UQAC
Université du Québec
à Chicoutimi

CONTINUOUS SYSTEMATIC LITERATURE REVIEW IN SOFTWARE ENGINEERING

BY BIANCA MINETTO NAPOLEÃO

THESIS PRESENTED TO THE UNIVERSITY OF QUEBEC AT CHICOUTIMI WITH A VIEW TO OBTAINING THE DEGREE OF PHILOSOPHIÆ DOCTOR (PH.D.) IN COMPUTER SCIENCE

QUÉBEC, CANADA

ABSTRACT

Context: Over the last two decades, Systematic Literature Reviews (SLRs) have been adopted as a research method to summarize evidence in Software Engineering (SE). However, scientific evidence continuously arises with advances in the SE field, leading to the need of updating SLRs. Outdated SLRs could lead researchers to obsolete conclusions or decisions about a research topic. **Goal:** We propose and evaluate the concept, process and guidelines to update SLRs in SE called Continuous Systematic Literature Review (CSLR). In addition, we explore automation alternatives to support the CSLR process execution. **Method:** A range of research methods were used for the construction and evaluation of this thesis. Firstly, two preliminary works were carried out to better understand the advances and existing needs in the area: an experience report on how to transfer the know-how of SLRs to facilitate their updates; and a cross-domain systematic mapping that identifies and summarizes the state-of-the-art on automation support for the two activities that triggers an SLR update, the search and selection of evidence. Secondly, to elaborate on the CSLR concept and process, we performed a synthesis of evidence by conducting a meta-ethnography, addressing knowledge from varied research areas. Thirdly, to build the guideline detailing and exemplifying the activities of the CSLR process, we carried out a systematic search and a narrative synthesis. Then we carry out an expert analysis with SE SLR experts to evaluate the CSLR guidelines and process and obtain feedback on the adoption of the CSLR process in practice. Lastly, we propose and evaluate an automation solution to support the search and selection of studies using Natural Language Processing and Machine Learning techniques. Also, we provide directions on the automation of the CSLR process. **Results:** The CSLR process and guidelines showed be beneficial in facilitating the identification if an SLR has been updated or not; assisting in the identification (search and selection) of potentially relevant evidence; promoting the sharing of potentially relevant evidence available in open repositories that are freely accessible by the SE community; supporting the decision on the need to update an SLR; and supporting SLR authors throughout the update process. The results from our prototype tool evaluation demonstrated the potential to reduce by at least 2.5 times the effort potentially reflecting on the researchers' time spent during the search and selection of studies to update SLRs in SE. As main future avenues for automation of the CSLR process, we encourage the development of a dedicated SLR repository in SE with the integration of the CSLR pipeline/workflow and exploration of recent technologies such as Large Language Models. **Conclusion:** The CSLR concept, process and guidelines provide a feasible and systematic way to continuously incorporate new evidence into SLRs, supporting trustworthy and up-to-date evidence for SLRs in SE. Moreover, they represent a valuable alternative to help keeping SLRs in SE up to date. We encourage further investigations in the direction of the automation of the CSLR process to assist the SE community keep SLRs up to date at the pace of the rapid increase of new evidence.

RÉSUMÉ

Contexte : Au cours des deux dernières décennies, les Revues Systématiques de la Littérature (RSL) ont été adoptées comme méthode de recherche pour résumer les preuves en Génie Logiciel (GL). Cependant, des preuves scientifiques apparaissent continuellement avec les progrès dans le domaine du GL, ce qui conduit à la nécessité de mettre à jour les RSL. Les RSL obsolètes pourraient conduire les chercheurs à des conclusions ou à des décisions obsolètes sur un sujet de recherche. **Objectif :** La présente thèse propose et évalue le concept, le processus et les lignes directrices pour mettre à jour les RSL de GL, appelé Revue Systématique Continue de la Littérature (RSCL). De plus, nous explorons des alternatives d'automatisation pour soutenir l'exécution du processus de RSCL. **Méthode :** Une gamme de méthodes de recherche a été utilisée pour la construction et l'évaluation de cette thèse. Tout d'abord, deux travaux préliminaires ont été réalisés afin de mieux comprendre les avancées et les besoins existants dans le domaine : un rapport d'expérience sur la manière de transférer le savoir-faire des RSL pour faciliter leurs mises à jour ; et une cartographie systématique interdomaine qui identifie et résume l'état de l'art en matière d'automatisation pour les deux activités qui déclenchent une mise à jour de RSL, la recherche et la sélection de preuves. Deuxièmement, pour élaborer sur le concept et le processus de RSCL, nous avons effectué une synthèse des preuves en effectuant une méta-ethnographie, abordant les connaissances de divers domaines de recherche. Troisièmement, pour construire la ligne directrice détaillant et illustrant les activités du processus de RSCL, nous avons effectué une recherche systématique et une synthèse narrative. Ensuite, nous effectuons une analyse experte avec des experts en matière de RSL en GL pour évaluer les lignes directrices et le processus de RSCL ainsi qu'obtenir des commentaires sur l'adoption du processus de RSCL dans la pratique. Enfin, nous proposons et évaluons une solution d'automatisation pour soutenir la recherche et la sélection d'études à l'aide de techniques de traitement automatique du langage naturel et d'apprentissage automatique. En outre, nous fournissons des instructions sur l'automatisation du processus de RSCL. **Résultats :** Le processus et les lignes directrices du RSCL se sont avérés bénéfiques pour faciliter l'identification si une RSL a été mis à jour ou non ; aider à l'identification (recherche et sélection) de preuves potentiellement pertinentes ; promouvoir le partage de preuves potentiellement pertinentes disponibles dans des référentiels ouverts librement accessibles par la communauté de GL; soutenir la décision sur la nécessité de mettre à jour une RLS ; et soutenir les auteurs d'une RSL tout au long du processus de mise à jour. Les résultats de l'évaluation de notre outil prototype ont démontré le potentiel de réduire d'au moins 2,5 fois l'effort, reflétant potentiellement le temps passé par les chercheurs lors de la recherche et de la sélection d'études pour mettre à jour les RSL. En tant que principales pistes futures d'automatisation du processus de RSCL, nous encourageons le développement d'un référentiel de RSL en GL dédié avec l'intégration du flux de travail de RSCL et l'exploration de technologies récentes telles que les « Large Language Models ». **Conclusion :** Le concept, le processus et les lignes directrices du RSCL fournissent un moyen faisable et systématique d'incorporer en permanence de nouvelles preuves dans les RSL, soutenant des preuves fiables

et à jour pour les RSL en GL. De plus, ils représentent une alternative précieuse pour aider à maintenir à jour les RSL en GL. Nous encourageons des enquêtes futures dans le sens de l'automatisation du processus de RSCL pour aider la communauté de GL à maintenir les RSL à jour au rythme de l'augmentation rapide des nouvelles preuves.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SE Software Engineering

EBSE Evidence-Based Software Engineering

SLR Systematic Literature Review

CSLR Continuous Systematic Literature Review

SM Systematic Mapping

RQ Research Question

IC Inclusion Criteria

EC Exclusion Criteria

DL Digital Library

CI Continuous Integration

CD Continuous Delivery

KM Knowledge Management

VTM Visual Text Mining

TM Text Mining

NLP Natural Language Processing

ML Machine Learning

TC Text Classification

SVM Support Vector Machines

STC Suffix Tree Clustering

HFSRM Hybrid Feature Selection Measure

VSM Vector Space Models

LSA Latent Semantic Analysis

DT Decision Trees

KNN K-Nearest Neighbor

BACA Blocked Adaptive Cross Approximation

LDA Latent Dirichlet Allocation

LMT Logistic Model Trees

WSS Work Saved over Sampling

3PDF Third-Party Decision Framework

AHRQ Healthcare Research and Quality

NIHR National Institute for Health Research

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SEGRESS Software Engineering Guidelines for REporting Secondary Studies

OSF Open Science Framework

DOI Digital Object Identifier

URL Uniform Resource Locator

CSV Comma Separated Values

LSVM Linear Support Vector Machine

SGD Stochastic Gradient Descent

NB Naïve Bayes

MNB Multinomial Naïve Bayes

AI Artificial Intelligence

LLM Large Language Models

GPT Generative Pre-trained Transformers

BERT Bidirectional Encoder Representations from Transformers

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God and Our Lady of Aparecida for always giving me strength during these years. My faith was an incredible reassurance during this learning process.

I would like to express my gratitude to my academic advisor, Sylvain Hallé, and my co-advisor, Fabio Petrillo. Their guidance, encouragement, and patience have been the driving force behind this research. Thank you for sharing your knowledge and shaping my academic and personal growth.

I extend my deepest appreciation to my family, especially to my lovely mother, Heloisa Minetto. Her sacrifices, encouragement, and unconditional love have been my rock for life. Thank you for instilling in me the values of perseverance and hard work. You are my number one source of inspiration. This achievement is as much yours as it is mine.

I want to extend a special thank you to my incredible conjoint, Xavier Mailhot. Throughout the challenging journey of completing this thesis, his unwavering support, understanding, and love have been my constant source of strength. His encouragement during late-night study sessions and his patience during times of stress have made this academic endeavor more manageable and joyful. Your love and support mean the world to me.

To all my friends, the long-time ones, and the ones that I have made here in Canada and Quebec. Thank you for your friendship, for the moments we shared, laughs, and kind words of encouragement. You have made my transition from Brazil to Canada smoother and funnier.

Lastly, I would like to thank everyone who, in some way, was part of my Ph.D. journey.

PREFACE

The results of this work have been either published or submitted to the appreciation of editorial boards of journals, conferences and workshops, according to the list of publications presented organized by categories (relation with this thesis) and ordered by publication year. My contributions to each publication are also listed.

PUBLICATIONS RESULTING FROM WORK RELATED TO THIS THESIS

- **1 – Napoleão, B. M.**, Felizardo, K. R., Hallé, S., Petrillo, F., & Kalinowski, M. (2023). *Guidelines to Perform Continuous Systematic Review in Software Engineering*. (Submitted - under review).
  Journal: Information and Software Technology (IST).
  Level of contribution: High – the Ph.D. candidate is the main investigator and conducted the work together with her contributors.

- **2 – Napoleão, B. M.**, Sarkar, R., Hallé, S., Petrillo, F., & Kalinowski, M. (2023). *Emerging Results on Automated Support for Searching and Selecting Evidence for Systematic Literature Review Updates*. (Submitted - under review).
  Workshop: ACM/IEEE 46th International Conference on Software Engineering (ICSE) – 1st International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE).
  Level of contribution: High – the Ph.D. candidate is the main investigator and conducted the work together with her contributors.

- **3 – Napoleão, B. M.**, Petrillo, F., Hallé, S. & Kalinowski, M. (2022). *Towards Continuous Systematic Review in Software engineering* (Napoleão *et al.*, 2022b).
  Conference: IEEE 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).
  Level of contribution: High – the Ph.D. candidate is the main investigator and conducted the work together with her contributors.

- **4 – Napoleão, B. M.**, Petrillo, F. & Hallé, S. (2021). *Automated Support for Searching and Selecting Evidence in Software Engineering: A Cross-domain Systematic Mapping*. (Napoleão *et al.*, 2021a).
  Conference: IEEE 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).
  Level of contribution: High – the Ph.D. candidate is the main investigator and conducted the work together with her contributors.

- **5 –** Felizardo, K. R., de Souza, E. F., Malacrida, T., **Napoleão, B. M.**, Petrillo, F., Hallé, S., Vijaykumar, N. L. & Nakagawa, E. Y. (2020). *Knowledge Management for Promoting Update of Systematic Literature Reviews: An Experience Report* (Felizardo

et al., 2020a).
Conference: IEEE 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).
Level of contribution: Medium – the Ph.D. candidate contributed with the paper writing and review. She also elaborated the presentation and presented the paper in the conference.

OTHER PUBLICATIONS – LOW RELATION TO THIS THESIS

- **6 –** Octaviano, F., Felizardo, K. R., Fabbri, S. C. P. F., **Napoleão, B. M.**, Petrillo, F. & Hallé, S. (2022). *SCAS-AI: A Strategy to Semi-Automate the Initial Selection Task in Systematic Literature Reviews* (Octaviano *et al.*, 2022).
  Conference: IEEE 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).
  Level of contribution: Medium – the Ph.D. candidate contributed with the paper organization and review. She also presented the paper in the conference.

- **7 – Napoleão, B. M.**, Felizardo, K. R., de Souza, E. F., Petrillo, F., Hallé, S., Vijaykumar, N. L. & Nakagawa, E. Y. (2021). *Establishing a Search String to Detect Secondary Studies in Software Engineering* (Napoleão *et al.*, 2021).
  Conference: IEEE 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).
  Level of contribution: High – the Ph.D. candidate is the main investigator and conducted the work together with her contributors.

INTRODUCTION

The last nearly 20 years of research and publications on Systematic Literature Reviews (SLRs) have been responsible for the Software Engineering (SE) community's adoption of SLRs as a tool to provide meaningful summaries of evidence on several topics for both SE academia and practitioners (Kitchenham *et al.*, 2015). Especially over the last few years, the number of SLRs in SE has increased substantially (Mendes *et al.*, 2020; Napoleão *et al.*, 2021).

SLR updates are also being conducted in SE area (e.g. Franca *et al.* (2011); Bezerra *et al.* (2015); Boyle *et al.* (2016); Vallon *et al.* (2018b); Pizzoleto *et al.* (2019); de A. Cabral *et al.* (2023)). According to Mendes *et al.* (2020), an SLR update is a more recent (updated) version of an SLR that includes new evidence (primary studies). They may also include new methods, such as new quality criteria to evaluate evidence, different search strategies to detect new evidence, and a repetition or remake (using a more recent synthesis method, for example) of the analysis of the original SLR.

PROBLEM STATEMENT AND RESEARCH MOTIVATION

One known challenge in evidence-based disciplines is to keep SLRs updated. As stated in Higgins *et al.* (2019), an SLR that is not maintained may become out-of-date or misleading. Furthermore, with the advance of the computer science field and the growth of research publications, new evidence continuously arises. This fact impacts directly the challenge of keeping SLRs up to date. An out-of-date SLR could lead researchers to obsolete conclusions or decisions about a research topic (Watanabe *et al.*, 2020).

In the field of medicine, SLR update has a consolidated process (Moher *et al.*, 2008; Shekelle *et al.*, 2011). Also in SE, there are several fragmented initiatives on SLR updates such

as establishing the SLR update process (Dieste *et al.*, 2008a; Mendes *et al.*, 2020), searching for new/updated evidence (Felizardo *et al.*, 2016; Wohlin *et al.*, 2020), selecting updated evidence (Watanabe *et al.*, 2020) and experience reports (Garcés *et al.*, 2017; Felizardo *et al.*, 2020a). Despite the effort of the SE community to keep SLRs updated, a Systematic Mapping (SM – a kind of lightweight SLR (Kitchenham *et al.*, 2015)) showed that only 22 SLRs in SE were updated since the start of SLR publication in 2004 (Nepomuceno & Soares, 2019). In February 2021, the SE area had more than 1000 SLRs and SMs published in several SE venues (Napoleão *et al.*, 2021). The need to update tertiary studies is also starting to gain the attention of the SE community. Two tertiary studies have been updated and published in SE venues (da Silva *et al.*, 2010; Barros-Justo *et al.*, 2021).

Creating and maintaining up-to-date SLRs demands a significant effort for reasons such as the rapid increase in the amount of evidence (Zhang *et al.*, 2018; Stol & Fitzgerald, 2015) and the limitation of available databases (Imtiaz *et al.*, 2013). Furthermore, the lack of detailed protocol documentation and data availability (Ampatzoglou *et al.*, 2019; Zhou *et al.*, 2015) makes the SLR update process even more difficult since most of the tacit knowledge from the SLR conduction is lost (Felizardo *et al.*, 2020a; Fabbri *et al.*, 2013).

Concerns on reducing resource consumption such as time and effort during the SLR conduction and update have been also highlighted by researchers through the proposal of a sustainability view (dos Santos *et al.*, 2021). The study of dos Santos *et al.* (2021) states the need for SE community efforts to promote (i) social aspects - researcher's communication and participation during the SLR conduction and update; (ii) economic aspects - resources reduction during the SLR conduction and update; and (iii) technical aspects - supporting tools and technologies to conduct and update the SLR.

Conventionally, SE SLRs are not updated or updated intermittently (Wohlin *et al.*, 2020). Periodic updating leaves gaps between updates, during which time the SLR may be missing crucial new research, placing it at risk of being inaccurate and wasting the potential contribution of new research to decision-making and evidence synthesis for answering research questions.

Fully or partially automating the SLR process is an alternative that has been explored by several SE researchers seeking to reduce time and effort over the years (Felizardo & Carver, 2020). Even problems and barriers regarding SLR automation as well as automation mitigation strategies have been considered, but many of the proposed mitigation strategies are relatively unknown by the SE community and are not applied in practice (Felizardo & Carver, 2020; dos Santos *et al.*, 2021). Regarding dedicated automation alternatives for SLR update, to the best of our knowledge, there are only two studies addressing this point: Felizardo *et al.* (2014) and Watanabe *et al.* (2020). Both studies are focused only on the study selection stage. Therefore, there is a lack of automation approaches dedicated to all stages of the SLR update.

More research efforts are needed to remedy the knowledge gaps previously described in updating SLRs and ascertain the potential benefits of continuous updating for SLRs.

RESEARCH GOALS

In the field of medicine, in order to mitigate the SLR updating issue, Elliott *et al.* (2017) introduced the concept of "Living Systematic Review" (LSR). An LSR is an SLR that is continually updated, incorporating relevant new evidence as it becomes available.

In software development, the DevOps concepts (mindset, practices, and tools) brought several benefits, such as the faster release of features, improved monitoring of systems in production, stimulation of collaboration among team members, and others. All of these

benefits lead to higher-quality software (Bass *et al.*, 2015). In addition, Continuous Integration (CI) and Continuous Delivery (CD), DevOps practices, aim to build, integrate and deliver all working versions of the software code, keeping the software updated. They include automation (tools) and a cultural mindset (Humble & Farley, 2010). On its side, maintaining an SLR up to date requires protocol changes, searching for new evidence, and management strategies to support the demand for updating. Thus, DevOps concepts are a potential solution to supporting continuous changes in SLRs to keep them up-to-date and with reliable results.

Moreover, an important trend in the SE community is the open science movement. It consists of making any research artifact available to the public following open access, open data, and open source practices (Mendez *et al.*, 2020). Open science approaches directly impact the SLR's conduction, not limited to the access and availability of primary evidence for the conduction of SLRs, but in the adoption of open science practices during the conduction of SLRs that reflects on their reproducibility and maintainability. Thus, concepts of open science must be addressed by researchers that conduct any evidence-based study.

Inspired by LSRs from evidence-based medicine and considering the DevOps concepts and open science practices, this thesis introduces the concept of **Continuous Systematic Literature Review (CSLR)** in SE, defines its process and proposes guidelines on its conduction. The construction of the CSLR concept, process and guidelines considers evidence from these three areas as well as concepts regarding SLR and SLR updates in SE. Additionally, automation solutions are proposed to support two trigger activities of the CSLR process. Our overall research goal is translated into three specific Research Goals (RG):

- **RG1 – Definition and evaluation of the CSLR concept and process**

  - **Evaluation Hypothesis:** The CSLR concept and process activities (steps) are practically feasible for updating SLRs in SE, and the CSLR process is efficient in helping mitigate the intermittent update issue in SLRs.
  - **Contribution:** Evidence supporting the practical feasibility and effectiveness of the CSLR concept and process activities in updating SLRs in SE, thereby offering a solution to help mitigate the intermittent update issue commonly observed in SLRs.

- **RG2 – Definition of CSLR guidelines and validation of the CSLR process and guidelines together**
  - **Evaluation Hypothesis:** The joint implementation of the CSLR process and guidelines demonstrates benefits to the SE community in supporting the update of SE SLRs.
  - **Contribution:** A set of systematic guidelines to the CSLR process describing details and examples on how to update SLRs in SE continuously. Improvements and observations on the pertinence of the CSLR process and guidelines based on SLR in SE experts' experience.

- **RG3 – Automation of two trigger activities (search and selection) of the CSLR process**
  - **Evaluation Hypothesis:** The automation solution contributes to reducing efforts and speeding up the search and selection activities of studies for the CSLR process execution.
  - **Contribution:** Automated solution (prototype) to support the search and selection of relevant new evidence for the CSLR process in SE and future directions on CSLR automation.

SUMMARY OF CONTRIBUTIONS

The concept of the CSLR introduced in this thesis brings a new paradigm for the conduction of SLRs in SE. Furthermore, this approach (concept, process and guidelines) may collaborate with the Evidence-Based Software Engineering (EBSE) area, enabling an innovative way to conduct SLRs supporting trustworthy and up-to-date evidence for SE SLRs.

In summary, this thesis presents the following contributions to the SE community:

- Definition of a concept and systematic process to support the update of SLRs in SE, contributing to keeping SLRs up to date – The CSLR concept and process;
- Proposal and evaluation of a guideline to support the CSLR process execution;
- Introduction and evaluation of automation solutions for two trigger activities (search and selection of studies) of the CSLR process as well as future directions on automation of the CSLR process.

The whole idea of this thesis comprises an "umbrella" framework. Therefore, several other projects can be derived from the definition of the CSLR process – for example, automating the proposed CSLR process completely, creating a continuous integration platform with the execution pipeline/workflow of the CSLR process including activities to support data extraction of studies and collaboration among SLR authors.

THESIS OUTLINE

The remainder of this thesis is organized as follows:

- **Chapter 1** provides a background on the main concepts addressed in the research that underlines this research thesis.

- **Chapter 2** comprises the results of two preliminary works conducted to better understand knowledge management in updating SLRs and to obtain an overview of the state-of-the-art through a systematic mapping of two trigger activities for updating SLRs: the search and selection of studies. These works provided evidence that motivate this thesis. The ideas discussed in this chapter are also presented in Publication 4 and 5 (see Preface).

- **Chapter 3** introduces the concept and process of Continuous Systematic Review (CSLR) in SE. The research detailed in this chapter is also described in Publication 3.

- **Chapter 4** exposes the validation of the CSLR process through the application of the CSLR process to published SLRs in SE. The work presented in this chapter is based on Publication 3.

- **Chapter 5** proposes guidelines to perform CSLR as well as presents an evaluation of the CSLR guidelines and process based on SE SLR experts' perceptions. The work presented in this chapter is based on Publication 1.

- **Chapter 6** proposes an automation alternative for the triggers activities of the CSLR process: search and selection of evidence. In addition, it describes future directions on CSLR process automation. The work presented in this chapter is reported mainly reported in Publication 2 with an indirect contribution from Publications 3, 6 and 7.

- The **Conclusion** concludes this thesis by highlighting the research contributions and summarizing limitations and directions for future work.

Figure 1 summarizes the overview of this thesis placing emphasis on the contribution of the research publications conducted during the Ph.D. program.

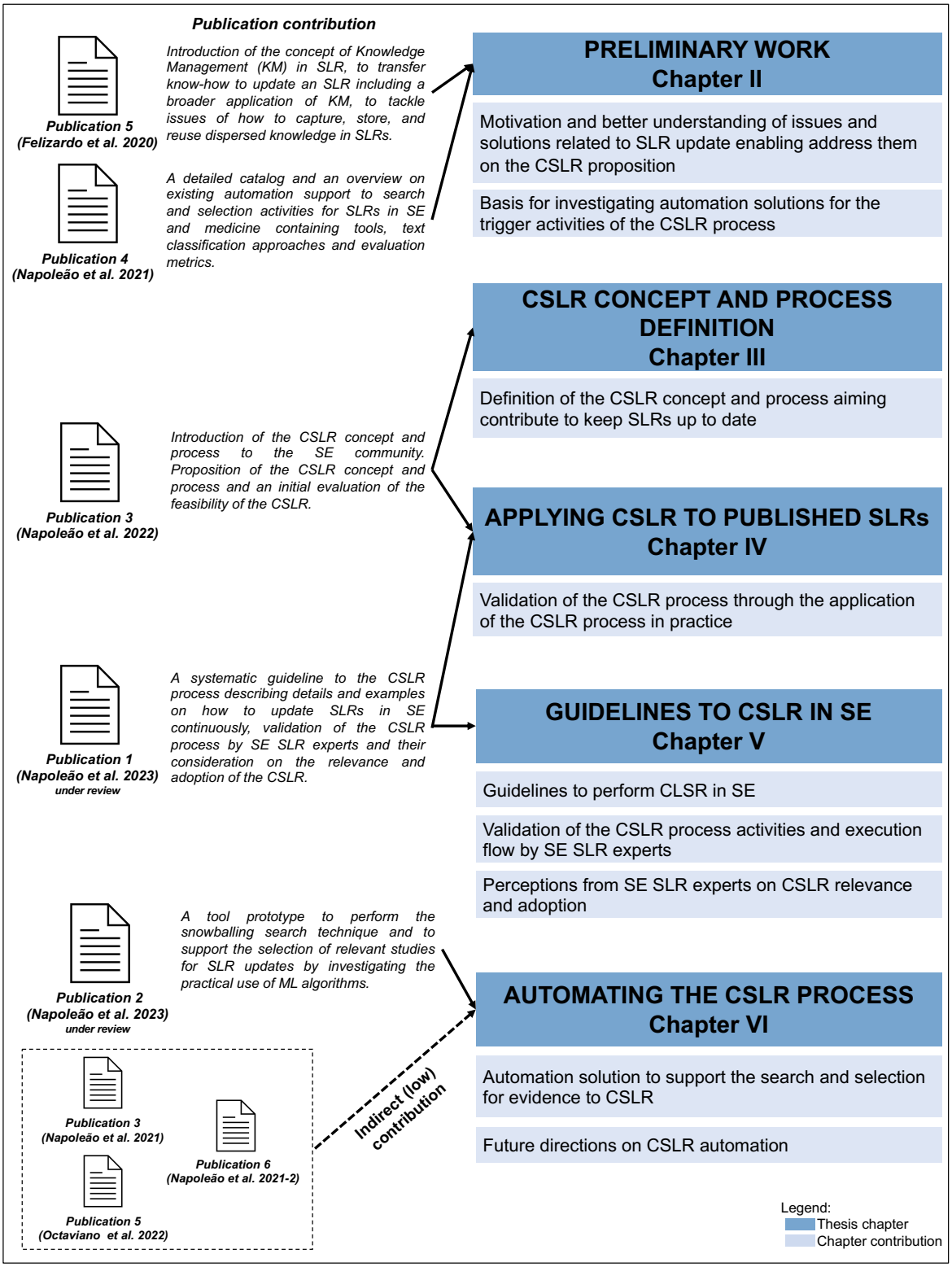


Figure 1 : Thesis overview. ©Bianca Minetto Napoleão.

CHAPTER I

BACKGROUND AND RELATED WORK

This chapter presents the concepts that underline this research thesis. It starts by describing the core concepts of the SLR process and SLR update in Section 1.1. Next, Section 1.2 introduces the concept of DevOps, the concept used as a metaphor for the creation of the CSLR process. Section 1.3 addresses the open science practices. Finally, Section 1.4 closes the chapter by describing the related works to this thesis.

1.1 FOUNDATIONS OF SLR AND SLR UPDATE

According to Kitchenham (2004), the use of SLRs in SE is encouraged to obtain the best current evidence about a topic of interest and integrate it into practice. As a result, from January 2004 to May 2016, more than 430 SLRs were published in SE (Mendes *et al.*, 2019). A more recent study (Napoleão *et al.*, 2021), with a search performed in February 2021, indicates that this number has more than doubled for SLRs and SMs in SE, surpassing 1000 studies.

Given the importance and popularity of SE SLRs, Section 1.1.1 briefly details SLR's main concepts and process from guidelines established by Kitchenham & Charters (2007) and further updated by Kitchenham *et al.* (2015). In addition, since keeping SLRs up to date is of interest to the SE community, Section 1.1.2 briefly presents the state of the art on SLR updates in SE.

1.1.1 FOUNDATIONS OF SLR

The objective of an SLR is to synthesize evidence regarding a specific topic of interest, providing a complete and fair evaluation of the state of the art about a given topic. It follows a systematic process in which its input is made of what are called *primary studies* (e.g., controlled experiments, case studies, surveys) (Kitchenham *et al.*, 2015).

An SLRs is a valuable instrument for studying research trends and answering specific questions in research projects. For example, given a particular scenario, it can be used to determine whether certain methods or practices are preferable to others or even to determine the advantages of using a tool in specific circumstances (Kitchenham *et al.*, 2015). Moreover, SLRs and SMs (a more lightweight form of SLR that follows the same process) can be used as the basis for a Ph.D. research since it is crucial that the Ph.D. student understands well the advances in the chosen research area, knows the most influential authors and gathers relevant evidence to conduct innovative research that fills a research gap (Kitchenham *et al.*, 2015). Felizardo *et al.* (2020b) surveyed the literature and SE researchers to provide an overview of the adoption of secondary studies in academia. They concluded that the use of SLRs and SMs by MSc. and Ph.D. students is valuable because, in addition to providing an overview of a research area, they provide answers for research questions that can be used as supporting arguments for a research grant application either/neither backing up decisions on a research project.

In the literature, there are well-established guidelines to conduct SLRs. They are: Kitchenham & Charters (2007) guidelines for SLRs and Petersen *et al.* (2015) guidelines specific to SM. Kitchenham *et al.* (2010) pointed out some differences between SLR and SM, but the process of conducting SLRs and SMs is the same. As a matter of fact, it is essential to highlight that Petersen *et al.* (2015) followed Kitchenham & Charters (2007) guidelines

to conduct their guidelines study. As mentioned before, in 2015 Kitchenham *et al.* (2015) published a book addressing a revision of the SLR and SM process and procedures.

The main difference between SLR and SM is that an SM is a more open form of SLR that focuses on providing a broader overview of a topic of interest (Petersen *et al.*, 2015). Other differences are related to the research questions objective, search strings, quantity of selected studies, data extraction, and synthesis (Kitchenham *et al.*, 2010). However, according to the study conducted by Napoleão *et al.* (2017), in practice, only the quality assessment of candidate studies is conducted differently between SLRs and SMs.

The SLR process involves three main phases: (i) planning, (ii) execution, and (iii) reporting the review (Kitchenham *et al.*, 2015). In Figure 1.1 we illustrate the SLR process activities, phases and possible interactions among them.

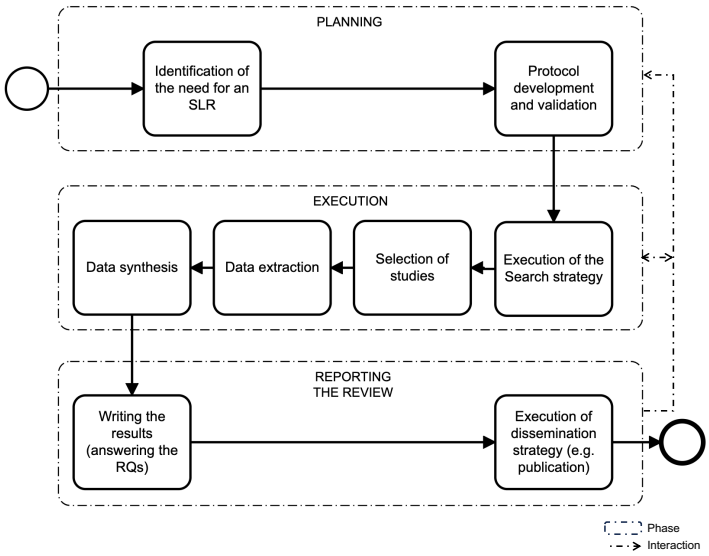


Figure 1.1 : Summary of the SLR process. Built from Kitchenham *et al.* (2015).

The three phases of the SLR process can be conducted iteratively; for example, during the execution, a need to modify the protocol can be identified, and then the process returns to

the planning phase. In the following, we provide a summary of each phase describing their activities according to Kitchenham *et al.* (2015) book.

PLANNING PHASE

This phase consists of the identification of the need for an SLR and the protocol development and validation. If other SLRs already exist on the same topic and have already answered the same research questions, there is no need to perform a new SLR. However, when the SLR is outdated, an SLR update should be performed.

In addition to verifying the existence of SLRs on the topic of interest, it is essential to consider whether conducting an SLR on a given topic will actually contribute to knowledge about the research topic as well as verifying whether it is possible to conduct the SLR with the available resources (e.g. availability of evidence and research team) (Kitchenham *et al.*, 2015). Examples of reasonable justifications for conducting an SLR and/or SM are presented in Kitchenham *et al.* (2015).

Another interesting point included in Kitchenham *et al.* (2015)'s book is the visualization of an SLR conduction as a project. The authors also advise considering review management activities such as: specifying the time schedule for the review, deciding which tools will be used in each review activity (automated or semi-automated tools, reference managers, spreadsheet tools, etc.) to support collaboration, organizing the protocol development and validation; and assigning roles to the team members (e.g. team leader – in a Ph.D. research project usually the Ph.D. student takes the lead role). Those steps are fundamental for smooth SLR conduction.

