funds grant.

REFERENCES

[1] Ramon Abilio, Flávio Morais, Gustavo Vale, Claudiane Oliveira, Denilson Pereira, and Heitor Costa. 2015. Applying Information Retrieval Techniques to Detect Duplicates and to Rank References in the Preliminary Phases of Systematic Literature Reviews. *CLEI Electronic Journal* 18, 2 (2015), 3–27.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*. Springer, 722–735. https://doi.org/10.1007/978-3-540-76298-0_52

[3] Muhammad A. Babar and He Zhang. 2009. Systematic Literature Reviews in Software Engineering: Preliminary Results from Interviews with Researchers. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 346–355. https://doi.org/10.1109/ESEM.2009.5314235

[4] Victor R. Basili, Gianluigi Caldiera, and Dieter H. Rombach. 1994. The Goal Question Metrics Approach. *Encyclopedia of Software Engineering* 1 (1994), 528–532.

[5] Andrew Booth, Anthea Sutton, and Diana Papaioannou. 2016. *Systematic Approaches to a Successful Literature Review*. SAGE.

[6] Ricardo Britto, Vitor Freitas, Emilia Mendes, and Muhammad Usman. 2014. Effort Estimation in Global Software Development: A systematic Literature Review. In *International Conference on Global Software Engineering (ICGSE)*. IEEE, 18–21. https://doi.org/10.1109/ICGSE.2014.11

[7] Rick Cattell. 2011. Scalable SQL and NoSQL Data Stores. *SIGMOD Record* 39, 4 (2011), 12–27. https://doi.org/10.1145/1978915.1978919

[8] Fabio Q.B. da Silva, André L.M. Santos, Sérgio Soares, A. César C. França, Cleviton V.F. Monteiro, and Felipe Farias Maciel. 2011. Six Years of Systematic Literature Reviews in Software Engineering: An Updated Tertiary Study. *Information and Software Technology* 53, 9 (2011), 899–913. https://doi.org/10.1016/j.infsof.2011.04.004

[9] Gabriel C. Durand, Anusha Janardhana, Marcus Pinnecke, Yusra Shakeel, Jacob Krüger, Thomas Leich, and Gunter Saake. 2018. Exploring Large Scholarly Networks with Hermes. In *Extending Database Technology (EDBT)*. ACM. https://doi.org/10.5441/002/edbt.2018.76

[10] Tore Dybå and Torgeir Dingsøyr. 2008. Empirical Studies of Agile Software Development: A Systematic Review. *Information and Software Technology* 50, 9 (2008), 833–859. https://doi.org/10.1016/j.infsof.2008.01.006

[11] Tore Dybå and Torgeir Dingsøyr. 2008. Strength of Evidence in Systematic Reviews in Software Engineering. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 178–187. https://doi.org/10.1145/1414004.1414034

[12] Sandra C. P. F. Fabbri, Cleiton Silva, Elis Hernandes, Fábio R. Octaviano, André Di Thommazo, and Anderson Belgamo. 2016. Improvements in the StArt Tool to Better Support the Systematic Review Process. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 21:1–21:5. https://doi.org/10.1145/2915970.2916013

[13] Katia R. Felizardo, Norsaremah Salleh, Rafael M. Martins, Emilia Mendes, Stephen G. MacDonell, and Jose C. Maldonado. 2011. Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 77–86. https://doi.org/10.1109/ESEM.2011.16

[14] Katia R. Felizardo, Simone do R. S. de Souza, and José C. Maldonado. 2013. The Use of Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews: A Replication Study. In *International Workshop on Replication in Empirical Software Engineering Research (RESER)*. 91–100. https://doi.org/10.1109/RESER.2013.9

[15] Edgar Hassler, Jeffrey C. Carver, David Hale, and Ahmed Al-Zubidy. 2016. Identification of SLR Tool Needs – Results of a Community Workshop. *Information and Software Technology* 70 (2016), 122–129. https://doi.org/10.1016/j.infsof.2015.10.011

[16] Salma Imtiaz, Muneera Bano, Naveed Ikram, and Mahmood Niazi. 2013. A Tertiary Study: Experiences of Conducting Systematic Literature Reviews in Software Engineering. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 177–182. https://doi.org/10.1145/2460999.2461025

[17] Yang Kiduk and Meho Lokman I. 2007. Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology* 43, 1 (2007), 1–15. https://doi.org/10.1002/meet.14504301185

[18] Barbara A. Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Technical Report TR/SE-0401. Keele University.

[19] Barbara A. Kitchenham, Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic Literature Reviews in Software Engineering – A Systematic Literature Review. *Information and Software Technology* 51, 1 (2009), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

[20] Barbara A. Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press.

[21] Barbara A. Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report EBSE-2007-01. Keele University and University of Durham.

