ÉRICA MOURÃO DA SILVA

EVALUATING THE EFFICIENCY OF HYBRID SEARCH STRATEGIES FOR SYSTEMATIC LITERATURE REVIEWS IN SOFTWARE ENGINEERING

Thesis presented to the Computing Graduate Program of the Fluminense Federal University in partial fulfillment of the requirements for the degree of Master of Science. Topic area: Systems and Information Engineering.

Advisor: Prof. D.Sc. Leonardo Gresta Paulino Murta

Niterói

2019

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

S586e Silva, Érica Mourão da
Evaluating the Efficiency of Hybrid Search Strategies for Systematic Literature Reviews in Software Engineering / Érica Mourão da Silva ; Leonardo Gresta Paulino Murta, orientador. Niterói, 2019.
74 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2019.

DOI: http://dx.doi.org/10.22409/PGC.2019.m.08102087765

1. Systematic Literature Review. 2. Hybrid Search Strategy. 3. Database Search. 4. Snowballing. 5. Produção intelectual. I. Murta, Leonardo Gresta Paulino, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

ÉRICA MOURÃO DA SILVA

EVALUATING THE EFFICIENCY OF HYBRID SEARCH STRATEGIES FOR SYSTEMATIC LITERATURE REVIEWS IN SOFTWARE ENGINEERING

Thesis presented to the Computing Graduate Program of the Fluminense Federal University in partial fulfillment of the requirements for the degree of Master of Science. Topic area: Systems and Information Engineering.

Approved on July 2019.

APPROVED BY

Prof. D.Sc. Leonardo Gresta Paulino Murta – Advisor

UFF

Prof. D.Sc. Marcos Kalinowski

PUC-Rio

D.Sc. Troy Costa Kohwalter

UFF

Prof. D.Sc. Daniel Cardoso Moraes de Oliveira

UFF

Prof. D.Sc. Rodrigo Pereira dos Santos

UNIRIO

Niterói

2019

This work I dedicate first to God who is essential in my life,
to my daughter Isabella and my Mourão family,
who has always believed in me.

ACKNOWLEDGMENTS

I would like to thank my Mourão family for supporting my decisions.

I am grateful for Marcos Kalinowski for providing me the topic of research, for all the help and for beliving in me.

This thesis would not have been possible without the patience, wisdom, and help of my advisor Leonardo Murta.

I could not thank Joao Felipe Pimentel enough for helping me in the academic laboratory of Systems and Information Engineering at IC-UFF.

I would also like to thank all members of the committee who agreed to participate in this thesis examination.

I am grateful for all graduate professors, students, and employees in the Computer Science department for promoting inovation and a welcoming academic environment.

Finally, I would like to acknowledge the financial, academic, and technical support of the Fluminense Federal University and CAPES.

“If I have seen further it is by standing on the shoulders of giants.” (Isaac Newton, 1675).

RESUMO

Ao realizar uma Revisão Sistemática de Literatura (RSL), os pesquisadores geralmente enfrentam o desafio de projetar um protocolo que equilibre adequadamente a qualidade dos resultados e os esforços da revisão. Por um lado, usar buscas em bibliotecas digitais (ou bancos de dados) ou *snowballing* sozinhas não seria suficiente para alcançar resultados de alta qualidade. Por outro lado, o uso conjunto de buscas em bibliotecas digitais e *snowballing* pode aumentar o esforço geral da revisão. O objetivo desta pesquisa é propor e avaliar estratégias de busca híbridas que combinam seletivamente buscas em bibliotecas digitais com *snowballing*. Propusemos quatro estratégias de busca híbridas combinando buscas em bibliotecas digitais com *snowballing* iterativo, paralelo, sequencial *backward* e *forward*, ou sequencial *forward* e *backward*. Simulamos as estratégias ao longo de três RSL existentes na Engenharia de Software que adotaram busca em bibliotecas digitais seguidas de *snowballing*. Nós analisamos sete estratégias de busca: quarto estratégias híbridas, busca em bibliotecas digitais, *snowballing* e busca em bibliotecas digitais seguidas de *snowballing*. Nós comparamos o resultado de busca em bibliotecas digitais, *snowballing*, busca em bibliotecas digitais seguidas de *snowballing* e as estratégias híbridas por meio de *precision, recall* e *F-measure* para investigar a eficiência de cada estratégia. Nossas investigações permitiram observar que, para as RSL analisadas, a combinação de buscas na biblioteca digital Scopus com *snowballing* paralelo ou sequencial alcançou o equilíbrio mais adequado entre *precision* e *recall*. Dependendo das metas da RSL e dos recursos disponíveis, usar uma estratégia híbrida, envolvendo busca em uma biblioteca digital representativa, seguida de *snowballing* paralelo ou sequencial, pode ser uma alternativa apropriada para balancear qualidade e esforço em RSL.

Palavras-chave: revisão sistemática da literatura, busca em biblioteca digital, busca em banco de dados, *snowballing*, engenharia de software, estratégia de busca híbrida.

ABSTRACT

When performing a Systematic Literature Review (SLR), researchers usually face the challenge of designing a protocol that appropriately balances result quality and review effort. On the one hand, using digital library (or database) searches or snowballing alone would not be enough to achieve high-quality results. On the other hand, using both digital library searches and snowballing together may increase the overall review effort. The goal of this research is to propose and evaluate hybrid search strategies that selectively combine digital library searches with snowballing. We proposed four hybrid search strategies combining searches in digital libraries with iterative, parallel, sequential backward and forward, or sequential forward and backward snowballing. We simulated the strategies over three existing SLRs in Software Engineering that adopted both digital library searches and snowballing. We analyzed seven search strategies: four hybrid strategies, database search, snowballing, and database searches in several digital libraries followed by snowballing. We compared the outcome of digital library searches, snowballing, database search followed by snowballing, and the hybrid strategies combining both by means of precision, recall, and F-measure to investigate the efficiency of each strategy. Our investigations allowed observing that, for the analyzed SLRs, combining searches in the Scopus digital library with parallel or sequential snowballing achieved the most appropriate balance of precision and recall. We put forward that, depending on the goals of the SLR and the available resources, using a hybrid search strategy involving a representative digital library and parallel or sequential snowballing tends to represent an appropriate alternative to balance quality and effort when searching for evidence in SLRs.

Keywords: systematic literature review, digital library search, database search, snowballing, software engineering, hybrid search strategy.

LIST OF FIGURES

LIST OF TABLE

LIST OF ACRONYMS AND ABBREVIATIONS

BS – Backward Snowballing

EBSE – Evidence-based Software Engineering

DL – Digital Library

FS – Forward Snowballing

QA – Quality Assurance

RQ – Research Question

SE – Software Engineering

SLR – Systematic Literature Review

SM – Systematic Mapping

TR – Tertiary Review

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION

1.1 MOTIVATION

Systematic Literature Reviews (SLR) aim at identifying, evaluating, and interpreting relevant research in a specific topic area (KITCHENHAM; CHARTERS, 2007). Kitchenham and Charters (2007) and Wohlin (2014) provide guidelines for searching for evidence when conducting an SLR in the Software Engineering (SE) domain.

The guideline provided by Kitchenham and Charters (2007) recommends composing a string to search on several digital libraries to find relevant studies. This SLR search strategy is known as database search. On the other hand, Wohlin (2014) recommends recursively identifying papers based on the reference list or the citations of papers that were found by an informal database search. This alternative SLR search strategy is known as Backward Snowballing (BS) (when searching within the reference lists) and Forward Snowballing (FS) (when searching within the citing papers).

The database search strategy is the most common and the first published recommendation for SLRs in the SE domain (KITCHENHAM; CHARTERS, 2007). However, it imposes several challenges to researchers. Some of them including the selection of appropriate digital libraries and the design of a specific search string for conducting searches within those libraries (WOHLIN, 2014). Furthermore, this strategy needs customization of the search string to allow the use in different digital libraries. Reported difficulties regarding digital libraries include the diversity of user interfaces, the limitation of operators, and not handling synonyms of terms (WOHLIN, 2014). Moreover, usually this strategy presents overlap of papers in different digital libraries, difficulties concerning the concatenations of keywords, and search execution inconsistencies within specific digital libraries (SINGH; SINGH, 2017).

Snowballing emerged as an attractive alternative to database search. Snowballing does not require searching in more than one digital library; the approach is more understandable and easy to follow (JALALI; WOHLIN, 2012). More especifically, snowballing is expected to be more efficient when relevant database search keywords include general terms, by reducing the amount of noise. However, it also has drawbacks, such as dependency on an appropriate seed (start) set of relevant papers (WOHLIN, 2014). Badamputi *et al.* (2015) discuss the organization of papers in the seed set into different categories, so that the seed set should have at least one paper in each category. Furthermore, Jalali and Wohlin (2012) discuss difficulties of judgments based on the title of the paper, when going backward and forward, which might result in missing papers with no relevant keyword in the title. Additionally, they report that a threat in snowballing is find several papers from the same authors. Hence, the results might be biased by over representing research from specific authors (JALALI; WOHLIN, 2012).

Another alternative consists in combining both database searches and snowballing, in a complete way (VASCONCELLOS *et al.*, 2017). E.g., the database search step retrieved 517 studies. After the duplicate elimination step, 495 studies belong to the initial dataset. Out of 495 studies, 22 were selected and used as a start set to detect other 29 relevant papers through snowballing. Finally, the 51 primary studies compose the selected papers. This alternative improves the quality of the results at the price of adding more effort to the review process. Moreover, it also retains some of the drawbacks of both database searches and snowballing.

Some previous works have faced the problem of balancing quality and effort in SLR. The studies reported by Badampudi *et al.* (2015) and Jalali and Wohlin (2012) compared the efficiency of the database and snowballing search strategies. The study reported by Badampudi *et al.* (2015) concludes that the efficiency of database searches and snowballing is comparable. Jalali and Wohlin (2012) argue that snowballing does not require searching in more than one database; and that database searches require more effort to refine the searches in order to identify relevant papers and discard irrelevant ones. Wohlin (2014) puts forward that different approaches to identifying relevant literature should be used to ensure the best possible coverage.

However, the literature investigating search strategies is scarce and there is a need to further evaluate the efficiency of the different strategies.

1.2 GOAL

The goal of this dissertation is to define and evaluate hybrid search strategies that combine specific aspects of databases searches and BS and FS. The defined hybrid search strategies regard complementing a single database search on an efficient digital library (Scopus) with four different snowballing variations. These variations are iterative BS and FS, parallel BS and FS, sequential BS and FS, and sequential FS and BS.

We evaluated the hybrid strategies by simulating their execution over three previously conducted SLRs identified to compose our corpus (SILVA, A. *et al.*, 2017; TARHAN; GIRAY, 2017; VASCONCELLOS *et al.*, 2017), which employed database search in several digital libraries followed by an iterative BS and FS. This simulation allowed us to assess the precision, recall, and F-measure of the hybrid strategies, should they were used in the original SLR. The results of those four hybrid strategies were compared against popular conventional SLR strategies: pure database search, pure snowballing, and completely combining database searches in several digital libraries followed by iterative BS and FS.

We used our simulation results to provide answers to the following research questions:

RQ1 - What is the efficiency of the database search in the published SLRs?

RQ2 - What is the efficiency of the snowballing in the published SLRs?

RQ3 - What is the efficiency of each hybrid search strategy in the published SLRs?

The results of our study show that hybrid search strategies are more efficient and may be an appropriate alternative compared with both database search and snowballing.

1.3 ORGANIZATION

This work is organized in four chapters, beside this introduction.

Chapter 2 describes the concepts and stages involved in conducting SLR. It discusses the search strategies for SLR and presents the works related to this dissertation. Chapter 3 details our research questions and introduces the hybrid search strategies. Furthermore, the chapter describes the corpus and the research method. Lastly, it outlines the supporting tool to simulate the search strategies. Chapter 4 presents the results for each research questions, together with some discussions. Finally, Chapter 5 concludes this work, listing contributions, threats to the validity, and future works.

CHAPTER 2 – SEARCH STRATEGIES IN SYSTEMATIC LITERATURE REVIEWS

2.1 INTRODUCTION

Kitchenham *et al.* (2004) proposed a framework for Evidence-based Software Engineering (EBSE) derived from medical standards. This framework compiles the best available evidence to address engineering research questions proposed by software engineers and empirical software engineering researchers. They recommended a Systematic Literature Review (SLR) as a methodology for aggregating all empirical studies in a particular topic. After that, Kitchenham customized the medical guidelines for SLRs to researchers and proposed a guideline for performing SLR in SE (KITCHENHAM; CHARTERS, 2007).

In this chapter, we present the concept of an SLR, its phases, activities and resources to execute the process. Therefore, the search strategies and the related works are presented.

2.2 SYSTEMATIC LITERATURE REVIEWS

An SLR is the main mean to summarize research evidence (KITCHENHAM *et al.*, 2009). In the last decade, the SLR has been more used in the SE area (DA SILVA *et al.*, 2011; KITCHENHAM *et al.*, 2009, 2010; ZHANG *et al.* 2011). Moreover, the thousands of citations over the years to the guideline for performing SLRs introduced by Kitchenham and Charters (2007) indicate an increasing interest in SE for SLR.

An SLR is a means to identify, evaluate and interpret available relevant research to a topic of research (KITCHENHAM; CHARTERS, 2007). SLR is referred to as a secondary study. The individual studies that it analyzes are called primary studies (KITCHENHAM; CHARTERS, 2007).

There are diferent types of secondary studies (KITCHENHAM; CHARTERS, 2007; KITCHENHAM *et al.*, 2010): Systematic Literature Review (SLR), Systematic Mapping (SM) and Tertiary Review (TR). The SLR aggregates results related to a specific research question or topic area. SM aims at identifying and classifying the primary studies in a topic area. The TR is an SLR of SLR, in order to answer wider research questions.

Brereton *et al.* (2007) reported that the summarization of results of primary studies through secondary studies can be very valuable in offering new insights in a topic area. It might also be important to identify where an issue might be clarified by additional primary studies in a future research.

2.2.1 SLR PROCESS

According to Kitchenham and Charters (2007), an SLR is conducted by a process with activities which can comprised of the following three phases: 1. *Planning the review*; 2. *Conducting the review*; and 3. *Reporting the review*.

The stages associated with the *Planning the review* phase are: 1.1. Identifying the need for a review; 1.2. Commissioning a review; 1.3. Specifying the research questions; 1.4. Developing a review protocol; and 1.5. Evaluating the review protocol. In this phase, the protocol allows transparency and reproducibility to the review.

The stages associated with the *Conducting the review* phase are: 2.1. Identifying the research; 2.2. Selecting primary studies; 2.3. Assessing the quality of the studies; 2.4. Extracting the data and monitoring this process; and 2.5. Synthetizing the data. In this phase, the review can be started according to the protocol evaluated in the planning the review phase.

The stages associated with the *Reporting the review* phase are: 3.1. Specifying dissemination mechanisms; 3.2. Formatting the main report; and 3.3. Evaluating the report. In this phase, the report can be specified according to the results obtained from the review phase.

After researchers ensure that a systematic review is necessary, it should be carried out in accordance with a review protocol. The review protocol aims at minimizing the bias that may be made by researchers (KITCHENHAM; CHARTERS, 2007).

Wohlin *et al.* (2013) conducted a study of two SM to evaluate the reliability and point out some challenges related to this type of study in SE. They concluded that the reliability of secondary studies cannot and should not be taken for granted. The comparison of the two SM shows that the decisions taken by researchers and the assessments made influence the outcome – not only which papers are found but also what the researchers conclude from their secondary studies.

One of the advantages of using an SLR is the methodology. The results of the literature are less likely to be biased due to it. However, there is no way to prevent publication bias in the primary studies. Publication bias refers to the problem that positive results are more likely to be published than negative results (KITCHENHAM; CHARTERS, 2007).

According to Kitchenham and Charters (2007), the major disadvantage of systematic literature reviews is that they require more effort than traditional literature reviews. Nevertheless, the conduction of a review based in a systematic approach results in a controled, rigorous, auditable, repeatable, and impartial process.

2.3 SEARCH STRATEGIES

According to Kitchenham and Charters (2007), SLRs must be undertaken by strictly following a predefined search strategy. This search strategy should be unbiased and allow assessing the completeness of the search. Kitchenham and Charters (2007) argue that the initial searches for primary studies can be conducted by using several digital libraries, but also indicate that other complementary searches should be employed (e.g., manual searches within proceedings and journals).