Next, the SLR protocol should be developed. The protocol is an essential element for SLR because it documents the process and the steps that will be executed in the study, including (Kitchenham *et al.*, 2015):

(i) background describing a summary of related SLRs in the research area and the justification for the need for SLR conduction;

(ii) research questions that translate the SLR objective and drive the whole SLR conduction;

(iii) search strategy used to detect relevant studies that provide evidence to answer the research questions. This protocol section must contain details on the search method(s) adopted. Among the research methods adopted to detect primary evidence for SLRs are:

automated search – search on digital libraries e.g. IEEEXplore, ACM Digital Library, ScienceDirect, etc. or indexing systems e.g. Scopus, Google Scholar, etc., that includes elaborating and adapting a search string. It is important that all the adapted search strings for each digital library chosen to be documented in the protocol to facilitate reproduction of the search (Felizardo *et al.*, 2020a);

manual search – search manually in well-known and relevant journals and conference proceedings of the research area of interest. The list of conferences and journals (including series, volume, issues, etc.) needs to be documented in the protocol with a justification;

snowballing – search method that uses a "seed set" of relevant studies already selected and analyzed their references (snowballing backward) and/or citations (snowballing forward) to detect other relevant studies (Wohlin, 2014); and

contacting experts researchers – the authors must contact well-known researchers in the research field of interest asking for their studies.

Usually, more than one research method is adopted as a search method for an SLR conduction (Kitchenham *et al.*, 2015).

(iv) study selection criteria used to include or exclude a study of the review process based on their contribution to answering the research questions as well as the steps that will be adopted to apply the criteria. For example, whether the inclusion and exclusion criteria will be applied first in the title, abstract and keywords of the studies and then, in a second moment the reviewers will consider the full text of the paper in the selection analysis. This step is usually executed by more than one researcher (or performed by one and a portion reviewed by a second one) and disagreements are solved by consensus. Besides documenting all inclusion and exclusion criteria in the protocol, it is recommended to document the results of the selection process including the borderline papers, resolution of disagreements and, most importantly, the list of included and excluded papers to facilitate future updates (Felizardo *et al.*, 2020a; Mendes *et al.*, 2020).

Another step performed during the study selection is also to assess the quality of the primary studies. This step is not mandatory for SMs, but for SLRs is highly recommended. The quality criteria is usually a checklist with elements that evaluate the quality of the research methods (e.g. surveys, case studies, experience reports, etc.) used in the primary study. In practice, the reviewers assess the quality of the studies individually and then through a consensus meeting, they solve disagreements. All adopted quality checklists, a justification for their adoption and the outcome of their application must be documented in the protocol (Kitchenham *et al.*, 2015).

(v) data extraction defines the strategy for extracting useful data from the selected papers. A common approach is to create a data extraction form that enables the extraction of bibliographical information from the studies and all the information necessary to answer the research questions. It is important to register in the protocol who is going to perform the data extraction (usually the research leader), and how the data will be recorded and stored (e.g. shared spreadsheets or using a supporting tool) (Kitchenham *et al.*, 2015).

(vi) data synthesis strategy for summarizing, combining, integrating and comparing the data to answer the proposed research questions. The data analysis can be quantitative or qualitative. Quantitative analysis is less common in SE SLR due to the heterogeneity of the available data. On the other hand, qualitative analysis methods such as narrative and thematic analysis are largely adopted to answer the SLR questions. In addition, graphics, diagrams and tables can be also used to report results (more popular in the conduction of SM) (Kitchenham *et al.*, 2015).

(vii) reporting the approach that will be used to disseminate the SLR results. An SLR is often reported as a research paper (journal or conference) and/or a chapter of a Ph.D. thesis. The protocol has to describe the dissemination approach defined by the review team.

(viii) limitations and management decisions should be registered in the protocol. The identified limitations of the SLR that cannot be addressed by the research design must be described, for example, reviewers' interpretation during the data extraction. Nonetheless, management decisions not described in the other protocol sections including role attributions, scheduling, adoption of tools, etc. should also be recorded in the protocol (Kitchenham *et al.*, 2015).

When the protocol is finished, it must be validated. The validation can be performed through a pilot test (Kitchenham & Charters, 2007) aiming to verify if any modifications

or refinements are necessary. Kitchenham *et al.* (2015) suggests reviewers to perform an internal validation which consists of trialing search strings, data extraction forms and synthesis methods. Moreover, they propose an external evaluation of the protocol by independent researchers in order to ensure that the SLR guidelines were followed. In Kitchenham *et al.* (2015) there are examples of questions that can be used to validate the SLR protocol.

Since the SLR process is interactive, the protocol is a living document that must be updated when changes are made during any phase of the SLR conduction (Kitchenham *et al.*, 2015).

EXECUTION PHASE

After protocol validation, the execution phase starts. It consists of executing the plan proposed in the study protocol. In summary, in the execution phase, the proposed search strategy is executed, then the studies are selected (inclusion, exclusion and quality criteria – if adopted – are applied), and then all necessary data to answer the study research questions are extracted and synthesized following the synthesis method defined in the SLR protocol (Kitchenham *et al.*, 2015).

According to Kitchenham *et al.* (2015), there are three factors that influence the search strategy: the first one is to decide if completeness is critical or not. Usually, completeness is a critical factor for SLRs, but for SM it might be less critical. The authors advise answering the following question to decide the criticality of completeness: "Can my research questions be answered adequately if some relevant papers are not detected by the search process?". If the answer is no, additional research methods should be considered to reach completeness. Otherwise, completeness is not a critical factor.

The second one is to make sure that the search process is validated. The validation process involves quantifying the number of studies identified and comparing it with a preset of known studies in the research area (this group of studies can be defined by preliminary reading, or suggested by an area expert). Again, if the validation process does not result in completeness and/or several relevant studies are missing, Kitchenham *et al.* (2015) suggest adding other search methods such as backward snowballing and/or refining the search string until the expected completeness is reached.

The third one is to decide on an appropriate mix of research methods. The decision of which search methods to adopt is often decided during the protocol construction and validation, but in practice, changes are made during the execution of the protocol. The most commonly adopted search methods in SE for SLRs is automated search as the main research method and a complementary research method such as snowballing (Kitchenham *et al.*, 2015). Mourão *et al.* (2020) demonstrate that the combination of automated search from the indexing system Scopus with forward and backward snowballing is the most appropriate combined strategy to search for evidence for SLRs in SE.

The next step in the execution of the protocol is to execute the search process defined previously. Basically, it consists in applying the defined inclusion and exclusion criteria. However, it is carried out in several stages. Firstly, the inclusion and exclusion criteria are applied to the set of candidate studies based on the title, abstract and keywords of these studies. This step is also known as the initial selection of studies. In practice, often the decision to include a study or not is made later in the review process due to the need for a deeper analysis of the study before taking a decision. Secondly, the selected candidate studies pass through a full-text analysis to verify if the study provides useful information to answer the research questions (Kitchenham *et al.*, 2015). Last but not least, if quality criteria were defined, they

are be applied at this last stage of the selection process. In Kitchenham *et al.* (2015) there are examples of inclusion, exclusion, and quality criteria for SLRs.

When more than one reviewer participates in the selection process, it is possible to validate it by verifying the participants' level of agreement through a Kappa analysis (Cohen *et al.*, 2010b). It is a method that was first adopted in medicine for evaluating the quality level of the selection process which has later been also adopted in SE. An example of an SE SLR that makes use of the Kappa analysis is presented in Marshall & Brereton (2013).

The search and selection of studies are two of the most labor-intensive and time-consuming activities in the process of conducting an SLR or SM (Al-Zubidy & Carver, 2019; Hassler *et al.*, 2014). They involve several challenges, for example (i) limitations of digital libraries and insufficient mechanisms to support the automated search for SLRs (Imtiaz *et al.*, 2013; Ghafari *et al.*, 2012), (ii) lack of standardization in the SE terminology, which makes the elaboration of effective search queries even more difficult (Ros *et al.*, 2017), (iii) extensive demand of human efforts to extract and analyze references and citations to perform the snowballing search strategy, and (iv) the large number of primary studies to be read and analyzed (Fabbri *et al.*, 2016). An alternative to overcome these challenges is to count on automated support (Kitchenham *et al.*, 2015). In this respect, Chapter 2 – Section 2.2 will present a preliminary work that consists of a cross-domain (medicine and SE) SM on existing automation support for searching and selecting studies for SLRs in SE.

After selecting the studies and assessing their quality (if opted), the last two steps consist of extracting the data and performing the synthesis analysis of that data to answer the proposed research questions. It is worth reminding that the chosen strategy to extract data from the selected studies, including the data extraction form, needs to be defined and justified in the protocol. In addition, the reviewers should stick to the data extraction form and to

the established extraction approach and tools (e.g. use of shared spreadsheets) to maintain consistency during the extraction process (Kitchenham *et al.*, 2015).

Last but not least in the execution phase, the reviewers must perform the synthesis of the data extracted following the synthesis method defined in the protocol (e.g. narrative synthesis, thematic synthesis, meta-analysis, etc.) to answer the research questions. Kitchenham *et al.* (2015) describe in their book detailed alternatives to perform data extraction and synthesis for quantitative and qualitative SLRs as well as SMs.

REPORTING THE REVIEW PHASE

After executing the protocol, with all data synthesized and analyzed, the SLR results are reported. Therefore, this phase focuses on writing the results to answer the proposed research questions and disseminating the results to reach potential interested researchers and practitioners (Kitchenham & Charters, 2007).

Kitchenham *et al.* (2015) highlights three points to be considered during the SLR or SM results reporting. They are:

(i) SLR readership – Who will be interested in the SM or SLR results? The main readerships of SMs and SLRs are researchers, but practitioners can also be readers. A researcher expects an SLR with a detailed description of the research method with clear traceability between the reported results, data and performed analysis. On the other hand, practitioners are focused on the results and implications for SE practice (Kitchenham *et al.*, 2015).

(ii) Report structure – Kitchenham *et al.* (2015) recommend the use of the basic elements of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-

Analyses) structure to report an SLR. It must contain a title, abstract, introduction, background, methods, results, discussion, conclusions, acknowledgments (if any) and appendices (any additional information such as primary studies excluded during the selection process, dataset external links, etc.). More recently, Kitchenham *et al.* (2023) presented an integrated set of guidelines called SEGRESS (Software Engineering Guidelines for REporting Secondary Studies) based on the PRISMA 2020 statement to report SMs and qualitative SLRs. The SEGRESS guidelines showed suitable to address reporting problems found in SE SLRs. Nonetheless, aiming to facilitate the adoption of SLR results in practice, Cartaxo *et al.* (2016) propose, evaluate and make available a template to produce evidence briefings, a designed one-page document to serve as a medium to transfer the knowledge acquired from SLRs to practice.

(iii) Validating the report – It is important that all authors review the final report carefully to ensure that: all research questions were clearly answered; the methodology is correctly described providing traceability among the research questions, data collection, synthesis and conclusions; and if the type of report fits the target readers. In addition, if possible inviting an external researcher to evaluate the SLR can be useful to bring independent insights to the SLR or SM (Kitchenham *et al.*, 2015). In practice, when the authors decide to publish the SLR in peer-reviewed sources (e.g. conference and/or journal) their reviewers end up acting as independent reviewers of the SLR or SM under evaluation.

1.1.2 FOUNDATIONS OF SLR UPDATE

Considering the importance of SLRs for SE, they should be updated to include new evidence that emerged after their conduction (Garcés *et al.*, 2017). The overall objective to performing an SLR update is to incorporate the new information into the existing SLR in order to keep the existing SLR useful and relevant for researchers and practitioners. Moreover, with

the emergence of new evidence, different answers for the explored research questions can be identified, which can lead to different findings and conclusions when compared to the original SLR. The findings and conclusions of the SLR can even be reaffirmed with the new evidence identified (Mendes *et al.*, 2020).

In summary, updating SLRs is important for reasons such as (i) to ensure that the SLR findings are still generalizable by providing an up-to-date view of the current state-of-the-art on a research topic; (ii) to improve the reliability and accuracy of results presented in the SLR since new evidence can reinforce the findings as well as reduce the risk of bias; and (iii) to identify new research trends which can inform future research and guide practice on a SE research topic (Higgins *et al.*, 2008; Higgins & Green, 2011; Mendes *et al.*, 2020).

The need to update SE SLRs is known by the SE community. Several studies address different aspects of SLR updates: searching for new/updated evidence (Felizardo *et al.*, 2016; Wohlin *et al.*, 2020), selecting updated evidence (Watanabe *et al.*, 2020), deciding on whether to update or not (Mendes *et al.*, 2020), and experience reports (Garcés *et al.*, 2017; Felizardo *et al.*, 2020a). See Section 1.4 for details on these studies.

The decision of whether to update or not an SLR is a research topic that was first explored and established in medicine. Garner *et al.* (2016) proposed a decision framework that includes three steps addressing questions that must be considered sequentially to guide the update decision. A few years later, Mendes *et al.* (2019) introduced Garner *et al.* (2016) framework to SE naming it 3PDF (Third-Party Decision Framework). In the following year, the study of Mendes *et al.* (2019) was extended as a journal paper (Mendes *et al.*, 2020). Figure 1.2 summarises the 3PDF including its steps and questions to be considered during the evaluation of the need for updating of SLRs in SE.

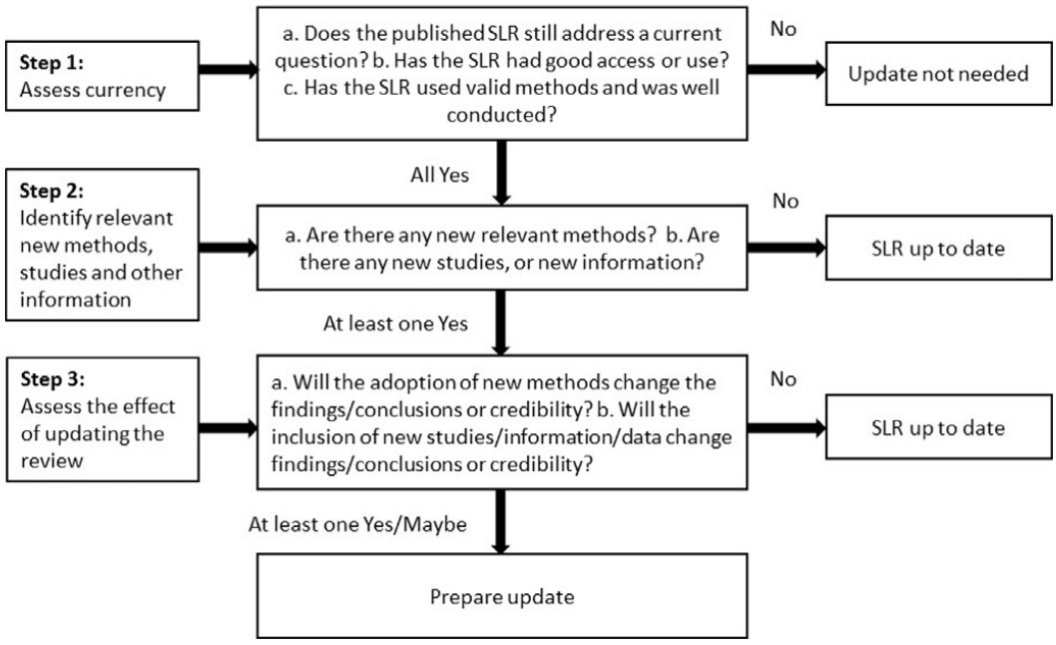


Figure 1.2 : 3PDF to assess SLRs for updating (Mendes *et al.*, 2020). ©2020 Elsevier.

As can be seen in Figure 1.2, the goal of *Step 1* is to assess how current and actual is the SLR under investigation by asking three questions (1.a, 1.b and 1.c). The answer for all three questions must be affirmative to move on to Step 2. Otherwise, the conclusion is that the SLR does not need an update. The goal of *Step 2* is to verify if there are any relevant new methods, studies and or information published after the SLR publication. At least one answer to the two questions (2.a or 2.b) asked in this step must be affirmative to pass to the last step. Lastly, *Step 3* assess the consequences/effects of updating the SLR. As long as one of the two questions (3.a or 3.b) of Step 3 is partially affirmative, it is recommended to proceed with the update. Nonetheless, regarding question 3.b, which is difficult to determine if the inclusion of new studies/information/data change findings/conclusions or credibility of the SLR under assessment, Mendes *et al.* (2020) recommend answering this question by performing an informal analysis based on reading title, abstract of potential candidate studies retrieved from the update search.

1.2 FOUNDATIONS OF DEVOPS

DevOps can be defined as a set of practices that combines software development and operations. The benefits of DevOps include encouragement of collaboration among team members, faster release of features, improved monitoring systems in production, among others (Humble & Farley, 2010). CI (Continous Integration) and CD (Continous Delivery), DevOps practices, aim to streamline the software development lifecycle delivering software faster, more frequently and with higher quality. These practices include automation (tools) and the cultural mindset of integrating changes constantly to help teams rapidly and reliably deploy and innovate for their customers. These tools should automate manual tasks, and teams manage and keep the control of complex environments (Humble & Farley, 2010; Meyer, 2014). Figure 1.3 summarizes the CI, CD and observability stages of the DevOps practice.

As illustrated in Figure 1.3, during the CI developers frequently merge their code changes into a central repository, next this process triggers an automated build and testing processes. The main objective of CI is to build software at every change. A build consists of a process for integrating all versions of the software code together and verifying if the software works as a cohesive and updated unit (Duvall *et al.*, 2007; Meyer, 2014).

A traditional CI scenario starts with a developer committing their code to the version control repository. Concurrently, the CI server installed on the integration build machine actively monitors the repository for any updates, checking at regular intervals. Once a new commit is detected, the CI server initiates the integration process. It retrieves the latest version of the code from the version control repository and proceeds to execute a build script (compiling code, bundling resources, and creating the necessary output files) for seamlessly integrating the software components (Duvall *et al.*, 2007).

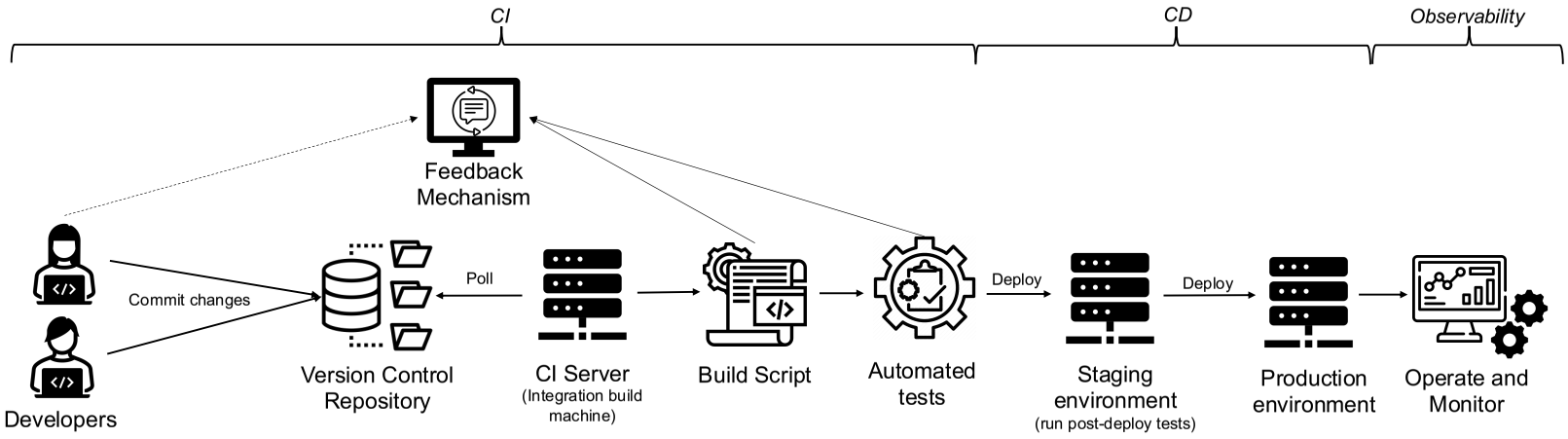


Figure 1.3 : DevOps main components. Adapted from Duvall *et al.* (2007) and Humble & Farley (2010).

Following the completion of the build process, the CI server generates feedback by sending notifications containing the detailed build results to the designated project members. The CI server's involvement does not cease after a single build. It continues to actively poll the version control repository for any further changes, ensuring that subsequent code commits are swiftly identified and follow the next steps of the CI process to be incorporated into the application (Duvall *et al.*, 2007). Once the build is complete and the application artifacts are generated, automated tests are executed (e.g. unit tests, integration tests, etc.) to ensure that the application remains reliable and consistent. Then the test results are shared with the designated project members (Humble & Farley, 2010; Meyer, 2014).

CD is the next stage after CI, once the changes pass the automated tests (CI) they are automatically deployed to a staging or production environment. A common practice is to deploy in a staging environment first and then run a "post-deploy test", i.e. user acceptance tests to validate aspects such as usability and performance. If the application passes the user acceptance testing and meets the quality criteria, the CD process automatically deploys the application to the production environment (Humble & Farley, 2010).

Last but not least, once the application is deployed, it enters the observability (operational and monitoring) phase where it is actively used by end-users. Continuous monitoring helps in tracking application performance, infrastructure health, and user experience. In addition, feedback is collected from users, stakeholders and monitoring systems to understand how the application performs in the real-world scenario. Finally, the process iterates starting again with improvements and new features through the commit of code changes (Humble & Farley, 2010).

Besides all the mentioned components in Figure 1.3, Humble & Farley (2010) highlight in their book the need of having a team with a mindset geared towards continuously applying

small merges, prioritizing the application working. Another important step is to choose a CI/CD integration tool. There are several established CI/CD tools available on the market, such as Travis CI[1], Jenkins[2], GitHub Actions[3], etc.

In the same way that the CI/CD process makes it possible to integrate source code in an automated and fast way, with the CSLR process, we intend to facilitate the detection and integration of new evidence into SLRs as they emerge. The general idea is to use the CI/CD concepts as a metaphor for building the CSLR process.

In a CI environment, a developer pulls code from a repository and works on it; at any time the code can be integrated into a CI server. After all modifications are performed, the source code is committed to the code repository. This action triggers the execution of the CI/CD tool (e.g. Jenkins pipeline or GitHub Actions workflow) (Jenkins, 2023; GitHub Docs, 2023). After the pipeline/workflow execution, the code becomes available in the staging and then in the production environment. Similarly, in the CSLR process, the inputs for the CSLR will be data from the SLR protocol. An automated monitoring system will run periodically searching and selecting potential new evidence. As soon as a piece of new evidence or a set of new evidence is detected, the trigger for CSLR pipeline/workflow execution will be initiated. After all the full pipeline/workflow execution, the SLR will be ready to go through the remaining steps and be completely updated. The whole CSLR process as well as its guidelines will be presented in Chapters 3 and 5 respectively.

[1] https://www.travis-ci.com

[2] https://www.jenkins.io

[3] https://github.com/features/actions

1.3 FOUNDATIONS OF OPEN SCIENCE

As stated by Mendez *et al.* (2020), open science can be defined as *"the movement of making any research artifact available to the public. It ranges from the disclosure of software source code (open source) over the actual data itself (open data) and the material used to analyze the data (such as analysis scripts and open material) to the manuscripts reporting on the study results (open access)"*.

Open science is essential to keep moving forward in the SE scientific research community. For example, how many times during an analysis of a research paper is it impossible to locate the source code of the proposed application or the data used to run the experiments? Examples like this place barriers on repeatability (capacity of the same team with the same experimental setup reliably repeat their experiment), replicability (capacity of a different team to be able to use the same experimental setup and obtain the same results), and reproducibility (capacity of a different team using a different setup obtain the same result using artifacts independently developed) of studies (Mendez *et al.*, 2020).

In the context of SLRs, it is expected that SLRs can be repeated and replicated. In addition, the guidelines of Kitchenham & Charters (2007); Kitchenham *et al.* (2015) emphasize the need to document in the protocol all the SLR steps in detail. One consequence of the lack of details in the SLR protocol is that it makes the update difficult (Garcés *et al.*, 2017; Felizardo *et al.*, 2020a).

According to Mendez *et al.* (2020), SE researchers are still facing challenges to apply several open science concepts in their research activities, such as performing open peer reviews. On the other hand, the open science movement is gaining force in the SE community. Recent conferences and journals had significant participation of authors disclosing their research data. Publishers such as ACM are providing open science badges for publications that meet

open science criteria to encourage authors to participate in open science initiatives. Some SE researchers are sharing their research preprints on arXiv[4], managing their research project on GitHub[5] or OSF (Open Science Framework)[6], and making their research data available on Zenodo[7] or Figshare[8].

Another relevant aspect regarding open science and SLRs is to make accessible and reusable all data generated by the researchers' analysis during the conduction of SLRs (Kitchenham *et al.*, 2015). It includes elements such as the set of included and excluded studies, the data (e.g., spreadsheets) created during the search strategy execution (e.g., studies returned from each digital library and the adapted search string for each digital library), the selection procedure and the data extraction form with the extracted data from the included studies. All the mentioned elements are essential to enable a smoother SLR update (Felizardo *et al.*, 2020a). Therefore, SE researchers can use external platforms (e.g., Github, Zenodo) to make the SLR data available.

For the concept of CSLR presented in this thesis, the open science practices mentioned for the construction of a repeatable and replicable protocol and data availability (from the original SLR and the SLR updated) are critical elements to enable the execution of the CSLR process. Furthermore, we expect that the adoption of open science practices in SLRs shall contribute to promoting, even more, the adoption of open science practices by the SE community.

[4] https://arxiv.org

[5] https://github.com

[6] https://osf.io

[7] https://zenodo.org

[8] https://figshare.com

1.4 RELATED WORK

The work of Mendes *et al.* (2020) brought and adapted the definition of SLR update as a new version of a published SLR that includes new primary studies that can come from new search strategies (e.g., snowballing, manual or database search). It also can include new methods, such as a more elaborated quality checklist to assess the new studies and/or a new approach to combine evidence (e.g., thematic analysis); and new analyses, such as a new research synthesis method. In the same work, the authors provided recommendations on when to update SLRs in SE. They proposed and evaluated their work using the 3PDF, a support decision mechanism to assess the need to update an SLR in SE. The results of this work were integrated as an activity of the CSLR process presented in this thesis (Napoleão *et al.*, 2022b).

In SE, the number of SLR publications is increasing year by year. In February 2021, the SE area had more than 1000 SLRs and SMs (Napoleão *et al.*, 2021). However, until 2019, only 20 SLRs in SE were updated (Mendes *et al.*, 2020). The need to update tertiary studies (an SM of SLRs and SMs (Kitchenham *et al.*, 2015)) is also starting to gain the attention of the SE community. Two tertiary studies have been updated and published (da Silva *et al.*, 2010; Barros-Justo *et al.*, 2021).

In more mature evidence-based areas such as medicine, SLR update has a consolidated process (Moher *et al.*, 2008; Shekelle *et al.*, 2011). Also, in medicine, there has been a concern for some years to keep SLRs up to date (Shojania *et al.*, 2007; Garner *et al.*, 2016). An alternative to deal with the consequences of SLR being outdated in medicine was the proposition of Living Systematic Review (LSR) (Elliott *et al.*, 2017). An LSR is an SLR type that is continually updated, incorporating relevant new evidence as it becomes available (Elliott *et al.*, 2017). Despite being a recent concept, the Cochrane community[9] has supported

[9] https://community.cochrane.org/review-production/production-resources/living-systematic-reviews

the conduction of LSRs, and even released guidelines for the production and publication of Cochrane LSR (Brooker *et al.*, 2019). The LSR principle was also a crucial instrument in defining the concept and process of CSLR in SE, the result of the doctoral research presented in this thesis (Napoleão *et al.*, 2022b).

In the SE area, several isolated initiatives address the SLR updates. Indeed, in 2008 the study by Dieste *et al.* (2008b) proposed an improved process to perform SLRs in SE by conducting an SLR and a subsequent update. As a result, the authors mention suggestions to facilitate the update of the SLR. For example: ensuring that the authors are involved in understanding all topics related to the review, updating the glossary of the SLR, and carefully checking contradictory results to avoid incorrect generalizations. This thesis proposes explanatory CSLR process guidelines with examples based on a summary of current research on updating SLRs.

Identifying new relevant evidence for updating an SLR is one of the initial steps in identifying the need and possibility of updating an SLR. If there is no new primary evidence, it is useless to update an SLR (Mendes *et al.*, 2020), indicating that the SLR research topic under evaluation has not evolved since the analysis performed by the published SLR. In SE, studies investigated search strategies dedicated to searching for evidence for updating SLRs. Felizardo *et al.* (2016) proposed the use of the forward snowballing technique (Wohlin, 2014) (i.e., citations analysis from the included studies in the original SLR – also known as "seed set") as a search strategy to update SE SLRs. Their results showed a reduction of more than five times the quantity of primary studies to be analyzed during an SLR update. Four years later, in 2020, Wohlin *et al.* (2020) further investigated the use of the forward snowballing technique to update the SE SLR process. Their study proposed and evaluated guidelines for the search strategy to update SLRs in SE. They demonstrated that using a single iteration of the forward snowballing technique, using Google Scholar as the search engine, and employing

the original SLR plus its primary studies as a "seed set", is the most cost-effective way to search for new evidence for updating an SLR. The contributions of these two studies were incorporated into the CSLR process and guidelines.

The selection of evidence resulting from the execution of the search strategy is a laborious and time-consuming activity during the SLR update process that demands knowledge about the research topic investigated (Fabbri *et al.*, 2013; Felizardo *et al.*, 2020a). To the best of our knowledge, only two studies specifically investigate the selection of new evidence for updating SLRs in SE: Felizardo *et al.* (2014) and Watanabe *et al.* (2020). In Felizardo *et al.* (2014), visual text mining is explored to support the selection of new evidence (primary studies) for SLR updates. The tool presented, called Revis, connects the new evidence with the evidence of the original SLR applying the KNN (K-Nearest Neighbor) Edges Connection technique and presenting the results in two different visualizations: content-map and Edge Bundles diagram. The results showed an increase in the number of studies correctly included compared to the traditional manual approach.

In Watanabe *et al.* (2020), the authors also take advantage of the fact that a published SLR that needs updating already has a list of included studies. They investigated the use of text classification techniques, including supervised machine learning algorithms (Decision Trees and Support Vector Machines), to make the initial selection of primary evidence (based only on the title, abstract, and keywords of the studies) to update SLRs. The authors achieved a median recall of 0.93 and precision of 0.92 and reduced the number of studies that need to be analyzed by reviewers by a factor of 0.62 on average. The study showed the potential of using automated techniques to reduce the efforts required to select studies for SLR updates.

Regardless of the two specific automated approaches to support the selection of studies for SLR updates, Wohlin *et al.* (2020) highlight the significance of having more than one

researcher selecting studies during the conduction of an SLR update in order to reduce bias. The CSLR process and guidelines systematically compile the results of these existing selection investigations for SLR updates. Furthermore, some alternatives applied to the SLR context could also have the potential to work for SLR updates. The work of Napoleão *et al.* (2021a) (also a preliminary work to this thesis – see Section 2.2) mapped existing SLR automation tools to support the search and selection of studies for SLRs.

In SE, there are experience reports on SLR updating (Garcés *et al.*, 2017; Felizardo *et al.*, 2020a). These reports present suggestions, lessons learned, and concerns during the performance of several SLR update activities. Garcés *et al.* (2017) reported four lessons learned based on the author's experience in updating SLRs. They are: (i) use automated tools to support the SLR update activities, (ii) obtain and maintain as much information from the original SLR, (iii) involve researchers who participated in the previous SLR conduction, and (iv) reuse the previous SLR protocol. Felizardo *et al.* (2020a) performed an experience report addressing how to transfer the know-how of SLRs to facilitate their updates (preliminary work of this thesis – see Section 2.1). They instantiated Nonaka-Takeuchi's knowledge management model to the SLR update scenario. They provided a list of the same lessons learned mentioned in Garcés *et al.* (2017) but added quantitative tables to support the data merging between the original SLR and the update. All these practical experiences and lessons learned highlighted by these experience reports will be considered during the CSLR process and guidelines construction.

Nepomuceno & Soares (2018, 2019, 2020) investigate, in three different studies, aspects related to SLR updates: In the first study (Nepomuceno & Soares, 2018), the authors conducted a survey with experienced researchers on SLR conduction to how SLRs can be maintained (updated), including the benefits and drawbacks of the maintenance process. As a result, almost 70% of the respondents mentioned interest in keeping their SLR up-to-date, but

they were concerned about the effort required to do it. In addition, 71% of the participants were optimistic about sharing their SLRs in repositories. The results presented in this study corroborate the concept of CSLR and reinforce the need for a structured process and guidelines to support SLR updates in SE.