[22] Barbara A. Kitchenham, Tore Dyba, and Magne Jorgensen. 2004. Evidence-Based Software Engineering. In *International Conference on Software Engineering (ICSE)*. IEEE, 273–281. https://doi.org/10.1109/ICSE.2004.1317449

[23] Sara Mahdavi-Hezavehi, Matthias Galster, and Paris Avgeriou. 2013. Variability in quality attributes of service-based software systems: A systematic literature review. *Information and Software Technology* 55, 2 (2013), 320–343. https://doi.org/10.1016/j.infsof.2012.08.010

[24] Viviane Malheiros, Erika Höhnr, Roberto Pinho, Manoel Mendonca, and José C. Maldonado. 2007. A Visual Text Mining Approach for Systematic Reviews. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE, 245–254. https://doi.org/10.1109/ESEM.2007.21

[25] Christopher Marshall and Pearl Brereton. 2013. Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*. 296–299. https://doi.org/10.1109/ESEM.2013.32

[26] Fábio Octaviano, Cleiton Silva, and Sandra Fabbri. 2016. Using the SCAS Strategy to Perform the Initial Selection of Studies in Systematic Reviews: An Experimental Study. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 25:1–25:10. https://doi.org/10.1145/2915970.2916000

[27] Fábio R. Octaviano, Katia R. Felizardo, José C. Maldonado, and Sandra C. P. F. Fabbri. 2015. Semi-Automatic Selection of Primary Studies in Systematic Literature Reviews: Is it Reasonable? *Empirical Software Engineering* 20, 6 (2015), 1898–1917. https://doi.org/10.1007/s10664-014-9342-8

[28] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. 2016. PaperQuest: A Visualization Tool to Support Literature Review. In *International Conference on Human Factors in Computing Systems (CHI)*. ACM, 2264–2271. https://doi.org/10.1145/2851581.2892334

[29] Giuseppe Rizzo, Federico Tomassetti, Antonio Vetró, Luca Ardito, Marco Torchiano, Maurizio Morisio, and Rapha'el Troncy. 2015. Semantic Enrichment for Recommendation of Primary Studies in a Systematic Literature Review. *Digital Scholarship in the Humanities* (2015), 1–14. https://doi.org/10.1093/llc/fqv031

[30] Jefferson S. Molléri and Fabiane B. V. Benitti. 2012. Automated Approaches to Support Secondary Study Processes: A Systematic Review. In *International Conference on Software Engineering & Knowledge Engineering (SEKE)*.

[31] Ivonne Schröter, Jacob Krüger, Philipp Ludwig, Marcus Thiel, Andreas Nürnberger, and Thomas Leich. 2017. Identifying Innovative Documents: Quo Vadis?. In *International Conference on Enterprise Information Systems (ICEIS)*. ScitePress, 653–658. https://doi.org/10.5220/0006368706530658

[32] Yusra Shakeel, Jacob Krüger, Ivonne von Nostitz-Wallwitz, Christian Lausberger, Gabriel C. Durand, Gunter Saake, and Thomas Leich. 2018. (Automated) Literature Analysis - Threats and Experiences. In *International Workshop on Software Engineering for Science (SE4Science)*. ACM, 20–27. https://doi.org/10.1145/3194747.3194748

[33] Yusra Shakeel, Jacob Krüger, Gunter Saake, and Thomas Leich. 2018. Indicating Studies' Quality based on Open Data in Digital Libraries. In *International Conference on Business Information Systems (BIS)*. Springer, 579–590. https://doi.org/10.1007/978-3-030-04849-5_50

[34] Federico Tomassetti, Giuseppe Rizzo, Antonio Vetro, Luca Ardito, Marco Torchiano, and Maurizio Morisio. 2011. Linked Data Approach for Selection Process Automation in Systematic Reviews. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. IET, 31–35. https://doi.org/10.1049/ic.2011.0004

[35] Sonali V. Gaikwad, Archana Chaugule, and Pramod Patil. 2014. Text Mining Methods and Techniques. *Journal of Computer Applications* 85, 17 (2014), 42–45.

[36] June M. Verner, Pearl Brereton, Barbara A. Kitchenham, Mark Turner, and Mahmood Niazi. 2012. Systematic Literature Reviews in Global Software Development: A Tertiary Study. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. IET, 2–11. https://doi.org/10.1049/ic.2012.0001

[37] Jane Webster and Richard T. Watson. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly* 26, 2 (2002), xiii–xxiii.

[38] Claes Wohlin. 2014. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 1–10. https://doi.org/10.1145/2601248.2601268

[39] Hong Xie. 2006. Evaluation of Digital Libraries: Criteria and Problems from Users' Perspectives. *Library and Information Science Research* 28, 3 (2006), 433–452. https://doi.org/10.1016/j.lisr.2006.06.002

[40] You Zhou, He Zhang, Xin Huang, Song Yang, Muhammad A. Babar, and Hao Tang. 2015. Quality Assessment of Systematic Reviews in Software Engineering: A Tertiary Study. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 14:1–14:14. https://doi.org/10.1145/2745802.2745815