One of the challenges of the database searches is to identify terms and formulate an appropriate search string to be used in the digital libraries. Furthermore, SE search engines are not designed to support SLRs (KITCHENHAM; CHARTERS, 2007) and an improper selection of search keywords or bugs related to features of the digital libraries could lead to missing relevant papers or retrieving irrelevant papers (SINGH; SINGH, 2017). Regarding the manual searches, they provides low reproducibility of SLRs and, consequently, are out of the scope of our investigation.

Wohlin (2014) provides guidelines for conducting snowballing as an SLR search strategy. The guidelines define, illustrate, and evaluate snowballing by replicating a published SLR that originally used a database search strategy. The snowballing approach has the challenge of identifying an appropriate seed set of papers. They concluded that snowballing represents an alternative search strategy to use when conducting SLRs, instead of searching in several different databases.

To mitigate the risk of missing relevant evidence, several researchers have combined both search strategies, starting with database searches in several digital libraries and then applying BS and FS iteratively on the set of papers that were selected based on the database searches. Examples of such SLRs include (SILVA, A. *et al.*, 2017; TARHAN; GIRAY, 2017; VASCONCELLOS *et al.*, 2017). Nevertheless, while helping to mitigate the risk of missing relevant evidence, the adoption of database searches in several digital libraries followed by iterative BS and FS might result in significant effort, involving analyzing more irrelevant research papers.

Having this in mind, in our previous work (MOURAO *et al.*, 2017), we defined and investigated a *hybrid search strategy* for selecting studies by combining searches in a specific digital library (Scopus) followed by parallel BS and FS. The proposed hybrid search strategy (MOURAO *et al.*, 2017) comprised the following four activities: *identify research using Scopus database search, select primary studies to compose the seed set, apply backward snowballing, and apply forward snowballing in parallel over the seed set.*

We assessed the efficiency of this hybrid search strategy over two previously conducted SLRs (DIESTE; JURISTO, 2011; MENDES *et al.*, 2014), which originally employed database searches. Our findings indicated that the proposed hybrid strategy was suitable for the investigated SLRs, providing similar results to using database searches on several digital libraries.

Nevertheless, this preliminary investigation had some significant limitations. First, we only compared the hybrid strategy against the database search strategy, missing comparisons against snowballing and against exhaustive combined database and snowballing searches. Second, we employed a specific snowballing strategy in which BS and FS are conducted in parallel over the seed set, i.e., the papers obtained by BS are not subject to FS, and vice-versa. This snowballing strategy was introduced as a tentative to increase precision without compromising recall (MOURAO *et al.*, 2017).

Therefore, in this dissertation we take the investigation on hybrid search strategies further, more precisely defining different hybrid search strategy possibilities (involving iterative, parallel, and sequential snowballing) and evaluating them against database searches, snowballing searches, and exhaustive combined database and snowballing searches. As a basis for comparisons, the evaluations were performed on three different SLRs (SILVA, A. *et al.*, 2017; TARHAN; GIRAY, 2017; VASCONCELLOS *et al.*, 2017) that employed exhaustive combined database and iterative snowballing searches. Finally, we use the metrics precision, recall, and F-measure to compare the efficiency of each search strategy.

2.4 RELATED WORK

Many other studies contrasted different SLR search strategies. In this chapter, we present some of these related works.

Jalali and Wohlin (2012) conducted a study to compare two different search approaches: the use of database search and the use of snowballing in the same SLR. Both SLR search strategies were conducted in the same topic and moment. They observed similar results from both SLR search strategies.

Wohlin (2014) proposes a guideline for snowballing and assesses such guideline through the replication of a published SLR that used only database search. As a conclusion, snowballing figures out as a potential alternative to database searches.

Badamputi *et al.* (2015) applied snowballing in a study, evaluated the efficiency and reliability of snowballing, and compared it with database search strategy. They concluded that the efficiency of snowballing is comparable to the efficiency of database searches.

Wohlin (2016) compared snowballing with a database search update (i.e., two similar database search SLR covering different time periods). They concluded that both approaches are comparable when it comes to which papers they find, although snowballing is more efficient.

Felizardo *et al.* (2016) compare outcomes of an SLR update using forward snowballing versus database search. Although database search reached higher recall, forward snowballing reached significantly higher precision. Consequently, forward snowballing has potential to reduce the effort in updating SLR in SE.

Mendes *et al.* (2019) also evaluate the use of different search strategies (e.g., database search and forward snowballing) for updating SLRs and provide specific recommendations for the SLR update context. They recommend that SLRs should be updated using a single iteration of forward snowballing, using both the results from Google Scholar and the primary studies of the original SLR as seed set.

Kitchenham *et al.* (2010) compared the use of manual search with broad automated searches. They found that broad automated searches were able to find more studies than manual searches, but eventually with poor quality.

Skoglund and Runeson (2009) presented a reference-based search strategy that checks whether papers cite together other papers. They evaluated their strategy over three published SLR and observed significant variation in the results.

Dieste and Padua (2007) analyzed the effects of adding few or many terms to queries on the sensitivity and precision of the SLR. They concluded that optimizing search strategies is not a straightforward task.

MacDonell *et al.* (2010) investigated how consistent are the process adopted in SLR, and the effects on the stability of outcomes. They compared the results of two independent SLR. They concluded that groups of researchers with similar domain experience could reach the same outcomes.

These related works in most cases compared database search with snowballing. The researchers do not investigate other search strategies, combining iterations of snowballing to conduct an SLR.

2.5 FINAL REMARKS

This chapter presented the concepts of an SLR, how to conduct an SRL into a process, and the advantages and disadvantages of using an SLR. The SLR must follow a predefined search strategy. It can be conducted by using a database search, snowballing, manual searches, complementing database search with snowballing, and hybrid search strategies. We also showed the published studies about SLRs in SE.

Differently from the aforementioned studies, which in most cases directly contrasted database search with snowballing, our study went further and dug into specific aspects of database search and snowballing. Regarding database search, we investigate the efficiency (both actual and potential) of different digital libraries and how they complement each other. Regarding snowballing, we investigated whether multiple iterations are necessary and how forward compares to backward. Finally, we also proposed four hybrid search strategies and contrasted their efficiency.

CHAPTER 3 – MATERIALS AND METHODS

3.1 INTRODUCTION

In this chapter, we provide details on how we conducted our evaluation. Therefore, we start decomposing our three research questions into concrete subquestions. Then, we describe the proposed strategies, how we selected the SLRs that were used to simulate the different strategies, the evaluation procedure, and the supporting tool implemented to automate the simulations.

3.2 RESEARCH QUESTIONS

Our first research question – *RQ1 - What is the efficiency of the <u>database search strategy</u> in the published SLRs?* – focuses on analyzing the efficiency of the database search component alone in the published SLRs. We decomposed this question into three sub-questions:

RQ1.1 - What was the efficiency of each digital library in the published SLRs? Digital libraries are different, some are more selective, returning only papers that are very adherent to the query. Others are more inclusive, returning many papers that may or may not fit to the researchers needs. In this RQ, we contrast the efficiency of digital libraries in terms of precision, recall, and F-measure. The precision is the percentage of papers retrieved by the digital library that were selected by the SLR. The recall is the percentage of selected papers of the SLR that were retrieved by the digital library. The F-measure is the harmonic mean between precision and recall, indicating a compromise between precision and recall.

RQ1.2 - How many papers of the published SLRs were indexed in the digital libraries? The search interface, the search engine, and the search string are not perfect. Sometimes, papers that are indexed in the digital library are not retrieved. In this RQ, we do a direct search for the title of each paper in the digital library to evaluate their recall regardless of the search string.

This measure indicates the percentage of the papers indexed by the digital library that could have been retrieved by the SLR.

RQ1.3 - How complementary or overlapping were the digital libraries in the published SLRs? The search in several digital libraries consumes a considerable amount of effort. Considering that some papers are indexed in more than one digital library, this overlap could represent unnecessary effort for researchers. In this RQ, we evaluate whether some digital libraries used in the published SLR are redundant, retrieving the same set of papers of other digital libraries.

Our second research question – *RQ2 - What is the efficiency of the snowballing search strategy in the published SLRs?* – focuses on analyzing the efficiency of the snowballing component alone in the published SLRs. It was also decomposed into three sub-questions:

RQ2.1 - What was the efficiency of complementing digital library searches with snowballing in the published SLRs? Automated search in digital libraries is the most common strategy for conducting SLRs. However, difficulties to create appropriate search strings and the quality of the search engines may jeopardize the SLR. Thus, snowballing over the reference list and citations of selected papers may help to identify other relevant studies and complement the digital library search. In this RQ, we measure the precision, recall, and F-measure of each forward and backward snowballing iteration.

RQ2.2 - How complementary or overlapping were backward and forward snowballing in the published SLRs? Both backward and forward snowballing are important for finding relevant papers. However, there is a chance to find the same set of papers in both. On the other hand, given the seed set, one could miss relevant papers by choosing to do just backward or forward snowballing alone. To identify if backward and forward snowballing are equally efficient, we first simulate forward and backward snowballing independently, to collect the set of papers that could have been obtained by each one of them. Then, we investigate the intersection between these sets.

Our third research question – *RQ3 - What is the efficiency of each hybrid search strategy in the published SLRs?* – focuses on analyzing the efficiency of the proposed hybrid search strategies in the published SLR. A hybrid strategy combines specific variations of database search and snowballing, eventually focusing on result quality in detriment of review effort, or vice versa. For this RQ, we measure the precision, recall and F-measure of each hybrid strategy, and contrast to the values obtained with the baseline strategies (digital library search, snowballing, and the exhaustive combination of both).

3.3 HYBRID SEARCH STRATEGIES

As previously commented, we proposed four hybrid search strategies and contrasted them with three baseline strategies – database search (KITCHENHAM; CHARTERS, 2007), snowballing (WOHLIN, 2014), and database search followed by exhaustive iterative BS and FS – to answer our research question. These hybrid strategies combine database search over one specific digital library with different variations of snowballing steps. When conceiving the hybrid strategies, we chose Scopus for the database search because it was the most efficient digital library when answering RQ1 (See CHAPTER 4 – RESULTS AND DISCUSSION). Moreover, the snowballing steps comprise running backward and forward snowballing in iterations (the strategy defined in the original snowballing guidelines by Wohlin (2014)), in parallel (the strategy defined in our previous work (MOURAO *et al.*, 2017)), in sequence with BS followed by FS, and in sequence with FS followed by BS. Table 1 lists all seven strategies analyzed in this dissertation. The first three represent the baseline strategies used for comparisons, while the remaining four represent the proposed hybrid strategies.

Table 1: Analyzed search strategies.

Strategy	Hybrid Strategy	Description
DB Search		Database search
SB Search (BS*FS)		Snowballing search = Google Scholar + iterative backward & forward snowballing
DB Search + BS*FS		Database search + iterative backward & forward snowballing
Scopus + BS*FS	x	Scopus database search + iterative backward & forward snowballing
Scopus + BS‖FS	x	Scopus database search + parallel backward & forward snowballing
Scopus + BS+FS	x	Scopus database search + sequential backward & forward snowballing
Scopus + FS+BS	x	Scopus database search + sequential forward & backward snowballing

DB Search: this strategy follows the usual database search guideline (KITCHENHAM; CHARTERS, 2007). The selected papers come from different queries over multiple digital libraries.

SB Search (BS*FS): this strategy follows the usual snowballing guideline (WOHLIN, 2014). It starts with an informal search in Google Scholar to compose a seed set. Then, multiple iterations of backward and forward snowballing find other papers from the seed set, recursively.

DB Search + BS*FS: this strategy combines both the search over all digital libraries and the full-fledged iterative snowballing, respectively described in the *DB Search* and *SB Search (BS*FS)* strategies. First, we perform searches over different digital libraries to compose our seed set. Then, we apply iterative backward and forward snowballing over the seed set and the results obtained by the snowballing, recursively.

Scopus + BS*FS: this strategy first runs a search over Scopus to compose a seed set. Then, other papers are obtained from the seed set via iterative backward and forward snowballing, recursively.

Scopus + BS||FS: this strategy also starts with a search over Scopus to compose a seed set. Then, backward and forward snowballing run in parallel over the same seed set. In other words, the papers obtained by backward snowballing are not subject to forward snowballing, and vice-versa. This strategy was first introduced by Mourão *et al.* (2017) as a tentative to increase precision without compromising recall.

Scopus + BS+FS: similar to the previous strategies, this strategy also starts with a search over Scopus to compose a seed set. Then, multiple iterations of backward snowballing occur over the seed set. After finishing all backward snowballing iterations, forward snowballing starts. In this strategy, the papers obtained by forward snowballing are not subject to backward snowballing.

Scopus + FS+BS: this strategy also starts with a search over Scopus to compose a seed set. Then, multiple iterations of forward snowballing occur over the seed set. After finishing all forward snowballing iterations, backward snowballing starts. In this strategy, the papers obtained by backward snowballing are not subject to forward snowballing.

All seven strategies analyzed are visually represented in Appendix C.

3.4 CORPUS

In order to answer the research questions, we needed high-quality SLRs that had all the necessary information for the intended simulations. We first performed a search in Scopus for SLR in SE which cite both database search and snowballing. This search was restricted to title, abstract, and keywords. We identified a set of seven candidate SLRs. We then evaluated the quality of the SLRs using the same quality criteria used by Kitchenham *et al.* (2010) in their tertiary study on SLRs:

- Are the review’s inclusion and exclusion criteria described and appropriate (QA1)?
- Is the literature search likely to have covered all relevant studies (QA2)?
- Did the reviewers assess the quality/validity of the included studies (QA3)?

- Were the basic data/studies adequately described (QA4)?

Additionally, we had to formulate an additional criterion to ensure that the selected studies had all the necessary information for simulating the hybrid strategies:

- Does the SLR combine database search with iterative backward and forward snowballing (QA5)?

We score the questions as suggested by Kitchenham *et al.* (2009):

QA1: Yes, the inclusion criteria are explicit; Partly, the inclusion criteria are implicit; No, the inclusion criteria are not defined.

QA2: Yes, the authors have either searched four or more digital libraries and included additional search strategies or identified and referenced all journals addressing the topic of interest. Partly, the authors have searched three or four digital libraries with no extra search strategies, or searched a defined but restricted set of journals and conference proceedings. No, the authors have search up to two digital libraries or an extremely restricted set of journals.

QA3: Yes, the quality criteria are explicit and they were applied to each primary study; Partly, the research question involves quality issues that are addressed by the study; No, the quality criteria are not defined.

QA4: Yes, information about each study is described; Partly, only summary information about the set of studies is described. No, the basic information about the studies were not described.

QA5: Yes, the authors applied database search and complemented it with iterative backward and forward snowballing. Partly, the authors applied database search and complemented it with either backward or forward snowballing. No, the authors applied database search and snowballing, but snowballing was not directly applied on the results of the database search, i.e., the snowballing was applied in the results of the manual search and database search.

The scoring procedure was Yes = 1, Partly = 0.5, and No = 0. Table 2 shows the score for each of the seven candidate SLRs.

Table 2: Quality evaluation of the candidate SLRs.

Study	Ref	QA1	QA2	QA3	QA4	QA5	Total score
S1	Vasconcellos *et al.* (2017)	1	1	1	1	1	5
S2	Silva *et al.* (2017)	1	1	1	1	1	5
S3	Tarhan and Giray (2017)	1	1	1	1	1	5
S4	Steinmacher *et al.* (2015)	1	1	1	1	0	4
S5	Calderón and Ruiz (2015)	1	1	1	1	0	4
S6	Munir *et at.* (2014)	1	1	1	1	0.5	4.5
S7	Nguyen *et al.* (2015)	1	1	1	1	0	4

The results of the quality assessment show that, while all studies scored 1 in questions QA1 to QA4, only three studies scored 1 in our additional question QA5.

Studies S4 (STEINMACHER *et al.*, 2015) and S5 (CALDERÓN; RUIZ, 2015) applied the database search strategy and manual searches, then, after merging the results of these strategies, applied backward snowballing alone. Study S6 (MUNIR *et al.*, 2014) was conducted using database search and complemented with backward snowballing alone. Study S7 (NGUYEN *et al.*, 2015) was conducted using database search and manual searches and, after merging the results of these strategies, complemented with backward and forward snowballing.

Studies S1 (VASCONCELLOS *et al.*, 2017), S2 (SILVA *et al.*, 2017), and S3 (TARHAN; GIRAY, 2017) refers to SLRs that were completely compliant to QA5 and that were selected to compose the corpus of our study. They combine a database search in several digital libraries with iterative backward and forward snowballing. We detail each of the selected SLRs, hereafter referred to as SLR1, SLR2, and SLR3, in the following paragraphs.