In the following year, in Nepomuceno & Soares (2019), the authors performed a systematic mapping to understand how SLRs have been updated in SE and a survey to obtain opinions from SE researchers about SLR updates. They concluded that the concerns about SLR updates are increasing due to the risk of losing their impact over time. Therefore, actions supporting the SLR updating issue in SE are highly relevant to EBSE. In addition, the study's results presented findings related to the ones described in Garcés *et al.* (2017) and Felizardo *et al.* (2020a) regarding lessons for performing SLR updates, such as having a researcher from the original SLR as a participant or consultant during the updating process, and performing protocol changes only if necessary and including a reasonable justification. These findings are also considered in the CSLR process and guidelines construction.

As an extension of this previous study, in 2020, Nepomuceno & Soares (2020) narrowed down their investigation on plagiarism aspects related to SLR updates. They concluded that plagiarism in SLR does not differ from other areas of research, but in the context of SLR updates, some artifacts such as the reuse of data and results can lead a researcher to commit plagiarism (or self-plagiarism – reuse of his/her own previously published work without proper citation or acknowledgment). Nepomuceno & Soares (2020) affirm that plagiarism might happen without proper care by the authors, but they also provided a list of good practices to avoid plagiarism in an SLR update. Unlike Nepomuceno & Soares (2020), the CSLR idea addresses open science concepts instead of plagiarism concerns.

CHAPTER II

PRELIMINARY WORK

This chapter presents two preliminary works (Felizardo *et al.*, 2020a; Napoleão *et al.*, 2021a) that motivate this thesis.

Section 2.1 reports on the first preliminary work (Felizardo *et al.*, 2020a) which introduces the concept of Knowledge Management (KM) into the SLR context to transfer know-how to update an SLR, including a broader application of KM to tackle issues of how to capture, store, and reuse dispersed knowledge in SLRs. This work contributes to a better understanding of existing issues and solutions related to SLR updates, enabling addressing them on the CSLR proposition (see Chapters 3, 4 and 5).

Section 2.2 presents the second preliminary work (Napoleão *et al.*, 2021a) which provides an overview of existing automation support to the search and selection activities for SLRs in SE and medicine. It covers tools, Text Classification (TC) approaches (including TM and ML) and evaluation metrics. This SM is used as a basis for investigating automation solutions for the two trigger activities (search and selection of studies) of the CSLR process presented in Chapter 6.

2.1 KNOWLEDGE MANAGEMENT FOR PROMOTING UPDATE OF SYSTEMATIC LITERATURE REVIEWS: AN EXPERIENCE REPORT

The documentation that details a given SLR should enable its update by any other researcher, but such documentation has not many times received enough attention – in the sense that updates can not be adequately conducted based only on such documentation. Lack of details about the SLR protocol, lack of information regarding the planning and execution

process, and even an ill-reported synthesis are some of the problems that jeopardize a smoother update. It is worth highlighting that performing SLR produces an important intellectual capital (tacit knowledge in the mind of researchers who conducted the SLR) and the updates of the previous SLR could benefit from such knowledge and experience previously acquired. Therefore, there is also a tacit knowledge not mentioned in the documentation (Fabbri *et al.*, 2013). Unlike explicit knowledge (knowledge that can be documented), tacit knowledge can easily be lost if it is not adequately externalized. According to Nonaka & Takeuchi (1997), tacit knowledge is undoubtedly valuable, but the fact is that it is hard to make it explicit, requiring innovative strategies to acquire and process it.

Motivated by the lack of initiatives in the literature that discuss Knowledge Management (KM) in SLR, in this section we present an experience report where we introduce the concept of KM in SLR, aiming at transferring know-how to the update of SLR. For this, we used the Nonaka and Takeuchi model of KM, also known as Knowledge Spiral or SECI process (Nonaka & Takeuchi, 1997). The Nonaka and Takeuchi model considers creating knowledge as a continuous and dynamic interaction between tacit and explicit knowledge.

2.1.1 EXPERIENCE REPORT

This section describes the difficulty in transferring know-how to update SLRs. We use two SLR updates conducted by us to illustrate some of the knowledge-capturing and sharing issues by employing the SECI process.

CONTEXT OF THE EXPERIENCE REPORTED

The main goal of KM is to promote knowledge storage and sharing, as well as the emergence of new knowledge (O'Leary & Studer, 2001). KM formally manages knowledge

resources to facilitate access and reuse (Zack & Serino, 2000). There are two main types of knowledge (Nonaka & Takeuchi, 1997): the first is *tacit knowledge* that typically remains only in people's minds and involves intangible factors, such as beliefs, perspectives, values and intuition, encompassing knowledge not articulated and associated to the senses, movement skills, physical experiences, intuition, or implicit rules of thumb; the second is *explicit knowledge* that can be documented and, hence, shared by several individuals.

The concept of knowledge conversion explains how tacit and explicit knowledge interact (Nonaka & Takeuchi, 1997), as shown in Figure 2.1. It presents the classic, well-known Knowledge Spiral (also known as the SECI process). The SECI process illustrates the four different modes of knowledge conversion:

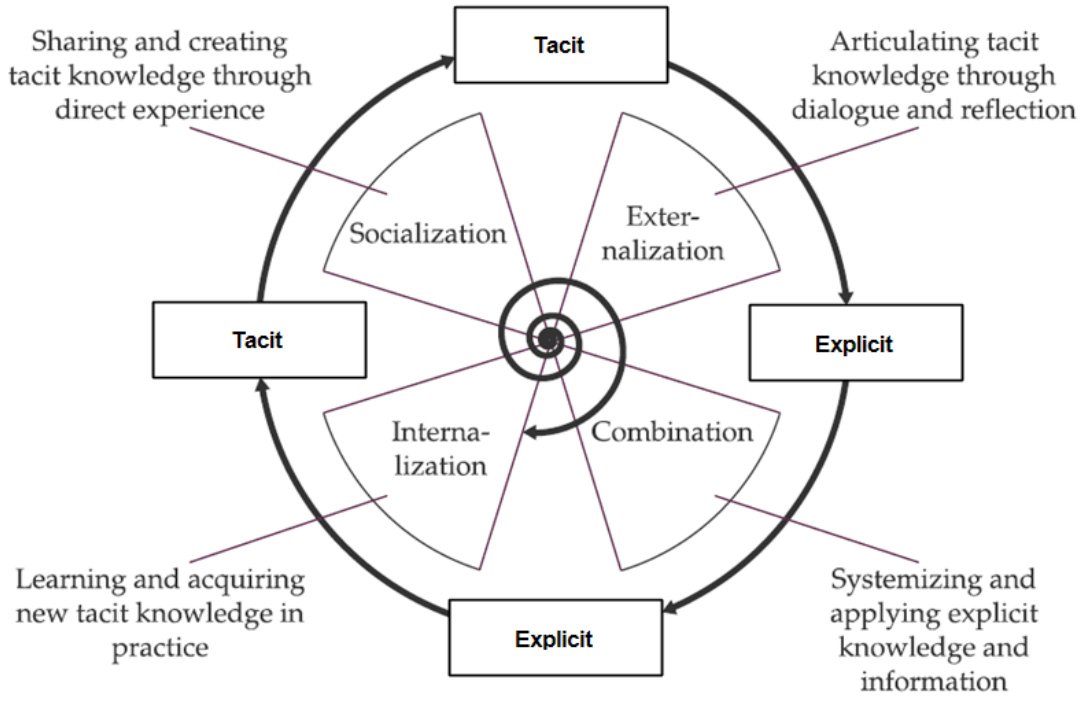


Figure 2.1 : SECI Process of Knowledge Spiral (Felizardo *et al.*, 2020a). Reproduced with the permission of IEEE.

(i) *Tacit to Tacit (Socialization)*: transmission of tacit knowledge from one individual to another;

(ii) *Tacit to Explicit (Externalization)*: transformation of tacit knowledge into explicit knowledge through symbolic representation;

(iii) *Explicit to Explicit (Combination)*: systematization and application of knowledge combining different sets of explicit knowledge to generate new explicit knowledge; and

(iv) *Explicit to Tacit (Internalization)*: learning and acquisition of new tacit knowledge from the incorporation of explicit knowledge.

Figure 2.2 illustrates the amount of knowledge acquired or kept by the teams in two moments: (i) while conducting an SLR (left side); and (ii) while updating an SLR with the same team (upper right) and with a new team (bottom right). The team that conducts SLR can accumulate implicit knowledge (referred to as "Tacit Knowledge") acquired from a large number of primary studies usually found in the first execution (reading and selecting studies), as well as the entire process and tasks through which the SLR is executed. This knowledge can ground tasks of the SLR update when the same team does that, e.g., new studies relevant to include during the update can be more easily recognized and possibly less time/effort is spent to the understanding of new evidence. However, even when documentation of SLR is available, researchers of a new team updating an SLR will have difficulty understanding such implicit knowledge.

Hence, the main problem is the difficulty in transferring SLR know-how from the original team (the knowledge providers) to the new team that updates SLR (the knowledge users). The main challenge is then to create a systematic way to capture and share knowledge between teams, enabling SLR update teams to benefit from previously accumulated knowledge.

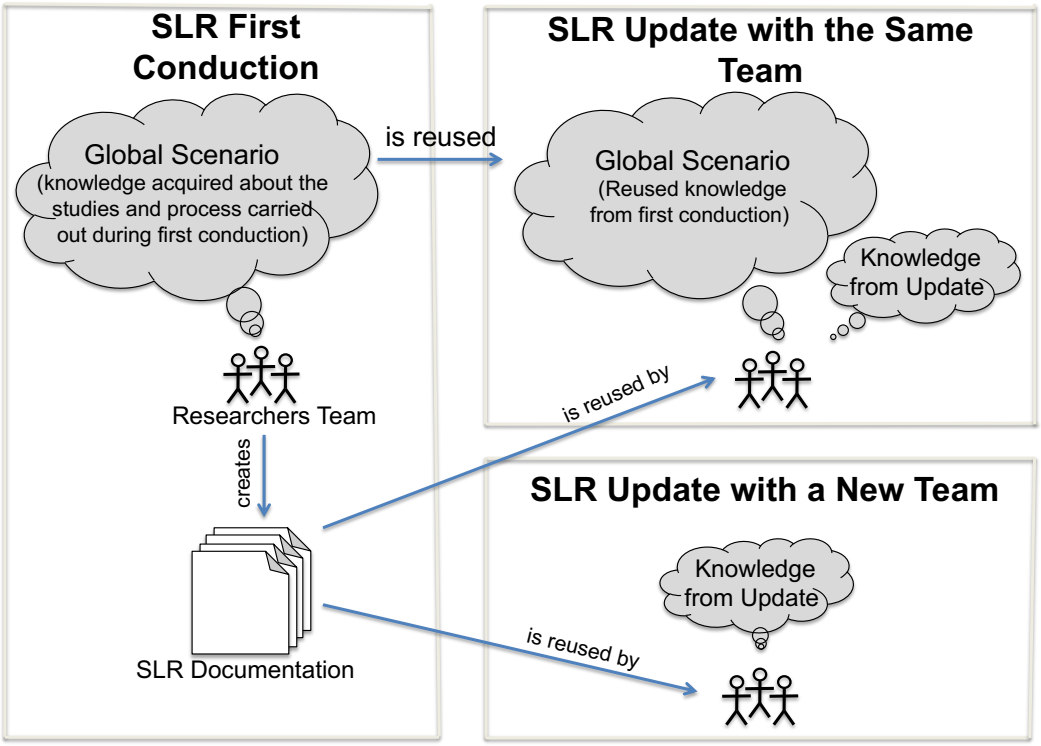


Figure 2.2 : Tacit knowledge acquired when conducting an SLR and reusing this knowledge when updating SLR involving the same team; otherwise, no tacit knowledge reused when updating SLR with a new team (Felizardo *et al.*, 2020a). Reproduced with the permission of IEEE.

In this scenario, our proposal is the adoption of KM principles, entailing formally easier access and reuse of such knowledge, as well as the transfer of best practices to other researchers.

APPLICATION OF SECI PROCESS IN SLR

This section reports the application of the SECI process in SLR to further facilitate their update. This section is also illustrated with two SLRs (hereafter referred to as SLR1 and SLR2), which were part of research projects, in which the first conduction was useful to

identify the state of the art and research opportunities on a given topic. In particular, SLR1 analyzed reference architectures and reference models for Ambient Assisted Living (AAL) (Garcés *et al.*, 2020) whereas SLR2 investigated the use of service orientation for robotic system development (de Oliveira, 2015). After 22 months (for SLR1) and 40 months (for SLR2), they were updated (referred to as SLR1' and SLR2') to verify the relevance and originality of the projects being conducted in our research group.

Figure 2.3 represents a model that puts together the Knowledge Spiral (SECI Process) and the processes of conducting and updating an SLR. This figure was created to facilitate the visualization of what SLR activities occur considering the phases of the SECI model. The SLR process phases (i.e., planning, conduction, and reporting) occur along the modes of knowledge conversion; e.g., during the planning phase, modes of knowledge conversion, "socialization" and "externalization" are identified. Besides, the first cycle corresponds to the first conduction of an SLR, while the next cycles represent an update. For each mode of knowledge conversion, we also established a set of activities to favor updates. In the following, we present how the four modes of knowledge conversion were treated in the conduction and update of the two SLRs.

Socialization – it involves sharing tacit knowledge among team researchers when planning the SLR update. The reuse of a protocol is recommended during the update of SLRs, since the establishment of a new SLR protocol usually consumes considerable time and effort and it is a complex task (Babar & Zhang, 2009). Even when reusing the protocols of SLR1 and SLR2, it was necessary to refine them, and during this refinement, new tacit knowledge was accumulated from tasks such as the following:

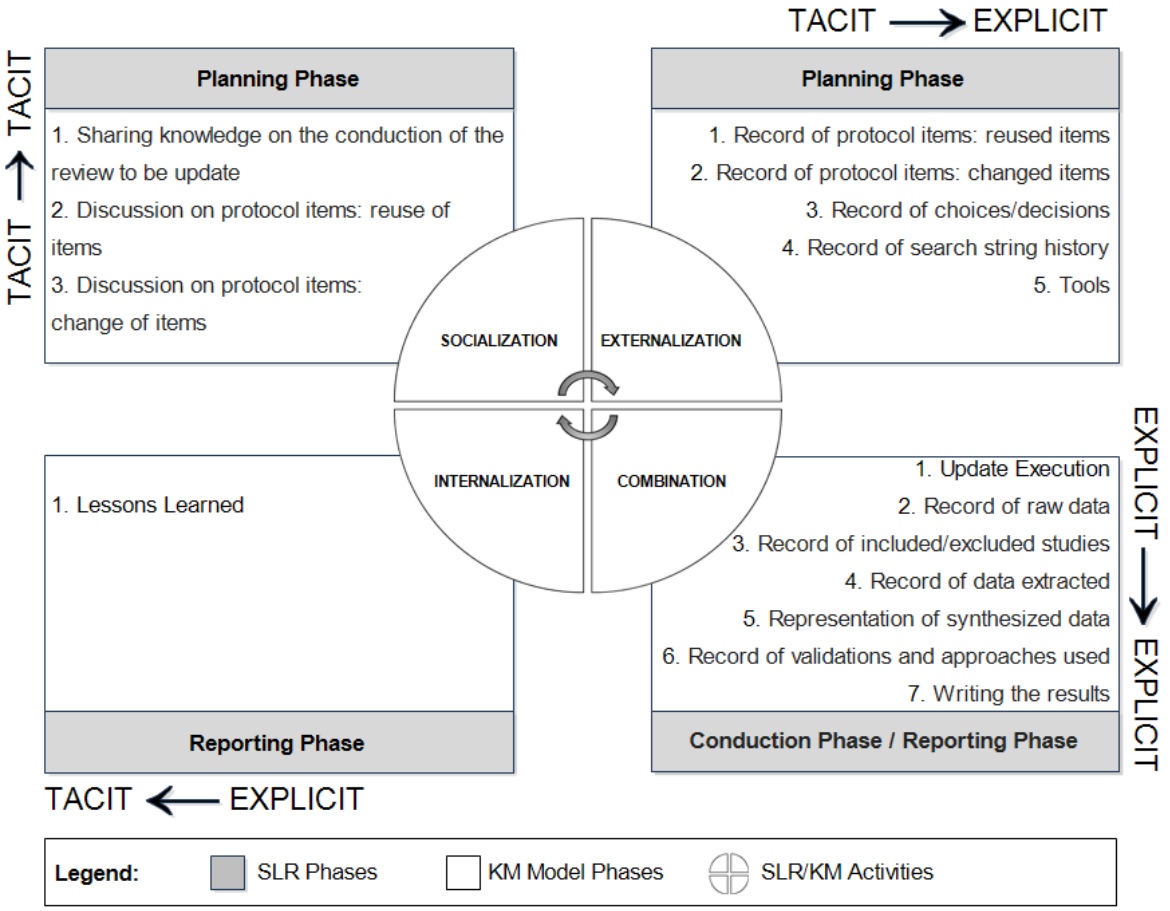


Figure 2.3 : Knowledge Spiral applied to the process of SLR conduction and update (planning, conduction, and reporting), where the internal cycle refers to the first conduction of an SLR and the next cycles refer to its updates (Felizardo *et al.*, 2020a). Reproduced with the permission of IEEE.

- *Trying to establish more adequate research questions* – The scope of the SLR can change during an update, by enlarging or reducing. RQs not answered due to the lack of evidence during the previous version could be included in the updated version. New RQs can also be defined. For answering these RQs, a reassessment of studies previously included (and even those excluded) must be conducted, together with the new studies found during an update. When updating, the protocols of both SLR1 and SLR2 were fully reused without considerable changes, as well as the RQs previously defined. RQs

of SLR1 and SLR2 were extensively refined during update planning, as researchers discussed and continually gained better knowledge about what could effectively be answered from the studies;

- *Calibrating the search string using diverse terms and synonyms* – New keywords/synonyms in the research topic of SLR1 and SLR2 did not emerge after the last conduction, therefore the search strings of SLR1 and SLR2 were reused. Some libraries and their search engines do not support SLRs completely and show limitations in Boolean expressions, e.g., the number of characters or terms is sometimes limited. In our case, SLR1 involved 18 terms in the search string and multiple substrings combining terms were required to solve this limitation. Therefore, the reuse of adapted strings (for saving time/effort) and the team's past experience was a good strategy, although some minor changes were required in the strings, once the search engines of some digital libraries have changed. The strings were readapted to each digital library and a test verified their feasibility and adequacy using a pre-selected set of relevant studies;

- *Choosing digital libraries* – For the search of studies in SLR1' and SLR2', the same digital libraries used in their previous version, (ACM Digital Library, IEEE Xplore, Springer, ScienceDirect, Compendex, Scopus, and Web of Science), were adopted. The inclusion of new search sources for the finding of more studies (expanded search) was not considered.

  In summary, the reuse of adapted search strings and their readaptation is easier than proposing new ones; therefore, such a practice should be adopted for updating SLRs. The history of these changes could be recorded in a new field in the protocol, named as: "Practical issues related to the handling of digital libraries".

- *Determining the search period* – It covers the period not covered by the last iteration and usually refers to the period between the last SLR and the current date. Some studies show a delay between their publication date and their access for downloading in the digital libraries. For instance, although a previous SLR covered the publication date of the studies, they may not be found. During the update, the starting date of the search should be anticipated by some months. In our case, we anticipated by six months;

- *Elaborating data extraction forms* – The forms were tailored to the new RQs for data collection and answering questions.

We have learned that time spent on checking the protocol was less than the time for the first preparation of the protocol, as also observed in SLR1' and SLR2', whose protocol checking consumed less than 1/5 of the time of the first protocol preparation.

Externalization – it involves the recording of protocol, where explicit and implicit choices of what should be written down are being made. It is in these implicit choices that a lot of important tacit information is lost.

SLRs usually involve intense data/file management and the adoption of supporting tools thus become indispensable. From the same perspective, these tools are also important in updates and offer advantages, such as ease and speed in the selection of new studies and increased trustworthiness of the results.

These tools are even more important for SLR updates when they store detailed information (e.g., authors, title, abstract, keywords, publication year, venue, and digital libraries where they were found) of all studies included and excluded and information on the SLRs (e.g., inclusion and exclusion criteria, digital libraries used, search execution date, search string used, years covered by the searches, data extraction forms, other search techniques

adopted, as manual search, contact of specialists, forward and/or backward snowballing, and gray literature searched).

In our SLRs, both conducted and updated, we adopted Jabref[10], as it can store several types of information, including detailed bibliographical information of the studies (in BibTeX files). The reuse of such information provides access to rich knowledge sometimes not available in the SLR documentation, such as the search string history.

Combination – it improves explicit knowledge through protocol evolution. In our case, the combination involved more externalization, such as changes to the first review that occurred during the execution of updates, records of raw data, analyses and results.

To select studies in SLR1' (and also in SLR2'), two researchers independently performed the screening on titles and abstracts for the inclusion of studies. Individual results were compared and disagreements were solved. The full texts of the included studies were then obtained and the same researchers read them for their final decision on inclusion or exclusion.

Only in SLR1' and SLR2', an approach based on VTM (Visual Textual Mining) (Felizardo *et al.*, 2014) was applied to avoid the exclusion of new studies that should have been included. VTM is a new explicit knowledge which enables reviewers to visualize similar studies regarding content (title and abstract) and their citation relationships. Similar studies are grouped into the same cluster. The main advantage of VTM is to facilitate the selection of new studies by showing studies included and excluded in the previous version of the SLR. For example, if a study was included in a previous SLR and is similar to a new study, it is a *clue* that this new study could be included. Moreover, if a new study cites or is cited by studies included in the previous SLR, it could also be included. On the other hand, a new study dissimilar to all studies previously included could be probably excluded and a new study

[10] https://www.jabref.org

that cites or is cited by studies previously excluded has a chance to be irrelevant. Therefore, the use of VTM in SLR1' and SLR2' increased confidence in the selection activity.

A researcher filled out the data extraction form in such updated SLRs. For validation purposes, 30% of the primary studies were randomly selected and had their data extracted by another researcher. Whenever the data extracted differed, differences were discussed until a consensus could be reached.

In SLR1 and SLR2, data extraction tables (i.e., quantitative tables) were used to summarize data collected previously through extraction forms. Tables were reused for the inclusion and merging of data during the update. In general, both effort and time consumed in the analysis, synthesis, and report in SLR1' were less painful than those in SLR1. The same occurred for SLR2' and we believe such an observation can be extended to other SLRs. Therefore, the use of the same quantitative tables (explicit knowledge) for the addition/merging of data is fundamental in the data analysis, synthesis, and conclusions towards updating SLRs (tacit knowledge).

Internalization – Some lessons learned (LL) were identified to update SLRs:

- *LL1* – to adopt tools to support the updating process: the use of tools is quite important to facilitate managing studies during updates. For example, tools store primary studies found in the previous version of the SLR and information about these studies is useful during the selection of new studies;
- *LL2* – to provide as much information as possible about the previous SLR: it is essential to keep all information related to that SLR, especially the set of studies returned from each database, the adapted search string for each database, and even the set of excluded studies;

- *LL3* – to involve researchers from the previous SLR: the effort to update an SLR is significantly higher if none of the researchers of the previous review are part of the team, providing direct access to the employed instruments;

- *LL4* – to reuse the protocol from preliminary SLR: reusing such protocol is undoubtedly interesting when updating SLRs. In our two experiences, we replicated the original protocols. In this context, we experienced the importance of having access to a complete and detailed protocol to guide our updates. The time spent in checking SLR protocols was reduced if compared to protocol elaboration to conduct the preliminary SLR. Considering the two SLRs under evaluation, the time consumed to check the protocol was less than 1/5 of the time consumed to elaborate a protocol from scratch. Therefore, reusing the protocol is in fact quite relevant;

- *LL5* – to use quantitative tables: a difficulty for updating SLRs refers to the merging of data from new studies into the data set of previous SLR. Therefore, previous data should be stored in a standardized format, e.g., tables in a spreadsheet that can be reused during the update. Besides supporting data merging, tables are the simplest type of data presentation, and can combine findings from different tables, which enable quantitative analysis and more complete and useful data syntheses.

Since the KM Model for SLRs in Figure 2.3 presents the main activities conducted in an SLR in each phase of the KM cycle, it is possible to enhance these activities through the application of KM practices. For example, the socialization phase is marked by the exchange of tacit knowledge (tacit $\rightarrow$ tacit) —that is, socialization occurs in the sharing of tacit knowledge experiences between individuals, similar to advisement, generating a shared understanding among the members of a group. Therefore, some KM practices that can be used

to enhance SLR activities in the socialization phase are: brainstorming, informal encounters, observation, imitation, provocative dynamics, among others.

With respect to the externalization phase, it is possible to say that this phase has as main objective the creation of new explicit concepts from the tacit knowledge. This is achieved by generating hypotheses, metaphors, or mental maps that allow individuals to intuitively grasp and symbolically represent their understanding. KM practices such as discussion forums, communities of practice, and groupware can help potentiate the activities that are conducted at that phase by considering an SLR. The use of tools, in particular, help in creating a common and unique knowledge base to other SLR team researchers about what was externalized.

In the combination phase, the structuring and application of knowledge (grouping for organization) is the main objective. The combination can also be characterized as an integration of explicit knowledge, and occurs in the manipulation of data by individuals, such as emails, documents or reports. This phase is characterized by the combination of different sets of explicit knowledge. This combination occurs through the formalization of documents; here the groupware tools continue to play a very important role in this formalization. In this step, a tool may help in file sharing, document management and task registry, to manage and share tasks scheduling. In an SLR, as presented earlier, the representation of data can be synthesized in tables (or other graphical representations). It is worth emphasizing that, in the combination phase, explicit knowledge can generate the creation of new knowledge.

Finally, in the internalization phase, the explicit knowledge is embedded into tacit knowledge. Practices such as e-learning can be used at this phase, since it corresponds to a non-face-to-face teaching model and provides the team with information that can be consulted online, as often as necessary. In an SLR, for example, given the amount of information to be

presented, it is possible to say that the tools can play a fundamental role in the automation and learning of each team member that conducts the SLR.

We can say that for generating team knowledge, it is necessary that the tacit knowledge accumulated by the individuals be socialized with other team members, concluding the knowledge generation cycle. Through this cycle, there is a continuous interaction between tacit and explicit knowledge until the knowledge generation is amplified and consolidated within the team that performs the SLR, as can be observed in Figure 2.3.

2.1.2 DISCUSSIONS AND FINAL REMARKS

SLR is a knowledge-intensive process and therefore can benefit from the use of the experience gained from past reviews. In this context, principles of KM can be applied to promote knowledge capture and sharing as well as the emergence of new knowledge for updates.

If the update is a future intention, a set of essential information of the SLRs, besides that contained in the protocol and results, should be detailed and documented. If it is intended that other teams update them, this essential information should be publicly available. For instance, while some information, such as the digital libraries used and selection criteria, is sometimes made available (in the protocol published), others are not, such as the set of studies excluded and the forms for data extraction.

The following question was elaborated for the identification of such essential information: "What information could reduce the effort and time for the update of SLRs?" To answer this question, we checked: (i) information that usually is or should be in the protocol; (ii) information that usually is or should be in the results reports; (iii) information usually known

by the team, e.g., the list of excluded studies, but sometimes not explicit (i.e., not written in documents); and (iv) implicit information only contained in the mind of reviewers.

This set of information enabled us to select those (Table 2.1) that positively answered our questions. The selected information is also a rich source for future verification: for instance, why a given study was included/excluded and who did that, why a specific digital library was not used, why the search string did not include a given term, and so on.

This set of information could promote a reduction in the overall effort and time spent on the update. However, the EBSE community has usually underrated it. SLR researchers have not kept a detailed record of decisions taken throughout the review process or other alternatives that might have been taken and their justifications. As a consequence, tasks already conducted during the first execution of the SLRs must be re-executed during an update.

A global analysis revealed that an SLR update seems to be easier in comparison to the conduction of its first version, mainly due to the reuse of knowledge (when an update is performed by the same team) and/or reuse of available information (especially the SLR protocol, primary studies included, and results). However, concrete/quantitative evidence is missing to assess how easy such updates are. Reuse should also be promoted for making also available more detailed information, such as data sheets with studies and decisions history. If SLRs are systematic, transparent (including the sharing of tacit knowledge accumulated during previous conduction), and replicable as expected (Kitchenham *et al.*, 2015), researchers can update them more easily.

The use of tacit knowledge can offer other gains. First, experience acquired during the reading of studies decreased the inevitable subjectivity present in study selection since each reviewer already understood and manipulated the selection criteria. Second, researchers can

Table 2.1 : Essential Information for SLR Updates (Felizardo *et al.*, 2020a). Reproduced with the permission of IEEE.

Information	Justification
Search process	Research teams could understand all steps performed for searching studies.
Search string history	Keywords and synonyms not considered, including their description and the reasons for not considering them, could avoid doubts regarding their choice.
Digital libraries	Research teams could understand the extent of the searches and the reason why such digital libraries were adopted or not. Particularities of each digital library, search dates and years covered are important.
Adapted search strings	Adaptations of the search string for each digital library are required, but are a laborious task; therefore, such adaptations need to be available.
Other search sources	Information of additional sources, including specialists, gray literature, authors' contacts (emails) and manual search, which were previously used, should be available, as their identification may be difficult.
Inclusion and Exclusion criteria	As selection criteria are the main support for an adequate selection of relevant studies, the same criteria should be used during update.
Set of studies returned	The entire set of studies returned from the digital libraries shows how the search was completed. It is also possible to verify the overlapping of studies with new studies identified during the update, reducing the effort to analyze repetitive studies.
Selection process	The same selection process used in the previous version should be used during the update, Therefore, its availability ensures its repetition, hence, reusing to some extent past experience. Moreover, history of consensus meetings, including the list of studies reviewed, researchers involved, and decisions, are important information.
Set of primary studies included	They could be reused and reanalyzed if necessary during the update and also used as a seed set for forward snowballing.
Set of primary studies excluded	Studies previously analyzed and excluded and the reasons for exclusion can be useful for the avoidance of reanalysis.
Data extraction form	The data extraction activity can be reproduced during an update, which facilitates comparisons and merges of extracted data (older and new ones).
Supporting tool	All tools and the way they were used can support their more effective use during update.
Threats to validity and limitations	Justifications on the way threats (including internal, external, and construction ones) and limitations were mitigated increase the confidence in the SLR; therefore, the same mitigation strategy could be used during the update.

more easily evaluate the quality level of new studies, comparing them to studies previously included. Finally, data are more easily extracted using already known forms for data extraction.

Since the same people participated in the first conduction and update of both SLR, the team reused all tacit knowledge. The amount of knowledge acquired through the SLR execution is useful for the update. This includes (i) terms related to the research topic that might be included in the search string; (ii) digital libraries that might be discarded, as they are not relevant to such SLR; (iii) identification of a study that is an updated version of a previously included study; (iv) grouping of similar studies for the facilitation of synthesis; (vi) particularities of each digital library are sometimes an implicit knowledge and documentation do not usually explicit such knowledge; and (vii) opportunities for future research.

It is important to mention that although in this experience report the focus was put on conducting and updating SLR, SMs follow the same process. Thus, we believe that all the results and lessons learned presented in this section are also valid for SMs.

We conclude that KM principles can be applied to manage the knowledge generated during the update of an SLR. KM principles can help facilitate the SLR update activities as it allows for the continuous iteration of knowledge among the team members conducting an SLR. However, EBSE and KM, although being shown together the promising areas for research, it is not yet possible to say that they are consolidated since there is a lack of studies that deeply and broadly discuss KM in SLR. In this sense, we believe that the major contribution of this study was to introduce the concept of KM in secondary studies to transfer know-how to update SLR. KM seems to be then a potential mechanism to give support to existing and ongoing research on reducing the time and effort in the SLR update.

2.2 A CROSS-DOMAIN SM ON AUTOMATED SUPPORT FOR SEARCHING AND SELECTING EVIDENCE FOR SLRS IN SE

Despite the existence of initiatives (Marshall & Brereton, 2013) and available tools (Marshall *et al.*, 2015) for automating or semi-automating activities of the SLR process. However, automation of the SLR activities is still missing (Marshall *et al.*, 2018; Al-Zubidy *et al.*, 2017). For example, the study conducted by Al-Zubidy *et al.* (2017) which prioritizes value-added requirements for SLR tool infrastructure, highlighted the need for automation for the search execution and study selection activities. However, it is not clear what are the existing automation approaches explored by researchers to support the activities of search and selection of studies for SLRs in SE. To the best of our knowledge, there is not an SM nor SLR focused on the automation of the search and selection of studies for the SLR process in SE.