SLR1 (VASCONCELLOS *et al.*, 2017) investigated evidence on approaches for the strategic alignment of software process improvement (SPI). It started with a digital library search on Springer, Scopus, Web of Science, Science Direct, Compendex, IEEE Xplore, and ACM Digital Library, followed by iterative backward and forward snowballing. It was published in 2017, with a database search conducted in August 2015, without limiting publication year. Snowballing was performed in July 2016. SLR1 selected 51 studies in total, where 22 came from the database search and 29 from snowballing. The study has an additional quality assessment step after snowballing, resulting in a final dataset with 30 papers. However, in terms of the traditional SLR process (KITCHENHAM; CHARTERS, 2007), the search strategy concerns the study identification and we used the study selection step (i.e., the application of the inclusion and exclusion criteria) as the basis for our assessment of precision and recall. Indeed, the search strategy has limited influence on the quality of the studies and the quality assessment is a separate and later step of the SLR process (KITCHENHAM; CHARTERS, 2007), typically influenced by specific SLR goals. We decided using the 51 papers obtained before this step of quality assessment for the evaluations of the search strategies. This decision is important because all 51 papers were subject of snowballing, having direct effect on the effort of the SLR. Not considering all of them would jeopardize the precision, recall, and F-measure metrics. The list of selected papers of SLR1 is available at companion website[1].

SLR2 (SILVA, A. *et al.*, 2017) aims at identifying and making a synthesis of the Definition of Done (DoD) criteria used in agile software development projects. It was published in 2017, with the database search conducted in June 2016 and the snowballing conducted in August 2016. The search strategy involved database search on ACM Digital Library, Engineering Village (Compendex), Science Direct, Scopus, Springer, Web of Science, and Wiley Online Library. After that, they performed snowballing on the set of papers selected as a result of the database search. SLR2 selected 20 research papers, where 16 came from the

[1] https://gems-uff.github.io/hybrid-strategies

database search and 4 from snowballing. Thereafter, the authors applied an additional quality assessment and ended with 8 papers. Following the same argumentation of SLR1, we decided to use the 20 papers obtained before this step of quality assessment for the evaluation of the search strategies. The list of selected papers of SLR2 is also available at our companion website.

SLR3 (TARHAN; GIRAY, 2017) investigates the use and usefulness of ontologies in Software Process Assessment (SPA). It was published in 2017, with database search and snowballing conducted in December 2016 and January 2017, respectively. The search strategy involved database search on ACM Digital Library, Google Scholar, IEEE Xplore, Science Direct, Scopus, Springer, Web of Science, and Wiley Online Library. Afterwards, they performed snowballing on the set of papers selected by the database search. SLR3 selected 14 research papers, being 11 during database search and 3 during snowballing. The list of selected papers of SLR3 is also available at companion website1.

We analyzed the digital libraries adopted by the SLRs that compose our corpus. Table 3 shows the use of the digital library in an SLR, denoted by “x”, and the absence, denoted by “-”. It is possible to observe that, while each SLR used a different set of digital libraries, they used at least seven different digital libraries each and have several digital libraries in common.

Table 3: Digital Libraries used for each SLR.

Digital Library	SLR1	SLR2	SLR3
ACM Digital Library	x	x	x
Compendex	x	x	-
Google Scholar	-	-	x
IEEE Xplore	x	-	x
Science Direct	x	x	x
Scopus	x	x	x
Springer	x	x	x
Web of Science	x	x	x
Wiley Online Library	-	x	x

Finally, as a summary of the study selection in the three SLRs, Table 4 shows that together they returned 2,803 papers in the database search. However, 891 were duplicates, leading to a total of 1,912 unique papers. Out of those 1,912 unique papers, 49 were selected from the database search and 36 were selected from snowballing.

Table 4: Study selection summary.

SLR	Papers with duplicates	Papers without duplicates	Seed set	Snowballing
SLR1	517	497	22	29
SLR2	1,715	935	16	4
SLR3	571	480	11	3

3.5 RESEARCH METHOD

After selecting the SLRs that compose our corpus, our evaluation procedure comprised three main activities: data extraction, strategies simulation, and analysis. Hereafter, we describe each of these activities.

3.5.1 DATA EXTRACTION

In this work, the data extraction was conducted considering the research questions. For each paper, we first extracted the available data in the paper. Then, we contacted the authors requesting additional data. Finally, if the data was not complete enough to reproduce the results, we rerun the queries over the digital libraries.

The authors provided us spreadsheets with the studies listed in the published SLR. From each published SLR, we extracted the number of visited papers in each digital library, number of primary studies selected in the SLR, the search string used, the list of digital libraries used, and the number of duplicated papers. For each paper, we identified the name, year of publication, authors, type of paper, publisher of the selected studies, the references list and the citations of the paper. Moreover, we tagged the papers, identifying them as seed set (provenient from database search) or as snowballing, and if it was selected in the SLR.

In some cases, the authors did not register duplicates. For example, a paper that was obtained from both ACM Digital Library and Scopus should have been registered as both ACM Digital Library and Scopus, and marked as duplicate. However, some authors just registered the first digital library that returned the paper. This could jeopardize our results, as we needed to count the visited and selected papers for each digital library to calculate the precision and recall. In these cases, we rerun the search in each digital library to check whether there were missing duplicates.

3.5.2 STRATEGIES SIMULATION

We simulate the hybrid search strategies using a snowballing supporting tool that we created (see Section 3.6) and applying it to the three published SLR of our corpus. The following steps compose the process applied to each SLR.

1. **Extract data from each paper.** A total of 2,803 papers has been identified based on the database search in all SLR (Table 4). We extracted the data of interest from each paper to insert it in the supporting tool.
2. **Insert data in the supporting tool.** We inserted the data of the paper and a tag with its provenance in the supporting tool. During this process, the tool identified and removed 891 duplicates. Nevertheless, the tool registers the set of digital libraries in which each paper was found and information if it was in the seed set of papers or retrieved through snowballing.
3. **Simulate search strategies.** We simulate each of the baseline and hybrid search strategies in the supporting tool. We simulated the strategies independently: *DB Search; SB Search; DB Search + BS*FS; Scopus + BS*FS; Scopus + BS||FS; Scopus + BS+FS; Scopus + FS+BS*. We

focused on reproducing equivalent results of the original SLRs, limited to identifying the papers that were selected in the original SLRs within the seed set and in the iterations of snowballing.

4. Calculate metrics. Finally, the tool calculates the metrics precision, recall, and F-measure based on the number of visited and selected papers for each search strategy. We compared these results to analyze the efficiency of each search strategy.

3.5.3 ANALYSIS (ANSWERING THE RQS)

Aiming at answering our RQs, we used the strategy as independent variable and precision, recall, and F-measure as dependent variables. Precision and recall are traditional information retrieval metrics (BAEZA-YATES; RIBEIRO-NETO, 1999) used to compare results with a predefined oracle. Originally, precision indicates the fraction of retrieved documents that are known to be relevant, recall indicates the fraction of known relevant documents that were effectively retrieved, and F-measure indicates the harmonic mean between precision and recall.

In our context, the oracle is the set of selected papers from the SLRs (strategy *DB Search + BS*FS*). Precision indicates the correctness of a given strategy in finding appropriate papers. For instance, a 100% precision would indicate that all papers visited by the strategy were actually selected by the SLR. Recall indicates the completeness of a given strategy. For instance, a 100% recall would indicate that all selected papers of the SLR were visited by the strategy. Finally, F-measure indicates the best compromise between precision and recall. Formally, we can define these metrics as follows:

$$Precision = \frac{Visited \cap Selected}{Visited}$$

$$Recall = \frac{Visited \cap Selected}{Selected}$$

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.6 SUPPORTING TOOL

Simulating a SLR is not trivial. Another researcher using the SLR protocol should be able to reproduce the same results. It requires keeping track of the visited papers, avoiding duplicates of them for avoiding rework. Simulating each strategy to compare them is even harder, as it requires managing the same papers multiple times manually.

To reduce this effort and potential errors, we used a snowballing supporting tool[2] and developed a set of Jupyter notebooks and Python scripts[3] to manage SLR papers and allow simulating the strategies. These scripts enable researchers to register papers without duplicates, find forward citations, and analyze the results of the snowballing. Moreover, after registering all papers obtained from the exhaustive database search followed by iterative snowballing, the scripts allow the simulation of all other strategies.

The usage of the scripts can be divided into two phases: registration and analysis. In the first phase, the researcher registers papers and then the tool provides support for performing both the backward and forward snowballing. In the second phase, the researcher analyzes the results. Since the scripts are able to collect the citations, they can generate citation graphs for the simulated strategies.

In the first phase, researchers can use Jupyter notebooks to register papers from the seed set and from backward snowballing into the tool. In our case, we had the snowballing results of the published SLRs. Thus, we inserted all studies included in the original SLR. In addition to these papers, we also performed backward snowballing on the selected papers and inserted the results. For the backward snowballing, we extracted the references to BibTeX and used the notebooks to insert the papers together with their citations.

The next step consists of applying forward snowballing over the selected papers and registering studies that cite those papers. Note that the scripts use only Google Scholar to find

[2] https://github.com/JoaoFelipe/snowballing
[3] https://gems-uff.github.io/hybrid-strategies

the citations during forward snowballing. We repeat all steps of the process for each selected study of each SLR.

In addition to registering the papers, we also included tags in each paper to indicate their provenance, i.e., which digital libraries return them, which papers belong to the seed set, which papers were selected by the SLRs, and which papers were obtained through snowballing. We used these tags afterwards to support the analyses.

In the second phase, researchers can use Jupyter notebooks to analyze papers' metadata, generate provenance and citation graphs, and describe the snowballing process through simulations of strategies' iterations based on citations. Thus, we automated the calculation of the measures for each hybrid search strategy. All Jupyter notebooks are available at our companion website.

3.7 FINAL REMARKS

This chapter presented our research questions and the proposed hybrid search strategies. We selected three SLRs to compose the corpus of our study. They combine database search in several digital libraries with iterative backward and forward snowballing. Then, after selecting our corpus, we described our evaluation procedure which comprised three main activities: data extraction, strategies simulation, and analysis.

In the data extraction phase, we considered the data in the paper to answer the research questions. In the strategies simulation phase, we presented the steps that compose the process applied to each SLR. In the analysis phase, we position the strategies as independent variables and precision, recall, and F-measure as dependent variables to answer the RQs. Finally, to reduce effort and potential errors, we adopted a supporting tool to simulate and compare each strategy.

CHAPTER 4 – RESULTS AND DISCUSSION

4.1 INTRODUCTION

In this chapter, we present and discuss the results of the research questions defined in Chapter 3.

4.2 WHAT WAS THE EFFICIENCY OF EACH DIGITAL LIBRARY IN THE PUBLISHED SLRS (RQ1.1)?

Our results show that the database search strategy visited 497 papers retrieved from digital libraries in the SLR1, 935 papers in the SLR2, and 480 in the SLR3. Moreover, among the visited papers we identified 51 studies selected by the SLR1, 20 studies in the SLR2, and 14 in the SLR3. Thereafter, as shown in Table 5, we calculate the efficiency of each digital library in terms of precision, recall, and F-measure.

Regarding precision, we can observe that Compendex provides the highest value for SRL2 and the third-highest value for SRL1 (SRL3 did not adopt Compendex), when compared to the other digital libraries. Although Web of Science showed the highest result for SLR1, it had very poor results for SLR2 and SLR3. Finally, Scopus consistently delivered high values for precision: the second-highest result for SLR1 and SLR2, and the highest result for SLR3.

Regarding recall, Scopus is again a prominent option – it provides the highest results for SLR1 and SLR2, and the second-highest for SLR3. No other digital library was consistently efficient regarding recall in all three SLRs. Google Scholar showed the highest recall for SLR3 (SLR1 and SLR2 did not adopt Google Scholar). This result is not surprising, considering that Google Scholar is not exactly a digital library, but a search engine that references multiple digital libraries. However, it also returns gray literature in the search results.

Finally, regarding F-measure, Scopus was undoubtedly the most prominent digital library, with the highest values for all three SLRs. Compendex also appears as a strong competitor, with the third-highest result for SLR1 and second-highest for SLR2 (almost tied with Scopus). Google Scholar also appears with the second-highest value for SLR3. No other digital library consistently provided high values in terms of F-measure. Figure 1, Figure 2, and Figure 3 complement Table 5 by showing a visual representation of the measures of efficiency of each digital library for SLR1, SLR2, and SLR3, respectively.

Table 5: Efficiency of digital libraries in the SLRs.

Digital Library	Precision (%)			Recall (%)			F-measure (%)		
	SLR1	SLR2	SLR3	SLR1	SLR2	SLR3	SLR1	SLR2	SLR3
ACM Digital Library	5.00 (5/100)	2.38 (5/210)	NAN (0/0)	9.80 (5/51)	25.00 (5/20)	0.00 (0/14)	6.62	4.35	0.00
Compendex	38.46 (5/13)	25.00 (2/8)	-	9.80 (5/51)	10.00 (2/20)	-	15.62	14.29	-
Google Scholar	-	-	2.36 (11/466)	-	-	78.57 (11/14)	-	-	4.58
IEEE Xplore	13.95 (6/43)	-	NAN (0/0)	11.76 (6/51)	-	0.00 (0/14)	12.77	-	0.00
Science Direct	0.51 (1/195)	2.01 (5/249)	0.00 (0/21)	1.96 (1/51)	25.00 (5/20)	0.00 (0/14)	0.81	3.72	0.00
Scopus	46.67 (7/15)	9.09 (7/77)	3.80 (3/79)	13.73 (7/51)	35.00 (7/20)	21.43 (3/14)	21.21	14.43	6.45
Springer	1.42 (2/141)	0.81 (1/124)	1.54 (1/65)	3.92 (2/51)	5.00 (1/20)	7.14 (1/14)	2.08	1.39	2.53
Web of Science	50.00 (5/10)	0.00 (0/3)	NAN (0/0)	9.80 (5/51)	0.00 (0/20)	0.00 (0/14)	16.39	0.00	0.00
Wiley Online Library	-	0.00 (0/295)	0.00 (0/15)	-	0.00 (0/20)	0.00 (0/14)	-	0.00	0.00

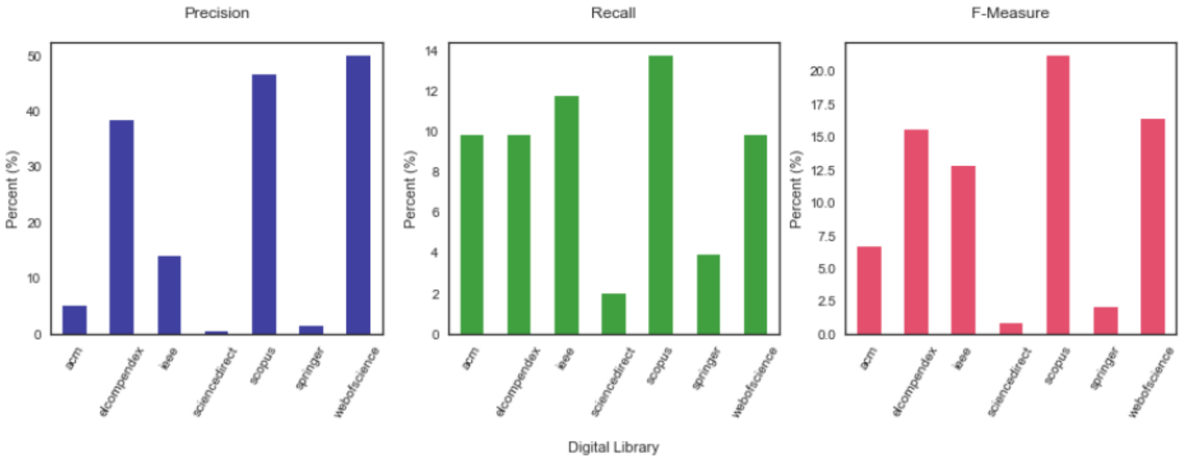


Figure 1: Overview of the efficiency of each digital library for SLR1.