In our SM, we provide a detailed synthesis and comparison of the existing approaches and tools to support the activities of search and selection of studies. Considering the establishment of the application of Text Classification (TC) approaches to support SLR automation (Cruzes & Dybå, 2010; Felizardo *et al.*, 2012; Ros *et al.*, 2017; Watanabe *et al.*, 2020; Olorisade *et al.*, 2019), we focused our investigation on the use of TC approaches presenting gaps and insights for future research on automation of the SLR search and selection activities. Furthermore, since there is an increasing acceptance of the use of TC in medicine (Miwa *et al.*, 2014; García Adeva *et al.*, 2014; Bekhuis *et al.*, 2014; O'Mara-Eves *et al.*, 2015), we expand our SM search to a cross-domain analysis also mapping available evidence from medicine to support the automation of the search and selection of studies for SLRs.

In the following, we describe the research questions investigated in our SM, the adopted search strategy used to detect relevant studies, the data extraction and analysis process, the

answer to our research questions (SM results), and finally we discuss our results and present our final remarks.

2.2.1 RESEARCH QUESTIONS

We translated our SM goal into three Research Questions (RQ):

RQ1: *What are the existing approaches and tools to support the search and selection of studies for SLRs in SE?*
This research question aims to investigate the automated approaches and tools already applied to facilitate the search and selection for SLRs in SE.

RQ2: *Which text classification approaches have been explored to automate the search and selection of studies for SLRs in SE and medicine?*
This research question investigates the application of text classification techniques to provide automated support during the search and selection activities through a cross-domain study (SE and medicine). We aim to map and compare the techniques that have been explored in each area.

RQ3: *What strategies and measurements are used to assess the performance of the applied text classification approaches?*
This research question aims to summarize the strategies and measurements used to evaluate the performance of the explored TC approaches.

2.2.2 SEARCH STRATEGY

The adopted search strategy includes a two-stage search: an automatic search and a snowballing search (Wohlin, 2014). Our two-stage search process and its results are illustrated in Figure 2.4.

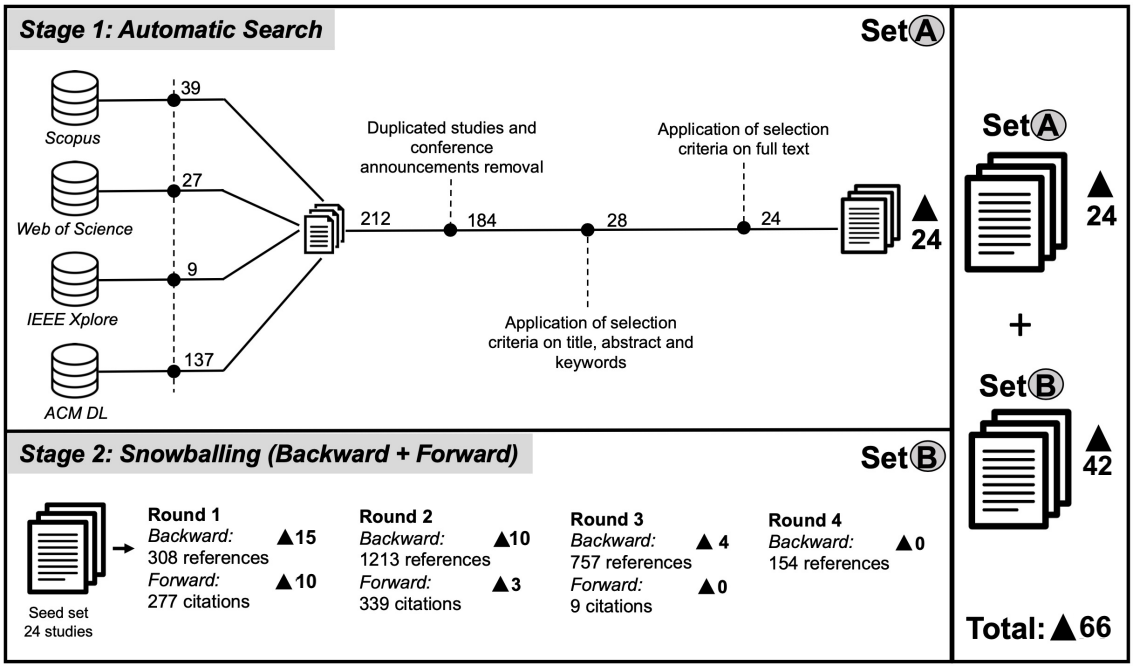


Figure 2.4 : Search strategy process. (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

To perform the automatic search, we developed a search query, and we ran a search pilot test as recommended by Kitchenham *et al.* (2015). Our search query is described next.

```
(((“systematic review automat*” OR “SLR tool” OR “literature review
automat*”))OR ((“text classification” AND “machine learning”) AND
(“systematic review” OR “literature review” OR “systematic mapping”)))
```

We chose to run our search query on what is considered the most renowned SE Digital Libraries (DLs) (Kitchenham *et al.*, 2015): IEEE Xplore[11], ACM Digital Library[12], Scopus[13] and Web of Science[14]. Scopus and Web of Science were chosen because they index studies of several international publishers, including Springer[15], Wiley-Blackwell[16], Elsevier[17], IEEE Xplore and ACM Digital Library; although not necessarily the most recent conference proceedings. Therefore, we opted for searching on IEEE Xplore and ACM Digital Library individually, because they are considered the two-key publisher-specific resources which together cover the most important SE and computer science conferences (Kitchenham *et al.*, 2015). We executed the search query in three metadata fields: title, abstract, and keywords. In addition, the search query was adapted to meet specific search criteria (e.g., syntax) of each DL.

The selection criteria are organized into three Inclusion Criteria (IC) and five Exclusion Criteria (EC):

- **IC1:** The study must present an automation approach or tool applied to support the activities of search and selection of studies; AND
- **IC2:** The study must be within the field of SE or medicine; AND
- **IC3:** The study must present results addressing automated approaches.
- **EC1:** The study is just published as an abstract; OR

[11] https://ieeexplore.ieee.org

[12] https://dl.acm.org

[13] https://www.scopus.com

[14] https://webofknowledge.com

[15] https://www.springer.com

[16] https://onlinelibrary.wiley.com

[17] https://www.elsevier.com

- **EC2:** The study is not written in English; OR
- **EC3:** The study is an older version of another study already considered; OR
- **EC4:** The study does not discuss approaches or strategies to automate the search and selection of studies; OR
- **EC5:** The study is not a primary study, (such as tutorials, keynotes, editorials, etc.).

The adoption of SLRs in SE emerged from the field of medicine (Kitchenham, 2004) because medicine has been adopting SLRs since long before SE, and it presents several advancements regarding SLR process automation. Therefore, we opted to consider in our analysis studies that address search and selection automation for SLRs from medicine to provide a systematic analysis of potential TC techniques employed in the medicine domain that could be explored in the SE context.

As illustrated in Figure 2.4 – Stage 1, a total of 212 items were returned from the automated search execution. Then, we removed all duplicated studies and conference announcements, totaling 184 studies. Next, we read the papers' title, abstract and keywords and applied the selection criteria (IC and EC) on these fields, which reduced our number to 28 candidate studies. Finally, the selection criteria were applied considering the reading of each study's full text, resulting in a set of 24 included studies from this stage. This step was performed by the Ph.D. candidate and revised by her co-advisor (100% of agreement).

The starting point of the snowballing technique is to define a "seed set" of relevant studies (Wohlin, 2014). We considered as "seed set" the 24 included studies from the automated search strategy. Next, we performed forward and backward snowballing, considering the citations and references list of the included studies, respectively. The citations were extracted with the support of search engines, such as *Google Scholar*, *ACM Digital Library* and *IEEE*

Xplore. We applied the IC and EC in each snowballing iteration, first on title, abstract, and keywords, and next on full text. We performed four backward snowballing iterations and three forward snowballing iterations, stopping their execution when no more relevant study was detected. The results from each snowballing iteration can be observed in Figure 2.4 – Stage 2. As a final result of the snowballing technique, 47 new studies were added to our included studies. Seventy-one studies were included (Stage 1: 24 studies + Stage 2: 42 studies). From the 66 included studies, 33 are from the SE domain and 33 are from the medicine domain. The final list of included studies is available online (Napoleão *et al.*, 2021b).

2.2.3 DATA EXTRACTION AND ANALYSIS

In order to answer our RQs, we created a data extraction form based on our RQs goals. The data extraction form contains all the fields necessary to analyze and synthesize the data extracted to answer the RQs impartially. In Table 2.2, we summarize the content of our data extraction form as well as the rationality of the extracted content.

Table 2.2 : Summary of the data extraction form (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

Category	Rationale	Addressed RQs
Study metadata	Identification and management of the study to detect the domain and publication data from the study.	RQ1, RQ2, RQ3
Search and selection automation approaches and tools in SE	Identification of automated approaches and tools that fully or partially support the activities of search and selection of studies for SLRs in SE.	RQ1, RQ2
Text classification approaches	Identification of approaches and measurements of text classification approaches for searching and selecting studies for SLRs in SE and medicine.	RQ2, RQ3

The data synthesis was performed through a combination of qualitative and quantitative analysis. The data synthesis results are presented as answers to our RQs in Section 2.2.4.

2.2.4 SM RESULTS

In the following are presented the answers to the proposed RQs.

RQ1: What are the existing approaches and tools to support the search and selection of studies for SLRs in SE?

To answer RQ1, we divided our analysis into two focuses: (i) proposed automation *approaches* to support the search and selection for SLR, and (ii) general and specific *tools* that automate the search and selection activities. As general tools, we considered tools that address the automation of several activities of the SLR process, including search and selection activities. As specific tools, we considered tools that address only individual challenges faced during the performance of the search and selection activities.

(i) Automated approaches: In the SE field, several approaches have been investigated to provide automated support to the activities of search and selection of studies for SLR.

As illustrated in Figure 2.5, the majority of automated approaches address the activity of selecting studies (14 studies – (Malheiros *et al.*, 2007; Tomassetti *et al.*, 2011; Felizardo *et al.*, 2012; Octaviano *et al.*, 2014; Sellak *et al.*, 2015; Ouhbi *et al.*, 2016; Rizzo *et al.*, 2017; Yu & Menzies, 2018; Marcos-Pablos & García-Peñalvo, 2020; Hassler *et al.*, 2018; Olorisade *et al.*, 2019; González-Toral *et al.*, 2019; Silva *et al.*, 2019; Watanabe *et al.*, 2020), followed by approaches that support the searching for studies (4 studies – (Ghafari *et al.*, 2012; Mergel *et al.*, 2015; Feng *et al.*, 2017; Souza *et al.*, 2017)).

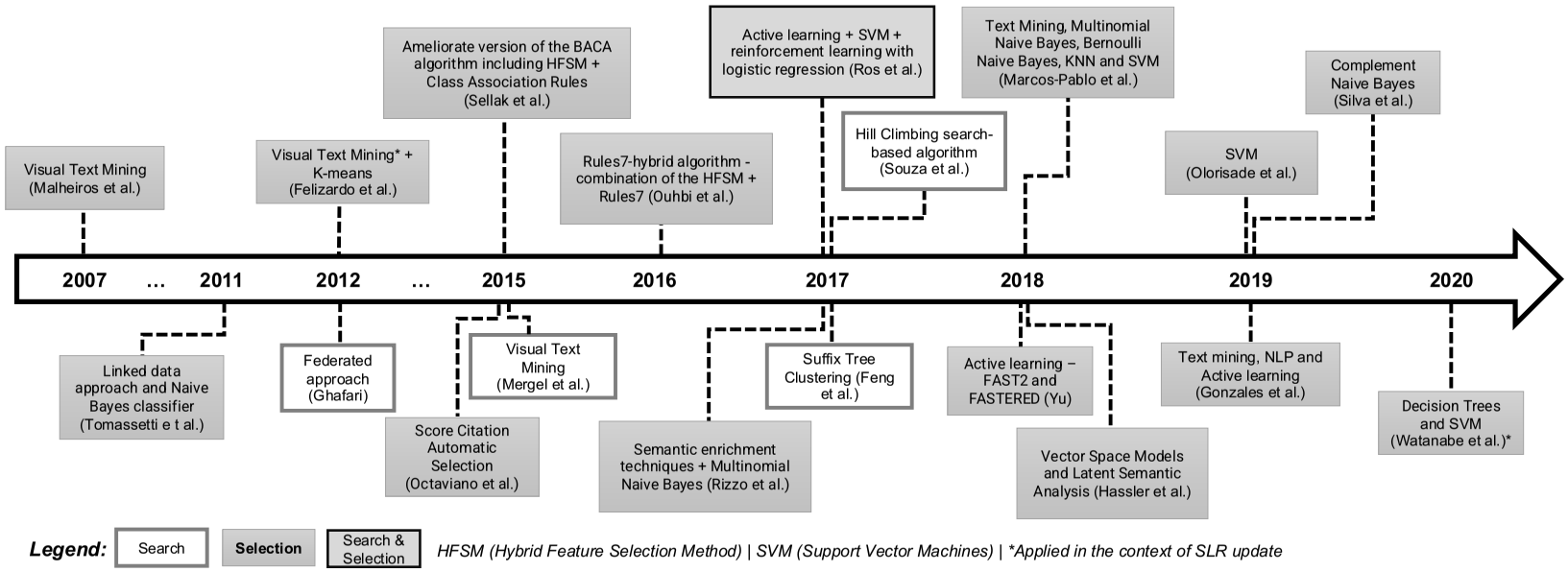


Figure 2.5 : Timeline of existing automated approaches for searching and selecting studies in SE up to 2020 (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

Only one study combined in an integrated approach a proposal of an automated solution to support both search and selection activities (Ros *et al.*, 2017).

Finding 1: *Integrated approaches addressing the activity of search and selection of studies together have not been explored by researchers.*

In an external document (Napoleão *et al.*, 2021b) we present a spreadsheet with data details from each approach presented in Figure 2.5 including a brief description of the approach, evaluation method, corpus considered in the evaluation method, results/conclusions, and future work.

Visual Text Mining (VTM) was first introduced in the SE field in 2007 by Malheiros *et al.* (2007) and further explored by Felizardo *et al.* (2011, 2012) to aid the selection activity. In addition, VTM is also investigated in the context of the search activity to aid the construction of search strings.

TC techniques addressing the use of Text Mining (TM), Natural Language Processing (NLP), and Machine Learning (ML) are strongly adopted by researchers to automate the selection of studies. The most adopted ML technique involves supervised ML models such as Support Vector Machines (SVM) (Ros *et al.*, 2017; Olorisade *et al.*, 2019; Marcos-Pablos & García-Peñalvo, 2020; Watanabe *et al.*, 2020) and active learning (Ros *et al.*, 2017; Yu *et al.*, 2018; Yu & Menzies, 2018; González-Toral *et al.*, 2019). Variations of the Naïve Bayes classifier has also been explored (Rizzo *et al.*, 2017; Marcos-Pablos & García-Peñalvo, 2020; Silva *et al.*, 2019), for example, Hybrid Feature Selection Method (HFSM) combined with other algorithms, including the hierarchical low-rank decomposition Blocked Adaptive Cross Approximation (BACA) (Sellak *et al.*, 2015) and the classical Rules7 (Ouhbi *et al.*, 2016). Besides VTM, Suffix Tree Clustering and the optimization local search algorithm Hill Climbing (Feng *et al.*, 2017) are employed to automate the search for studies. However, unlike

other approaches, Ghafari *et al.* (2012) propose a federated search approach integrating search mechanisms across well-known SE DLs automatically.

Finding 2: *SVM and active learning are the most recent adopted approaches that show promising results in their application to support the automation of the selection of studies activity.*

(ii) General and Specific tools in SE – In Table 2.3, we detail general and specific SLR tools that present some automation of the activities of search and selection. All presented tools partially support the search and selection activities; none of them fully automate any of these activities.

Regarding the specific tools that directly address automation of the activities of search and selection (see Table 2.3), we detected two specific tools that provide automation support to SLR search activity (Mergel *et al.*, 2015; Feng *et al.*, 2017) and three specific tools to support the SLR study selection activity (Malheiros *et al.*, 2007; Felizardo *et al.*, 2014; Yu & Menzies, 2018). In addition, three of the five identified tools use VTM as an automation technique (Malheiros *et al.*, 2007; Felizardo *et al.*, 2014; Mergel *et al.*, 2015).

RQ2: Which text classification approaches have been explored to automate the search and selection of studies for SLRs in SE and medicine?

Over the past 15 years, TC has gained significant attention in the context of SLR. As one of the critical Text Mining activities (also known as document classification), TC can be defined as automatically assigning semantic labels to texts given a set of fixed semantic categories or classes (Joachims, 1999). The most common TC techniques combine TM

Table 2.3 : General and Specific SE SLR tools addressing the search and selection activities (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

Tool	Search & Selection support	Year
	General Tools	
SLR-TOOL (Hinderks *et al.*, 2020)	Refinement of search using text mining; clustering studies thought similarities among them; exportation of data and references on EndNote, Bibtex and Ris formats.	2010
SLuRp (Bowes *et al.*, 2012)	Execution of search terms on some DLs; Semi-automatic extract and store studies' full text .pdf (if there are appropriate permissions); recording bibliographical data in Bibtext and Ris format; recording of assessment from reviewers as well as managing reviewer's selection and exclusions of studies.	2012
Slrtool (Barn *et al.*, 2014)	Automatic extraction of the Bibtex data from the located studies and automatic download of full-text studies .pdf (subject to permission of the host institutions); definition of the search criteria independent of target resource database; possibility of categorize studies and perform the management of the application inclusion and exclusion.	2014
SESRA (Molléri & Benitti, 2015)	Importation of search results from SE DLs (i.e. IEEE Xplore, IET Digital Library and SpringerLink) or through a Bibtex file; support on the consensus decision on the inclusion or exclusion of one study.	2015
StArt (Fabbri *et al.*, 2016)	Support to the main online search databases, including Scopus, IEEE, ACM and Web of Science; automated calculation of a study's score based on keywords occurrences on title, abstract and keywords and number of citations; automatic detection of duplicated and similar studies; semi-automation of the snowballing technique (under development).	2016
SLR Toolkit (Götz, 2018)	Simple literature filtering; design of a taxonomy; classification of studies; analysis of the classification by generated diagrams.	2018
SLR-Tool (Hinderks *et al.*, 2020)	Importation of search results from DLs and evaluation of the quality of the search results; Management of search results by including or excluding each paper.	2020
	Specific Tools	
PEx (Malheiros *et al.*, 2007)	Projection Explorer (PEx) tool uses VTM to increase study selection efficiency and allow researchers to broaden their search algorithms to create a larger corpus, since the tool quickened the identification of irrelevant studies.	2007
ReViS (Felizardo *et al.*, 2014)	ReViS uses VTM to support the selection task in systematic reviews.	2014
SLR.qub (Mergel *et al.*, 2015)	Automated support the researcher by suggesting new terms for the string using VTM algorithms.	2015
SLRPSS (Feng *et al.*, 2017)	Unified search engine wrapper for the SLR DLs: IEEEXplore, the ACM Digital Library, the Web of Science, Science Direct, Scopus, and Google Scholar.	2017
FAST2 (Yu & Menzies, 2018)	Automated support to studies selection that helps further minimize researcher efforts by using keywords to identify and rank relevant studies.	2018

approaches and ML algorithms to automatically learn and categorize new data from previously categorized data (García Adeva *et al.*, 2014).

We summarize in Table 2.4 the TC approaches identified in the selected primary studies categorizing them according to application field (SE and medicine), respectively. Overall, the SE field explored more diverse TC approaches. On the other hand, medicine is more consolidated on exploring the Naïve Bayes and SVM approaches. According to our selected studies, SE researchers have not explored approaches such as Rocchio, Latent Dirichlet Allocation (LDA), Logistic Model Trees (LMT), and neural networks to address challenges related to the search and selection of studies. In the medicine field, approaches and models such as Suffix Tree Clustering (STC), Hybrid Feature Selection Measure (HFSRM), Vector Space Models (VSM), Latent Semantic Analysis (LSA), reinforcement learning, Decision Trees (DT), Rules7, Blocked Adaptive Cross Approximation (BACA), and VTM have not been explored yet.

Finding 3: *SE and medicine research reports potential results on automated search and selection of studies. SE researchers should explore TC approaches (alone or in combination) already applied to medicine and not yet explored in SE; and vice versa.*

Regarding the TC approaches mentioned in Tables 2.4 and 2.5, we considered studies with different features and evaluation corpus. This fact prevents the performance comparison between the identified approaches. In addition, different variants of algorithms were adopted, for example, the following variants of the Naïve Bayes algorithm: Complement Naïve Bayes (Silva *et al.*, 2019; Miwa *et al.*, 2014), Multinomial Naïve Bayes (Marcos-Pablos & García-Peñalvo, 2020) and Bernoulli Naïve Bayes (Marcos-Pablos & García-Peñalvo, 2020).

Although the greater variety of approaches explored came from SE, the results presented by the field of medicine show more consolidated and systematically-evaluated results. This can

Table 2.4 : Text Classification approaches explored in SE and medicine field (Part 1) (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

TC approach	SE studies	medicine studies
Naïve Bayes	(Tomassetti *et al.*, 2011; Silva *et al.*, 2019; Marcos-Pablos & García-Peñalvo, 2020)	(García Adeva *et al.*, 2014; Almeida *et al.*, 2016; Aphinyanaphongs *et al.*, 2005; Bekhuis & Demner-Fushman, 2012; Popoff *et al.*, 2020; Frunza *et al.*, 2011; Marcos-Pablos & García-Peñalvo, 2020)
Support Vector Machine (SVM)	(Marcos-Pablos & García-Peñalvo, 2020; Olorisade *et al.*, 2019; Ros *et al.*, 2017; Watanabe *et al.*, 2020)	(García Adeva *et al.*, 2014; Almeida *et al.*, 2016; Marcos-Pablos & García-Peñalvo, 2020; Olorisade *et al.*, 2019; Aphinyanaphongs *et al.*, 2005; Bannach-Brown *et al.*, 2019; Bekhuis & Demner-Fushman, 2010; Bekhuis & Demner-Fushman, 2012; Götz, 2006; Cohen *et al.*, 2009, 2010a; Kim & Choi, 2012; Miwa *et al.*, 2014; Olorisade *et al.*, 2019; Popoff *et al.*, 2020; Timsina *et al.*, 2015; Wallace *et al.*, 2010; Timsina *et al.*, 2016)
K-Nearest Neighbor (KNN)	(Marcos-Pablos & García-Peñalvo, 2020; Olorisade *et al.*, 2019; Felizardo *et al.*, 2012; Ros *et al.*, 2017; Watanabe *et al.*, 2020; Ros *et al.*, 2017)	(Marcos-Pablos & García-Peñalvo, 2020; Bekhuis & Demner-Fushman, 2012; García Adeva *et al.*, 2014; Almeida *et al.*, 2016)
Rocchio	–	(García Adeva *et al.*, 2014)
Suffix Tree Clustering (STC)	(Feng *et al.*, 2017)	–
Active Learning	(González-Toral *et al.*, 2019; Feng *et al.*, 2017; Ros *et al.*, 2017; Yu & Menzies, 2018)	(Miwa *et al.*, 2014)
Label spreading	(Timsina *et al.*, 2016)	(Liu *et al.*, 2018)
Label propagation	(Timsina *et al.*, 2016)	(Liu *et al.*, 2018; Kontonatsios *et al.*, 2017)
Hybrid Feature Selection Measure (HFSRM)	(Ouhbi *et al.*, 2016; Sellak *et al.*, 2015)	–
Vector Space Models (VSM)	(Hassler *et al.*, 2018)	–
Latent Semantic Analysis (LSA)	(Hassler *et al.*, 2018)	–
Latent Dirichlet allocation (LDA)	–	(Bannach-Brown *et al.*, 2019)

Table 2.5 : Text Classification approaches explored in SE and medicine field (Part 2 - cont.) (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

TC approach	SE studies	medicine studies
Unsupervised K-means	(Felizardo *et al.*, 2012)	(Xiong *et al.*, 2018)
Logistic Model Trees (LMT)	–	(Almeida *et al.*, 2016)
Reinforcement Learning	(Ros *et al.*, 2017)	–
Decision Trees (DT)	(Ros *et al.*, 2017; Watanabe *et al.*, 2020)	–
Rules7	(Ouhbi *et al.*, 2016)	–
Blocked Adaptive Cross Approximation (BACA)	(Sellak *et al.*, 2015)	–
Neural Network	–	(Kontonatsios *et al.*, 2017; Götz, 2006)
Visual Text Mining (VTM)	(Malheiros *et al.*, 2007; Felizardo *et al.*, 2012; Mergel *et al.*, 2015)	–

be observed by analyzing the medicine studies mentioned in Tables 2.4 and 2.5 is reinforced by the fact that well-established tools have been used and evaluated by the medical community. During our study, we identified 7 different studies that directly report evaluation of well-established SLR selection (screening) tools: Abstrackr (Gates *et al.*, 2020, 2019; Rathbone *et al.*, 2015), RobotAnalyst (Gates *et al.*, 2019; Przybyła *et al.*, 2018), DistillerSR (Gates *et al.*, 2019; Hamel *et al.*, 2020), RelRank (Saha *et al.*, 2016), and SWIFT-Review (Howard *et al.*, 2016). In contrast, in the SE field, there is one tool, StArt (Fabbri *et al.*, 2016) that has reported on its practical use. It is worth mentioning that medicine has online available tools that support not only search and study selection activities, but the entire SLR process. DistillerSR[18] and Convidence[19] are examples of web-based commercial tools available online for managing all stages of the SLR process.

[18] https://www.distillersr.com/

[19] https://www.covidence.org/

Finding 4: *The most observed type of future work observed in SE and MED studies is a validation of the results of the proposed approach with a more extensive set of data from the same area or different areas; followed by the variation of parameters of the adopted ML models in order to improve results.*

RQ3: What strategies and measurements are used to assess the performance of the applied text classification approaches?

This section presents the strategies and measures adopted to assess and the TC techniques presented in the selected primary studies.

Strategies to assess the results from TC techniques:

From the 66 selected primary studies, 46 studies assessed the proposed TC approach or technique. Considering that most of the studies presented results from the application of ML techniques, the two most adopted approaches to assess the results from the applied TC techniques were cross-validation followed by experiments.

Cross-validation is performed by dividing simple data into subsets considering the analysis performed on a unique subset (training set) while other subsets (testing sets) are kept for subsequent use to validate the analysis (García Adeva *et al.*, 2014). The most adopted cross-validation present in our selected studies is N-fold cross-validation, which divides the dataset into N equally-sized mutually-exclusive "folds" with one fold serving as the test set and the remaining N-1 folds to form the training set. This process is repeated until each fold is used once as the training set. 10-fold cross validation was the predominate type of cross validation (García Adeva *et al.*, 2014; Bekhuis & Demner-Fushman, 2012; Bekhuis & Demner-Fushman, 2010; Kontonatsios *et al.*, 2020; Ouhbi *et al.*, 2016; Ros *et al.*, 2017; Sellak *et al.*, 2015) followed by 5-fold cross validation (Matwin *et al.*, 2010; Götz, 2006;

Bannach-Brown *et al.*, 2019; Cohen *et al.*, 2011; Olorisade *et al.*, 2017) and 7-fold cross validation (Marcos-Pablos & García-Peñalvo, 2020). A different type of cross-validation called Monte-Carlo cross-validation (Picard & Cook, 1984) was adopted by Hassler *et al.* (2018); it consists of randomly selecting a portion of the data as a training set, while the rest of the data is used as a test set. This process is repeated several times.

Another highly adopted form of assessment for text classification techniques are experiments (also referred as case studies) considering data from published SLRs performed manually (González-Toral *et al.*, 2019; Olorisade *et al.*, 2019; Kim & Choi, 2012; Silva *et al.*, 2019; Popoff *et al.*, 2020; Rathbone *et al.*, 2015; Saha *et al.*, 2016; Timsina *et al.*, 2016; Tomassetti *et al.*, 2011; Tsafnat *et al.*, 2018; Wallace *et al.*, 2010; Watanabe *et al.*, 2020; Xiong *et al.*, 2018; Yu & Menzies, 2018; Yu *et al.*, 2018; Hamel *et al.*, 2020; Howard *et al.*, 2016). In these studies, the authors usually have two or more groups of participants to emulate the search or selection process using the proposed automated approach and compare its results against the manually performed search or selection process (Felizardo *et al.*, 2012; Feng *et al.*, 2017; Malheiros *et al.*, 2007; Mergel *et al.*, 2015).

Adopted performance metrics

In order to present the results of the proposed TC techniques, studies used several metrics to describe their results. Tables 2.6 and 2.7 describe each adopted metric, its definition, applied context (SE or medicine, or both) and the studies that adopted the respective metric.

The most adopted metrics to evaluate the performance of the automation techniques are recall, precision, and F-measure. 21 of 46 SE and medicine studies (45.65%) presented an assessment approach using these metrics. On the other hand, *Work Saved over Sampling* (WSS), a measure defined by Götz (2006), was adopted just in 11 (23.91%) medicine studies

and only one SE study. In addition, the Area Under the Curve (AUC), Burden, Yeld, and Utility were applied only in medical studies.

Finding 5: *For searching and selecting studies adopting TC approaches, cross-validation and experiment are the most chosen form of assessment considered. Recall, precision and F-measure were shown to be the most frequently used performance metrics.*

Table 2.6 : Assessment metrics for text classification approaches (Part 1). Adapted from O'Mara-Eves *et al.* (2015) (Napoleão *et al.*, 2021a). Reprod. with the permission of IEEE.

Metric	Definition	Context	Studies
Accuracy	Ratio of included and commonly excluded studies with the combination of included and excluded ones.	MED and SE	(Bannach-Brown *et al.*, 2019; Kim & Choi, 2012; Popoff *et al.*, 2020; Ros *et al.*, 2017)
Precision	Ratio of correctly identified relevant studies to all of those predicted as relevant.	SE and MED	(García Adeva *et al.*, 2014; Ananiadou *et al.*, 2009; Aphinyanaphongs *et al.*, 2005; Bannach-Brown *et al.*, 2019; Götz, 2006; Bekhuis & Demner-Fushman, 2012; Bekhuis & Demner-Fushman, 2010; Frunza *et al.*, 2010, 2011; Hassler *et al.*, 2018; Liu *et al.*, 2018; Marcos-Pablos & García-Peñalvo, 2020; Olorisade *et al.*, 2019; Ouhbi *et al.*, 2016; Popoff *et al.*, 2020; Sellak *et al.*, 2015; Silva *et al.*, 2019; Timsina *et al.*, 2015; Timsina *et al.*, 2016; Tomassetti *et al.*, 2011; Tsafnat *et al.*, 2018; Watanabe *et al.*, 2020)
Recall (or Sensitivity)	Ratio of correctly predicted relevant studies to all relevant ones.	SE and MED	(García Adeva *et al.*, 2014; Ananiadou *et al.*, 2009; Aphinyanaphongs *et al.*, 2005; Bannach-Brown *et al.*, 2019; Götz, 2006; Bekhuis & Demner-Fushman, 2012; Bekhuis & Demner-Fushman, 2010; Frunza *et al.*, 2010, 2011; Hassler *et al.*, 2018; Liu *et al.*, 2018; Marcos-Pablos & García-Peñalvo, 2020; Olorisade *et al.*, 2019; Ouhbi *et al.*, 2016; Popoff *et al.*, 2020; Sellak *et al.*, 2015; Silva *et al.*, 2019; Timsina *et al.*, 2015; Timsina *et al.*, 2016; Tomassetti *et al.*, 2011; Tsafnat *et al.*, 2018; Watanabe *et al.*, 2020; Yu & Menzies, 2018)

Table 2.7 : Assessment metrics for text classification approaches (Part 2 - cont.). Adapted from O'Mara-Eves *et al.* (2015) (Napoleão *et al.*, 2021a). Reprod. with the permission of IEEE.