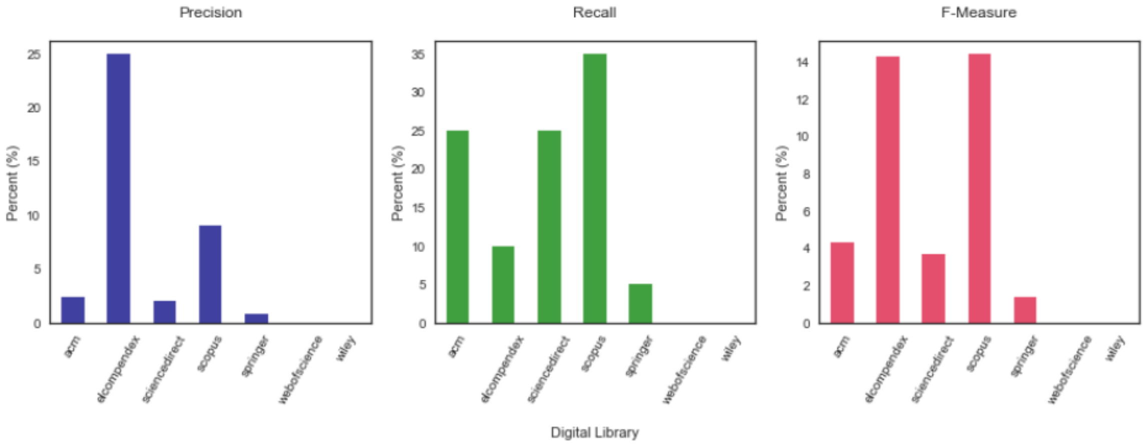


Figure 2: Overview of the efficiency of each digital library for SLR2.

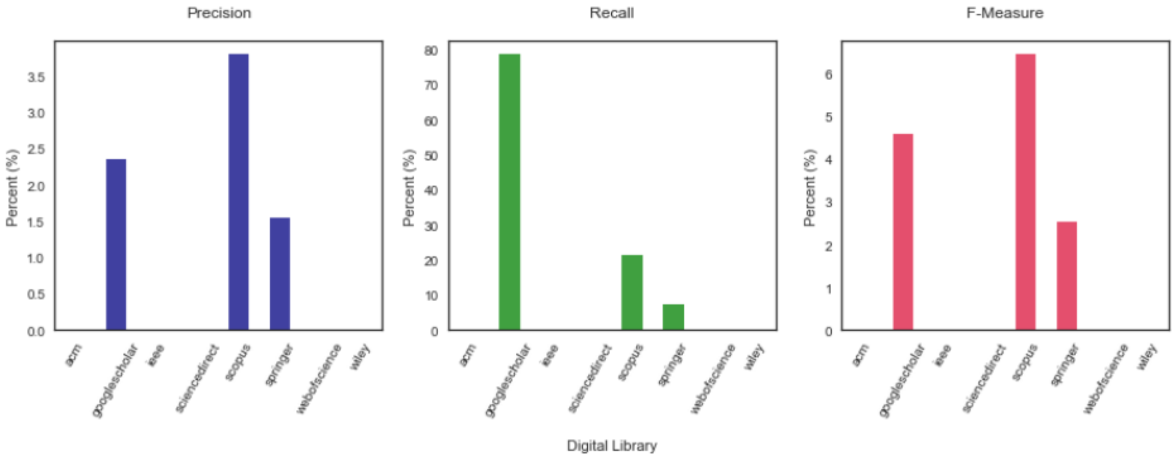


Figure 3: Overview of the efficiency of each digital library for SLR3.

Answer to RQ1.1: Scopus and Compendex (only SLR 1 and 2) are prominent options in terms of precision. Scopus and Google Scholar (only SLR3) are prominent options in terms of recall. When considering precision and recall together (F-measure), Scopus and Compendex stood out, with Scopus clearly ahead.

Implications: Scopus is a consistent option, but it finds just a limited amount (13% to 35%) of relevant papers alone. Thus, complementing Scopus with other digital libraries or snowballing is necessary.

4.3 HOW MANY PAPERS OF THE PUBLISHED SLRS WERE INDEXED IN THE DIGITAL LIBRARIES (RQ1.2)?

After querying the title of each selected paper from the SLRs on each digital library used by the SLRs, we could observe in Table 6 that Compendex delivered the highest recall result for SLR2 (a tie with Scopus) and the second-highest for SLR1 (SLR3 did not adopt Compendex). Google Scholar provides the highest value for SLR3 (SLR1 and SLR2 did not adopt Google Scholar). Finally, Scopus delivered again the highest result for SLR1, SLR2 (tied with Compendex), and second-highest result for SLR3.

Most of the digital libraries had a substantial increase from the concrete recall shown in Table 5: to the potential recall shown in Table 6. However, Wiley Online Library did not provide any increase in results for SLR2 and SLR3 (SLR1 did not adopt Wiley Online). Science Direct provided a small increase for all SLRs. Moreover, ACM Digital Library showed a very subtle increase for SLR1 and did not show any increase for SLR2 and SLR3.

Table 6: Recall of papers indexed in the published SLR.

Digital Library	Recall of indexed (%)		
	SLR1	SLR2	SLR3
ACM Digital Library	11.76 (6/51)	25.00 (5/20)	0.00 (0/14)
Compendex	68.63 (35/51)	95.00 (19/20)	-
Google Scholar	-	-	100.00 (14/14)
IEEE Xplore	27.45 (14/51)	-	21.43 (3/14)
Science Direct	3.92 (2/51)	30.00 (6/20)	7.14 (1/14)
Scopus	82.35 (42/51)	95.00 (19/20)	50.00 (7/14)
Springer	31.37 (16/51)	10.00 (2/20)	21.43 (3/14)
Web of Science	52.94 (27/51)	55.00 (11/20)	35.71 (5/14)
Wiley Online Library	-	0.00 (0/20)	0.00 (0/14)

Answer to RQ1.2: Scopus and Compendex are prominent options in terms of potential recall, with Scopus ahead. A side note is for Google Scholar that reached 100% recall in the only SLR that adopted it (SLR3).

Implications: Scopus could have found alone 50% to 95% of the papers. This motivates extra effort on the elaboration of search strings, considering the gap between the potential recall and the concrete recall delivered by the SLRs. Finally, as the sum of potential recall for each SLR surpasses 100%, a wise choice of digital library is necessary to avoid rework (many duplicates).

4.4 HOW COMPLEMENTARY OR OVERLAPPING WERE THE DIGITAL LIBRARIES IN THE PUBLISHED SLRS (RQ1.3)?

To answer this question, as shown in Table 7, we compared the set of selected papers among each pair of digital libraries. For instance, all 10 papers selected by ACM Digital Library are unique. Consequently, ACM Digital Library complements other digital libraries for SLR1 and SLR2 (no paper was identified for SLR3). On the other hand, six papers out of seven selected by Compendex are duplicate. Consequently, Compendex was mostly redundant.

We highlight in bold all cases with more than 50%. Google Scholar was able to retrieve all papers retrieved by Scopus and Springer. However, as previously mentioned, just SLR3 adopted Google Scholar and this data lacks more evidence. Moreover, Scopus was able to retrieve most of the papers retrieved by Compendex (85%) and Web of Science (80%). Finally, Compendex was able to retrieve 60% of the papers retrieved by Web of Science.

Table 7: Complementing versus overlapping digital libraries in the published SLR.

Row contains Column (diagonal indicates unique papers)	ACM Digital Library	Compendex	Google Scholar	IEEE Xplore	Science Direct	Scopus	Springer	Web of Science	Wiley Online Library
ACM Digital Library	**10/10**	0% (0/7)	0% (0/11)	0% (0/6)	0% (0/6)	0% (0/17)	0% (0/4)	0% (0/5)	NAN (0/0)
Compendex	0% (0/10)	1/7	-	16% (1/6)	0% (0/6)	42% (6/14)	33% (1/3)	**60% (3/5)**	NAN (0/0)
Google Scholar	NAN (0/0)	-	**7/11**	NAN (0/0)	NAN (0/0)	**100% (3/3)**	**100% (1/1)**	NAN (0/0)	NAN (0/0)
IEEE Xplore	0% (0/5)	20% (1/5)	0% (0/11)	**5/6**	0% (0/1)	10% (1/10)	0% (0/3)	0% (0/5)	NAN (0/0)
Science Direct	0% (0/10)	0% (0/7)	0% (0/11)	0% (0/6)	**5/6**	5% (1/17)	0% (0/4)	0% (0/5)	NAN (0/0)
Scopus	0% (0/10)	**85% (6/7)**	27% (3/11)	16% (1/6)	16% (1/6)	**6/17**	25% (1/4)	**80% (4/5)**	NAN (0/0)
Springer	0% (0/10)	14% (1/7)	9% (1/11)	0% (0/6)	0% (0/6)	5% (1/17)	2/4	0% (0/5)	NAN (0/0)
Web of Science	0% (0/10)	42% (3/7)	0% (0/11)	0% (0/6)	0% (0/6)	23% (4/17)	0% (0/4)	1/5	NAN (0/0)
Wiley Online Library	0% (0/5)	0% (0/2)	0% (0/11)	NAN (0/0)	0% (0/5)	0% (0/10)	0% (0/2)	NAN (0/0)	**NAN (0/0)**

Table 8 complements Table 7 by showing the set of selected papers of each digital library. The selected paper of the digital library are denoted by “P_1”, “P_2”, or “P_3”, respectively for each SLR, and the absence is denoted by “-”.

Compendex, Scopus, and Web of Science retrived the same papers “$P_1$7”, “$P_1$8”, and “$P_1$10” for SLR1. On the other hand, Compendex and Scopus retrieved the same papers “$P_2$6” and “$P_2$7” for SLR2. Finally, Google Scholar and Scopus retrieved the same papers “$P_3$2”, $P_3$5”, and “$P_3$8” for SLR3. This table shows the redudance in the digital libraries. We highlighted these papers in bolt in Table 8.

All selected papers from the digital libraries are listed in Appendix A.

Table 8: Set of selected papers of each digital library.

Digital Library	SLR1	SRL2	SLR3
ACM Digital Library	$P_1$1, $P_1$2, $P_1$3, $P_1$4, $P_1$5	$P_2$1, $P_2$2, $P_2$3, $P_2$4, $P_2$5	0
Compendex	$P_1$6, **$P_1$7**, **$P_1$8**, $P_1$9, **$P_1$10**	**$P_2$6**, **$P_2$7**	-
Google Scholar	-	-	$P_3$1, **$P_3$2**, $P_3$3, $P_3$4, **$P_3$5**, $P_3$6, $P_3$7, **$P_3$8**, $P_3$9, $P_3$10, $P_3$11
IEEE Xplore	$P_1$9, $P_1$11, $P_1$12, $P_1$13, $P_1$14, $P_1$15	-	0
Science Direct	$P_1$16	$P_2$8, $P_2$9, $P_2$10, $P_2$11, $P_2$12	0
Scopus	**$P_1$7**, **$P_1$8**, $P_1$9, **$P_1$10**, $P_1$17, $P_1$18, $P_1$19	**$P_2$6**, **$P_2$7**, $P_2$8, $P_2$13, $P_2$14, $P_2$15, $P_2$16	**$P_3$2**, **$P_3$5**, **$P_3$8**
Springer	$P_1$20, $P_1$21	$P_2$6	$P_3$10
Web of Science	**$P_1$7**, **$P_1$8**, **$P_1$10**, $P_1$17, $P_1$22	0	0
Wiley Online Library	-	0	0

Answer to RQ1.3: ACM Digital Library is complementary to other digital libraries and Scopus contains most of the results provided by Compendex and Web of Science. Google Scholar contained both Scopus and Springer, but the observation came from just one SLR.

Implications: When Scopus is adopted, adopting also Compendex and Web of Science does not seem to be highly relevant. On the other hand, ACM Digital Library should be considered. However, Scopus and ACM Digital Library were not able to reach all papers. Thus, complementing digital library search with snowballing is recommended.

4.5 WHAT WAS THE EFFICIENCY OF COMPLEMENTING DIGITAL LIBRARY SEARCHES WITH SNOWBALLING IN THE PUBLISHED SLRS (RQ2.1)?

As shown in Table 9, one iteration of snowballing was able to provide 100% of recall for SLR2 and SLR3 with high values of precision. In the first iteration for SLR1, the precision falls by 16%. On the other hand, the recall increases109% (from 43.14% to 90.20%). Note that this seems to be an interesting trade-off, especially considering that high recall is important in an SLR. Finally, regarding F-measure, we could observe a drop of 11% to 31% after complementing database search with one iteration of snowballing. Note that stopping the SLR in the database search phase would have retrieved just 43% to 80% of the papers. An iteration of snowballing increased recall to 90% to 100% of the papers. We highlight in bold the seed set and the union values.

Figure 4, Figure 5, and Figure 6 complement Table 9 by showing a visual representation (generated by our supporting tool) of the whole search process for SLR1, SLR2, and SLR3, respectively. They capture the number of visited and selected papers for each one of the digital libraries and the snowballing iterations. These visualizations reinforce the need of at least one snowballing iteration, especially when a limited number of digital libraries is adopted.

Table 9: Efficiency of complementing digital libraries with snowballing for SLRs.

Iteration	State	Accumulated Precision (%)			Accumulated Recall (%)			Accumulated F-measure (%)		
		SLR1	SLR2	SLR3	SLR1	SLR2	SLR3	SLR1	SLR2	SLR3
0	**seed set**	**4.43 (22/497)**	**1.71 (16/935)**	**2.29 (11/480)**	**43.14 (22/51)**	**80.00 (16/20)**	**78.57 (11/14)**	**8.03**	**3.35**	**4.45**
1	forward	3.97 (34/856)	1.52 (17/1116)	2.18 (12/551)	66.67 (34/51)	85.00 (17/20)	85.71 (12/14)	7.50	2.99	4.25
	backward	4.28 (38/887)	1.46 (20/1374)	1.99 (14/703)	74.51 (38/51)	100.00 (20/20)	100.00 (14/14)	8.10	2.87	3.91
	union	**3.72 (46/1238)**	**1.29 (20/1555)**	**1.81 (14/773)**	**90.20 (46/51)**	**100.00 (20/20)**	**100.00 (14/14)**	**7.14**	**2.54**	**3.56**
2	forward	3.27 (49/1497)	1.27 (20/1569)	1.57 (14/894)	96.08 (49/51)	100.00 (20/20)	100.00 (14/14)	6.33	2.52	3.08
	backward	3.05 (47/1541)	1.27 (20/1576)	1.73 (14/811)	92.16 (47/51)	100.00 (20/20)	100.00 (14/14)	5.90	2.51	3.39
	union	**2.78 (50/1796)**	**1.26 (20/1590)**	**1.50 (14/932)**	**98.04 (50/51)**	**100.00 (20/20)**	**100.00 (14/14)**	**5.41**	**2.48**	**2.96**
3	forward	2.82 (51/1806)	-	-	100.00 (51/51)	-	-	5.49	-	-
	backward	2.71 (50/1848)	-	-	98.04 (50/51)	-	-	5.27	-	-
	union	**2.74 (51/1858)**	**-**	**-**	**100.00 (51/51)**	**-**	**-**	**5.34**	**-**	**-**
4	forward	2.74 (51/1860)	-	-	100.00 (51/51)	-	-	5.34	-	-
	backward	2.73 (51/1871)	-	-	100.00 (51/51)	-	-	5.31	-	-
	union	**2.72 (51/1873)**	**-**	**-**	**100.00 (51/51)**	**-**	**-**	**5.30**	**-**	**-**

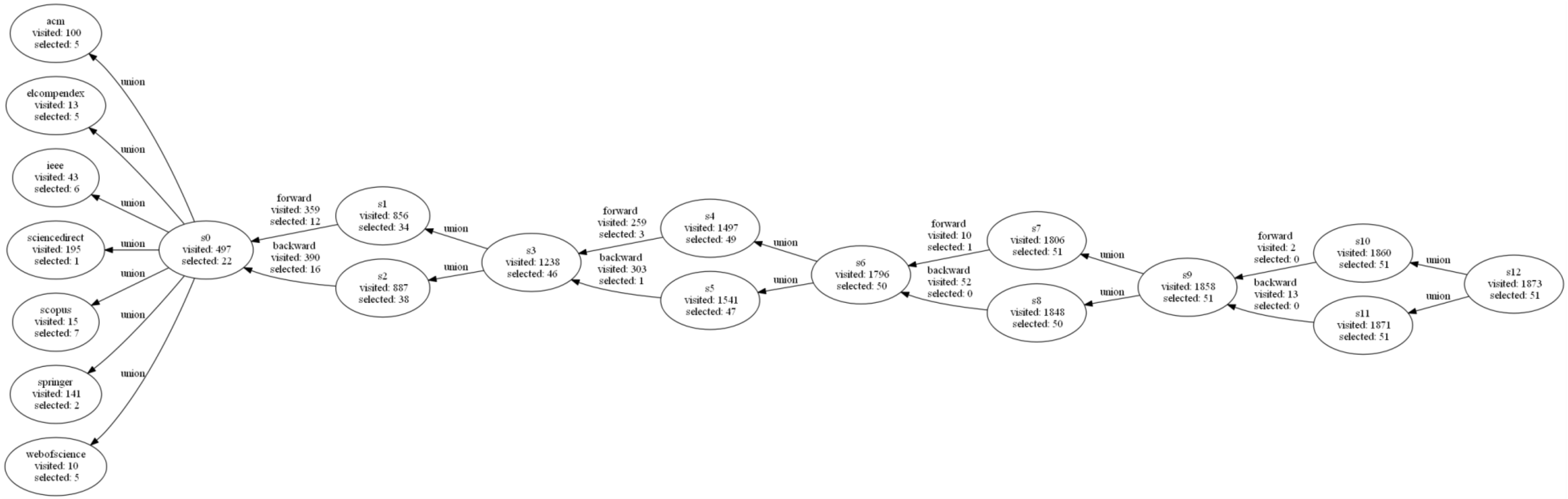


Figure 4: Overview of the database search followed by iterative snowballing for SLR1.

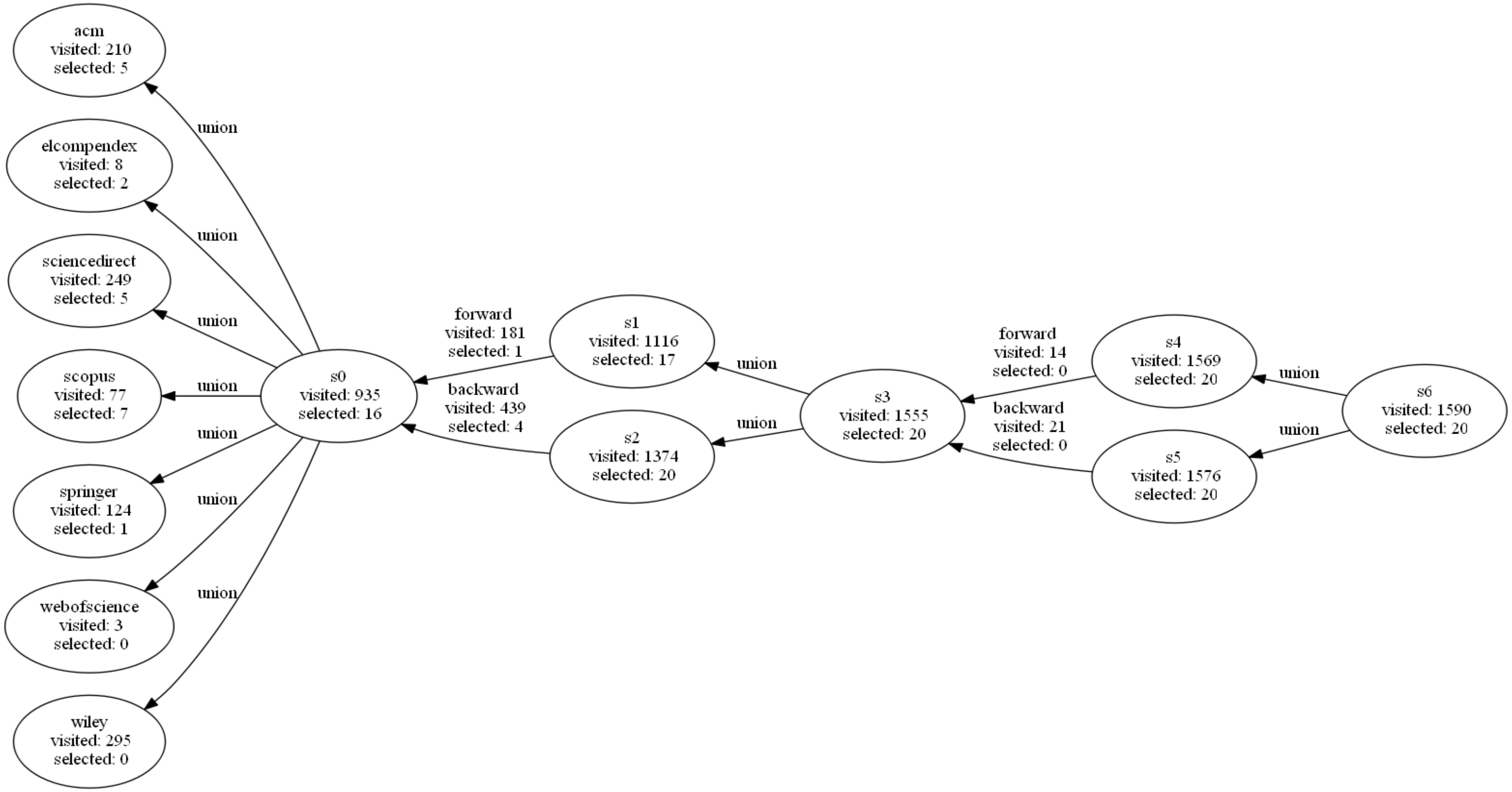


Figure 5: Overview of the database search followed by iterative snowballing for SLR2.

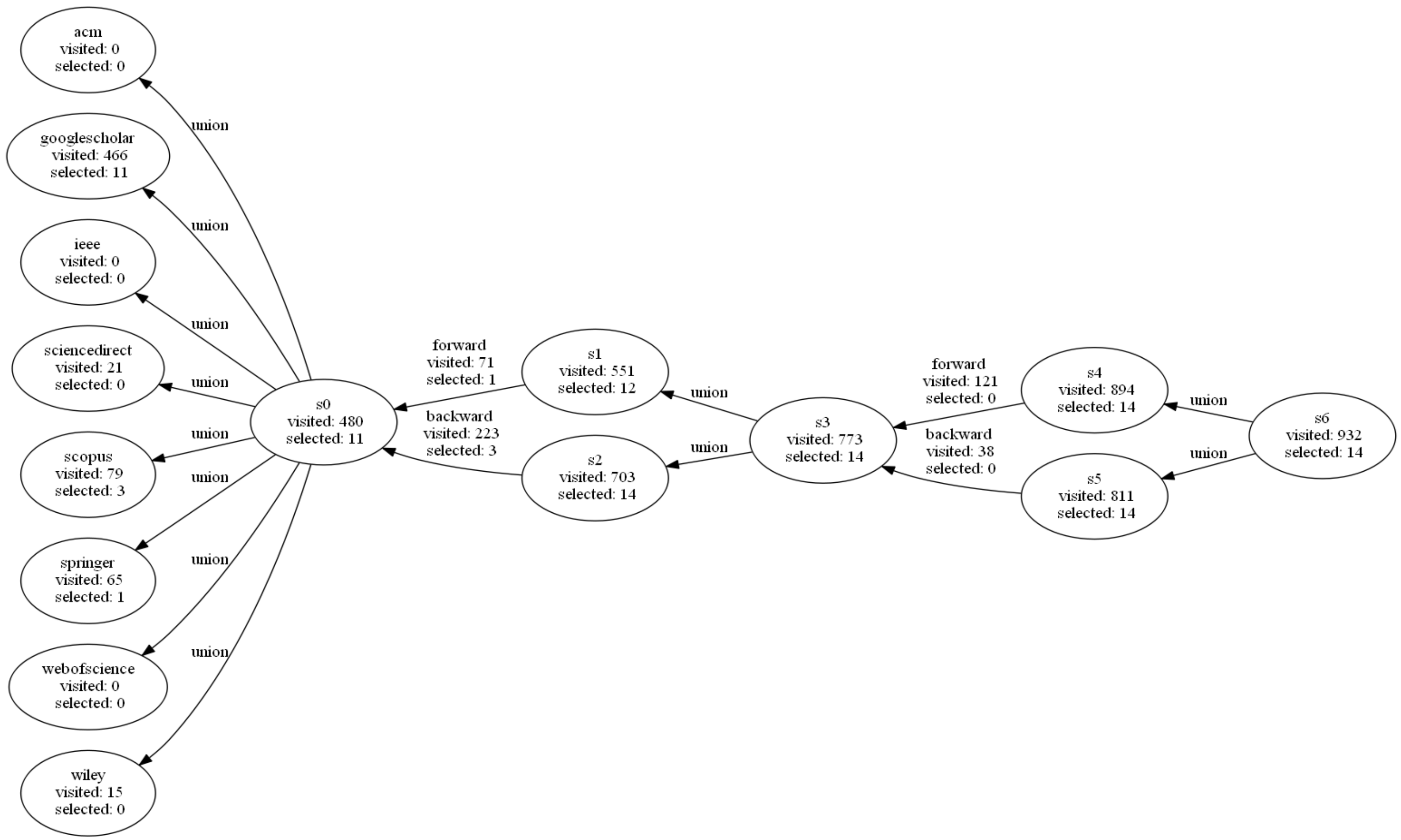


Figure 6: Overview of the database search followed by iterative snowballing for SLR3.

Answer to RQ2.1: A single iteration of backward and forward snowballing complemented the database search by providing 100% of recall for SLR2 and SLR3, and 90% recall for SLR1.

Implications: Although the first iteration of snowballing is clearly positive, we could only observe a 100% of recall for all three SLR after three iterations of snowballing, with negative consequences for precision. Thus, analyzing the interplay of backward and forward snowballing may help on devising more efficient hybrid strategies.

4.6 HOW COMPLEMENTARY OR OVERLAPPING WERE BACKWARD AND FORWARD SNOWBALLING IN THE PUBLISHED SLRS (RQ2.2)?

As show in Table 9, regarding precision, we can observe that forward snowballing provides the highest value in the iterations concerning the total of selected papers. However, the difference among forward and backward is low. Regarding recall, in the first iteration, backward snowballing provides the highest value for all SLRs. Forward snowballing provides the highest value in the following iterations for SLR1 after SLR2 and SLR3 reached 100%.

Figure 7 visually contrasts the sets of papers retrieved using forward and backward snowballing. For SLR1, the final set of papers includes 17 found via backward and 16 via forward with 4 identical papers. This means a 24% overlap for backward and 25% for forward snowballing. Thereby, backward and forward snowballing are complementary. For SLR2, the final set of papers includes 4 found via backward, 1 of them also found via forward snowballing. For SLR3, the final set includes 3 found via backward, 1 of them also found via forward snowballing. In the cases of SLR2 and SLR3, backward contains all papers found via forward snowballing.

All selected papers retrieved using backward and forward snowballing are listed in Appendix B.

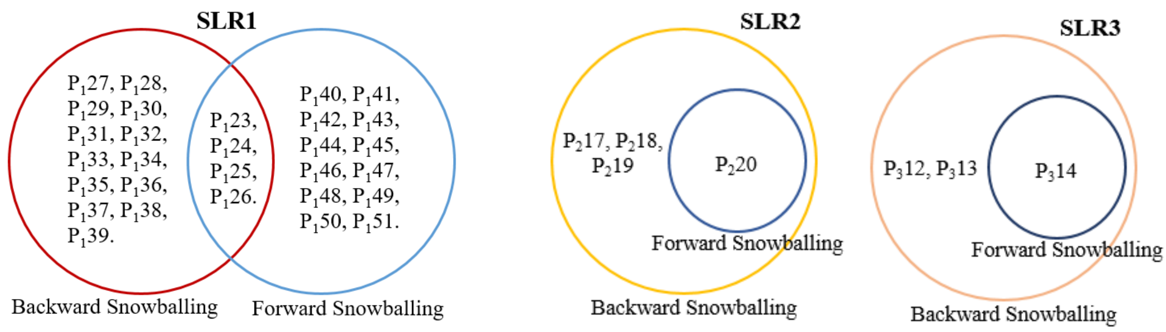


Figure 7: Venn diagram contrasting BS and FS.

Answer to RQ2.2: Forward snowballing is a prominent option in terms of precision. Backward snowballing is a prominent option in terms of recall. In the SLR1, backward and forward were complementary. In the SLR2 and SLR3, the backward snowballing contained all papers retrieved by the forward snowballing.

Implications: Both backward and forward snowballing contribute to the precision and recall of SLR. However, one or two iterations may suffice to reach relevant recall values.

4.7 WHAT IS THE EFFICIENCY OF EACH HYBRID SEARCH STRATEGY IN THE PUBLISHED SLRS (RQ3)?

As shown in Table 10, regarding precision, we can observe that *Scopus + BS+FS* provides the highest value for SLR2 (tied with *Scopus + BS||FS*), second highest value for SLR1, and third highest value for SRL3, when compared to the other strategies. *Scopus + FS+BS* provides the second highest value for SLR2 and SLR3, and third highest value for SLR1. However, the strategy *Scopus + BS||FS* consistently provides the highest value for SLR1, SLR2, and SLR3.

On the other hand, regarding recall, the strategies *Scopus + BS||FS*, *Scopus + FS+BS*, and *Scopus + BS+FS* show only the third or the fourth highest values. *Scopus + BS*FS* provides the second-highest recall for SLR1, SLR2, and SLR3. Of course, the highest recall is achieved by *DB Search + BS*FS* for all three SLRs, as it refers to the most complete strategy, and the one used in the baseline of our corpus to identify the overall selected papers.

Finally, regarding F-measure, the Scopus + BS||FS provides the highest value for SLR2 and SLR3, and *Scopus + BS+FS* the highest values for SLR1 and SLR2 (the later one tied with *Scopus + BS||FS*).

Figure 8, Figure 9, and Figure 10 complement Table 10 by showing a visual representation of the measures of efficiency of search strategies for SLR1, SLR2, and SLR3, respectively.

Table 10: Efficiency of strategies for SLRs.

Strategy	Precision (%)			Recall (%)			F-measure (%)		
	SLR1	SLR2	SLR3	SLR1	SLR2	SLR3	SLR1	SLR2	SLR3
DB Search	4.43 (22/497)	1.71 (16/935)	2.29 (11/480)	43.14 (22/51)	80.00 (16/20)	78.57 (11/14)	8.03	3.35	4.45
SB Search (BS*FS)	3.35 (36/1076)	1.26 (6/478)	2.25 (11/489)	70.59 (36/51)	30.00 (6/20)	78.57 (11/14)	6.39	2.41	4.37
DB Search + BS*FS	2.72 (51/1873)	1.26 (20/1590)	1.50 (14/932)	100.00 (51/51)	100.00 (20/20)	100.00 (14/14)	5.3	2.48	2.96
Scopus + BS*FS	3.75 (44/1174)	1.89 (11/581)	2.19 (11/502)	86.27 (44/51)	55.00 (11/20)	78.57 (11/14)	7.18	3.66	4.26
Scopus + BS||FS	6.51 (19/292)	2.65 (10/378)	3.72 (9/242)	37.25 (19/51)	50.00 (10/20)	64.29 (9/14)	11.08	5.03	7.03
Scopus + BS+FS	6.19 (35/565)	2.65 (10/378)	2.59 (11/424)	68.63 (35/51)	50.00 (10/20)	78.57 (11/14)	11.36	5.03	5.02
Scopus + FS+BS	5.81 (24/413)	1.89 (11/581)	3.27 (9/275)	47.06 (24/51)	55.00 (11/20)	64.29 (9/14)	10.34	3.66	6.23

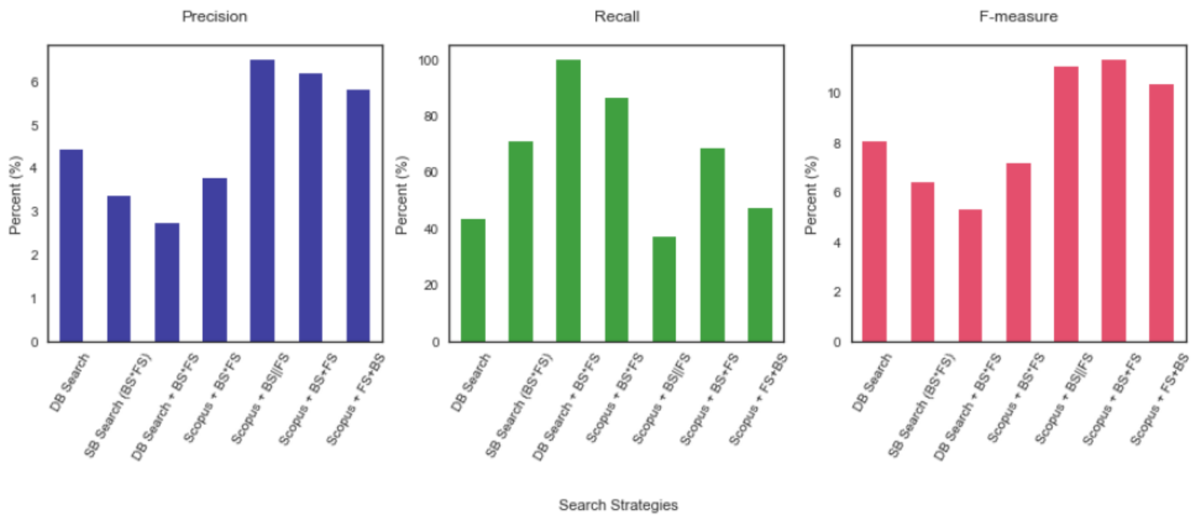


Figure 8: Overview of the efficiency of search strategies for SLR1.