Metric	Definition	Context	Studies
F-Measure	Combines Precision and Recall values. It corresponds to the harmonic mean of Precision and Recall.	SE and MED	(García Adeva *et al.*, 2014; Ananiadou *et al.*, 2009; Aphinyanaphongs *et al.*, 2005; Bannach-Brown *et al.*, 2019; Götz, 2006; Bekhuis & Demner-Fushman, 2012; Bekhuis & Demner-Fushman, 2010; Frunza *et al.*, 2010, 2011; Hassler *et al.*, 2018; Liu *et al.*, 2018; Marcos-Pablos & García-Peñalvo, 2020; Olorisade *et al.*, 2019; Ouhbi *et al.*, 2016; Popoff *et al.*, 2020; Sellak *et al.*, 2015; Silva *et al.*, 2019; Timsina *et al.*, 2015; Timsina *et al.*, 2016; Tsafnat *et al.*, 2018; Watanabe *et al.*, 2020)
WSS@95% (Work Saved over Sampling)	The percentage of studies that the reviewers do not have to read because they have been screened out by the classifier considered at 95% recall.	MED and SE	(Bannach-Brown *et al.*, 2019; Cohen *et al.*, 2011; Howard *et al.*, 2016; Kontonatsios *et al.*, 2020; Matwin *et al.*, 2010; Olorisade *et al.*, 2019; Przybyła *et al.*, 2018; Timsina *et al.*, 2015; Timsina *et al.*, 2016; Yu & Menzies, 2018; Yu *et al.*, 2018)
Area Under the Curve (AUC)	Area under the curve obtained by graphing the true positive rate against the false positive rate; 1.0 is a perfect score and 0.5 is equivalent to a random ordering.	MED	(Bannach-Brown *et al.*, 2019; Cohen *et al.*, 2009, 2010a; Miwa *et al.*, 2014)
Burden	The fraction of the total number of studies that a human must screen.	MED	(Kontonatsios *et al.*, 2017; Wallace *et al.*, 2010; Miwa *et al.*, 2014; Hamel *et al.*, 2020)
Yield	The fraction of studies that are identified by a given screening approach.	MED	(Kontonatsios *et al.*, 2017; Wallace *et al.*, 2010; Miwa *et al.*, 2014)
Utility	It is a weighted sum of Yield and Burden. It represents the relative importance of Yield in comparison to Burden.	MED	(Kontonatsios *et al.*, 2017; Wallace *et al.*, 2010; Miwa *et al.*, 2014)

2.2.5 DISCUSSIONS AND FINAL REMARKS

Despite the several search and selection approaches identified, the lack of automation is still present in these SLR activities. Efforts have been applied to reduce the search and selection workload and time spent, but it still needs to reduce the human effort required to search and select relevant studies for inclusion in SE SLRs.

The majority of the proposed search and selection automated approaches presented some validation. Cross-validation and case studies are the most adopted types of validation (see RQ3 results). However, each study validated their proposed approaches considering a limited number of sources (e.g., search only in one database such as Scopus or IEEE Xplore) and population (e.g., reduced number of SLRs studies from different SE and medicine domains). These factors prevent an accurate comparison of efficiency and workload reduction among the proposed approaches. Therefore, large-scale and exhaustive validation is needed to support results obtained through preliminary analysis and demonstrate the real applicability and benefits of the proposed approaches in the SE field.

The existing automation approaches have not been applied in practice by researchers who conduct SLRs in SE. As mentioned in RQ2, only one study reported feedback on the practical adoption of the proposed tool. Some of the reasons that we have assumed for the low use of the automated search and selection approaches were that the automated approach or tool is: (i) presented only as a prototype; (ii) not available online (e.g., broken access links); (iii) insufficiently documented; or (iv) not easy to use. Therefore, there is a need for better dissemination and the use of SE researchers' search and selection approaches to develop an evidence base about their usage and more insight into their relative advantages.

The combination of TM and ML applied to automate or semi-automate the SLR search and selection activities provides cost savings and allows replicability (Ros *et al.*, 2017).

However, one known difficulty concerning the use of TC approaches is that most supervised learning approaches used in these studies rely on a data set for training the model (Watanabe *et al.*, 2020). Considering this fact, **in the scenario of SLR update where the dataset for training is already known (original SLR selected studies), TC techniques can be promising**.

Our results highlight the extensive adoption of TC techniques to support SLR's search and selection activities. However, selecting the most appropriate ML algorithm, related methods, and text sections (e.g., title, abstract, keywords, references) is fundamental to achieve high recall and precision.

Integrated solutions to automate the search and selection process for SLR using TM and ML approaches is the most suitable combination of approaches since they can bring several benefits to facilitate the SLR execution, such as: (i) reviewers do not need to construct search strings; (ii) as a consequence, reviews can have better recall when it is not dependent on the recall of an initial search string; (iii) the set of included papers can be updated automatically by the tool once the classifier is sufficiently trained; (iv) the approach can be implemented with an efficient interface that the reviewer can use until the search and selection are made, reducing cognitive load; and (v) the process is well suited for complex reviews where search strings are hard to elaborate.

Furthermore, our results (see RQ3 results) show that quantitative and qualitative approaches have been used to demonstrate the efficiency of the proposed automated solutions. Since recall and precision (consequently F-measure) and workload and time-saving are the most adopted parameters to analyze the efficiency of automated proposals, we encourage the adoption of these metrics in all studies addressing SLR automation, especially to enable comparison results from different research. Our observations corroborate with Olorisade *et al.*

(2016) considering the TC adoption scenario: it is fundamental that the authors make available the data used in their evaluation and the replication package of the study to enable detailed comparisons and study replication.

CHAPTER III

CSLR CONCEPT AND PROCESS DEFINITION

Considering the growing increase in publications of SE SLRs over the last years (Napoleão *et al.*, 2021; Mendes *et al.*, 2020) the importance of investigating solutions that contribute to SLR update increases. Searching for new evidence, selecting evidence, deciding upon updating, and enacting the update process following lessons learned from experience reports are important pieces of a problem that has not yet been explicitly addressed as a whole: the problem of leaving gaps between the SLR publication and possible updates.

To address the problem of leaving gaps between the SLR publication and possible updates, we introduce the CSLR concept. **CSLR comprises a continuous and systematic surveillance and analysis of potential new relevant evidence for published SLRs, contributing to keeping SLRs up to date.** We designed a CSLR process by applying meta-ethnography, considering elements of the traditional SE SLR process (Kitchenham *et al.*, 2015), concepts from studies addressing supporting activities involved in updating SLRs in SE (Wohlin *et al.*, 2020; Mendes *et al.*, 2020; Felizardo *et al.*, 2014), LSR from medicine (Elliott *et al.*, 2017), metaphors from DevOps (Humble & Farley, 2010; Bass *et al.*, 2015) and open science practices in SE (Mendez *et al.*, 2020).

As a contribution of this chapter, we highlight the proposition of the CSLR concept in SE and the CSLR process definition. In the following, we summarize the study design approach (Section 3.1) and in Section 3.2 we report on the application of the meta-ethnography method and its results.

3.1 STUDY DESIGN

Towards reaching the RG1 of this thesis (*Definition and evaluation of the CSLR concept and process*), we started by conducting a meta-ethnography in order to define the CSLR process. Figure 3.1 illustrates a summary of the study design presented in this chapter.

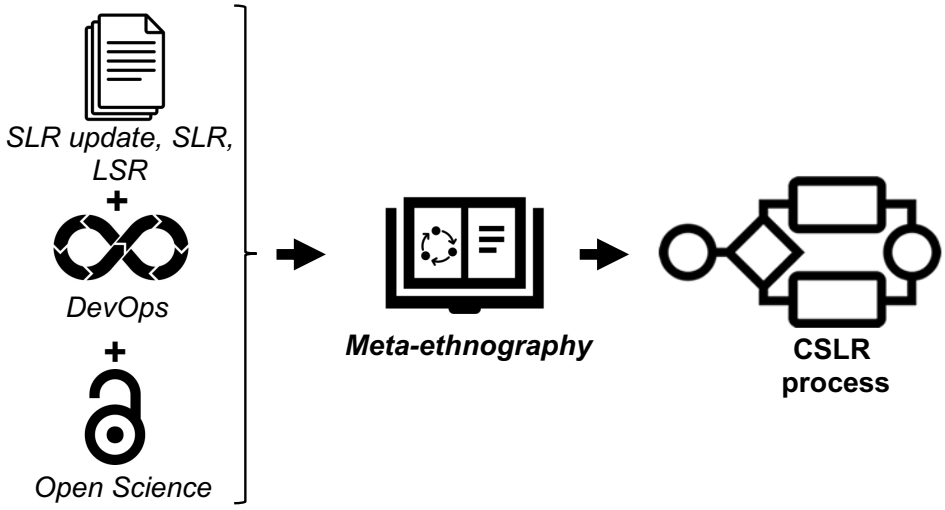


Figure 3.1 : Study design summary (Napoleão *et al.*, 2022b). Reproduced with the permission of IEEE.

We opted to use as our research method one called meta-ethnography (Noblit & Hare, 1988), since it enables a systematic qualitative synthesis and detailed understanding of how topics and studies are related. According to Hannes & Lockwood (2011), the meta-ethnography method provides a picture of the whole phenomenon under investigation from studies of its parts. In this sense, the resulting picture of our meta-ethnography is the CSLR process and its steps. The meta-ethnography method, in summary, involves researchers selecting, analyzing and interpreting qualitative studies through a process of translation, which provides an interpretation of the entire topic, in order to answer questions, gain new insights and/or build knowledge on a specific topic (Noblit & Hare, 1988).

3.2 APPLICATION OF THE META-ETHNOGRAPHY METHOD

In this section, we present the seven stages of the meta-ethnography (Noblit & Hare, 1988) with a brief description and the results of the practical application of each one of the stages.

3.2.1 STAGE 1 – GETTING STARTED

During this first stage, the goal is to identify a topic of interest to be qualitatively explored and define a Research Question (RQ) that represents the topic and guides the research (Noblit & Hare, 1988). In this sense, we aim to understand how concepts from the SLR process, SLR update, LSR, DevOps, and open science are related and how they can be integrated to help mitigate intermittent SLR update issues in SE. Therefore, we translated our study goal into a Research Question (RQ) to guide us to evidence of the relations in these areas:

How are the SLRs process, SLR update, LSR, DevOps and open science concepts related and how can they be integrated to help mitigate intermittent SLR update issues in SE?

To answer our RQ, we synthesized intersections and relationships of these concepts to create the CSLR process.

3.2.2 STAGE 2 – DECIDING WHAT STUDIES ARE RELEVANT TO THE TOPIC OF INTEREST

In the second stage, the goal is to search and select relevant studies on the topic of interest to be analyzed (Noblit & Hare, 1988). Firstly, we performed a systematic search on Scopus Digital Library (DL) using the terms: *((“systematic literature review process” OR “systematic review process” OR “systematic literature review guideline” OR “systematic review*

update" OR "SLR Update") AND "software engineering") to detect studies that address the SLR process and SLR update strategies. Secondly, we searched for studies and guidelines that address LSR on Scopus using the term: *("living systematic review")* and also in two medicine databases: PubMed[20] and Cochrane[21]. We chose Scopus because it indexes several relevant Computer Science publishers (Kitchenham *et al.*, 2015), Pubmed and Cochrane because they are two of the most renowned medicine DLs. We performed both searches in February 2022. Thirdly, since our goal is to synthesize information from DevOps practices, we opted to use as a base for our analysis two widely adopted books that describe the DevOps process (Bass *et al.*, 2015) and the CI/CD pipeline (Humble & Farley, 2010). Finally, for open science practices, we considered the book chapter of Mendez *et al.* (2020) since it presents a recent overview of open science practices in SE.

In order to select relevant studies on SLR, SLR update and LSR, we defined three Inclusion Criteria (IC) and four Exclusion Criteria (EC):

- **(IC1)** The study proposes or discusses a process or elements of a process on SLR Update in SE or LSR in medicine; *OR*
- **(IC2)** The study is a guideline for the conduction of SLR, SLR Update in SE or LSR in medicine; *OR*
- **(IC3)** The study is an experience report on SLR update in SE.
- **(EC1)** The study is an SLR, SLR update or LSR, but it does not discuss any step or aspect of the SLR, SLR update and LSR processes; *OR*
- **(EC2)** The study is an experience report on SLR conduction; *OR*

[20] https://pubmed.ncbi.nlm.nih.gov

[21] https://www.cochranelibrary.com

- **(EC3)** The study is an older version of another study already considered; *OR*
- **(EC4)** The study is not written in English.

A total of 298 studies addressing SLR, 41 SLR updates and 257 studies LSR were retrieved during the search process. First, we excluded duplicated studies, and then we applied the IC and EC on the title, abstract and keywords of these studies. As a result, we selected 17 candidate studies in this stage. Next, we read and applied the IC and EC on the full text of the candidate studies, and selected 12 studies in this stage. We also performed an iteration of the backward and forward snowballing (Wohlin, 2014) to identify additional studies through the list of references and citations of the selected studies. As a result, we selected three more studies. Thus, a total of 15 studies compose our final set of studies. Table 3.1 compiles in a list all 18 studies (S1 – S18), selected for the following stages of the meta-ethnography process execution. It includes our final set of 15 selected studies categorized by their main topic and the books and books' chapter previously selected (+3).

We opted to not perform an exhaustive search and selection of studies because the meta-ethnography guidelines and process do not demand it (Noblit & Hare, 1988). In fact, having a significant quantity of studies to be selected and analyzed can be challenging and lead to a poor-quality analysis (Fu *et al.*, 2019).

3.2.3 STAGE 3 – READING THE STUDIES

During the third stage, we read the set of selected studies and performed the data extraction (Noblit & Hare, 1988). In the context of this analysis, a requirement was to be familiar with the SLR process and the software development process. Therefore, the Ph.D. candidate and collaborators experience and their interactions facilitated the understating of the content of the selected studies.

Table 3.1 : Selected studies for the next stages of the meta-ethnography (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

ID	Main topic	Publication Year	Reference
S1	SLR update	2020	Watanabe *et al.* (2020)
S2	SLR update	2020	Wohlin *et al.* (2020)
S3	SLR update	2020	Mendes *et al.* (2020)
S4	SLR update	2020	Felizardo *et al.* (2020a)
S5	SLR update	2019	Nepomuceno & Soares (2019)
S6	SLR update	2018	Felizardo *et al.* (2018)
S7	SLR update	2017	Garcés *et al.* (2017)
S8	SLR update	2016	Felizardo *et al.* (2016)
S9	SLR update	2008	Dieste *et al.* (2008a)
S10	SLR process	2017	Kuhrmann *et al.* (2017)
S11	SLR process	2015	Kitchenham *et al.* (2015)
S12	SLR process	2007	Kitchenham & Charters (2007)
S13	LSR	2022	Simmonds *et al.* (2022)
S14	LSR	2019	Brooker *et al.* (2019)
S15	LSR	2017	Elliott *et al.* (2017)
S16	DevOps	2015	Bass *et al.* (2015)
S17	DevOps	2010	Humble & Farley (2010)
S18	Open Science	2020	Mendez *et al.* (2020)

The data extraction form was built to obtain (i) the general objective of the study; (ii) its main results and contributions; (iii) if the study proposes an approach or the use of a technique, its description and whether it has been validated; and (iv) if the study presents a process, and the description of each process activity including roles, inputs, processing, and outputs.

Analyzing the 18 selected studies, as shown in Table 3.1, the majority of the selected studies address the SLR Update process in SE. Since our goal was to mitigate intermittent SLR update issues in SE, it requires a deep understanding of the advancements on SLR updates in SE over the past years. Notably, SLR updates have recently gained the attention of the SE community. In 2020, four studies proposing improvements in conducting SLR updates in SE were published (S1, S2, S3 and S4). Watanabe *et al.* (2020) (S1) propose and evaluate the

use of text classification to provide automated support in the SLR update studies selection activity. Wohlin *et al.* (2020) (S2) propose guidelines on the search strategy to update SLRs in SE. Mendes *et al.* (2020) (S3) recommend using a decision framework when to update SLRs in SE. Finally, Felizardo *et al.* (2020a) (S4) present an experience report on how to transfer the know-how of SLRs to facilitate their updates through the instantiation of a knowledge management model.

Five other studies addressing SLR updates were selected (S5, S6, S7, S8 and S9). Besides S4, two other experience reports on SLR updates were considered in our study (S5 and S6). Nepomuceno & Soares (2019) (S5) present a systematic mapping and survey on how researchers are evolving their SLRs and what they think about SLR updates. Garcés *et al.* (2017) (S7) relate the authors' experience in updating two SLRs using automated techniques based on VTM (Visual Text Mining). Felizardo *et al.* (2016) (S8) introduce the adoption of forward snowballing (Wohlin, 2016) to search for studies to update SLRs in SE. Two years later, Felizardo *et al.* (2018) (S6) evaluate the use of different electronic databases for applying forward snowballing to update secondary studies. Study S2 presents a combined and more recent investigation addressing the approaches described in S5 and S6. Last but not least, Dieste *et al.* (2008a) (S9) propose a process to perform SLR updates in SE, taking into account lessons learned from updating an SLR.

Regarding the traditional SLR process, we considered the well-known Kitchenham & Charters (2007) guidelines (S12) as well as Kitchenham *et al.* (2015)' book (S11), which contains an extended description of the SLRs process in SE as well as the update of the guidelines proposed in S12. Kuhrmann *et al.* (2017) (S10) present an experience-based guideline to aid researchers in designing SLRs in SE, with emphasis on the studies search and selection procedures.

LSRs were introduced in the medical field in 2017 by Elliott *et al.* (2017) (S15), aiming to incorporate relevant new evidence to SLRs as it becomes available. Two years later, the Cochrane DL embraced the LSR concept and published its guidelines for the production and publication of Cochrane LSRs (Brooker *et al.*, 2019) (S14). More recently, Simmonds *et al.* (2022) (S13) described the general principles of LSR in a book chapter, when they might be of particular value, and how its procedure differs from conventional SLRs.

The last three studies included in our analysis are the book of Bass *et al.* (2015) (S16), which describes the DevOps concept through a software architecture perspective detailing each step of the DevOps process. The book of Humble & Farley (2010) (S17) presents the whole DevOps, and explains in detail the CI/CD practices. Finally, the recently published book chapter by Mendez *et al.* (2020) (S18) on open science for SE includes the open science definition, why SE researchers should engage in it, and how they should do it.

The Ph.D. candidate extracted the data by first carefully examining each selected study individually and extracting text passages that contain the information requested in the data extraction form. Once the analysis of each study was done, we asked the research collaborator to review the data. The reviewed data was conserved in our data extraction form available online (Napoleão *et al.*, 2022a) in order to make it possible to determine how these concepts are related (stage 4).

3.2.4 STAGE 4 – DETERMINING HOW THE STUDIES ARE RELATED

In this fourth stage, using the extracted data from stage 3 and revisiting the selected studies when needed, we extracted the metaphors of each selected study. These metaphors were keywords, phrases, ideas, and concepts that could be relevant to detecting relationships or connections among the studies. It is worth highlighting that a metaphor may be associated with

a keyword, even if it doesn't exactly match. It can be interpreted as similar to this keyword based on existing evidence. Next, we highlighted the main keywords and metaphors used in each area (SLR Update, SLR Process, DevOps, SLR and open science). We used spreadsheets to support this stage. In Tables 3.2 and 3.3 we present examples of metaphors associated with "versions".

When we started to extract the relationships among the studies, we noticed that all studies have some process phases and concepts that directly impact the updating of an SLR. In this sense, we organized our keywords and metaphors based on process factors. The list of organized metaphors is available online (Napoleão *et al.*, 2022a). We summarize the relationships among the selected studies in Table 3.4. The process factors are described in the first column of Table 3.4.

As shown in Table 3.4, we detected studies related to specific SLR update planning factors: study selection and search strategy. Despite the base of the SLR update process being the original SLR process (planning, executing and reporting), i.e. re-execute (and adapt, if necessary) the original review protocol, we identified that the DevOps concepts and process could be seen as a metaphor for building a process of continuous integration and delivery of evidence to support keeping SLRs up-to-date. Monitoring, a DevOps process factor (S16-S17), is connected to the LSR studies (S13-S14-S15) which are connected to the other three SLR and SLR update factors (planning, executing and reporting). In addition, the open science study (S18) is connected to the reporting factor since it addresses the availability of information.

3.2.5 STAGE 5 – TRANSLATING THE STUDIES INTO ONE ANOTHER

The main goal of this stage is to compare the metaphors (keywords and text fragments) extracted from the studies. Unfortunately, Noblit & Hare (1988) do not describe how to

Table 3.2 : Examples of metaphors associated with "versions" (Part 1). Build from the respective studies listed in Table 3.1.

Topic	Metaphors
SLR update	*"...the area is rapidly changing and old SLRs can lead researchers to obsolete directions", "...many SLRs are not kept up-to-date, which shortens their lifespan."* ***(S1)*** *"...two SLRs (SLR2 and SLR6) include both an original conference publication and an extended journal version of the original SLR", "...the extended journal version, which is published later, covers the same search span used in the conference paper version", "Google Scholar found both versions. Thus, only the final version is included"* ***(S2)*** *"Both the original secondary study and its update are available, and provide the data that is needed for comparison...", "The updated study is a previous version of a more recent update.", "...whenever there are different versions of the same primary study, the one to use is the most complete description.", "It is the oldest version of study S19"* ***(S3)*** *"and their updated versions, SLR1' and SLR2',", "... lack of evidence during the previous version could be included in the updated version.", "It covers the period not covered by the last version", "studies included and excluded in the previous version of the SLR", "...tools store primary studies found in the previous version of the SLR...", "All tools and the way they were used (source, version, required configuration)."* ***(S4)*** *"...what kinds of changes in SLRs artifacts would generate a new SLR", "changes to a certain extent could still characterize an SLR update", "Reusing a protocol of the original SLR as a base..."* ***(S5)*** *"...strings used in previous versions of the review", "...included in earlier versions of a secondary study."* ***(S6)*** *"the first version of both SLR1 and SLR2...", "...primary studies found in the previous version of the SLR..."* ***(S7)*** *"...contain the studies included in the previous version of this SLR...", "Included studies in a previous SLR"* ***(S8)*** *"...some time later, we carried out an update of that review.", "...initial SR and a later update."* ***(S9)***
SLR process	*"Use a version control system (VCS)."* ***(S10)*** *"...there is a conference version of the paper followed by an extended journal version of the paper.", "found by the previous reviews that are relevant to your topic area.", "Previous literature reviews (systematic or not) are extremely valuable for identifying known primary studies and validating your search process.", "If any previous systematic reviews or mapping studies were kept separate for validation purposes...", "...primary studies reported by the previous reviews...."* ***(S11)*** *"necessary to consult all versions of the report to obtain all the necessary data.", "...this light version of a systematic review...", "...the most complete version of the survey will be used.", "...results compare with previous reports?", "Summary of previous reviews"* ***(S12)***
LSR	*"...through explicit versioning of the review publication... ", "... monthly update generated a new publication each time.", "...findings of the meta-analyses may change between updates."* ***(S13)*** *"(i.e.a new version published)", "...an updated version of an existing Cochrane Review.", "...incorporated in the next version of the LSR.", "...different versions of the LSR.", "...effective version control of data, documents and other files is crucial.", "...peer reviewers for all versions of the LSR...", "An updated version of the LSR may not be up to date...", "The fact that a new version of the review has been published...", "...parts of the LSR that have changed from the previous version can ease the burden...", "...a 'compare version' can assist in identifying changes for a new version,...", "...while the baseline version of the review is being produced and published.", "...the most recent version of the review that included these methods)."* ***(S14)*** *"...the first version of the review should be published;", "...through explicit versioning of the review publication...", "...standard update of a pre-existing review..."* ***(S15)***

Table 3.3 : Examples of metaphors associated with "versions" (Part 2 cont.). Build from the respective studies listed in Table 3.1.

Topic	Metaphors
DevOps	*"...managing distinct versions of a system that are simultaneously in production,...", "Releasing a new system or version of an existing system to customers...", "...releasing a new version opens the possibility of incompatibilities..", "... be under version control and subject to examination for corrections.", "For example, a version control system is a form of automated coordination that keeps various developers from overwriting each other's code.", "...deploying new versions of software.", "Each contains slightly different versions of the same system.", "Project hardware typically includes integration servers and version control servers,...", "Are all features of the old version supported in the new version?"* ***(S16)*** *"Keep Everything in Version Control", "... delivering new versions of your software to users.", "...the right version of the code, sure, but also the correct version of the database schema,...", "...the version control system will alert you...", "Every change committed to version control is supposed to enhance the system that we are working on.", "The deployment of your application can be implemented using a fully automated process from version control.", "Although version control systems are the most obvious tool in configuration management, the decision to use one (and every team should use one, no matter how small)...", "...the aim of a version control system is twofold: First, it retains, and provides access to, every version of every file that has ever been stored in it.", "...the ability to step back to a recent, known-good version of your artifacts, it is important to check in frequently.", "The benefits of version control are enhanced when you commit regularly.", "Promoting a new version of your application from one environment to another.", "...your project in a version control repository."* ***(S17)***
Open Science	*"Open access can take several forms. The form depends on which version of the article is made public and at which point of the academic writing process.", "The work is called preprint if it reflects a version of their manuscript that has not yet been accepted for publication at a scientific venue.", "For version control, in our project, we decide to use Git...", "That version control system allows us and our collaborating partners to trace the versions of all produced text documents in an organised fashion.", "For our work to be reproducible in a long-term manner, we need to further document the versions of the software used.", "If the content of the own produced work is identical to the content of the accepted publication, it is called postprint."* ***(S18)***

Table 3.4 : Relationships among the selected studies (Napoleão *et al.*, 2021a). Reproduced with the permission of IEEE.

Process factor	Related studies
Planning	S3, S4, S5, S7, S9, S10, S11, S12, S13, S14, S15, S16, S17
Study selection	S1
Search strategy	S2, S6, S8
Executing	S1, S2, S3, S4, S5, S6, S7, S9, S10, S11, S12, S13, S14, S15, S16, S17
Reporting	S4, S5, S9, S10, S11, S12, S3, S14, S15, S16, S17, S18
Monitoring	S13, S14, S15, S16, S17

perform this stage, but one suggestion is to compare the synthesis of each study progressively (Silva *et al.*, 2013).

We analyzed the metaphors observed in stage 4, comparing them and analyzing their relationships. To construct the relations among the studies, we opted to guide this construction based on the original SLR process activities as demonstrated in Table 3.4 because it is also the basis process for updating an SLR. In addition, the LSR process follows the same base structure as the SLR process. In this sense, first, we build a table based on the elements of the relationships detected in stage 4. Second, we separated each metaphor into (i) *general metaphors*: fragments that address the relationship among the studies and procedural elements generally; (ii) *specific metaphors*: fragments that address specific relations among each activity necessary to perform an SLR update. Finally, we checked our list of metaphors to see if any of them directly mitigated the SLR intermittent update issues. Therefore, we checked if the relationships were found to converge to the objective of our study. We also added the references that support our findings. The table resulting from this stage of the meta-ethnography is available online at (Napoleão *et al.*, 2022a). The results of Stage 4 were revised and discussed through consensus meetings with other collaborators, i.e., any divergence between a relationship and/or citations was discussed and solved through consensus.

Analyzing the identified relationships (Napoleão *et al.*, 2022a), it is clear that the selected studies' practices, processes, activities, and recommendations are strongly interconnected. Hence, the systematization of the evidence found in this stage in the form of a process can contribute to the establishment of a dedicated process for mitigating the intermittent SLR update problem.

3.2.6 STAGE 6 – SYNTHESIZING THE TRANSLATIONS

During the sixth stage, we constructed a translation synthesis. The translations of studies result in many metaphors. They are compared to verify similarities and/or if some metaphors can encompass others. The result of this stage is usually represented as diagrams or figures (Noblit & Hare, 1988). In the context of our study, we compared and systematized the translations mapped in stage 5 as a continuous process for assessing new evidence and evaluating the need of update SLRs (Continuous Systematic Review – CSLR). Figure 3.2 illustrates the three stages of the CSLR process (Integration, Delivery and Observability) with their phases and activities. The stages and components of the DevOps practices (*cf.* Chapter 1, Figure 1.3) were included in the CSLR process in order to facilitate the correlation of the DevOps metaphor with the CSLR process proposition. The CSLR BPMN diagram is also available online (Napoleão *et al.*, 2022a). In the following we describe the CSLR process detailing each activity of it.

The CSLR process starts with the *Integration* stage and *Version Control* phase (see Figure 3.2). In this phase, the first activity is to verify if the SLR has an update or replication published. If yes, these studies must be linked and considered in the next process activity.

In the second phase of the *Integration* stage, the *Build* phase, the first activity is to obtain the protocol information of all considered studies (original SLR, updates, replications – if it exists) and store them in a database. This protocol information includes: addressed research questions, period covered by the SLR execution, list of included studies, inclusion and exclusion criteria, and quality criteria (if adopted).

In the next activity of the *Build* phase, once a day, using the original SLR and its list of included studies, one iteration of the forward snowballing using the Google Scholar as DL is

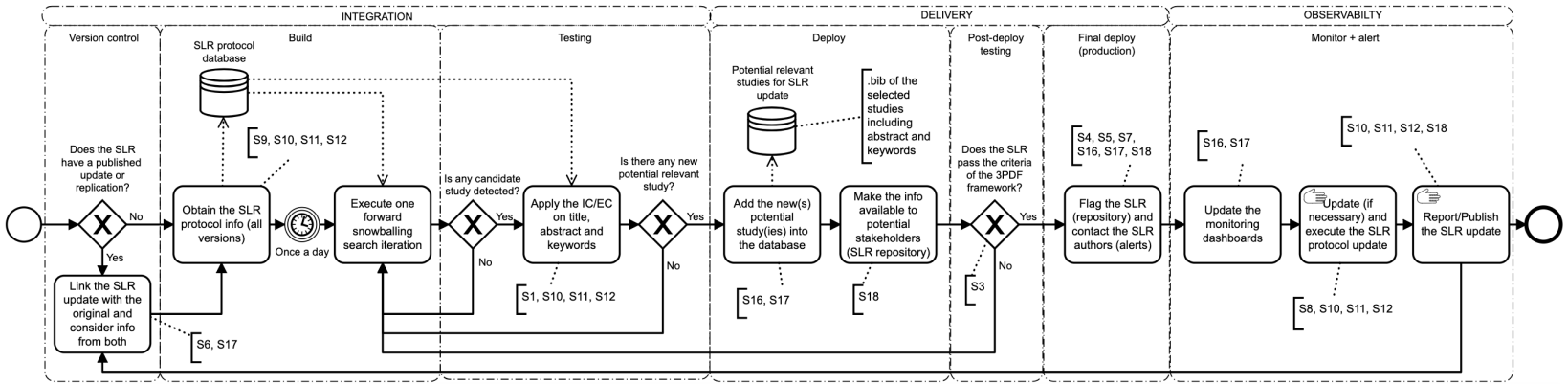


Figure 3.2 : CSLR process. ©Bianca Minetto Napoleão.

performed (Wohlin *et al.*, 2020). If there is an SLR update and/or replications linked to this SLR, these studies and their list of included studies must be considered in the snowballing execution.

The last phase of the *Integration* stage, the *Testing* phase, starts checking if any study was detected by the snowballing process (candidate studies). If no study was detected, the process returns to forward snowballing execution activity. If one or more studies were detected, the IC and EC criteria must be applied to the candidate studies' title, abstract, and keywords. The study of Watanabe *et al.* (2020) proposed an automated alternative for this stage using text classification techniques. The next activity is to verify if there is any potential relevant study according to the IC and EC. If not, the process returns to the activity of snowballing. If yes, the process moves to the *Delivery* stage.

The *Delivery* stage starts with the *Deploy* phase. The first activity in this phase is to add the new potential relevant studies to the database. In order to deal with automation, and to allow using reference management tools (such as Jabref and Zotero), the studies should be stored in the database in BibTeX format. The next activity is to make these potential relevant studies available to potential stakeholders. For that, we suggest uploading this information into a public repository such as Zenodo or ArXiv (Mendez *et al.*, 2020).

The next phase, the *Post-deploy testing*, has one activity: apply the 3PDF proposed in Mendes *et al.* (2020) to analyze if the SLR needs to be updated. The 3PDF consists in answering 7 questions (steps) as described next and further detailed in Chapter 1 (Mendes *et al.*, 2020):

- *Step 1.a – Does the published SLR still address a current question?*
- *Step 1.b – Has the SLR had good access or use?*

- *Step 1.c – Has the SLR used valid methods and was well-conducted?*
- *Step 2.a – Are there any new relevant methods? (e.g. more detailed quality checklists to assess evidence, new search strategy, new forms of aggregate evidence such as thematic analysis, etc.)*
- *Step 2.b – Are there any new studies or new information?*
- *Step 3.a – Will the adoption of new methods change the findings, conclusions or credibility?*
- *Step 3.b – Will the inclusion of new studies/information/data change findings, conclusions or credibility?*

From an automation perspective, two of these questions can benefit from automated support: Step 1b. (citation analysis) and Step 2a. (results from the application of the IC and EC activity).

If the results of the 3PDF show that the SLR is suitable for an update, in the *Final deploy* phase the SLR must be flagged as "outdated" and the authors of the original SLR, update and/or replication (if it exists) will be contacted. Otherwise, the process returns to the activity of executing the snowballing forward diary.

The last stage of the CSLR process is the *Observability* stage. This stage has a unique phase: *Monitor + Alert*. The first two activities of this stage are conditioned to process automation, i.e., the development of a dedicated SLR repository. It consists of updating the monitoring dashboards that report the SLR update data gathered during the CSLR process execution. The last two activities consist of performing the SLR update. It includes updating protocol items (if necessary) (Mendes *et al.*, 2020) and executing it; and reporting the results

of the SLR update (e.g., publishing) (Kitchenham *et al.*, 2015). Finally, with the first process cycle finished, the new and completed SLR updated must be linked to its other version(s), re-starting the CSLR process from its beginning.