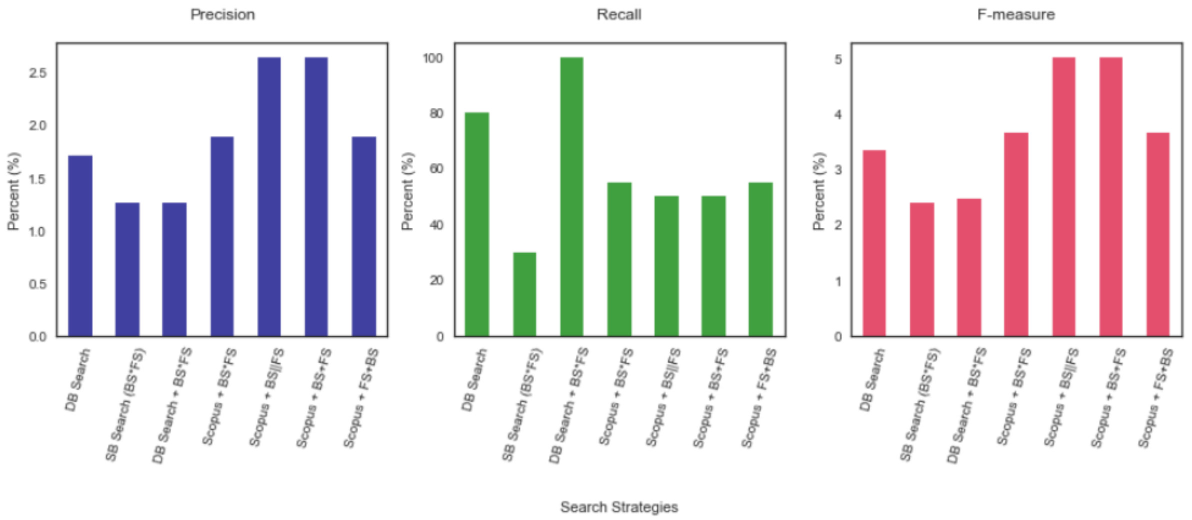


Figure 9: Overview of the efficiency of search strategies for SLR2.

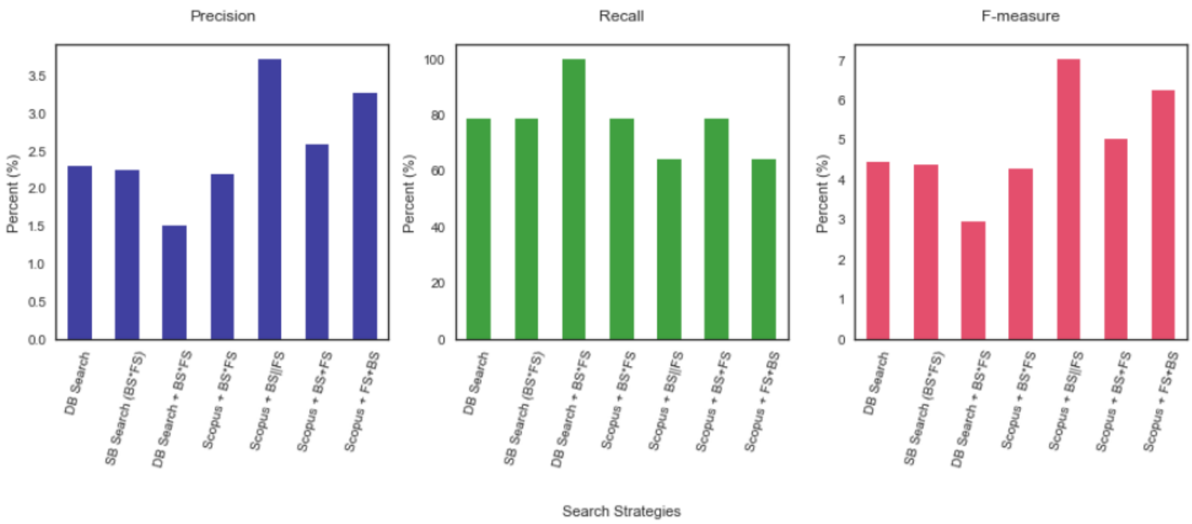


Figure 10: Overview of the efficiency of search strategies for SLR3.

Answer to RQ3: *Scopus + BS||FS* and *Scopus + BS+FS* strategies are prominent options in terms of precision. *DB Search + BS*FS* and Scopus + *BS*FS* are prominent options in terms of recall. When considering precision and recall together (F-measure), *Scopus + BS||FS* and *Scopus + BS+FS* stood out. Only *DB Search + BS*FS* provides 100% of recall, but with low precision, typically requiring significant effort.

Implications: *Scopus + BS||FS*, *Scopus + BS+FS*, and *Scopus + FS+BS* demand less effort than plain *DB Search* and *DB Search + BS*FS*, but with a price in recall. Depending on the goals of the SLR and the resources available, one of these hybrid strategies may be an appropriate alternative. However, no other strategy besides *DB Search + BS*FS* could guarantee consistently high levels of recall.

4.8 FINAL REMARKS

This chapter presented and discussed our results adressing each research question. The RQ1.1 was answered by calculating the efficiency of each digital library in terms of precision, recall, and F-measure. When considering precision and recall together, Scopus and Compendex stood out, but Scopus presented the greater value. When answering the RQ1.2, once more, Scopus and Compendex were prominent options in terms of potential recall with Scopus presenting the higher value. In the RQ1.3, Scopus contained most of the results provided by Compendex and Web of Science. ACM Digital Library was complementary to other digital libraries. When answering the RQ2.1, we identified that a single iteration of BS and FS complemented the database search providing 100% of recall for SLR2 and SLR3, and 90% of recall for SLR1. For RQ2.2, both BS and FS contributed to precision and recall. Finally, for RQ3, when the (F-measure) was considered, *Scopus + BS||FS* and *Scopus + BS+FS* stood out.

CHAPTER 5 – CONCLUSION

5.1 CONTRIBUTIONS

In this work, we proposed four hybrid search strategies that combine database searches with snowballing and contrasted them with three baseline strategies. In total, we evaluated seven search strategies.

We used the snowballing steps, which comprise backward and forward snowballing, in iterations (the strategy defined by Wohlin (2014)), in parallel (the strategy defined in our previous work (MOURAO *et al.*, 2017)), in sequence with FS followed by BS, and in sequence with BS followed by FS.

We used a snowballing supporting tool and developed a set of Jupyter notebooks and Python scripts to manage the SLR papers and allow simulating the strategies.

We could observe that Scopus is the most consistent option in achieving high recall, but it found just from 13% to 35% of the relevant papers alone. Complementing Scopus with ACM Digital Library is an appropriate choice. Scopus and ACM Digital Library together found from 23% to 60% of the relevant papers. Moreover, investing an extra effort on the elaboration of the search string is also worth it – the recall of Scopus could have raised from the range of 13% to 35% to the range of 50% to 95% depending on the search string.

Nevertheless, we could also observe that a single iteration of backward and forward snowballing complemented the database search providing 90% to 100% of recall. When choosing among backward and forward snowballing, one should keep in mind that forward snowballing is a prominent option to improve precision and backward snowballing is a prominent option to improve recall.

In general, *Scopus + BS||FS*, *Scopus + BS+FS*, and *Scopus + FS+BS* are efficient hybrid search strategies in comparison to plain *DB Search* strategy and *DB Search + BS*FS*.

Depending on the goals of the SLR and the resources available, one of these hybrid strategies may be an appropriate alternative.

5.2 THREATS TO VALIDITY

Although we aimed at reducing the threats to validity of our study, some decisions may have affected the results. We discussed the threats to validity based on the types presented by Wohlin *et al.* (2012) hereafter.

Regarding **construct validity**, we adopted precision, recall, and F-measure to assess the efficiency of the SLRs. The adoption of other metrics could have reached different results. Nevertheless, precision and recall are commonly used in other studies that investigate search strategy, allowing a more direct comparison of results.

Moreover, we are computing the recall of each strategy based on the total number of articles retrieved by the SLR. Although we cannot guarantee that all possible articles were correctly obtained by each SLR, we assessed the quality of the SLR to minimize this threat. Additionally, all three SLR were published in relevant venues, that adopt a serious peer-review process.

Regarding **internal validity**, papers that describe SLRs not having enough information to allow a precise reproduction of the results. For instance, some do not list duplicates or indicate the digital library that returned each article. To mitigate this threat, we contacted the authors and requested the complete review package containing the list of articles returned by each digital library. The paper describing the SLR1 indicates that 517 articles were obtained through the database search and 495 articles remained after duplicate removal. However, we identified in the spreadsheet provided by the authors two other articles that were not duplicate. Therefore, the total count of articles after removing the duplicates was in fact 497. In the SLR2, we also found one non-duplicate paper in a spreadsheet provided by the authors. Thus, the total count of articles was 935 instead of 934. Finally, the SLR3 listed a total of 571 papers including duplicate, but two of them were not present in the spreadsheet. After removing duplicates, the total count of articles was 480. Moreover, the SLR3 did not list duplicates for all digital libraries. In fact, it just indicated the first digital library that returned the articles. To mitigate this problem, we reproduced the original search in all other digital libraries to identify which of them also would have returned duplicates of the articles.

When running the *SB Search* strategy, we followed the guideline provided by Wohlin (2014). However, that guideline has a subjective step of informal Google Scholar search, which would compromise the reproducibility of our study. Aiming at mitigating this threat, we always considered the top-60 results provided by Google Scholar. The order used in the Google result was the default, by relevance. Since the SLRs have at most 51 selected articles, this number would be big enough to accommodate all selected articles (recall = 100%).

SLR1 adopted a quality control step in the end of the process, removing 18 studies and 3 not available articles from the already 51 selected articles. Although the final count reported in their paper is 30 articles, we decided to consider all 51 in our analyses. Working with the smaller set would incorrectly affect the precision and recall, as many of the visited articles in the snowballing were there because of the removed articles in the quality control step. Similarly, the SLR2 also adopted a quality control step at the end, removing 12 articles from the 20 already selected articles. We again considered all the 20 selected articles to compute precision and recall, for the same reasons of SLR1.

Regarding **conclusion validity**, we did not adopt statistical tests during data analysis due to the size of our sample (three SLRs). Consequently, although our results allow observing the efficiency of the different search strategies on the selected SLRs, they are not conclusive.

Finally, regarding **external validity**, we searched for published SLRs in the SE area that has used database search and snowballing to compose the corpus. Due to the small sample, consisting of only three SLRs in the field of software process, our results may not be generalizable to all other fields of software engineering. We suggest as future work additional replications of this study over a greater number of SLRs from other fields of software engineering.

5.3 FUTURE WORK

Future work concerns the replication of this study over additional SLR. We also suggest the inclusion of other search strategies in future analyses, such as manual search, author-based search, and venue-based search.

Another possible research option regards identifying how relevant is the choice of the start set for the published SLRs. The start set is the starting point of either forward or backward snowballing. Researchers could identify the amount of iterations or steps needed to reach other papers from each of the selected paper.

Another area for future work is the visualization of the citation graph. One situation is to show the citation graph indicating the paper that is more cited and analyse the impact on other papers. Moreover, researchers could apply different algorithms to buid the citation graph and analyze the number of steps from the seed set to reach selected papers.

It is also possible to build a complete effort estimation model for improving the metrics precision, recall, and F-measure. This model would consider varying efforts for different types of paper classifications. For example, included studies, excluded studies, noise, and duplicates would lead to different efforts.

Moreover, we also envison an improvement in the supporting tool used in this work. Currently, it does not automatically extract the references from the pdf files of the selected papers. With such automation, users would focus on just answering which papers are relevant or not, without needing to manually search for the pdf files of the papers.

In this work, we only used the provenance graph with few elements to support the visualizations of the search strategies. However, other interesting provenance elements, such as the authors and the publisher of the selected papers, could be used to analyze and compare the search strategies.

Finally, we also believe that the proposed analyses can be replicated for other areas, besides SE.

BIBLIOGRAPHY

BADAMPUDI, D.; WOHLIN, C.; PETERSEN, K. Experiences from using snowballing and database searches in systematic literature studies. In: PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING - EASE ’15, 2015. Nanjing, China: ACM Press, 2015. p. 1–10. Disponível em: <http://dl.acm.org/citation.cfm?doid=2745802.2745818>. Acesso em: 25 nov. 2018.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. 1st. ed. New York, USA: Addison Wesley, 1999.

BRERETON, P.; KITCHENHAM, B.; BUDGEN, D.; TURNER, M.; KHALIL, M. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, abr. 2007. , p. 571–583.

CALDERÓN, A.; RUIZ, M. A systematic literature review on serious games evaluation: An application to software project management. *Computers & Education*, v. 87, p. 396–422, set. 2015.

DA SILVA, F. Q. B.; SANTOS, A. L. M.; SOARES, S.; FRANÇA, A. C. C.; MONTEIRO, C. V. F.; MACIEL, F. F. Six years of systematic literature reviews in software engineering: An updated tertiary study. *Information and Software Technology*, v. 53, n. 9, p. 899–913, set. 2011.

DIESTE, O; JURISTO, N. Systematic review and aggregation of empirical studies on elicitation techniques. *IEEE Transactions on Software Engineering*, v. 37, n. 2, p. 283–304, mar. 2011.

DIESTE, O.; PADUA, A. G. Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews. In: FIRST INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT (ESEM 2007), set. 2007. Madrid, Spain: IEEE, set. 2007. p. 215–224. Disponível em: <http://ieeexplore.ieee.org/document/4343749/>. Acesso em: 25 abr. 2019.

FELIZARDO, K. R.; MENDES, E.; KALINOWSKI, M.; SOUZA, É. F.; VIJAYKUMAR, N. L. Using Forward Snowballing to update Systematic Reviews in Software Engineering. In: PROCEEDINGS OF THE 10TH ACM/IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT - ESEM ’16, 2016. Ciudad Real, Spain: ACM Press, 2016. p. 1–6. Disponível em: <http://dl.acm.org/citation.cfm?doid=2961111.2962630>. Acesso em: 25 nov. 2018.

JALALI, S.; WOHLIN, C. Systematic literature studies: database searches vs. backward snowballing. In: IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT (ESEM), 2012, Lund University, Sweden. Lund University, Sweden: ACM, 2012.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. , EBSE 2007-001.Technical Report EBSE 2007-001. UK: Keele University and Durham University Joint Report., 2007. Disponível em: <http://community.dur.ac.uk/ebse/resources/guidelines/Systematic-reviews-5-8.pdf>.

KITCHENHAM, B.; DYBA, T.; JORGENSEN, M. Evidence-based software engineering. In: PROCEEDINGS. 26TH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING, 2004. Edinburgh, UK: IEEE Comput. Soc, 2004. p. 273–281. Disponível em: <http://ieeexplore.ieee.org/document/1317449/>. Acesso em: 25 nov. 2018.

KITCHENHAM, B.; BRERETON, O. P.; BUDGEN, D.; TURNER, M.; BAILEY, J.; LINKMAN, S. Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, v. 51, n. 1, p. 7–15, jan. 2009.

KITCHENHAM, B.; PRETORIUS, R.; BUDGEN, D.; BRERETON, O. P.; TURNER, M.; NIAZI, M.; LINKMAN, S. Systematic literature reviews in software engineering – A tertiary study. *Information and Software Technology*, v. 52, n. 8, p. 792–805, ago. 2010.

KITCHENHAM, B.; BRERETON, O. P.; TURNER, M.; NIAZI, M.; LINKMAN, S. PRETORIUS, R.; BUDGEN, D. Refining the systematic literature review process—two participant-observer case studies. *Empirical Software Engineering*, v. 15, n. 6, p. 618–653, dez. 2010.

MACDONELL, S.; SHEPPERD, M.; KITCHENHAM, B.; MENDES, E. How Reliable Are Systematic Reviews in Empirical Software Engineering? *IEEE Transactions on Software Engineering*, v. 36, n. 5, p. 676–687, set. 2010.

MENDES, E.; FELIZARDO, K. R.; WOHLIN, C.; KALINOWSKI, M. Search Strategy to Update Systematic Literature Reviews in Software Engineering. In: 45TH EUROMICRO CONFERENCE ON SOFTWARE ENGINEERING AND ADVANCED APPLICATIONS - SEAA 2019, 2019. Kallithea, Greece: IEEE Press, 2019.