3.2.7 STAGE 7 – EXPRESSING THE SYNTHESIS

In this last stage, the synthesis findings are disseminated to interested parties. In our case this concerns making the CSLR process available as a direction for future research on continuously keeping SLRs in SE up-to-date. We performed this stage through the publication of the results of the meta-ethnography in the *48th Euromicro Conference on Software Engineering and Advanced Applications* (Napoleão *et al.*, 2022b), a target conference that has a specific track for SLRs in which several relevant studies in the area of SLR were published (e.g. Mendes *et al.* (2019)).

3.3 THREATS TO VALIDITY

In the following, we report the main threats to validity associated with the research described in this chapter and the mitigation strategies employed to address them.

Construct validity. We followed well-known guidance and advice on designing and conducting meta-ethnography studies (Noblit & Hare, 1988).

Internal validity and reliability. Since meta-ethnography is an interpretive approach to synthesis, we addressed the validity and reliability of our synthesis by performing discussions among the authors during the conduction of all seven steps of the method, and we evaluated the outcome (CSLR process) through a case study (Chapter 4).

3.4 CHAPTER FINAL REMARKS

In this chapter, we proposed the Continuous Systematic Literature Review (CSLR) concept and process to support SLRs updates in SE. We structured the CSLR process by synthesizing evidence through a meta-ethnography integrating knowledge from varied research areas.

After conducting the seven stages of the meta-ethnography and defining the CSLR concept and process, in Chapter 4, we shall evaluate the CSLR process by performing a case study with a well-known SLR in SE to evaluate the CSLR process observing its contributions to mitigating the SLR intermittent update issues.

CHAPTER IV

APPLYING CSLR TO A PUBLISHED SLR

In this chapter, we evaluate the feasibility of the CSLR process through a participative case study. As a contribution, the main findings of this chapter indicate that the CSLR concept and process provide an innovative and systematic way that can be applied to help maintaining SLRs, supporting continuously, trustworthy and up-to-date evidence for SLRs in SE.

4.1 CASE STUDY CONDUCTION

In order to evaluate the CSLR process, we selected a suitable SLR as an instrument to execute the process through the conduction of a participative case study (Baskerville, 1997). Our goal is to perform an initial evaluation of the feasibility of applying the proposed CSLR process and observe its contributions to mitigating the SLR intermittent update issues. We translated our case study goal into two Research Questions (RQs):

RQ1: *How do the steps of the proposed CSLR process perform in practice?*

RQ2: *Can the CSLR process help mitigate the intermittent SLR update issue in SE?*

We follow the five main steps for conducting case studies proposed by Runeson *et al.* (2012): design, preparation, collecting data, analysis and reporting. These steps are described hereafter.

4.1.1 DESIGN

Our design consists in selecting an SLR to be the instrument (input) of the CSLR process and then executing the CSLR process steps manually.

We chose the SLR by Kitchenham *et al.* (2007) which addresses the topic of cross-company vs. within-company effort estimation, for the following reasons: (i) The SLR is published as a conference paper (Kitchenham *et al.*, 2006), and after as a journal paper (Kitchenham *et al.*, 2007) in renowned SE venues; (ii) The SLR has being used as an evaluation instrument by several other studies (Wohlin *et al.*, 2020; Mendes *et al.*, 2020; Felizardo *et al.*, 2016; Wohlin, 2016); (iii) The SLR was last updated in 2014 (Mendes *et al.*, 2014), over nine years ago – with this, we can also evaluate the CSLR process when the SLR already has a published update; and (iv) The SLR update has as co-author one external collaborator with this Ph.D. research.

4.1.2 PREPARATION

We distributed the conduction of the participative case study between the Ph.D. candidate and the external collaborator. Both have experience in conducting SLRs and updates. The external collaborator is a co-author of the SLR Update (Mendes *et al.*, 2014). According to SLR experience reports (Felizardo *et al.*, 2020a; Nepomuceno & Soares, 2019; Garcés *et al.*, 2017), the participation of a member who already participated in the previous review can facilitate the update process besides contributing to avoiding bias.

4.1.3 COLLECTING DATA

Data collection is based on the SLR chosen for the CSLR process evaluation. The CSLR process has several data collection activities that must be carried out according to the process execution (e.g. check if there is a published SLR update, obtain the list of included studies, etc.).

4.1.4 ANALYSIS

We report our analysis on how the process steps are performed in practice by describing the case of applying each CSLR process activity to the selected SLR.

The first activity of CSLR involves verifying if our SLR candidate has a published update. For that, we checked the citations of the SLR on Google Scholar (428 citations). As a result, we identified an update (Mendes *et al.*, 2014) (SLR-Update) published in 2014, as well as two other studies published in 2016 that replicated the SLR update investigating different search strategies: Wohlin (2016) (SLR-UR1) and Felizardo *et al.* (2016) (SLR-UR2). Therefore, SLR-Update, SLR-UR1 and SLR-UR2 were selected in the version control phase (*Integration* stage) of the CSLR process execution. It is important to mention that these same three studies were also identified in Wohlin *et al.* (2020), who used this same SLR as an investigation instrument.

The next phase of the CSLR process (*Build*) begins with obtaining protocol information from the studies selected in the version control phase. With the support of a spreadsheet, we extracted the following information from the original SLR, SLR-Update, SLR-UR1 and SLR-UR2:

- *Research questions* – all studies investigated the same research questions;
- *IC, EC and quality criteria* – the IC, EC and quality criteria were the same for all studies;
- *Search strategy* – the original SLR performed an automated search on six SE DLs, a manual search on individual journals and conference proceedings and reference checking (a.k.a backward snowballing). The SLR-Update used the same method as the original SLR except for adding Scopus as an extra DL and not performing a manual search.

SLR-UR1 adopted backward and forward snowballing (Wohlin, 2016) and SLR-UR2 only forward snowballing;

- *Search strategy coverage period* – the original SLR covered the period from 1990 to November 2006, the SLR-Update from December 2006 to end 2013 and SLR-UR1 and SLR-UR2 both from 2006 to 2013; and
- *List of included studies* – The original SLR included ten primary studies, the SLR-Update has 11 additional primary studies. The SLR-UR1 has not included two studies included in the SLR-Update, but it included three other studies that were not included in the SLR-Update. The SLR-UR2 included all the studies included in the SLR-Update except for one study, and it also identified two other studies. Considering this scenario, for the CSLR, we considered all unique included studies from the original SLR, SLR-Update, SLR-UR1 and SLR-UR2, totaling a final list of 25 included studies published by the end of 2013.

The second activity in the *Build* phase is to execute one interaction of forward snowballing search technique on Google Scholar to identify new potential relevant studies. Using the 25 included studies and the original SLR, SLR-Update and SLR-UR1 and SLRU-R2 as seed, we performed an interaction of the forward snowballing technique we obtained a total of 2392 returned studies. Since the SLR-Update and both replications covered the search until 2014, we limited our search results from 2014 to February 2022, resulting in 858 returned studies. We exported the bibliographical data, including keywords and abstracts of all studies in BibTeX and CSV format.

Moving to the *Testing* phase, the first step is to perform the initial cleaning of our set of returned studies. Since this process has been executed manually, we used the .csv file to remove duplicated studies and conference announcements. Thus, we arrived at a list of 444

unique candidate studies. Next, we applied the inclusion and exclusion criteria on the title, abstract and keywords of each study arriving in a set of 24 potential candidate studies. It is worth mentioning that all 24 studies contain at least one or more keywords that would allow an automated selection method such as Watanabe *et al.* (2020) to identify these studies. The Ph.D. candidate and the external collaborator carefully analyzed the title, abstract and keywords of each study, and in a synchronous meeting, they decided through consensus what studies have a strong potential for inclusion. As a result, five new studies showed significant potential to be included in the SLR, and other five studies presented a new trend of investigations using cross-company and within-company mixed together to estimate software project effort (which could lead to updating the SLR protocol, including its research questions, to properly consider this new identified trend). The list with the 24 potential candidate studies and the ten selected potential studies to be included in a new update is available online (Napoleão *et al.*, 2022a).

The *Deploy* phase starts by adding the potential selected studies into a database and next making them available online in a repository. Since we are performing both activities manually, we created a .bib file with the ten potential included studies. We made them available at Zenodo (Napoleão *et al.*, 2022a), an open dissemination research data repository.

The *Post-deploy testing* phase consists of using the 3PDF (Mendes *et al.*, 2020) to verify if the SLR needs to be updated. Thus, we performed the seven steps (questions) of the 3PDF method. The answer to each question of the 3PDF is described next. As a result, the SLR needs to be updated.

- *Step 1.a – Does the published SLR still address a current question?* Yes. Effort estimation in software projects is still an open challenge in SE. Researchers are still addressing the investigation of using cross-company versus within-company data. Nevertheless, the combination of both types of data to estimate effort has been explored too.

- *Step 1.b – Has the SLR had good access or use?* Yes. The original SLR has over 428 citations where more than half of these citations (234) were after 2014. In 2021 the SLR received 32 citations. In addition, the SLR Update has received 25 citations since its publication in 2014, and six of them are over the last year.

- *Step 1.c – Has the SLR used valid methods and was well-conducted?* Yes. The SLR, SLR-Update, and both replication used were well-conducted using valid methods and published in SE in renowned venues. Besides, the original SLR made available its protocol in an external file where all the performed steps were detailed, and the update counted the participation of an author of the original SLR.

- *Step 2.a – Are there any new relevant methods? (e.g. more detailed quality checklists to assess evidence, new search strategy, new forms of aggregate evidence such as thematic analysis, etc.).* Yes. A new SLR update can be performed by adopting the guidelines for search strategy to update SLRs in SE presented in Wohlin *et al.* (2020) as well as following our proposed CSLR process (which includes the search strategy proposed in Wohlin *et al.* (2020)).

- *Step 2.b – Are there any new studies or new information?* Yes. We identified five new potential studies to be added in a new SLR update. In addition, we identified five other studies that can add change to the research direction on the SLR topic.

- *Step 3.a – Will the adoption of new methods change the findings, conclusions or credibility?* No. We based our decision for this question on the inclusion of new methods in the SLR-Update and both replications that resulted in no changes in the studies' findings.

- *Step 3.b – Will the inclusion of new studies/information/data change findings, conclusions or credibility?* Maybe. As mentioned by Mendes *et al.* (2020), it is a challenge to answer since it requires preliminary searches to evaluate the risk of performing an

update. Observing the potential five included studies and the five other studies that could address a new research direction for a new update, there is a considerable probability of changing the SLR findings.

In the *Delivery* stage, during the final deployment, we must flag the SLR with the status "SLR suitable for update" and notify the authors of SLR and SLR-Update (if exists) about the findings. As this is an initial feasibility study, we did not contact anyone with regard to these results yet.

The last stage of the CSLR process is the *Observability* stage. The first activity described in the *Monitor + alert* phase, (flag the SLR and contact the authors), is also suitable for performing manually. The activities of performing and publishing the SLR update can be performed without these activities. However, updating the SLR is out of the scope of this study. Hence, since the SLR is not updated yet, the process returns to the *Build* phase to search for new potential relevant studies.

4.1.5 REPORTING

RQ1: *How do the steps of the proposed CSLR process perform in practice?*

In general, the CSLR process flow seemed to be coherent. The stages, phases and activities were manually applicable in practice, addressing relevant aspects to help mitigate intermittent SLR update issues in SE.

However, when it comes to automation aspects, while some activities are clear candidates for automation (e.g., the whole CSLR systematic surveillance and analysis of new potentially relevant studies), not all of them seem to be ready for complete automation yet. For instance, the application of some decision steps of the 3PDF still requires human intervention

and reasoning. Also, the update itself, including the rigorous full-text-based assessment, application of the quality assessment criteria, data extraction, and research synthesis, is seen as a manual activity to be conducted by researchers, as depicted in Figure 3.2.

RQ2: *Can the CSLR process help to mitigate the intermittent SLR update issue in SE?*

During the execution of the participative case study, (manually) applying the CSLR process activities showed being feasible and enabled systematic surveillance and analysis of new potentially relevant studies. Furthermore, its application revealed the need to update an already published SLR. Hence, we conclude that the CSLR process can help to mitigate the intermittent SLR update issue.

With proper automation (e.g., including SLR repositories, continuous integration facilities, and availability of information in an open and accessible way, monitoring dashboards), CSLR could systematically support the identification of the need or not for SLR updates. Consequently avoiding unnecessary updates (Mendes *et al.*, 2020). Hence, we perceive making the CSLR process available as an important element to foster research in this direction.

4.2 DISCUSSIONS ON THE CSLR CONCEPT AND PROCESS

The motivation for the elaboration of the CSLR process is to provide a systematic process to help mitigating the intermittent SLR update problem, contributing to avoiding missing potential new relevant research in evidence-syntheses or decision-making. According to Nepomuceno & Soares (2019), actions to keep SLR updated are of great importance to the SLR research field. The CSLR process unifies isolated activities and metaphors, integrating them into a systematic approach to continuously assessing new evidence and evaluating the need of updating SLRs in SE.

One may question the similarities between LSR and CSLR. Indeed, both have the same aim of keeping the SLR up to date (Elliott *et al.*, 2017; Simmonds *et al.*, 2022). In summary, LSR is a medicine review approach to update an SLR continually (most manually) requiring authors to make explicit commitments as to the frequency of search and screening execution as well as publication updates (Brooker *et al.*, 2019; Elliott *et al.*, 2017). On the other hand, inspired in SE DevOps and open science practices, the CSLR process explores automation opportunities based on study findings investigated in the SE area on SE SLR updates (e.g. Wohlin *et al.* (2020); Mendes *et al.* (2020)) aiming to provide continuous and systematic surveillance and analysis of potential new relevant evidence. Furthermore, the idea of making the SLR protocol information and intermediate results openly available is not explored within the LSR context. LSR relates to continual SLR updates, while CSLR aims at continuous SLR updates. In Table 4.1 and 4.2 we further detail the differences and similarities between LSR and CSLR based on several aspects.

In a more general view, the CSLR could also be an instrument to direct the SE community on research subjects that have been investigated – if an SLR is often cited, it gives evidence that the subject of study addressed is constantly evolving. This fact leads to questions such as: Does the SLR remain relevant? Is the SLR up to date? Does it needs to be updated? As observed in our participative case study, new research trends can be identified, leading to the proposition of other research directions (questions) on a research subject.

Besides, the CSLR process showed to be effective during our evaluation and opens avenues for automating its activities and pipeline. Researchers have investigated automation of the SLR process over the years (Felizardo & Carver, 2020). However, to the best of our knowledge, only two studies are addressing automation alternatives for SLR updates (Felizardo *et al.*, 2014; Watanabe *et al.*, 2020). Moreover, both studies are focused only on the study selection activity. Therefore, there is a lack of approaches that automate or semi-automate

Table 4.1 : Differences and similarities between LSR and CSLR (Part 1). Adapted from Brooker *et al.* (2019) and Elliott *et al.* (2017).

Aspect	LSR	CSLR
Suitability evaluation	An SLR can emerge as an LSR, or it can go through a suitability assessment process to transition to a living mode. The decision to start a new LSR or transition to a living mode rests with the Cochrane editorial base.	There is no suitability assessment, as there is no change in the status of the SLR (e.g., from SLR to LSR). Any and all SLRs can go through the CSLR process.
Role assignment and resource management	There is an assessment of the availability, capacity, motivation, and commitment of a team of authors to maintain the LSR.	There is no assessment of resource availability and capacity, as CSLR is based on collaboration and open science principles. There are no role definitions or assignments as well as no need for authors to commit to keeping an SLR up to date.
Planning	If a new SLR emerges as an LSR, the LSR protocol, which includes planning the LSR methods (search method and frequency, integration of new information, retention of legacy information, and LSR transition to a living mode), must be published and reviewed. In the case of an existing SLR transitioning to an LSR, authors should include the LSR plan as an appendix to the revision update. Before searching for new evidence, producing and publishing an LSR baseline is mandatory.	There is no a priori planning step to start the execution of the CSLR process. That is, there is no need to pre-establish requirements such as the search method and frequency. The search decision and evidence integration are pre-defined by the CSLR process. The retention of SLR information and its updates and replications is permanent and also open to the community. Since there is no definition of LSR "status", there is no need for a status transition plan. The SLR and its replications and updates (if any) are the initial requirements of the CSLR process, conducting a baseline release is not necessary.
Evidence monitoring (search for evidence)	There is active monitoring of new evidence at a predefined frequency. LSR authors define the search frequency and commit to carrying it out.	Active monitoring of new evidence also occurs in CSLR. The process mainly uses the snowballing forward technique once a day and focuses on automation, eliminating the need for authors' commitment to perform the search.

Table 4.2 : Differences and similarities between LSR and CSLR (Part 2 cont.). Adapted from Brooker *et al.* (2019).

Aspect	LSR	CSLR
Evidence selection	The complete selection activity is performed at the same frequency established for the search. Cochrane LSRs can count on automated help during the screening process offered by the Cochrane Register of Studies (CRS-Web) platform.	The selection of studies is performed in two stages. First, a screening is performed based on the title, abstract and keywords of the studies retrieved from the search, and then these potentially relevant studies are submitted to a full-text analysis. The CSLR process is designed to count with automation to support this initial screening of evidence. Furthermore, the goal of CSLR is to make these potentially selected studies openly available to the community through an SE SLR dedicated repository.
Evidence integration	After planning the search, LSR authors assess the impact of new evidence on the LSR and decide whether to integrate it immediately or later based on its impact on the review conclusion or a fixed-interval schedule approved by the editorial team.	Similar to LSR, the CSLR process evaluates the impact of new studies on the SLR. However, the decision to update the review is not solely based on the importance of these impacts on the SLR findings, but on the results of the execution of the 3PDF that assesses whether an SLR needs updating.
Monitor and alert	Readers receive alerts about the LSR status at a pre-established frequency, such as whether the LSR is up to date, if there is an ongoing update, or if new evidence has been identified and the LSR requires an update. In addition, Cochrane LSR has a support team and a managing editor to monitor the update process of all LSRs.	Similar to LSR, the CSLR process offers monitoring and alerting options for the community. It suggests contacting the authors of the SLR under investigation and provides the option to flag the SLR in a repository to indicate its need for an update. Unlike the LSR context, the CSLR process includes activities focused on maintaining monitoring dashboards openly available for an overview of the SLRs, their replications, updates, and artifacts by the SE community. Implementing and populating a dedicated repository is necessary to carry out these activities.

other SLR update activities or approaches that integrate the update activities. In Chapter 6 we propose an automation alternative for the two trigger activities (search and selection of new evidence) of the CSLR process.

4.3 THREATS TO VALIDITY

In this section, we present the primary threats to validity and mitigation actions associated with the research described in this chapter.

Construct validity. The execution of the participative case study followed the established guidelines and concepts for conducting case studies in SE as (Runeson *et al.*, 2012) and (Baskerville, 1997). Additionally, a meticulous selection process was employed to choose the appropriate SLR to apply CSLR (*cf.* Section 4.1.1).

External validity. Similar case studies could have been applied to more SLRs to improve the generalizability of our results. However, manually applying the process involves significant effort. Nevertheless, we suggest as future research an extension of this case study with more SE SLRs, counting on the participation of external researchers collaborators to further investigate the practical application of the CSLR process.

4.4 CHAPTER FINAL REMARKS

The results of our evaluation suggest that CSLR can help mitigate the relevant SLR intermittent update issue for SLRs in SE. Besides, the CLSR process provides a systematic pipeline towards automating and managing SLR updates activities.

Bearing in mind the aforementioned benefits of the CSLR process as well as the RG2 (*Definition of the CSLR guidelines and validation of the CSLR process and guidelines together*) of this thesis, in Chapter 5 we extend our research by proposing guidelines for the CSLR

process. In addition, we analyze the joint performance and suitability of the CSLR process and guidelines through a SE SLR expert analysis.

CHAPTER V

GUIDELINES TO CSLR IN SE

The SE area has several guidelines to support the conduction of secondary studies, including SLRs (Kitchenham & Charters, 2007), systematic mappings (Petersen *et al.*, 2008, 2015), Rapid Reviews (Cartaxo *et al.*, 2020; Rico *et al.*, 2020), and Multivocal Literature Reviews (Garousi *et al.*, 2019). However, there are no unified guidelines for performing SLR *updates* in SE. In this sense, the goal of this chapter is to extend the research presented in Chapters 3 and 4 by proposing guidelines for the CSLR process and validating the CLSR process and guidelines through an evaluation by SE SLR experts. We used expert evaluation to obtain quality feedback on the CSLR process and guidelines descriptions as well as on their suitability.

The main contributions of the chapter include: (i) guidelines for the CSLR process describing details and examples on how to update SLRs in SE continuously; (ii) validation of the CSLR process through a SE experts evaluation resulting in improvements of both, the CSLR process flow and activities, and guidelines; and (iii) considerations of the SE SLR community on the relevance and potential adoption of the CSLR.

5.1 RESEARCH DESIGN

In order to achieve the goal of this chapter, which is to detail guidelines for the CSLR process proposed in Chapter 3 as well as in Napoleão *et al.* (2022b), the first step is to extend the systematic search initially performed to obtain all studies related to SLR updates in SE.

Our base study described in Chapter 3, Napoleão *et al.* (2022b), systematically searched Scopus DL to obtain the main SE studies addressing the SLR update process. It selected nine

studies: Wohlin *et al.* (2020); Mendes *et al.* (2020); Watanabe *et al.* (2020); Felizardo *et al.* (2020a); Nepomuceno & Soares (2019); Felizardo *et al.* (2018); Garcés *et al.* (2017); Felizardo *et al.* (2016); Dieste *et al.* (2008b). As mentioned in Chapter 3, since the meta-ethnography process does not consider mandatory the conduction of an exhaustive search (Noblit & Hare, 1988), it was not performed. Hence, we opted to extend the search and selection of studies aiming to ensure that all relevant studies regarding SLR updates published in SE are considered during the construction of the CSLR guidelines. We used these nine selected studies during the meta-ethnography process as a "seed set" to perform two iterations of the forward and backward snowballing techniques (Wohlin, 2014). Forward snowballing is an approach that considers the analyzed studies' citations, while backward snowballing considers each study's reference list aiming to find other relevant studies (Wohlin, 2014). We defined three IC and five EC to select relevant studies on SLR updates in SE:

(IC1) The study proposes or discusses a process or elements of it to update SLRs; *OR*

(IC2) The study proposes approaches, guidelines, or alternatives to improve or facilitate the SLR update process; *OR*

(IC3) The study is an experience report on SLR updates; *OR*

(EC1) The study is an SLR update that does not discuss any step or aspect of the SLR update process; *OR*

(EC2) The study is not peer-reviewed and not published in SE venues; *OR*

(EC3) The study is an older version of another study already considered; *OR*

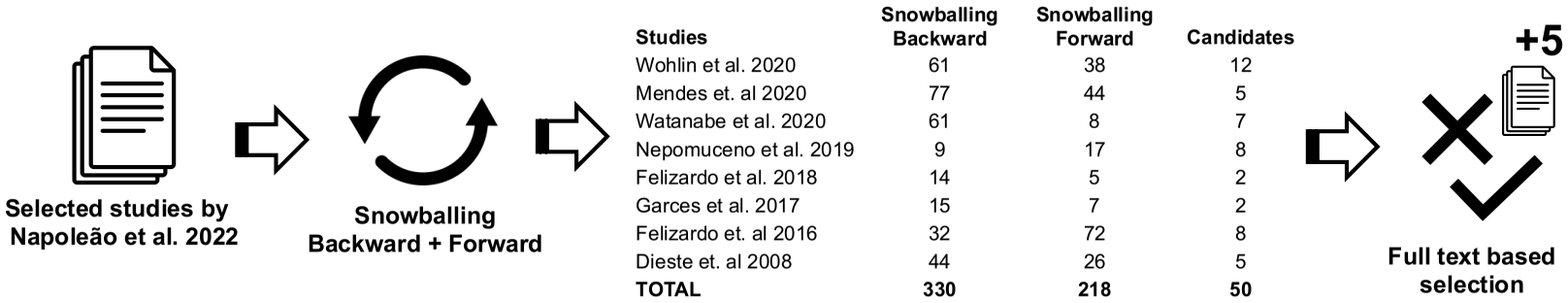
(EC4) The study is not in the SE domain; *OR*

(EC5) The study is not written in English.

Firstly, we extracted the references of each study and then their citations with the support of the Google Scholar search engine. Figure 5.1 illustrates the number of references and citations of each study. Next, we applied IC and EC to each retrieved study's title, abstract, and keywords. We called this group of studies "candidates" (see Figure 5.1). We selected a total of 50 candidate studies. The following step was to unify these 50 candidate studies in a single list, remove duplicates and apply the IC and EC criteria based on full text. The Ph.D. candidate performed the snowballing technique, and doubts about the inclusion or exclusion of a study were decided through consensus with two experienced SLR researchers contributors. As a result of both snowballing techniques, we added five more relevant studies to our included studies, totaling 14. The complete list of included studies ordered by publication year and describing the study origin (seed set or snowballing) is presented in Table 5.1.

In order to detail the CSLR guideline, we performed a narrative synthesis (Popay *et al.*, 2006) considering data extracted from the selected 14 studies. In Section 5.2, we describe the results of the narrative synthesis with each study's respective reference.

In addition to the 14 selected studies used as a basis to describe the CSLR guidelines, in the descriptions of the CSLR process activities, we provide examples of SE SLR updates that support the activity. For that, we looked for SLR updates published in SE. The study of Mendes *et al.* (2020) provides a list of SLRs updates in SE published until 2019. We updated this list by adding studies published between 2019 and 2022. Only one SLR update was retrieved by our search (de A. Cabral *et al.*, 2023). These examples provide the researcher performing CSLR process activities a practical reference of the CSLR process activity under execution. Later, in Section 5.2, we detail and exemplify each CSLR process activity. Finally, in Section 5.3 we performed an expert evaluation to evaluate and improve the CSLR process activities, execution flow and guidelines as well as to obtain perceptions on the relevance and potential adoption of the CSLR by the SE community.



Studies	Snowballing Backward	Snowballing Forward	Candidates
Wohlin et al. 2020	61	38	12
Mendes et. al 2020	77	44	5
Watanabe et al. 2020	61	8	7
Nepomuceno et al. 2019	9	17	8
Felizardo et al. 2018	14	5	2
Garces et al. 2017	15	7	2
Felizardo et. al 2016	32	72	8
Dieste et. al 2008	44	26	5
TOTAL	**330**	**218**	**50**

Figure 5.1 : Research Design. ©Bianca Minetto Napoleão.

Table 5.1 : Final list of included studies. ©Bianca Minetto Napoleão.

Title	Reference	Year	Origin
Guidelines for the search strategy to update systematic literature reviews in software engineering	Wohlin *et al.* (2020)	2020	Seed set
When to update systematic literature reviews in software engineering	Mendes *et al.* (2020)	2020	Seed set
Knowledge Management for Promoting Update of Systematic Literature Reviews: An Experience Report	Felizardo *et al.* (2020a)	2020	Seed set
Reducing efforts of software engineering systematic literature reviews updates using text classification	Watanabe *et al.* (2020)	2020	Seed set
Avoiding plagiarism in systematic literature reviews: An update concern	Nepomuceno & Soares (2020)	2020	Snowballing
On the need to update systematic literature reviews	Nepomuceno & Soares (2019)	2019	Seed set
Evaluating Strategies for Forward Snowballing Application to Support Secondary Studies Updates: Emergent Results	Felizardo *et al.* (2018)	2018	Seed set
Maintaining Systematic Literature Reviews: Benefits and Drawbacks	Nepomuceno & Soares (2018)	2018	Snowballing
An Experience Report on Update of Systematic Literature Reviews	Garcés *et al.* (2017)	2017	Seed set
A Visual Analysis Approach to Update Systematic Reviews	Felizardo *et al.* (2014)	2014	Snowballing
Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering	Wohlin (2014)	2014	Snowballing
Using Forward Snowballing to update Systematic Reviews in Software Engineering	Felizardo *et al.* (2016)	2016	Seed set
Formalizing a Systematic Review Updating Process	Dieste *et al.* (2008b)	2008	Seed set
Experimenting with a multi-iteration systematic review in software engineering	Ferrari & Maldonado (2008)	2008	Snowballing

5.2 GUIDELINES TO CSLR IN SE

In this Section, we propose the CSLR guidelines (see Sections from 5.2.1 to 5.2.9) detailing the CSLR process execution in order to guide SE researchers on how to perform the activities of the CSLR process.

It is important to mention that although several activities of the CSLR process can have automated support, the process can also be conducted manually as presented in Chapter 4. Since SE does not have specific guidelines for updating SLRs, in the following sections, we present step-by-step guidelines with examples to update SLRs in SE continuously.

5.2.1 VERIFYING IF THE SLR HAS A PUBLISHED UPDATE OR REPLICATION

The first activity to conduct the update of an SLR is to identify whether or not there are updates or replications of this published SLR (Mendes *et al.*, 2020; Felizardo *et al.*, 2018).

The objective of this activity corroborates the guidelines proposed by Kitchenham & Charters (2007) and (Kitchenham *et al.*, 2015) since, in the same way, Kitchenham & Charters (2007) guidelines highlighted that the first step before conducting an SLR is to identify the need to conduct an SLR in order to avoid wasting time and effort. In the context of SLR updates, it is crucial to identify whether there are any SLR updates or even replications to avoid conducting an unnecessary update. In addition, the protocol information of all versions of the SLR should be considered, since to they are crucial for decision-making in the update process; for example: (i) it is necessary to know the period covered by the original SLR and its respective updates in order to be able to perform a search for new evidence accurately; (ii) it is necessary to know if the SLR was replicated, the differences between the protocol information, and the result of the original SLR and its replication.

By definition, SLR updates and replications are published after the original SLR, and usually, they cite the original SLR. In order to identify if there is a published SLR update or a replication of the SLR under investigation, we suggest locating the SLR under review on Google Scholar Digital Library[22] and checking its citations, aiming to find updates or replications.

One example in SE that illustrates the need for identifying all versions of an SLR is the SLR performed by Kitchenham *et al.* (2006). This SLR was published as a conference paper in 2006 (Kitchenham *et al.*, 2006), and after as a journal paper in 2007 (Kitchenham *et al.*, 2007). Also, this SLR was updated in 2014 by (Mendes *et al.*, 2014), and its update has been replicated twice by two independent studies (Felizardo *et al.*, 2016; Wohlin, 2016) that aimed to investigate search strategies. As a result, together the SLR update and its replications identified in total 15 studies. However, only nine studies were in common among them. In situations like that, it is essential to consider all uniquely identified studies as the basis to perform a new SLR update, as well as other important protocol-related information such as (i) the search period coverage of the original SLR, SLR update and replication (Mendes *et al.*, 2020) in order to choose an adequate search period for a new update; (ii) the SLR update and its adopted replication search strategies to retrieve studies, since they could influence the results (Wohlin *et al.*, 2020); (iii) the participation of authors in common with the original SLR, SLR update and replication in order to facilitate the experience and knowledge transfer from the SLR versions (Felizardo *et al.*, 2020a).

[22] https://scholar.google.com

5.2.2 OBTAINING THE SLR PROTOCOL INFORMATION OF ALL EXISTING VERSIONS

After identifying all published versions of an SLR, its updates and replications (if they exist), the next activity is to extract the protocol information of all considered studies and organize and store them in a database. The protocol information includes the following:

Research Questions: The original SLR and updates addressed a set of clearly defined research questions. Usually, the SLR updates and replications address the same research questions as the original SLR. For instance, Paula & Carneiro (2016); de Aguiar Beninca *et al.* (2015); França *et al.* (2011); Vallon *et al.* (2018a); Mendes *et al.* (2014); Dantas *et al.* (2018); Guo *et al.* (2017) addressed the same research questions investigated in the original SLR. However, an SLR update could address similar questions or even have an additional question according to a new emergent trend, for example Mendes *et al.* (2020). Examples of SLR updates that added a new research question in the SLR update are Guo *et al.* (2017); Alabool *et al.* (2018); Nair *et al.* (2014); Hummel (2014). In counterpart, in SE, it is not common for an SLR update to discard some research questions presented in the original SLR, but we are aware of two occurrences of it in SE: Dantas *et al.* (2018) and (de A. Cabral *et al.*, 2023).

Search Strategy: The search methods adopted by the original SLR, updates, and replications, for instance, automated search on DLs and manual search and snowballing. For automated searches, it must include the search string (also the adapted search strings for each DL, if available) and the chosen DLs (Felizardo *et al.*, 2020a). For manual searches, the list of venues where the manual search was performed (and the web page links of these venues, if available). Finally, for snowballing, the list of studies considered as "seed set", the snowballing performed (backward or forward), and the number of iterations performed.