MENDES, E.; KALINOWSKI, M.; MARTINS, D.; FERRUCCI, F.; SARRO, F. Cross- vs. within-company cost estimation studies revisited: an extended systematic review. In: THE 18TH INTERNATIONAL CONFERENCE, 2014. London, England, United Kingdom: ACM Press, 2014. p. 1–10. Disponível em: <http://dl.acm.org/citation.cfm?doid=2601248.2601284>. Acesso em: 25 nov. 2018.

MOURAO, E.; KALINOWSKI, M.; MURTA, L.; MENDES, E.; WOHLIN, C. Investigating the Use of a Hybrid Search Strategy for Systematic Reviews. In: 2017 ACM/IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT (ESEM), nov. 2017. Toronto, ON: IEEE, nov. 2017. p. 193–198. Disponível em: <http://ieeexplore.ieee.org/document/8170102/>. Acesso em: 25 nov. 2018.

MUNIR, H.; MOAYYED, M.; PETERSEN, K. Considering rigor and relevance when evaluating test driven development: A systematic review. *Information and Software Technology*, v. 56, n. 4, p. 375–394, abr. 2014.

NGUYEN, P. H.; KRAMER, M.; KLEIN, J.; LE TRAON, Y. An extensive systematic review on the Model-Driven Development of secure systems. *Information and Software Technology*, v. 68, p. 62–81, dez. 2015.

SILVA, A.; ARAÚJO, T.; NUNES, J.; PERKUSICH, M.; DILORENZO, E.; ALMEIDA, H.; PERKUSICH, A. A systematic review on the use of Definition of Done on agile software development projects. In: PROCEEDINGS OF THE 21ST INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING

- EASE’17, 2017. Karlskrona, Sweden: ACM Press, 2017. p. 364–373. Disponível em: <http://dl.acm.org/citation.cfm?doid=3084226.3084262>. Acesso em: 26 nov. 2018.

SINGH, P.; SINGH, K. Exploring Automatic Search in Digital Libraries: A Caution Guide for Systematic Reviewers. In: PROCEEDINGS OF THE 21ST INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING - EASE’17, 2017. Karlskrona, Sweden: ACM Press, 2017. p. 236–241. Disponível em: <http://dl.acm.org/citation.cfm?doid=3084226.3084275>. Acesso em: 29 jan. 2019.

SKOGLUND, M.; RUNESON, P. Reference-based search strategies in systematic reviews. In: PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING - EASE’09, 20 abr. 2009. UK: ACM Press, 20 abr. 2009. Disponível em: <https://dl.acm.org/citation.cfm?id=2227044>.

STEINMACHER, I.; SILVA, M. A. G.; GEROSA, M. A.; REDMILES, D. F. A systematic literature review on the barriers faced by newcomers to open source software projects. *Information and Software Technology*, v. 59, p. 67–85, mar. 2015.

TARHAN, A.; GIRAY, G. On the Use of Ontologies in Software Process Assessment: A Systematic Literature Review. In: PROCEEDINGS OF THE 21ST INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING - EASE’17, 2017. Karlskrona, Sweden: ACM Press, 2017. p. 2–11. Disponível em: <http://dl.acm.org/citation.cfm?doid=3084226.3084261>. Acesso em: 26 nov. 2018.

VASCONCELLOS, F. J. S.; LANDRE, G. B.; CUNHA, J. A. O.; OLIVEIRA, J. L.; FERREIRA, R. A.; VINCENZI, A. M. Approaches to strategic alignment of software process improvement: A systematic literature review. *Journal of Systems and Software*, v. 123, p. 45–63, jan. 2017.

WOHLIN, C.; RUNESON, P.; HÖST, M.; OHLSSON, M. C.; REGNELL, B.; WESSLÉN, A. *Experimentation in Software Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. Disponível em: <http://link.springer.com/10.1007/978-3-642-29044-2>. Acesso em: 26 nov. 2018.

WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: PROCEEDINGS OF THE 18TH INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING - EASE ’14, 2014. London, England, United Kingdom: ACM, 2014. p. 38. Disponível em: <https://dl.acm.org/citation.cfm?id=2601268>.

WOHLIN, C.; RUNESON, P.; NETO, P. A. D. M. S.; ENGSTRÖM, E.; do CARMO MACHADO, I.; De ALMEIDA, E. S. On the reliability of mapping studies in software engineering. *Journal of Systems and Software*, v. 86, n. 10, p. 2594–2610, out. 2013.

WOHLIN, C. Second-generation systematic literature studies using snowballing. In: PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING - EASE ’16, 2016. Limerick, Ireland: ACM Press, 2016. p. 1–6. Disponível em: <http://dl.acm.org/citation.cfm?doid=2915970.2916006>. Acesso em: 25 nov. 2018.

ZHANG, H.; BABAR, M. A.; TELL, P. Identifying relevant studies in software engineering. *Information and Software Technology*, v. 53, n. 6, p. 625–637, jun. 2011.

APPENDIX A – SET OF SELECTED PAPERS FROM DATABASE SEARCH

PID	Title	Authors	Year
$P_1$1	Exploring the use of the cynefin framework to inform software development approach decisions	O Connor, Rory V and Lepmets, Marion	2015
$P_1$2	Utilizing GQM+ Strategies for business value analysis: An approach for evaluating business goals	Mandić, Vladimir and Basili, Victor and Harjumaa, Lasse and Oivo,Markku and Markkula, Jouni	2010
$P_1$3	Strategic alignment of software process improvement programs using QFD	Becker, André Luiz and Prikladnicki, Rafael and Audy, Jorge Luis Nicolas	2008
$P_1$4	Strategically balanced process adoption	Esfahani, Hesam Chiniforooshan and Yu, Eric and Annosi, Maria Carmela	2011
$P_1$5	Applying and adjusting a software process improvement model in practice: the use of the IDEAL model in a small software enterprise	Kautz, Karlheinz and Hansen, Henrik Westergaard and Thaysen, Kim	2000
$P_1$6	Defining and monitoring strategically aligned software improvement goals	Barreto, Andrea Oliveira Soares and Rocha, Ana Regina	2010
$P_1$7	Software process improvement: Supporting the linking of the software and the business strategies	Albuquerque, Adriano Bessa and Rocha, Ana Regina and Lima, Andreia Cavalcanti	2009
$P_1$8	An approach to support the strategic alignment of software process improvement programs	Becker, A.L. and Audy, J.L.N. and Prikladnicki, R.	2008
$P_1$9	Application of GQM+ Strategies{ extregistered} in the Japanese space industry	Kaneko, Tatsuya and Katahira, Masafumi and Miyamoto, Yuko and Kowalczyk, Martin	2011
$P_1$10	Software engineering strategies: aligning software process improvement with strategic goals	Plösch, Reinhold and Pomberger, Gustav and Stallinger, Fritz	2011
$P_1$11	Utilizing GQM+ Strategies for an organization-wide earned value analysis	Mandic, Vladimir and Basili, Victor and Oivo, Markku and Harjumaa, Lasse and Markkula, Jouni	2010
$P_1$12	Measuring and improving software process in China	Wang, Qing and Li, Mingshu	2005

$P_1$13	A business goal-based approach to achieving systems engineering capability maturity	Waina, RB	2001
$P_1$14	Managing process inconsistency using viewpoints	Sommerville, Ian and Sawyer, Peter and Viller, Stephen	1999
$P_1$15	Linking software development and business strategy through measurement	Basili, Victor R and Lindvall, Mikael and Regardie, Myrna and Seaman, Carolyn and Heidrich, Jens and Münch, Jürgen and Rombach, Dieter and Trendowicz, Adam	2010
$P_1$16	Business-oriented process improvement: practices and experiences at Thales Naval The Netherlands (TNNL)	Trienekens, Jos JM and Kusters, Rob J and Rendering, Ben and Stokla, Kees	2005
$P_1$17	Integration of strategic management, process improvement and quantitative measurement for managing the competitiveness of software engineering organizations	Guzmán, Javier García and Mitre, Hugo A and Amescua, Antonio and Velasco, Manuel	2010
$P_1$18	A Low-overhead method for software process appraisal	Wilkie, F George and Mc Caffery, Fergal and McFall, Donald and Lester, Neil and Wilkinson, Emmanuel	2007
$P_1$19	SPI: 'I can't get no satisfaction' - directing process improvement to meet business needs	Reiblein, Susan and Symons, Andy	1997
$P_1$20	Entropy based software processes improvement	Trienekens, Jos J.M and Kusters, Rob and Kriek, Dirk and Siemons, Paul	2009
$P_1$21	An elicitation instrument for operationalising GQM+ Strategies (GQM+ S-EI)	Petersen, Kai and Gencel, Cigdem and Asghari, Negin and Betz, Stefanie	2015
$P_1$22	ProPAMet: a Metric for process and project alignment	Martins, Paula Ventura and Da Silva, Alberto Rodrigues	2008
$P_2$1	Better planning via tasking as a team	O'Connor, Christopher P	2010
$P_2$2	Pragmatic Approach for Managing Technical Debt in Legacy Software Project	Gupta, Rajeev Kumar and Manikreddy, Prabhulinga and Naik, Sandesh and Arya, Kavi	2016
$P_2$3	Quantitative assessment with using ticket driven development for teaching scrum framework	Igaki, Hiroshi and Fukuyasu, Naoki and Saiki, Sachio and Matsumoto, Shinsuke and Kusumoto, Shinji	2014

$P_2$4	Using formal software development methodologies in a real-world student project: an experience report	Mitra, Sandeep and Bullinger, Thomas A	2007
$P_2$5	Letters from the edge of an agile transition	O'Connor, Christopher P	2010
$P_2$6	Definition of ready: An experience report from teams at cisco	Power, Ken	2014
$P_2$7	Social contracts, simple rules and self-organization: a perspective on agile development	Power, Ken	2014
$P_2$8	The impact of inadequate customer collaboration on self-organizing Agile teams	Hoda, Rashina and Noble, James and Marshall, Stuart	2011
$P_2$9	Aligning codependent Scrum teams to enable fast business value delivery: A governance framework and set of intervention actions	Vlietland, Jan and Van Solingen, Rini and van Vliet, Hans	2016
$P_2$10	Towards a governance framework for chains of Scrum teams	Vlietland, Jan and van Vliet, Hans	2015
$P_2$11	Integrating usability work into a large inter-organisational agile development project: Tactics developed by usability designers	Wale-Kolade, Adeola Yetunde	2015
$P_2$12	Exploring ScrumBut—An empirical study of Scrum anti-patterns	Eloranta, Veli-Pekka and Koskimies, Kai and Mikkonen, Tommi	2016
$P_2$13	Securing Scrum for VAHTI.	Rindell, Kalle and Hyrynsalmi, Sami and Leppänen, Ville	2015
$P_2$14	Architecting for Large Scale Agile Software Development: A Risk-Driven Approach	Ozkaya, Ipek and Gagliardi, Michael and Nord, Robert L	2013
$P_2$15	Driving quality improvement and reducing technical debt with the definition of done	Davis, Noopur	2013
$P_2$16	Quality criteria for just-in-time requirements: just enough, just-in-time?	Heck, Petra and Zaidman, Andy	2015
$P_3$1	Automated software engineering process assessment: supporting diverse models using an ontology	Grambow, Gregor and Oberhauser, Roy and Reichert, Manfred	2013
$P_3$2	A CMMI Ontology for An Ontology-Based Software Process Assessment Tool	Gazel, Sema and Tarhan, Ayca and Sezer, Ebru	2009
$P_3$3	Towards automated process assessment in software engineering	Grambow, Gregor and Oberhauser, Roy and Reichert, Manfred	2012

P34	Representing Software Process in Description Logics: n Ontology pproach for Software Process Reasoning and Verification	Kabaale, Edward and Wen, Lian and Wang, Zhe and Rout, Terry	2016
P35	An ontology based infrastructure to support CMMI-based software process assessment	Gazel, Sema and Sezer, Ebru Akcpinar and Tarhan, Ayca	2012
P36	An ontology-based approach to express software processes	Liao, Li and Qu, Yuzhong and Leung, H	2005
P37	A software process ontology and its application	Liao, Li and Qu, Yuzhong and Leung, H	2005
P38	A strategic test process improvement approach using an ontologicaldescription for MND-TMM	Ryu, Hoyeon and Ryu, Dong-Kuk and Baik, Jongmoon	2008
P39	Building process definition with ontology background	Stolfa, Svatopluk and KoZusznik, Jan and Kosinár, Michal and Duzí, Marie and Cíhalová, Martina and Vondrák, Ivo	2010
P310	Project assets ontology (PAO) to support gap analysis for organization process improvement based on CMMI v. 1.2	Rungratri, Suwanit and Usanavasin, Sasiporn	2008
P311	A partial formalization of the CMMI-DEV—A capability maturity model for development	Soydan, Gokhan Halit and Kokar, Mieczyslaw M	2012

APPENDIX B – SET OF SELECTED PAPERS FROM SNOWBALLING

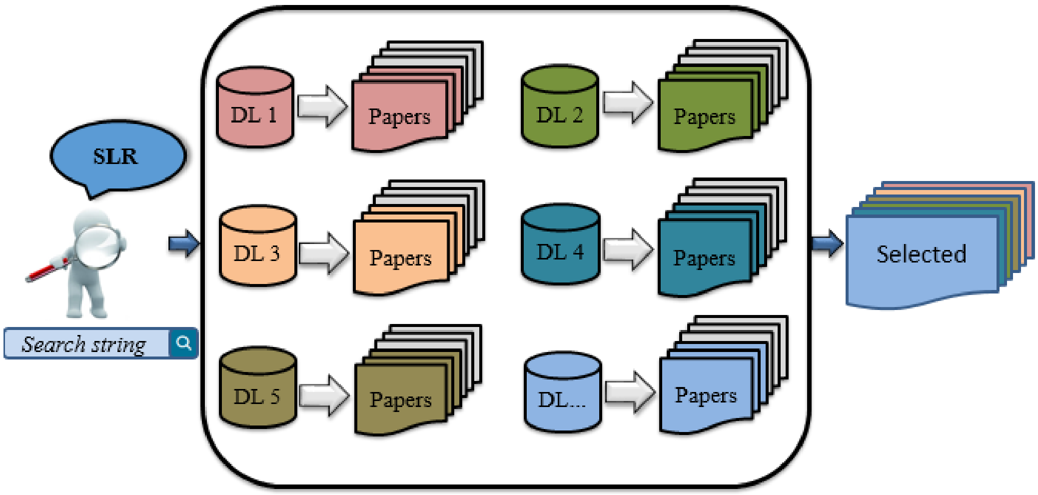
PID	Title	Authors	Year
$P_1$23	Experiences and insights from applying gqm+ strategies in a systems product development organisation	Münch, Jürgen and Fagerholm, Fabian and Kettunen, Petri and Pagels, Max and Partanen, Jari	2013
$P_1$24	SAS: A tool for the GQM+ strategies grid derivation process	Mandić Vladimir and Oivo, Markku	2010
$P_1$25	Aligning software projects with business objectives	Trendowicz, Adam and Heidrich, Jens and Shintani, Katsutoshi	2011
$P_1$26	The effects of gqm+ strategies on organizational alignment	Münch, Jürgen and Fagerholm, Fabian and Kettunen, Petri and Pagels, Max and Partanen, Jari	2013
$P_1$27	A goal-oriented approach for managing software process change	Hinley, DS and Reiblein, S	1995
$P_1$28	Interfacing three complementary technologies: strategic planning, process modeling, and system dynamics	McCoy, WL	1998
$P_1$29	Better Software Practice for Business Benefit: Principles and Experiences	Messnarz, Richard and Tully, Colin J	1999
$P_1$30	Linking software process improvement to business strategies: experiences from industry	Debou, Christophe and Kuntzmann-Combelles, Annie	2000
$P_1$31	Blending CMM and Six Sigma to meet business goals	Murugappan, Mala and Keeni, Gargi	2003
$P_1$32	QFD application in software process management and improvement based on CMM	Liu, Xiaoqing Frank and Sun, Yan and Kane, Gautam and Kyoya, Yuji and Noguchi, Kunio	2005
$P_1$33	Business-oriented software process improvement based on CMM using QFD	Liu, X. and Sun, Y. and Kane, G. and Kyoya, Y. and Noguchi, K.	2006
$P_1$34	Bridging the gap between business strategy and software development	Basili, Victor and Heidrich, Jens and Lindvall, Mikael and Münch, Jürgen and Regardie, Myrna and Rombach, Dieter and Seaman, Carolyn and Trendowicz, Adam	2007

$P_1$35	GQM+ Strategies: A comprehensive methodology for aligning business strategies with software measurement	Basili, Victor and Heidrich, Jens and Lindvall, Mikael and Münch, Jürgen and Regardie, Myrna and Rombach, Dieter and Seaman, Carolyn and Trendowicz, Adam	2007
$P_1$36	GQM+Strategies - Aligning business strategies with software measurement	Basili, Victor and Heidrich, Jens and Lindvall, Mikael and Munch, Jurgen and Regardie, Myrna and Trendowicz, Adam	2007
$P_1$37	ProPAM: SPI based on Process and Project Alignment	Martins, Paula Ventura and Silva, AR	2007
$P_1$38	Aligning software-related strategies in multi-organizational settings	Kowalczyk, Martin and Münch, Jürgen and Katahira, Masafumi and Kaneko, Tatsuya and Miyamoto, Yuko and Koishi, Yumi	2010
$P_1$39	A Cynefin based approach to process model tailoring and goal alignment	Lepmets, Marion and O Connor, Rory V and Cater-Steel, Aileen and Mesquida, Antoni Lluis and McBride, Tom	2014
$P_1$40	Aggregating viewpoints for strategic software process improvement—a method and a case study	Karlström, Daniel and Runeson, Per and Wohlin, Claes	2002
$P_1$41	Business objectives as drivers for process improvement: Practices and experiences at Thales Naval the Netherlands (TNNL)	Trienekens, Jos JM and Kusters, Rob J and Rendering, Ben and Stokla, Kees	2004
$P_1$42	Business-oriented software process improvement based on CMMI using QFD	Sun, Yan and Liu, Xiaoqing Frank	2010
$P_1$43	Determining organization-specific process suitability	Armbrust, Ove	2010
$P_1$44	The Rosetta Stone Methodology - A benefits-driven approach to software process improvement	McLoughlin, Fionbarr and Richardson, Ita	2010
$P_1$45	The Rosetta Stone Methodology - A benefits-driven approach to SPI	McLoughlin, F. and Richardson, I.	2010
$P_1$46	Which Processes Are Needed in Five Years? Strategic Process Portfolio Management at the Japan Aerospace Exploration Agency (JAXA)	Armbrust, Ove and Katahira, Masafumi and Kaneko, Tatsuya and Miyamoto, Yuko and Koishi, Yumi	2010

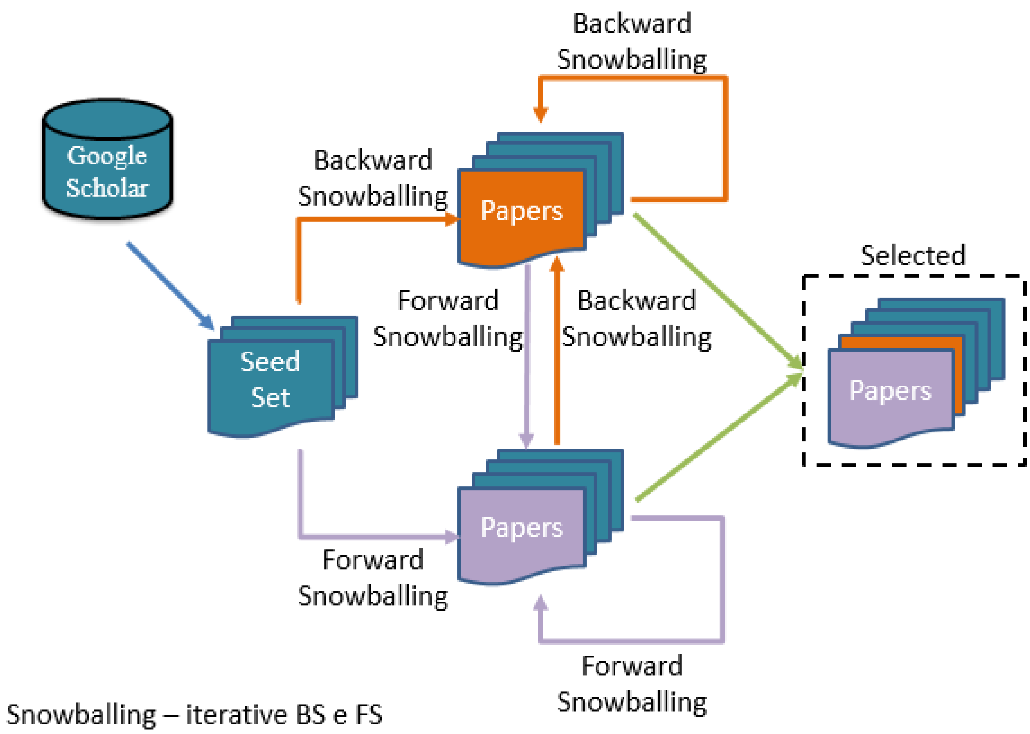
$P_1$47	A framework for systematic evaluation of process improvement priorities	Birkholzer, Thomas and Dickmann, Christoph and Vaupel, Jurgen	2011
$P_1$48	Development and Evaluation of Systems Engineering Strategies: An Assessment-Based Approach	Stallinger, Fritz and Plösch, Reinhold and Neumann, Robert and Horn, Stefan and Vollmar, Jan	2013
$P_1$49	Aligning Organizations Through Measurement	Trendowicz, Victor Basili Adam and Heidrich, Martin Kowalczyk Jens and Münch, Carolyn Seaman Jürgen and Rombach, Dieter	2014
$P_1$50	Application of GQM+ Strategies in a small software development unit	Cocozza, Francisco and Brenes, Enrique and Herrera, Gustavo López and Jenkins, Marcelo and Martínez, Alexandra	2014
$P_1$51	A conceptual framework for SPI evaluation	Unterkalmsteiner, Michael and Gorschek, Tony and Islam, AKM and Cheng, Chow Kian and Permadi, Rahadian Bayu and Feldt, Robert	2014
$P_2$17	All-out organizational Scrum as an innovation value chain	Barton, Brent	2009
$P_2$18	Scaling agile product ownership through team alignment and optimization: a story of epic proportions	Saddington, Peter	2012
$P_2$19	Scrum Anti-Patterns--An Empirical Study	Eloranta, Veli-Pekka and Koskimies, Kai and Mikkonen, Tommi and Vuorinen, Jyri	2013
$P_2$20	When agile meets the enterprise	Van Waardenburg, Guus and Van Vliet, Hans	2013
$P_3$12	Ontology-based intelligent decision support agent for CMMI project monitoring and control	Lee, Chang-Shing and Wang, Mei-Hui and Chen, Jui-Jen	2008
$P_3$13	Ontology-based computational intelligent multi-agent and its application to CMMI assessment	Lee, Chang-Shing and Wang, Mei-Hui	2009
$P_3$14	An OWL ontology for representing the CMMI-SW model	Soydan, Gokhan Halit and Kokar, M	2006

APPENDIX C – SET OF ANALYZED SEARCH STRATEGIES

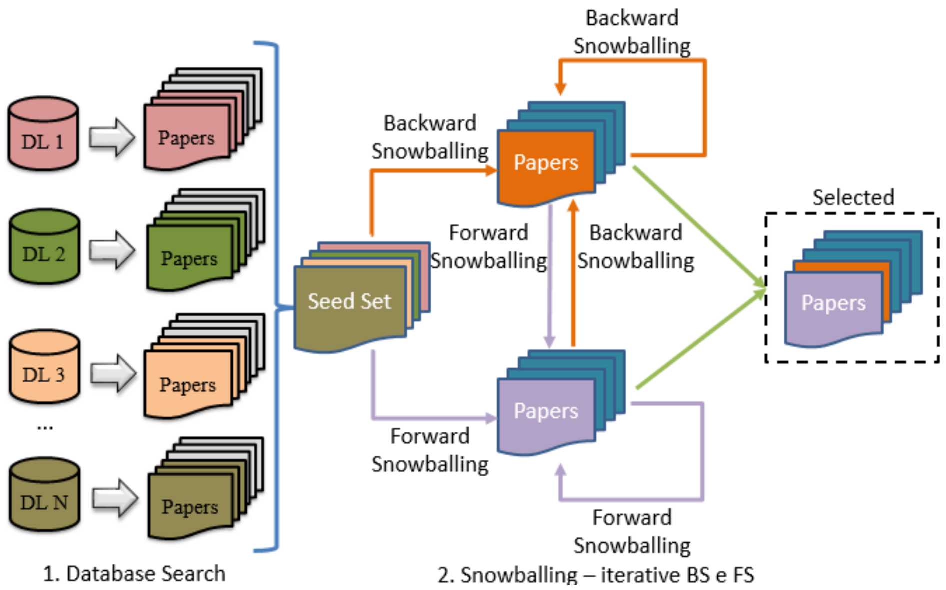
1. Database Search



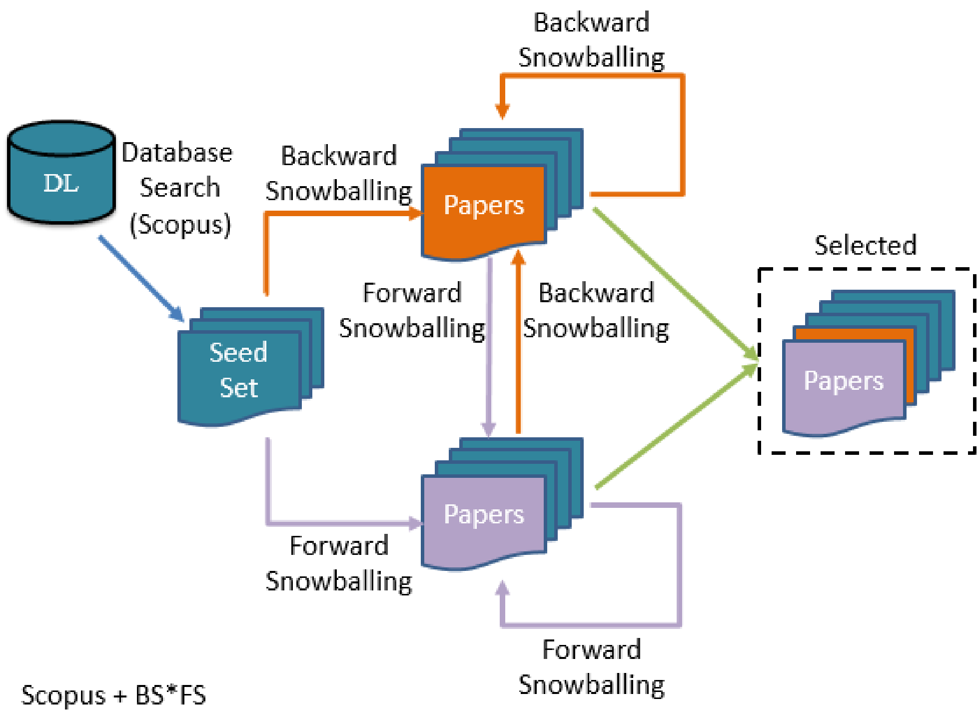
2. Snowballing (Google Scholar + iterative BS and FS)



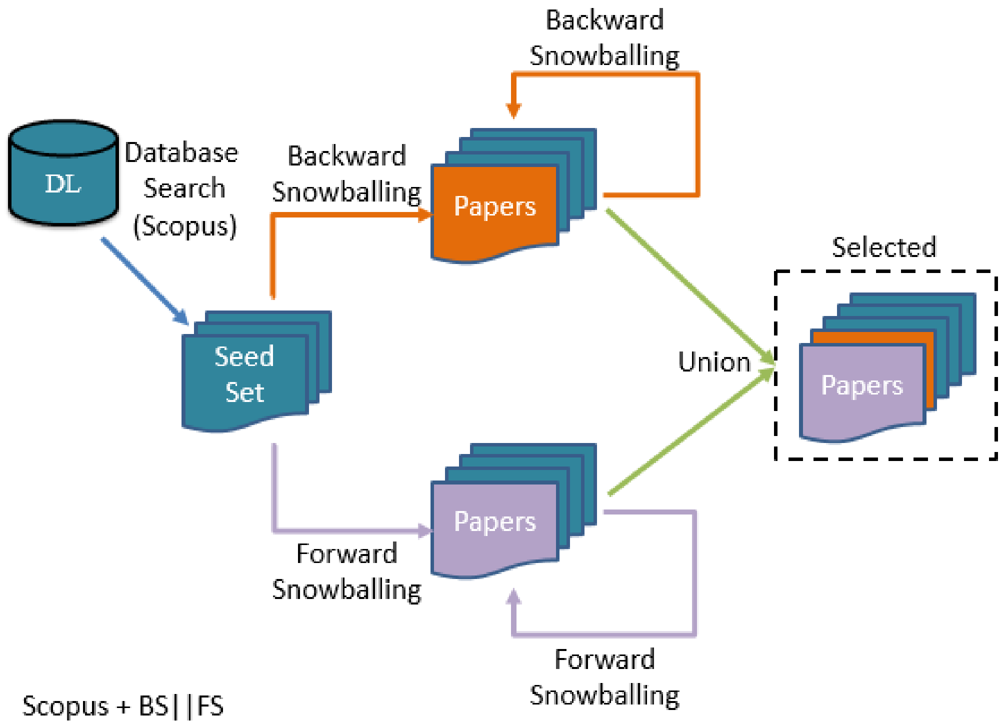
3. Database Search + Snowballing



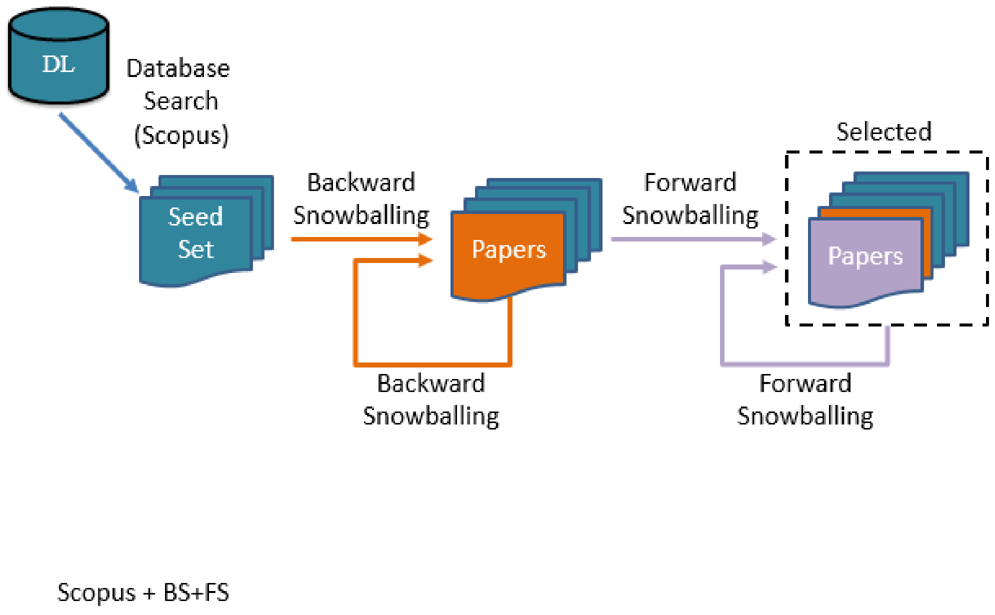
4. Scopus Database Search + iterative BS and FS



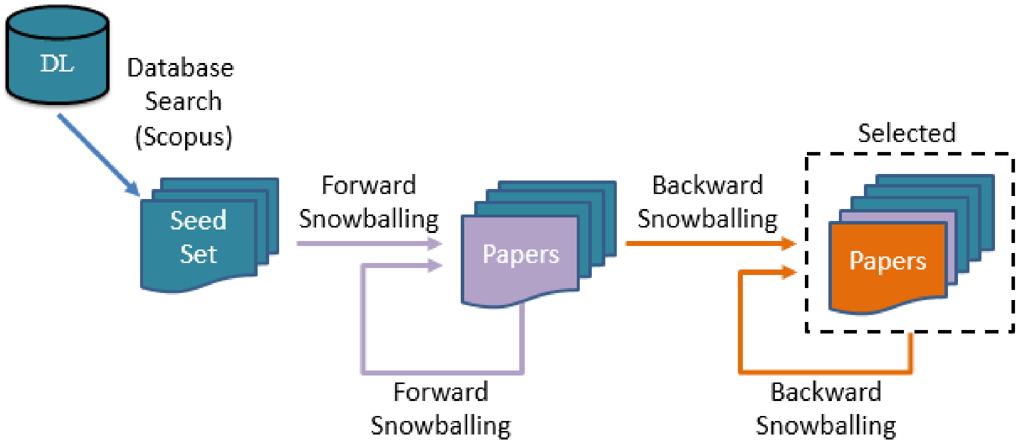
5. Scopus Database Search + parallel BS and FS



6. Scopus Database Search + sequential BS and FS



7. Scopus Database Search + sequential FS and BS



Scopus + FS+BS