Over the years, SE SLR updates tend to use the same search strategy adopted in the original update. Examples of SLR updates that conserved the same search strategy as the original SLR is presented in Guo *et al.* (2017); Vallon *et al.* (2018a); de Aguiar Beninca *et al.* (2015); Riaz (2012); Pizzoleto *et al.* (2019); Dantas *et al.* (2018); Hoisl & Sobernig (2016); Alabool *et al.* (2018). On the other hand, some SLR updates perform small changes on the SLR update search strategy, such as modifying the search string (Jiang *et al.*, 2015) or reducing or increasing the number of DLs considered in the update (Boyle *et al.*, 2016; Nair *et al.*, 2014; Mendes *et al.*, 2014). Adopting the snowballing technique (Felizardo *et al.*, 2016; Wohlin *et al.*, 2020) as a mechanism to detect new relevant evidence for SLR updates is exemplified in Dantas *et al.* (2018); Ameller *et al.* (2016).

Search Period: The covered search period information is crucial to determine the years the search was performed by the original SLR, SLR update, and replications and what period the new update must cover (Felizardo *et al.*, 2020a). Also, in this step, we recommend elaborating a timeline with the years covered by the considered SLR studies. The timeline enables a time view of the years considered by the SLRs and the gap between updates or replications.

We recommend that the search period for a (new) update covers at least one month retroactive of the stated search period. For example, if an SLR stated the performed search period on February 2020, we recommend executing the new search covering studies published between January 2020 and the current execution date. Unfortunately, several published SLRs and updates do not mention the search period. Sometimes this information is omitted due to limited space for a detailed protocol description (e.g. in conference papers). In these cases, the search period to be considered is one year retroactive to the year of publication of the SLR. For example, if the SLR was published in 2020, the search period must cover all publications since 2019. These temporal measures are suggested to mitigate the risk of losing relevant

evidence. In order to facilitate future detection of the search period covered by the new SLR update in progress, we suggest adding the covered search period as an inclusion or exclusion criterion statement, as performed in Jiang *et al.* (2015); Nair *et al.* (2014); Boyle *et al.* (2016).

Selection Criteria: According to Felizardo *et al.* (2020a), the same selection criteria used by the original SLR should be used in updates. The selection criteria comprise inclusion criteria, exclusion criteria, and quality criteria. In practice, most SLR updates published in SE use the same inclusion and exclusion criteria adopted in the original SLR. Examples of it can be observed in Guo *et al.* (2017); Mendes *et al.* (2014); Boyle *et al.* (2016); Hummel (2014); Vallon *et al.* (2018a); Pizzoleto *et al.* (2019); França *et al.* (2011); Leyh & Sander (2011); Sulayman & Mendes (2011). Controversially, we identified three SLR updates that added new inclusion and exclusion criteria in the SLR update (Ameller *et al.*, 2016; Alabool *et al.*, 2018) or made minor changes to the criteria (de A. Cabral *et al.*, 2023). However, these three studies with changes in the inclusion and exclusion criteria also modified or added new research questions in the SLR update. Therefore, when adding a new RQ in an SLR update, it is recommended to review the selection criteria and adapt them in order to gather evidence to answer all RQs addressed in the study.

The recommendation in Kitchenham *et al.* (2015)'s guidelines is to analyze the quality of the candidate studies by establishing quality criteria. In practice, the study of Napoleão *et al.* (2017) showed that almost half of the SLRs analyzed in this study assessed the quality of the candidate studies for selection. In addition, several SLR updates adopted the same quality criteria as the original SLR, for instance, Nair *et al.* (2014); Boyle *et al.* (2016); Mendes *et al.* (2014); Manikas (2016); França *et al.* (2011); Hoisl & Sobernig (2016); Sulayman & Mendes (2011); Dantas *et al.* (2018). We identified only two SLR updates that adopted different quality criteria presented in the original SLR (Guo *et al.*, 2017; Ameller *et al.*, 2016) and one that improved the quality criteria assessment by adding additional criteria (Vallon *et al.*, 2018a).

In summary, we recommend applying the inclusion and exclusion criteria and quality criteria (if available) presented in the original SLR, replications, and updates on the new SLR update. Nonetheless, there are divergences among the criteria presented in the SLR versions. In that case, they should be analyzed and discussed by the SLR authors of the update in progress, and through consensus, make decisions and document them explicitly in the new update.

List of Included Studies: The list of included primary studies by the original SLR, updates, and replications is a vital asset to be used as a "seed set" for performing the forward snowballing technique for a new update (Wohlin *et al.*, 2020). In addition, it could be used as an instrument of reanalysis for the authors conducting the new update to acquire knowledge on the investigated SLR topic (Felizardo *et al.*, 2020a).

Synthesis Method: Kitchenham *et al.* (2015) assert in their book that SLRs can be sub-classified according to their synthesis method: quantitative or qualitative. Quantitative SLRs usually use as inputs experiments or quasi-experiments, and their data synthesis includes a tabulation of different outcomes through statistical and meta-analysis methods. On the other hand, qualitative SLRs are the most adopted type in SE, using as inputs usually textual data from case studies and ethnographic studies, for example. Their data synthesis includes narrative synthesis or even a set of classification schemes. According to Mendes *et al.* (2020), an SLR update can include a more recent meta-analysis or research synthesis method or even follow the same synthesis method adopted by the original SLR.

By analyzing some SLR updates published in SE, we noticed that most updates compare results between the original and the updated SLR. This approach aims to present changes and advances in the research questions investigated. Among examples of SLR updates that performed comparisons between previous and updated versions are Guo *et al.* (2017); Mendes

et al. (2014); Boyle *et al.* (2016); Hummel (2014); Vallon *et al.* (2018a); Alabool *et al.* (2018); França *et al.* (2011); Riaz (2012); Ameller *et al.* (2016).

It is essential to mention that all this extracted protocol information should also be reported in the current update, as this information will be crucial for the success of future updates (Felizardo *et al.*, 2020a). Unfortunately, transferring know-how from the SLR conduction is still tricky, even with the protocol data available. However, external repositories such as Github, Zenodo, or ArXiv are options to mitigate the lack of space and problems linked to SLR protocol information availability (Mendez *et al.*, 2020).

5.2.3 SEARCHING FOR NEW EVIDENCE THROUGH THE EXECUTION OF ONE FORWARD SNOWBALLING ITERATION

The snowballing technique was introduced in SE as a search strategy instrument in 2014 by Wohlin (2014). Considering the advantage of having the included studies by the original SLR, two years later, Felizardo *et al.* (2016) proposed using the forward snowballing technique (citation analysis) to update SLRs in SE using as "seed set" the included studies by the original SLR. An example of an SLR update that adopted this strategy can be observed in Dantas *et al.* (2018).

The promising results of using forward snowballing to update SLRs in SE were also recently investigated by Wohlin *et al.* (2020). This study demonstrated that performing only one iteration of the forward snowballing technique, using as "seed set" the included studies of the original SLR with the Google Scholar search engine, was the most cost-effective approach to search for new evidence for SLR updates. Therefore, we suggest conducting this approach to search for new evidence to update SLRs in SE.

5.2.4 SELECTING NEW CANDIDATE STUDIES

The selection of studies can be divided into two parts: First, the selection of new candidate studies, which consists of the initial selection of studies retrieved by the search activity analyzing only the studies' title, abstract, and keywords. Next, the final selection of studies consists of analyzing the selected candidate studies by performing a full-text analysis (Kitchenham *et al.*, 2015).

Differently from this consecutive two-part traditional selection process proposed for SLRs by Kitchenham Kitchenham *et al.* (2015), we propose for SLR updates to perform only the first part: select candidate studies based only on title abstract and keywords. The second part of the process should be performed during the SLR update execution activity (Section 5.2.8). We chose this approach because we aim to reduce efforts during the selection activity. At this stage of the CSLR process, it is still not known if the SLR needs to be updated. This need is evaluated when verifying if the SLR or the SLR update needs to be updated (see Section 5.2.6).

5.2.5 MAKING EVIDENCE AVAILABLE TO POTENTIAL STAKEHOLDERS

As already mentioned, over the last few years the open science movement has gained importance in SE. It consists of making research artifacts available to the public addressing open access, open data, and open-source practices (Mendez *et al.*, 2020). Open science practices directly impact the conduction of an SLR update. The availability of primary evidence positively affects the maintainability of SLRs. Thus, researchers should adopt open science practices in any EBSE study.

In this sense, in this activity, we recommend making all the information extracted from all SLR protocol versions (see Section 5.2.2) as well as the list of candidate studies (see

Section 5.2.4) available online. While there is no dedicated repository to manage this data, we suggest adding them to open dissemination repositories such as Zenodo or Github, where this data can be freely accessed and updated, if necessary (Mendez *et al.*, 2020).

5.2.6 VERIFYING IF THE SLR NEEDS TO BE UPDATED

Before proceeding with an SLR update, deciding whether an SLR needs updating is essential. Mendes *et al.* (2020) proposed and evaluated the adoption of a framework to support this decision in their study. The 3PDF is a three-step method composed of seven questions to be answered based on the SLR candidate to update.

In this activity, the researcher aims to perform an SLR update and must submit the SLR to all questions and decision sieves described in the 3PDF. As a complement to the description of the questions from the 3PDF steps described in Mendes *et al.* (2020), we propose as support to answer two specific questions:

- (i) *Step 1.b Has the SLR had good access or use?* – One way to evaluate the impact (access/use) of an SLR is by observing the number of citations of the SLR under evaluation. However, considering only citations is not a silver-bullet solution. A more recent SLR could be due to an update (in SE, some areas evolve rapidly) and fewer citations than an older SLR. To mitigate this threat, we recommend adopting the formula introduced by Octaviano *et al.* (2022), which proposes a coefficient based on publication year and the number of citations to decide the relevance of an SLR.
- (ii) *Step 2.b Are there any new studies or information?* This question can be answered with the support of the results obtained in selecting new candidate studies (Section 5.2.4). The number of candidate studies and possible analysis of these studies provide a significant indication of new relevant studies and information.

Examples of SLR updates assessed under the 3PDF are described in Mendes *et al.* (2020). In this study, the authors evaluated the need for 20 SLR updates in SE; they concluded that 14 of the 20 SLRs did not need updating. These results underscore the importance of verifying the need to update an SLR before conducting an update, thus avoiding a waste of time and effort by SE researchers.

Nevertheless, the results obtained by executing the 3PDF should be shared in an open repository, preferably in the same repository where all the other information about the SLR update under investigation is stored.

5.2.7 REPORTING THE NEED TO UPDATE THE SLR

This activity reports the results obtained by conducting the previous activities to potential stakeholders, especially by verifying the need to update the SLR (Section 5.2.6). Studies (Nepomuceno & Soares, 2019; Felizardo *et al.*, 2020a) highlight the importance of having the participation of at least one author of the previous version of the SLR in its update process. In practice, the following SLR updates included at least one author of the original SLR in the updated version: Jiang *et al.* (2015); Manikas (2016); Nair *et al.* (2014); Boyle *et al.* (2016); Alabool *et al.* (2018); Mendes *et al.* (2014); Guo *et al.* (2017). Moreover, a few SLR updates have exactly the same authors as the original SLR: Leyh & Sander (2011); Paula & Carneiro (2016); Sulayman & Mendes (2011).

We suggest contacting potential stakeholders (i.e., authors of the SLRs' older versions) to disseminate the results of this analysis. If the need for updating the SLR is positive and the authors intend to update it, it is valuable to verify the possible interest of the original SLR authors in contributing or collaborating with the SLR update under evaluation. Even if the authors do not intend to proceed with an update, we advise sharing the need for an SLR

update with the original SLR authors informing them of the need for an update. No contact is necessary if the need for updating the SLR is negative.

5.2.8 UPDATING (IF NECESSARY) AND EXECUTING THE SLR UPDATE PROTOCOL

After concluding that an SLR needs to be updated (see Section 5.2.6), the next activity is to perform the full SLR update. In this activity, the gathered protocol information (see Section 5.2.2) should be revisited to elaborate the new SLR protocol update (Nepomuceno & Soares, 2019, 2018; Felizardo *et al.*, 2020a; Mendes *et al.*, 2020). The goal of visiting the protocol is to verify if the SLR update needs to address different protocol steps from the original SLR (i.e. different search method). However, it is worth mentioning that some protocol information is already defined, such as the search strategy using forward snowballing due to its proven suitability and efficiency (Wohlin *et al.*, 2020).

With the protocol revised and ready, the next step is to execute it. To perform the protocol execution, we suggest following the SLR execution activities described in Kitchenham & Charters (2007) including applying the selection criteria on the full text of candidate studies, data extraction, and synthesis and finally, answering the SLR research questions.

5.2.9 REPORTING/PUBLISHING AND MAINTAINING AN SLR UPDATED

The last activity of the SLR process is to report and disseminate the SLR results Kitchenham *et al.* (2015). In the SLR update scenario, these two activities are crucial too, but we propose to complement them by adding the maintenance activity.

SLR updates can be reported similarly: through descriptive analysis of technical reports, Ph.D. theses, and peer-reviewed publications. The most common form of dissemination is

through peer-reviewed publications in SE venues (journals and conferences). In SE, there are several examples of SLR updates publications: the study of Mendes *et al.* (2020) presents 20 examples of SLR updates published in SE between 2008 and 2019. From our analysis, the original venue where the original SLR was published is a suitable source for submitting the SLR update.

Regarding the maintenance of an SLR update, we suggest the authors that proposed the new SLR update keep periodically feeding the repository where the data and information about the SLR were stored to facilitate future updates. The survey performed by Nepomuceno & Soares (2019) report that more than 2/3 of the survey participants would be willing to share their SLR artifacts in a common repository.

With the publication of the updated SLR, the CSLR process starts all over again. The new updated version of the SLR must be linked to the other version(s) and continuously follow the CSLR process activities.

5.3 PERCEPTIONS ON CLSR PROCESS AND GUIDELINES

In this section, we perform an expert evaluation to obtain feedback on the CSLR process activities flow and guidelines, as well as on perceptions on the relevance and potential adoption of the CSLR by the SE community.

5.3.1 DESIGN AND EXECUTION

In order to increase the number of expert participants in this study, we conducted it through a questionnaire. We based its design on the exact phases proposed by Pfleeger (1995) for survey design. All phases and design decisions taken are described below.

Phase 1: Setting the objectives – As previously mentioned, our objective is to obtain feedback from the SE community on the CSLR process activities, their execution flow, and the proposed CSLR guidelines (proposed in Section 5.2). Also, we intend to obtain perceptions on the relevance and potential adoption of the CSLR idea by the SE community.

Phase 2: Designing the expert evaluation and defining participants – We developed a cross-sectional questionnaire (detailed in Phase 3) asking the participants to carefully analyze the CSLR process and guidelines and based on their analysis and experience in the SE SLRs domain, provide feedback on the CSLR process activities, execution flow, and relevance. Since our goal is to obtain perceptions from SE SLR expert researchers, the experts' participants must follow these three criteria: (i) be researchers who performed research on the SE SLR domain; *AND* (ii) be familiar with the SLR conduction process in practice – having conducted and published at least three SLRs or SMs; AND (iii) has no involvement as a contributor to this doctoral research. The Ph.D. candidate and the two contributors of this doctoral research created a suggestion list of potential participants as an initial step. Next, we assessed the participants' eligibility following the three established criteria. In total, 13 participants were selected and invited to participate in our study.

Phase 3: Developing the expert evaluation – The questionnaire is divided into three sections. The first section contains the consent form of participation and four closed demographic questions that aim to understand the participants' background and experience with conducting and updating SLRs.

The second section evaluates the CSLR process activities and guidelines in detail. The CSLR process and its activities were highlighted to facilitate comprehension. The experts were questioned about their level of agreement with the process activity and execution flow. We used the Likert Scale method (Likert, 2010) varying from "Strongly agree" to "Strongly

disagree". The final median score represents the overall agreement toward the subject matter. In addition to each question, the participants were asked to justify their answers and provide comments.

The third and last section contains three open questions. The participants were asked to share their observations about the relevance and impact of the CSLR in SE. In addition, they were invited to share any additional comments (e.g., suggestions and concerns). An online version of the questionnaire is available online (Napoleão *et al.*, 2023a).

Phase 4: Evaluating the questionnaire – We conducted a pilot questionnaire evaluation with two participants with the same population profile. We discussed their opinion to ensure that questions were understood as suggested in Wagner *et al.* (2020). Two modifications to the questionnaire are: (i) a synthesis to reduce the CSLR guidelines description and (ii) a graphical improvement in the BPMN process image of the CSLR process indicating directly in the process the activity asked.

Phase 5: Obtaining valid data – The selected 13 expert participants were invited to participate in our study by a direct email request from the authors. We used the Limesurvey[23] tool to develop our questionnaire since it was suggested by the first author's university ethical guidelines. The questionnaire was available from December 8th to 23rd, 2022. Seven days after the invitation was sent to the participants, a reminder email was sent with the goal of capturing more entries.

Phase 6: Analyzing the data – All the answers submitted by the participants were stored in the *Université du Québec à Chicoutimi (UQAC)* LimeSurvey server. A total of 6 experts answered the questionnaire totaling a 46% response rate.

[23] https://www.limesurvey.org

The results of the expert evaluation are described in detail in Section 5.3.2 and the thematic analysis applied to elaborate an inventory of improvements for the CSLR process and guidelines are presented in Section 5.3.3.

5.3.2 RESULTS

We start the analysis of our results by summarizing the profile of the SE SLR experts participating in our study. All six expert participants of our study identified themselves as professors or senior SE researchers. Also, all of them claimed that they published four or more SLRs in journals or conference proceedings.

Regarding the experts' experience with the conduction of SLR updates, Figure 5.2 illustrates that only one expert has not conducted an SLR update yet. However, he/she claimed to have conducted four or more SLRs in SE. All the other five experts have conducted at least one SLR update. Four of our experts have conducted one or more SLR updates. More specifically, one expert conducted five SLR updates, publishing three of them. Another expert conducted three SLR updates, publishing two of them. Two experts conducted two SLR updates, but only one was published. Finally, one expert mentioned conducting one SLR update but not publishing it in a conference or journal. Overall, the experts' experiences with SLRs and SLR updates lead us to conclude that all participants have valuable knowledge and practice in the SLR process.

Moving on to the CSLR process and guidelines perceptions, next presents the SE SLR experts' evaluation of each activity and guidelines description of the CSLR process. In the sequence, we described the experts' observations on the execution flow of the CSLR process. Lastly, we report the experts' observations on the relevance of the CSLR concept and process to the SE area.

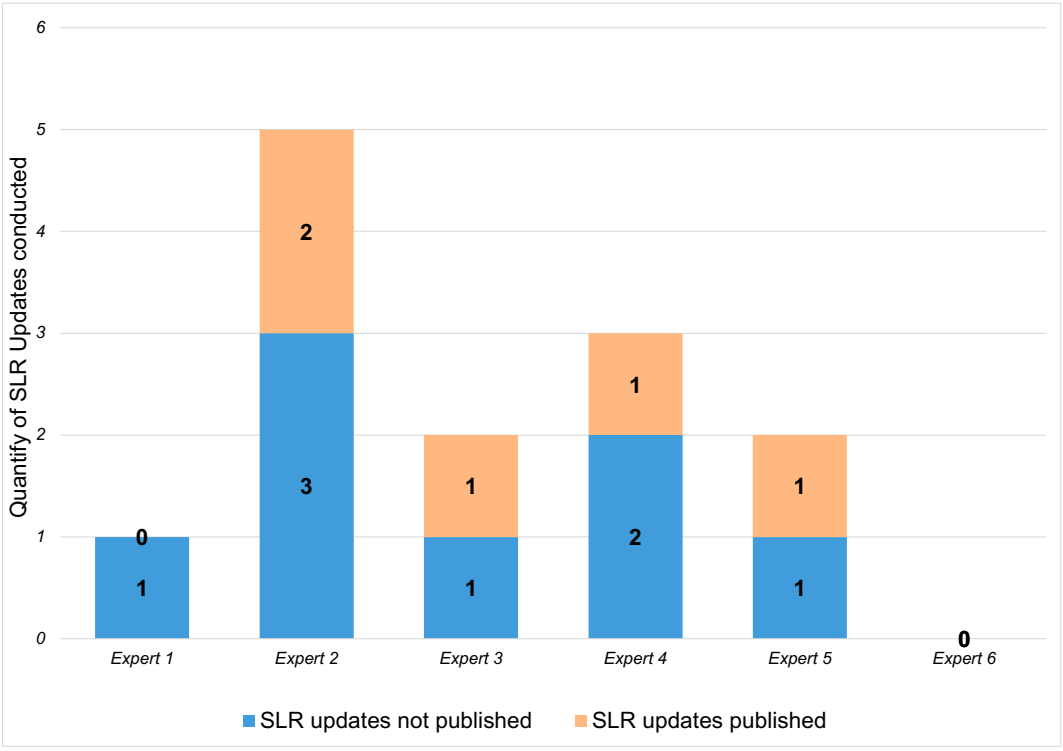


Figure 5.2 : Overview of the SLR updates conducted by experts. ©Bianca Minetto Napoleão.

PERCEPTIONS ON CSLR PROCESS ACTIVITIES

As demonstrated in Chapter 4 and Napoleão *et al.* (2022b), the CSLR process is compatible with manual execution and is able to achieve its goal without the support of automated tools. However, some CSLR process activities were designed to benefit from automation. Thus, we chose to exclude activities that rely on automation from individual evaluation by the experts. As shown in Figure 5.3, which illustrates the CSLR process shared with the expert participants, the activities highlighted in light gray did not undergo an individualized evaluation in the evaluation questionnaire. They were only considered in the evaluation of the process execution flow. We now detail our expert evaluation of the CSLR process activities and execution flow.

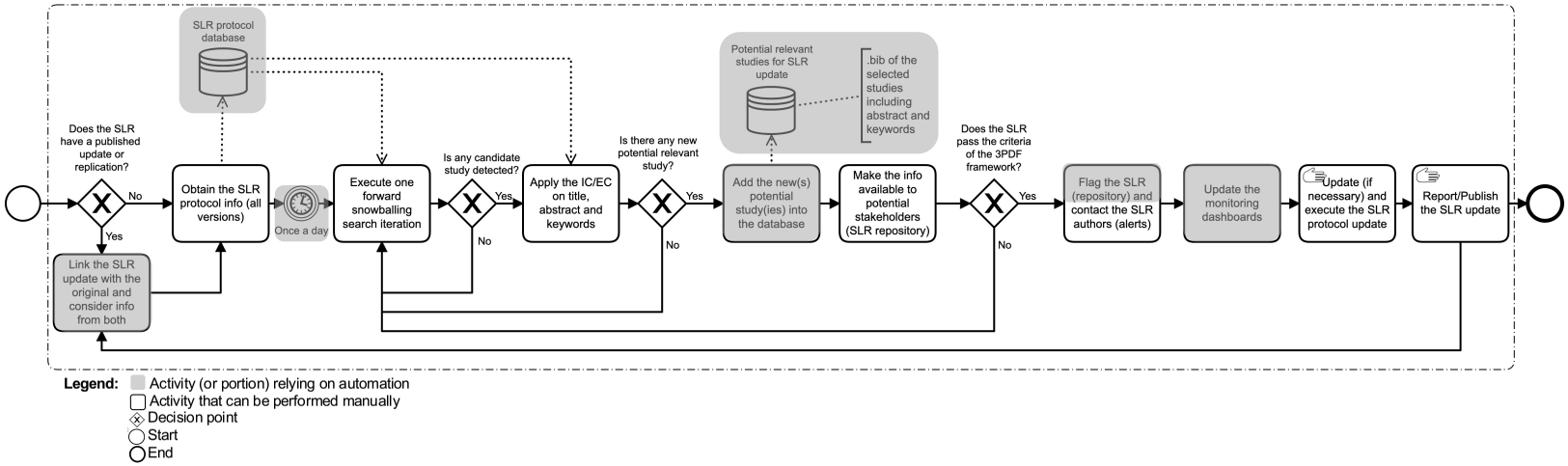


Figure 5.3 : CSLR process shared with the experts for evaluation. Adapted from Napoleão *et al.* (2022b).

The level of agreement of experts with each activity, including decision points of the CSLR process, is presented in Figure 5.4. Overall, the expert respondents seem to perceive the CSLR process positively. Most respondents either strongly agreed or agreed with the different decision points and activities. However, some respondents had neutral or negative responses to specific activities and decision points, indicating potential areas of improvement for the CSLR process. We now detail the experts' answers on each CSLR process activity and decision point.

The CSLR process begins with the decision point that questions **"Does the SLR have a published update or replication?"** aiming to identify all existing versions of the SLR under investigation. All the experts agreed with this point, of which 5/6 strongly agree, and 1/6 agree. Also, they recognized the need to identify SLR updates or replications before proceeding with the update of an SLR in order to avoid an unnecessary update. The experts mentioned two suggestions for improvements in the CSLR guideline:

> *(i) "The update may have been published many years ago or recently, and it makes a big difference. It should, of course, always be integrated with the original SLR, but it matters whether or not it is reasonable to update. I am aware that you are aiming for continuous, but I doubt that researchers will do the work if it is not publishable since the delta is too small."*

> *(ii) "I think you should better motivate the reason and rationalities that lead you to suggest locating the SLR under analysis using Google Scholar DL to assess if there are any updates to the SLR either it be an update or a pure replication."*

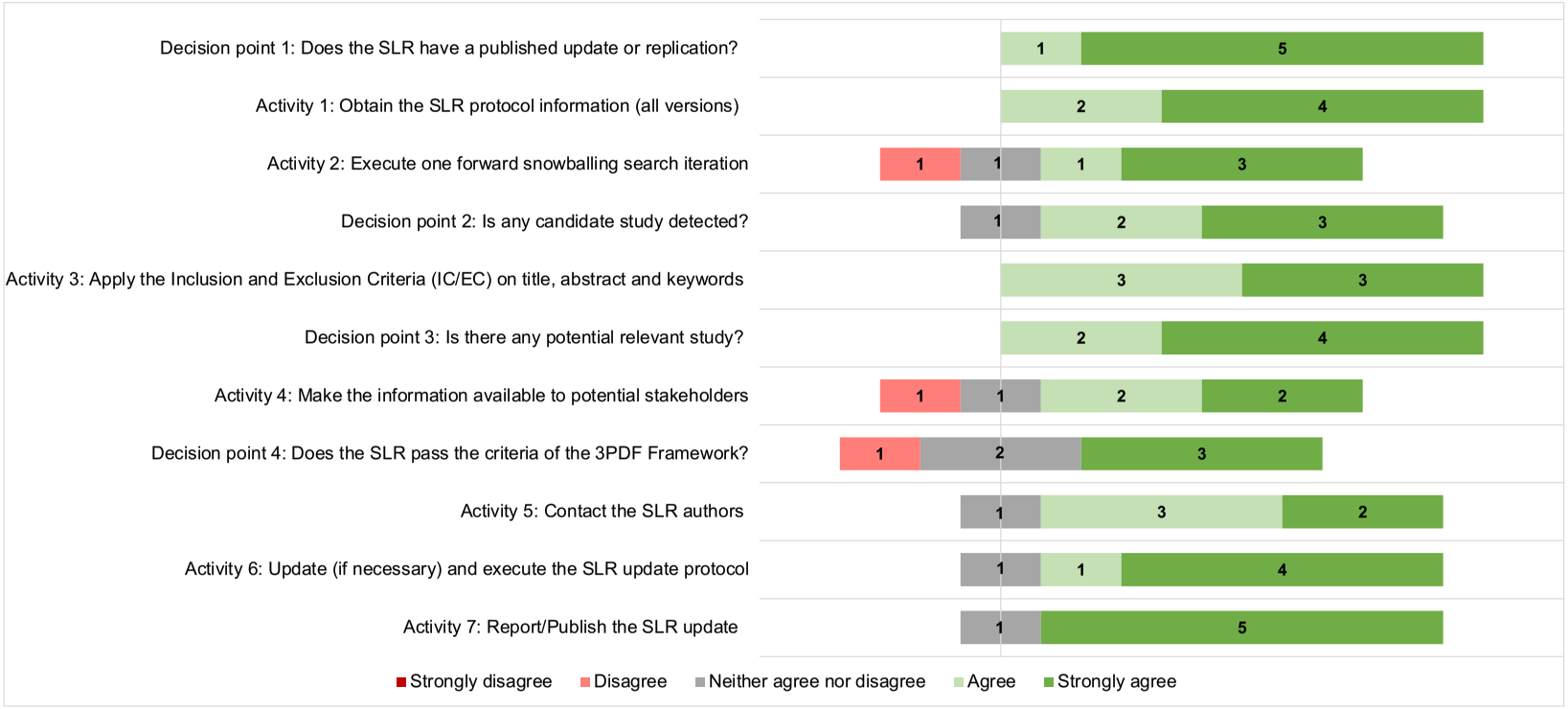


Figure 5.4 : Expert answers to CSLR process activities and decision points. ©Bianca Minetto Napoleão.

As shown in Figure 5.4, the **Activity 1: Obtain the SLR protocol information (all versions)** had all positive responses.

Despite the recognition of the importance of obtaining the protocol information of the original SLR, existing replications and updates, 5/6 experts expressed their concern about the availability of the protocol information by adding:

> *"The consistency in terms of replicating the protocol as closely as possible (and documenting/justifying deviations well) can only be achieved when the original protocol is available in a detailed format."*

> *"Currently, SLR process is permeated by many decisions that sometimes are not documented correctly in the protocol/final reports..."*

We corroborate this concern, but introducing the open science movement in SE could be a valuable ally to help change the mentality of making SLR artifacts openly available. Indeed, one of the expert respondents mentioned that creating better ways to store and retrieve SLR protocol information could catalyze SLR updates. One of the goals of the CSLR process is to unify SLR protocol information in an open repository.

One expert suggested addressing in the CSLR guidelines how to manage the obtained protocol information from the existing SLR, its updates, and replications (if they exist). The expert finishes his/her remark by adding *"So although I certainly agree with the statement. I think it is important to consider the risk and complexity of the CSLR once it merges and puts together the data from all the existing SLRs, its updates, and replications (if they exist). Probably at this point in SE, there are not so many SLRs that have been updated and extended, so the problem may not be so critical yet (we should however reflect on it)."*. This expert com-

ment corroborates with one of the objectives of our study: to contribute to the maintainability of the SLRs in SE.

Regarding the **Activity 2: Execute one forward snowballing search iteration**, 4/6 experts agree with this activity (3 strongly agree, one agrees). They also added that essential studies have already demonstrated that only one iteration of the forward snowballing technique is enough to update a review (the most time and effort-effective alternative). We had one neutral and one disagreeing response both justifying that, in their opinion, the forward snowballing technique cannot be replaced or complemented by a database search. Adopting forward snowballing as an initial search step is an alternative that could indicate that the area of the SLR under evaluation is evolving and that the SLR results may be becoming outdated.

The goal of **Decision point 2: Is any candidate study detected?** is to verify if any study is returned from the execution of the forward snowballing activity (Activity 2). 3/6 experts strongly agree with this decision point described in the CSLR guidelines while 2/6 experts agree with this decision point. One expert highlighted the combination of tool support and human efforts that could be added to this step to maintain the CSLR continuously. The neutral response pointed out *"...the problem is really with which ICs/ECs are used. If you can give guidance on this, your process will be more useful."* His/her statement directly refers to the next activity (Activity 3) of the CSLR process addressed below and also mentioned by the expert to consider his/her statement for the next activity.

The **Activity 3: Apply the Inclusion and Exclusion Criteria (IC/EC) on title, abstract, and keywords** had all positive answers. One expert justified his/her answer by adding "*given the dimensions and the complexity of the CSLR since it puts together information/data from all previous SLRs and its updates/replications, it is surely worth reducing effort at this point and pushing the more intense work to the later activities."* Two experts suggested

keeping in mind the SLR quality aspects in this activity by clarifying what IC/EC should be employed to maintain the quality of the studies during the execution of the quality assessment in later process activity. One expert even added on the importance of considering the evidence quality by mentioning *"Stratifying papers on their quality and level of evidence are key and few SLRs do."*

Concerning the **Activity 4: Make the information available to potential stakeholders**, overall, there is a general agreement that making the SLRs information available is important for maintaining the review, increasing transparency, and enabling reproducibility. Moreover, one expert highlights *"Unfortunately, heavily underrated by the SE community."*

One expert agrees with the activity but expresses concerns about potential second-order effects, such as researchers being hesitant to share data that could interfere with their publication process.

The one expert who agrees, the neutral, and the one who disagrees (3/6 experts) shared a common suggestion. While making research artifacts available is essential, it should not be done at this stage as the SLR is still being evaluated for a potential update.

> *"Making research artifacts available is very important. Indeed, many initiatives, such as Artifact Evaluation committees, have increasingly characterized conference tracks. Generally, supporting any publication with artifacts that underlay the research is highly recommended. The study is supported by evidence of the outputs of Activity 1 and Activity 3, i.e., the artifacts. But artifacts are also what comes out of the activities that are carried out after Activity 3, i.e., decision point 4 (guidelines state: the results obtained by executing the 3PDF should be shared in an open repository, preferably in the same repository where all the other information about the SLR update under investigation is stored.) Furthermore, at*

this stage, we are still determining if the SLR is worth updating (which is what we decide in Activity 6). So is it worth publishing information/artifacts in a public repository if the information is incomplete and does not lead to an actual update of the SLR?"

"I would therefore certainly publish the artifacts, but not at this stage..." "...access should be provided once the analysis is complete."

Regarding the **Decision point 4: Does the SLR pass the criteria of the 3PDF?**, two experts are unfamiliar with the 3PDF and therefore cannot fully judge its use at decision point 4. One expert agrees with its use, but disagrees with the placement of this decision point. However, due to the inputs of the 3PDF (Mendes *et al.*, 2020), it is impossible to perform this analysis earlier in the process.

On the other hand, two experts strongly agree on the value of the 3PDF. One of them emphasizes that the information coming from the 3PDF should be shared in the same repository as all other CSLR data.

As illustrated in Figure 5.4, **Activity 5: Contact the SLR authors** presents 5/6 positive answers demonstrating a general agreement that involving at least one author of the previous version of the SLR in the update process is beneficial. However, the experts raised some concerns about the practicality of involving authors if they are reluctant or difficult to reach, particularly in the case of older SLRs. One expert, in particular, suggested considering this limitation on reaching SLR authors and detailing in the guidelines the role of the original SLR authors in undertaking the SLR update.

One expert highlighted *"... a step unfortunately heavily underrated by the community (not only as a basis for collaborations but also as a collegial step to informing the original*

authors about updates on the research of their interest).” Another expert emphasized *“It is important to involve previous authors because they have knowledge of the protocol and the entire review process”*. Moreover, involving the original authors can provide additional insights and expertise that could improve the quality of the updated SLR.

In summary, involving at least one author of the previous version of the SLR in the update process is beneficial. However, it is crucial to acknowledge that practical difficulties may arise, especially as the number of SLRs increases and authors of older studies become harder to contact.

Concerning **Activity 6: Update (if necessary) and execute the SLR update protocol**, again, 5/6 positive answers were provided by the experts, but among these five positive answers, four strongly agreed with the activity.

The expert who neither agreed nor disagreed mentioned his/her impression that we assumed the extraction/synthesis methods would stay fixed over time. In this sense, he/she recommended that the extraction and synthesis methods might need to be updated as the topic evolves. Another expert suggests the need for automation tools to facilitate this task in the future, a point already envisaged by the CSLR idea. Both suggestions are under consideration for future work.

Finally, we present remarks on the last activity of the CSLR process, **Activity 7: Report/Publish the SLR update**. 5/6 experts strongly agree with reporting and disseminating (e.g., through publications) the SLR update results. Two experts emphasized the need to publish the SLR update results: *“Not publishing updated results is like shouting into an empty forest. It might feel good but provides little value.” “This is the driving force. If not publishing our results, we do not get sufficient "credit", and the research is less visible...”*.

Two experts mentioned that a central (open data and publication) repository would benefit the community. The repository represents an up-to-date documented SE SLR body of knowledge, including links among the SLR, SLR updates and replication done, and verification of the current state of an SLR update.

PERCEPTIONS ON CSLR PROCESS EXECUTION FLOW

The experts were asked about their level of agreement with the execution flow of the CSLR process, including activities dependent on automation. In Table 5.2, we present the answers regarding the experts' agreement level and justifications about the CSLR process as a whole.

As described in Table 5.2, overall, the CSLR process execution flow is well-received and seen as a valuable approach to updating SLRs by the participants. However, there are also concerns about:

(i) automation support to the CSLR process – two participants mentioned the relevance of automation support to reducing efforts during an SLR update. However, one participant highlighted that activities in the CSLR process that count on automation could also be performed manually by a researcher. The case study presented in Chapter 4 demonstrates the feasibility of applying the CSLR process manually;

(ii) usability of the CSLR in a continuous way – indeed, the CSLR will be used by a researcher in order to verify the state of an SLR. Nevertheless, the idea of CSLR is that the researcher uses the outputs of the process as well as contributes with inputs to keep the process running. Furthermore, the CSLR process and guidelines can also be used to perform a spot update of an SLR; and

Table 5.2 : Experts' comments on CSLR execution flow as a whole. ©Bianca Minetto Napoleão.

Agreement level	Expert's comments
Strongly agree	*"The flow is natural and makes absolute sense."*
Strongly agree	*"I agree with the flow of the CSLR process. It is important to highlight that the SLR needs strategies to reduce effort and automate tasks in the update. I see this in several process CSLR activities. This is essential to succeed."*
Strongly agree	*"Totally agree, but I believe that are barriers to be transposed. For example, the state of tools is currently immature. Another problem is that CSLR proposes deep changes in how reviews are done, hence, to be successful the agreement and cooperation of publishers (IEEE, ACM, etc.) are essential to disseminate this idea and make it possible. Another suggestion is to detach the CSLR process to automation since the tools to automate SLR are still premature. In my opinion, researchers could perform the activities tagged as to be executed by automation. This permit CSLR to be used even if technical aspects are not completely satisfied."*
Agree	*"Yeah, looks good. I'm not convinced it addresses the harder questions around the quality evaluation of studies, extraction and synthesis which is not always done in a good way in existing SE SLRs."*
Agree	*"In general, it is good. However, I wonder if researchers truly will work continuously. It is more likely that they look into it regularly; for example, once every quarter, half a year or yearly, we look into the state of an SLR and decide if an update is needed."*
Agree	*"It is clearly useful. I think it will work easily on SLRs that are planned and organized with this CSLR framework in mind, where the information is gathered and published as artifacts and the protocol is clear and made available from the beginning. It will be much more complex and cumbersome for existing SLRs that have not been planned with an intent of being updated, or have material shared, or have available authors."*

(iii) application of the CSLR process in published SLRs – the update of an SLR that considers the elements of the CSLR process will be smoother. On the other hand, the artifacts requested by the CSLR process are the same ones described in the SLR guidelines (Kitchenham *et al.*, 2015) that should be documented and available. The CSLR process emphasizes the idea of open science by seeking to draw attention from researchers to make information from their SLRs available openly. The CSLR process can be performed without all available artifacts from a published SLR. In the case of unavailable artifacts, the authors

can be contacted, and in the worst case, these artifacts can be estimated to enable the execution of the process. The critical point is that the new SLR update must be well-documented, and its artifacts must be openly available for future iterations of the CSLR process.

Next, the participants were asked if they would change the order of execution of any activity of the CLSR process. They were also asked to justify their answers. As a result, 3/6 participants mentioned that they did not change the execution order of the activities of the CSLR process. One participant suggested moving the dissemination to stakeholders at the final step of the CSLR process. Another participant suggested making all the CSLR artifacts available only at the end of the process. The last participant suggested moving the application of the 3PDF earlier in the process. However, given that the antecedent activities of the process serve as inputs for the execution of the 3PDF, the proposed alteration is not viable.

When the participants were asked about their opinion on adding, removing or modifying any activities of the CSLR process, 4/6 participants mentioned that no removal, addition, or modification was needed. On the contrary, one participant highlighted again (already mentioned in his/her feedback on the respective process activity) his/her concerns about the exclusive adoption of the forward snowballing technique. Another participant emphasized adding more activities to judge the quality of studies and their evidence and activities that bring insights into the extraction and synthesis of information from studies.

Last but not least, the participants were questioned if they would make their SLR data (e.g., protocol, data extraction form, included/excluded study list) available in an open repository (e.g., Zenodo, Github, or a dedicated platform). Except for one participant, all the others answered “yes”. Among the comments of the positive participants, they highlighted the importance of maintaining the SLR and keeping its information in an open repository to keep it visible and disclosing SLR manuscripts data and metadata. One participant mentioned ‘...

critical with openness and transparency. Many choices are involved in doing an SLR, which is rarely shared. If this is to be done continuously, the relative importance of this will increase.".

The single participant who answered "no" to this question mentioned that he/she would instead make the SLR information available through a publication and then keep the SLR supplementary data on his/her institution's web pages. Opposite this vision, one participant mentioned in his/her comments the relevance of having a dedicated platform for storing SLR data. SLR repository initiatives are strongly present in the medical field, such as the *Cochrane Library*[24], which contains a dedicated database of SLRs to support health-care decision-making; *PROSPERO*[25] an international database of prospectively registered SLRs funded by the United Kingdom National Institute for Health Research (NIHR); and *SDSR Systematic Review Data Repository*[26] freely accessible repository of SLRs supported by the Agency for Healthcare Research and Quality (AHRQ) – United States. We believe that a dedicated repository can benefit the SE area by providing a big picture of the body of knowledge on SE research summarized by SLRs, besides the benefits related to avoiding duplication of existing studies and facilitating the SLR update.

OBSERVATIONS ON THE RELEVANCE OF THE CSLR CONCEPT AND PROCESS

The last section of the questionnaire aims to obtain observations from the SLR SE expert participants on the relevance and impacts (negative and positive) of the CSLR process and guidelines. Lastly, they were asked to leave any additional comments (e.g., concerns and/or suggestions) about the CSLR proposition.

[24] https://www.cochranelibrary.com

[25] https://www.crd.york.ac.uk/prospero

[26] https://srdr.ahrq.gov

All participants recognized the relevance of the CSLR process and guidelines. Participants pointed out that the lack of a process will result in non-standardization, rework, and much effort. Therefore, in their opinion having a process and guidelines can benefit SE in ways such as:

Optimizing working time and reducing errors: the CLSR process has its activities pre-defined, supporting the SLR update steps. Also, the CSLR guidelines are essential to define a reliable standard and avoid wasting researchers' time performing unnecessary rework.

Providing a defacto reference: the CSLR process and guidelines provide a reference guide to SE researchers who want to update SLRs on what to do and how to do it. One participant that emphasized the need for support and assurance of commonality among researchers left the following reflection: *"we expect software developers to follow a process, so why should we not?"*.

On the one hand, one participant stated *"I think that at this stage of the progress of a community, we are dependent on such guidelines and, thus, I think that these are very relevant."*. On the other hand, another participant expressed his/her concerns that the CSLR process does not address an innovative summarizing technique to produce insights from it.

The positive and negative impacts (challenges) mentioned by the expert participants are summarized in Table 5.3. In summary, the experts mentioned more positive aspects than negative ones.

In the additional comments question, only two participants left comments. Both comments acknowledged and appreciated the authors' efforts in the CSLR research.

Table 5.3 : Positive and negative impacts of the CSLR process and guidelines. (The number in parentheses refers to the number of experts who mentioned the respective impact). ©Bianca Minetto Napoleão.

Positive Impacts	Negative Impacts (Challenges)
– Optimization of the working time to conduct an SLR update (2). – Ease assessment of submitted updates (e.g., as reviewers and external authors relying on updates) (1). – Terminological consistency across different SLRs research projects (2). – Constantly updated evidence avoiding that SLR outdated results mislead researchers (1). – Support primarily for novice researchers on updating SE SLRs (3). – Stimulate maintenance of the connection with authors/researchers of the original SLR (1). – Collection of material and make it available in an artifacts package that supports/backup the reported/published/updated SLR (2).	– Change in the mindset of the SE researchers to accept the CSLR idea (2). – Lack of flexibility of the CSLR process (e.g., a researcher be able to adapt activities that he/she does not fully agree with) (1). – Complexity to apply on more dated SLRs that may not have much information publicly available (same risk encountered when we replicate older empirical studies) (2). – High effort demand to recover existing information of SLR and its existing updates/replications (1).

5.3.3 DISCUSSION: IMPROVEMENTS ON THE CSLR PROCESS AND GUIDELINES BASED ON EXPERTS EVALUATION

The goal of collecting experts' perceptions of the CSLR process is to perform improvements to the CSLR process and guidelines based on their experience conducting and updating SLRs in SE. We performed a thematic analysis (Cruzes & Dyba, 2011) to organize and synthesize the experts' perceptions into improvements for the CSLR process. We followed the steps recommended in Cruzes & Dyba (2011). Firstly, we performed an initial reading of the experts' perceptions and then we identified segments of the perceptions (text) that are improvement suggestions for the CSLR process. Secondly, we performed a coding analysis by labeling the segments of the perceptions and transforming them into codes. Thirdly, we

translated the codes into themes by increasing the level of abstraction and generalizability. Lastly, we performed a second cycle of the translation of the themes in order to create a high-level categorization of the improvement suggestions. In the following, we further detail the performed coding process and the translation of the codes into themes.

Coding is a method that enables the categorization, organization and grouping of similar data into categories using tacit knowledge. (Cruzes & Dyba, 2011). According to Linneberg & Korsgaard (2019), the coding process essentially generates an inventory of the data, enabling deep, comprehensive, and thorough insights from the data. A code can be a label or tag that represents descriptive of inferential information put together (Cruzes & Dyba, 2011).

As the first step of the coding process, we established a research question to guide the coding analysis (Linneberg & Korsgaard, 2019; Saldana, 2012): *What do the experts emphasize as potential improvement elements to integrate into the CSLR process and guidelines?* The second step of the coding analysis consists of reading or reviewing relevant research literature to acquire knowledge about the subject under investigation. Considering the experience in SLR updates of the Ph.D. student and collaborations of this research, we only reviewed some specific concepts on SLR update research. The third step involves collecting the data in a format that allows systematic analysis. In our case, we unified in a spreadsheet file all the experts' perceptions previously identified during the first step of the thematic analysis to facilitate the categorization and analysis of the data.

The process of coding can be performed according to three approaches (Cruzes & Dyba, 2011): *deductive*, which uses a provisional "start list" of the codes previously established; *inductive* where the codes emerge entirely from the data under analysis; and *integrated*, which is a pathway between the deductive and inductive approaches. We adopted the deductive approach since we have as bases the CSLR process elements (i.e., activities, decision points

and execution flow) that could be used as initial codes (see Figure 5.5). Although a "start list" of codes can facilitate the development of new inquiries by leveraging previous insights, we are not limited to constraining the data into these predefined categories because the existence of a code may lead to forced categorization of the data, potentially distorting the findings. The results of our coding analysis were evaluated by two researcher collaborations in order to mitigate errors and misinterpretations before moving on to translating these codes into themes.

Cruzes & Dyba (2011) suggest the use of visual representations to support the coding translation into themes. Therefore, Figure 5.5 illustrates the resulting map of our thematic analysis including the identified codes and the results of two cycles of the translation of themes performed.

As shown in Figure 5.5, most outputs from the code translation into themes led to improvements in the CSLR process and guidelines. Nevertheless, the code *Placement of the 3PDF in the CSLR process* could not be translated as an improvement suggestion theme due to the limitation of the required inputs of the 3PDF (Mendes *et al.*, 2020). Thus, it is impossible to perform this analysis earlier in the process.

Lastly, our second cycle of translations into themes led to high-level categories of improvement suggestions: (i) Improvement points for future work – We considered future work the theme *Acceptance/adhesion by the SE community* and all elements linked to *Automation*, and *CSLR Repository* since they require a maturity time to obtain a realistic view of the SE community and to develop the repository and automated support tools; and (ii) Accepted improvements – We divided the accepted improvements point into two sub-categories, improvements of the CSLR execution flow and clarification in the CSLR guidelines (see Figure 5.5). Both cycles of the translations into themes were reviewed and validated by an experienced research collaborator in the SLR process domain. The thematic analysis presented in

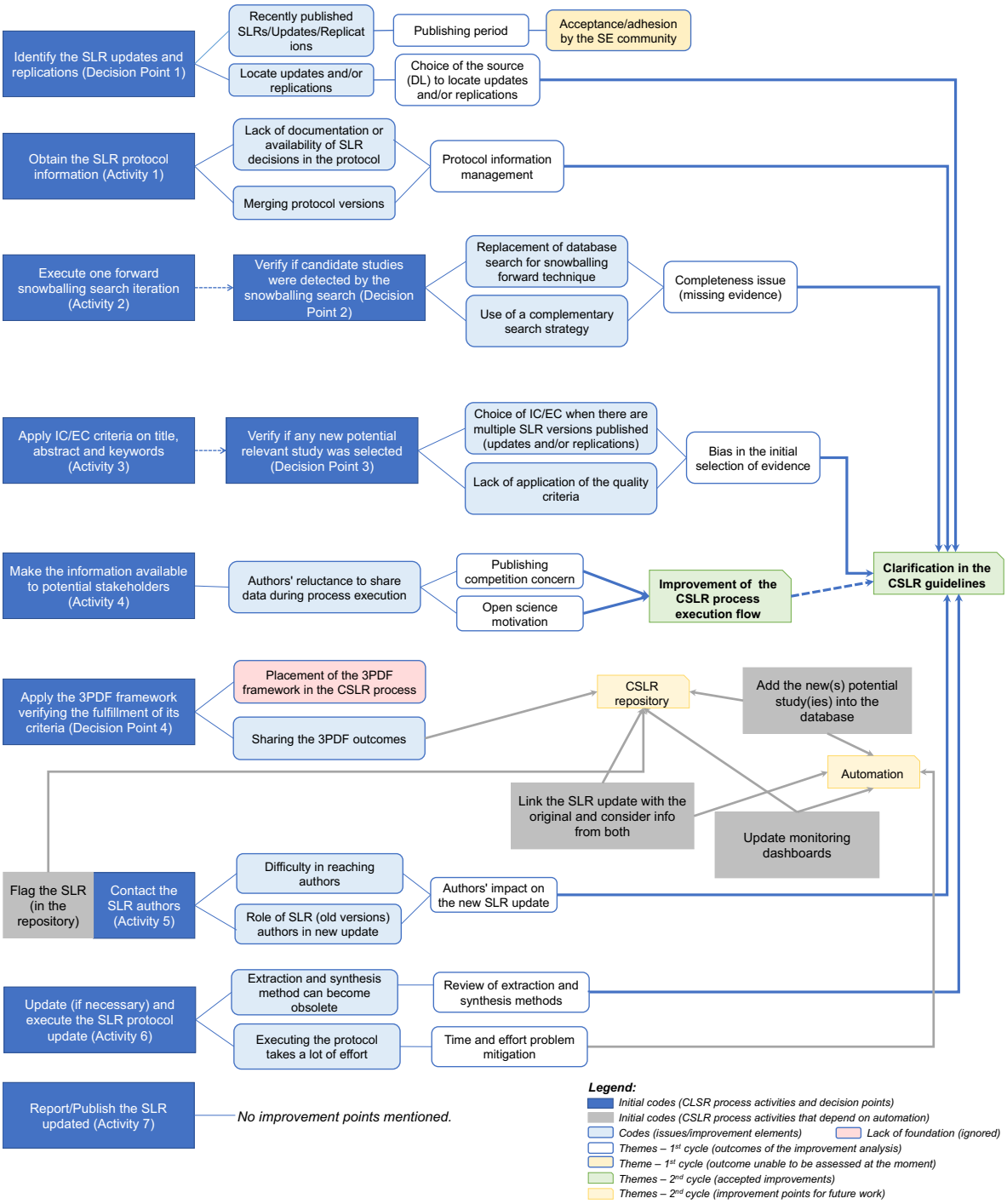


Figure 5.5 : Resulting map from the thematic analysis. ©Bianca Minetto Napoleão.

this Section as well as the improvements described afterward are the results of this peer-review validation.

The improvement of the execution flow in the CSLR process comes from concerns about *when the activity of making the information available to potential stakeholders should be done*, particularly concerning the timing and feasibility of continuous updating and potential second-order effects on the publication process. Following the experts' suggestions, we relocated this activity to the last activity of the CSLR process integrating it with the activity of Report/Publish the SLR update (see the highlighted green portion in Figure 5.6). The improved version (final version) of the CLSR process is illustrated in Figure 5.6 and also available online (Napoleão *et al.*, 2023a).

Next, we clarify the identified improvement points according to the results of our thematic analysis by adding the protocol improvements in each respective activity or decision point of the CSLR guidelines.

Verifying if the SLR has a published update or replication – The year of publication of the SLR, its updates, and replications (if exist) could have an impact on the SE researcher's adhesion to pursue the CSLR process. In cases where the publication period is short (e.g. one or two years), we still encourage the author to follow through with his/her analysis since besides enabling the process's next steps, identifying all existing versions is essential to understand the current state of the art about the research topic under investigation (Kitchenham *et al.*, 2015).

To verify the existence of an SLR update(s) and/or replication(s), we suggest locating the SLR under review on the Google Scholar DL and checking its citations to find updates or replications because usually SLR updates and replications cite the original study (Mendes *et al.*, 2020).

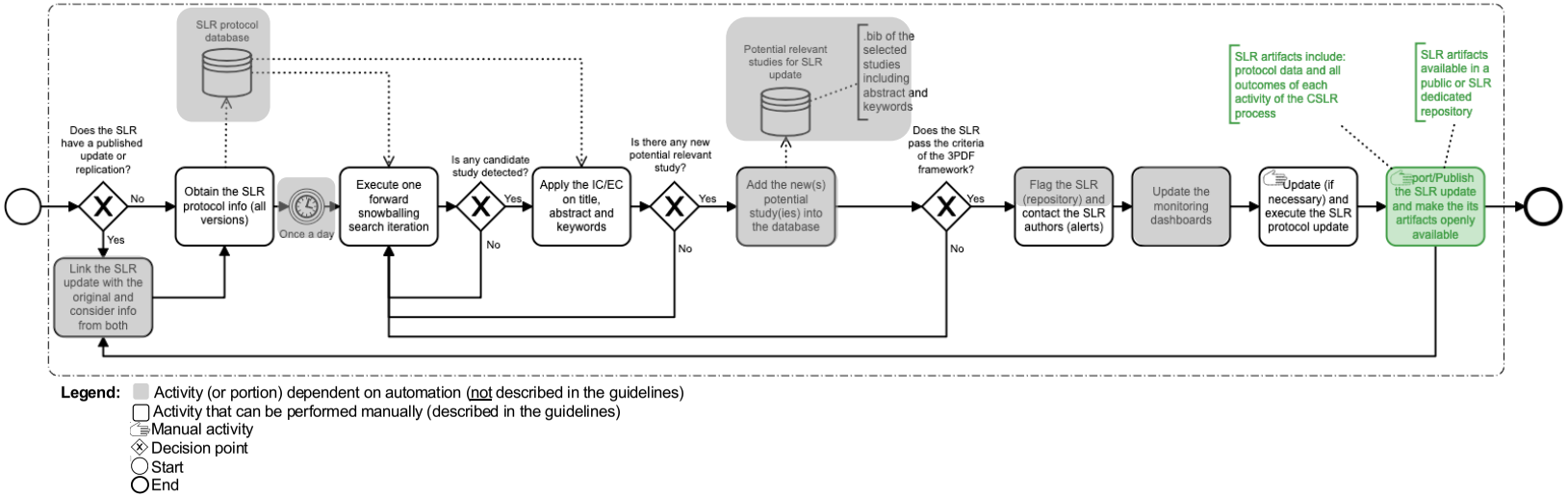


Figure 5.6 : Improved CSLR process (final version). ©Bianca Minetto Napoleão.

We suggest Google Scholar because it is a search engine that index studies from different publishers efficiently. Also, there are studies that compared the efficiency of Google Scholar versus another search engine (e.g. Scopus) and the results demonstrated a higher capability of Google Scholar in indexing SE studies (Wohlin *et al.*, 2020; Mourao *et al.*, 2017).

Obtaining the SLR protocol information of all existing versions – Our proposition to remedy this concern is at first to extract the protocol information of all versions of the SLR study as described in the guidelines (Section 5.2.2), then build a comparative timeline of the evolution of each protocol item (e.g. research questions, search strategy, etc.) and finally base the decisions of the new SLR update on the conclusions of this timed analysis.

Searching for new evidence through the execution of one forward snowballing iteration – The adoption of the forward snowballing as an initial search step is an alternative that could provide indications that the area of the SLR under evaluation is evolving and that possibly the SLR results may be becoming outdated. There is consistent evidence showing that an iteration of the forward snowballing showed being the most cost-effective practice to search for studies for SLR updates (Felizardo *et al.*, 2016; Wohlin *et al.*, 2020). If completeness is a concern for the authors and they are not satisfied with the forward snowballing results, during the protocol update and execution activity we suggest complementing the search with a database search.

Selecting new candidate studies (applying IC/EC criteria on title, abstract and keywords of the studies) – Questions regarding which IC/EC criteria should be used when more than one version of the SLR is published may arise. We suggest (i) building a comparison table in order to compare the similarities and differences among the lists of IC and EC, then (ii) performing a careful analysis of the SLR included studies' from the SLR past version seeking to identify the main evolution of the research topic between the versions, and finally

(iii) using this knowledge base to choose the IC and EC that make sense for the context of the current research (new update).

It is worth mentioning that at this stage of the selection of studies, the goal is to perform an initial analysis based on the title, abstract and keywords of the studies aiming to understand if the SLR under investigation needs an update. A deeper analysis of the studies including the application of quality criteria checklists (if necessary) requires a detailed full-text analysis (Kitchenham *et al.*, 2015). Thus, it will be performed later in the process (during the SLR protocol update and execution).

Verifying if the SLR needs to be updated (Applying the 3PDF) - We strongly suggest the authors document the outcomes from the application of the 3PDF describing the reasoning behind the decisions taken (Mendes *et al.*, 2020). This documented data must be shared in a public and open repository with other artifacts when the SLR is reported/published.

Reporting the need to update the SLR (Contact the SLR authors) - The objective of contacting the authors of the previous versions of the SLR publication is threefold: (i) investigate the possible collaboration of one (or more) authors in the current SLR update to mitigate bias during the update process; (ii) demand extra material of the published SLR version (if necessary) – a useful step especially when the SLR was published a long time ago and crucial protocol information (e.g. list of included studies) is not available; and (iii) disseminate the results of the new update analysis on the field of their interest aiming to have them supporting the dissemination of results, once they are produced and available. Respectively, the role of the authors of previous SLR versions can be (i) author of the current SLR update; (ii) supplementary information support; and (iii) promoter/results monitor of the SLR update results.

We strongly recommend trying to involve at least one author of a previous SLR version. However, this is not always a practical task, mainly because authors may be reluctant to the idea or even because of difficulties in reaching the authors, particularly in the case of older SLRs. If it is not possible to involve the old authors, the CSLR process can move on to the next activities.

Updating (if necessary) and executing the SLR update protocol - One important point to be considered during the protocol analysis is the choice of extraction and synthesis method. Over time the extraction and synthesis methods might need to be updated as the topic evolves. As mentioned by Mendes *et al.* (2020), in an SLR update authors can freely choose new methods for extracting and synthesizing data. Suggestions of SLR synthesis methods can be found in (Kitchenham *et al.*, 2015).

Reporting/Publishing and maintaining an SLR updated – In this last activity of the CSLR process cycle, not only the new SLR update results need to be reported and published, but all the SLR update artifacts, including all the protocol data and the outcomes of each CSLR process activity (e.g. snowballing search results, application of the 3PDF, etc.). These artifacts must be available in a public and open repository (Mendez *et al.*, 2020) (preferably in a dedicated repository for SLRs, if any).

The improvements suggested by the experts allowed us to enhance the CSLR process including its execution flow and activities, as well as to add relevant options and clarifications to the description of the guidelines. As a summary of our expert evaluation, the CSLR process and guidelines demonstrated to be useful to support SE authors throughout the SLR update process, especially contributing to the decision to update an SLR and assisting in the identification and selection of relevant evidence.

5.4 THREATS TO VALIDITY

Hereafter we present the primary threats to the validity associated with the research outlined in this chapter, along with the strategies implemented to mitigate them.

Reliability. The CSLR process was previously proposed and evaluated through a case study (Napoleão *et al.*, 2022b). This Chapter is an extension of Chapters 3 and 4 as well as of this publication, in which the guidelines were compiled from additional evidence to update SLRs that were systematically identified. In addition, the existing CSLR process was reevaluated and improved with the proposed guidelines through an expert evaluation. Regarding the expert evaluation, there is the risk that experts' opinions are not representative or become biased, given their vested interest in the CSLR process and guidelines. While we have only six experts, they are SLR researchers selected based on strict criteria active in independent research groups. Furthermore, they performed independent and anonymous assessments of CSLR, without communication between experts. They also provided free comments on their concerns, and we used them to improve the CSLR process and guidelines. To improve the reliability of the thematic analysis, its conduction was fully peer-reviewed. Although our extended study does not allow us to make strong generalizability claims, the converging results increase the confidence in the CSLR process and guidelines, indicating the importance to share this study with the community.

Construct Validity. A potential threat to the validity of the expert evaluation in this study is using a questionnaire to gather data. To mitigate this threat to validity, the authors took several steps. First, they carefully followed the design phases proposed by Pfleeger (1995), including defining the questionnaire objectives and participants (SE SLR experts), following appropriate methods, and pre-testing the evaluation instrument. In addition, the authors performed a pilot study of the questionnaire with two experienced SLR researchers

to identify and address any potential face and content validity issues or biases. Despite these precautions, bias in the data collected through the questionnaire is still possible. For instance, respondents may have provided incomplete or inaccurate information, or their personal beliefs or experiences may have influenced their responses. Therefore, future research could consider alternative data collection methods, such as interviews or observations, to provide a more comprehensive understanding of the expert's perspectives.

5.5 CHAPTER FINAL REMARKS

In this chapter, we proposed guidelines for the CSLR process. Moreover, based on an expert evaluation, we evaluated the proposed guidelines and the existing CSLR process, resulting in improvements to the CSLR process and refinements to the CSLR guidelines. The finding of the expert evaluation revealed encouraging outcomes, suggesting that the guidelines are promising to support SE authors throughout the SLR update process, especially contributing to the decision to update an SLR and assisting in the identification and selection of relevant evidence.

Our evaluation opened avenues for automating the activities and pipeline of the CSLR process. Researchers have investigated automation of the SLR process over the years (Felizardo & Carver, 2020). However, to our knowledge, only two studies address automation alternatives for SLR updates (Felizardo *et al.*, 2014; Watanabe *et al.*, 2020). Moreover, both studies are focused only on the study selection activity. From this perspective, in Chapter 6 we explore the RG3 by developing an automation solution for the two trigger activities (search and selection) of the CSLR process and discuss potential avenues for future research addressing automation of the CSLR process.

CHAPTER VI

AUTOMATING THE CSLR PROCESS

As described in Chapters 3 and 5, the CSLR process can benefit from automation support in several activities to speed it up and reduce efforts. In fact, the evaluation presented in this thesis opens avenues for automating the activities and pipeline/workflow of the CSLR process. In this regard, in this chapter, we address automation solutions to support the execution of CSLR process.

Considering that the two trigger activities of the CSLR process are search and selection of studies, the SM presented in Chapter 2 (Section 2.2) of this thesis served to map the existing approaches and solutions for the search and selection of studies for SLRs in SE and medicine. Taking into account the existence of only two studies in SE that address automation solutions for SLR updates (more specifically selection of studies), in Section 6.1 we propose and evaluate a prototype tool that provides automation support for both trigger activities of the CSLR process (see RG3 – Introduction). Besides, in Section 6.5 we briefly discuss future directions on CSLR automation.

6.1 TOOL DEVELOPMENT

In this section, we present details about our prototype tool developed to support the snowballing search activity and the selection of studies activity for SLR updates.

We started with the development of a prototype tool to perform both snowballing techniques, forward (citation analysis) and backward (references analysis) (Wohlin, 2014). Even though the main objective of this study is to investigate automation support for searching and selecting studies for the SLR updates, which do not require backward snowballing, during

development we noticed that the design of the forward solution could easily be adapted for the backward solution. Therefore, we opted for the development of a snowballing tool supporting both types.

The execution flow of our proposed algorithm for developing the snowballing prototype tool is shown in Figure 6.1. The snowballing automation is preceded by inputs from the user in the form of Digital Object Identifier (DOI) URLs (Uniform Resource Locator) of papers in the "seed set". Following this, the implementation code is run and the user is asked to specify the number of snowballing iterations he/she wants to run and whether they wish to proceed with either backward or forward snowballing, or both. We chose to add these two inputs because, for the context of SLR updates, a single iteration of forward snowballing is enough to return the relevant studies (Wohlin *et al.*, 2020; Felizardo *et al.*, 2016).

The snowballing solution (backward and forward) is implemented by querying the Semantic Scholar API (Semantic Scholar, 2023) based on the DOI of studies. The metadata returned as the query results are employed to extract the DOIs of citations and references cited in the queried study. In the case of studies without DOIs, the CrossRef API (CrossRef, 2023) is queried for DOIs by providing the input as a reference string generated using the keys 'authors', 'title', 'venue', and 'year' returned in the metadata by Semantic Scholar (Semantic Scholar, 2023).

Next, the acquired DOIs are passed through a redundancy check (for subsequent iterations) to ensure that the extraction has not been done for them in previous iterations. The main part of the data acquisition starts with getting the bibliographical metadata of the references and citations in the BibTeX format, by making a request to the DOI using the Urllib library (Python Team, 2023).

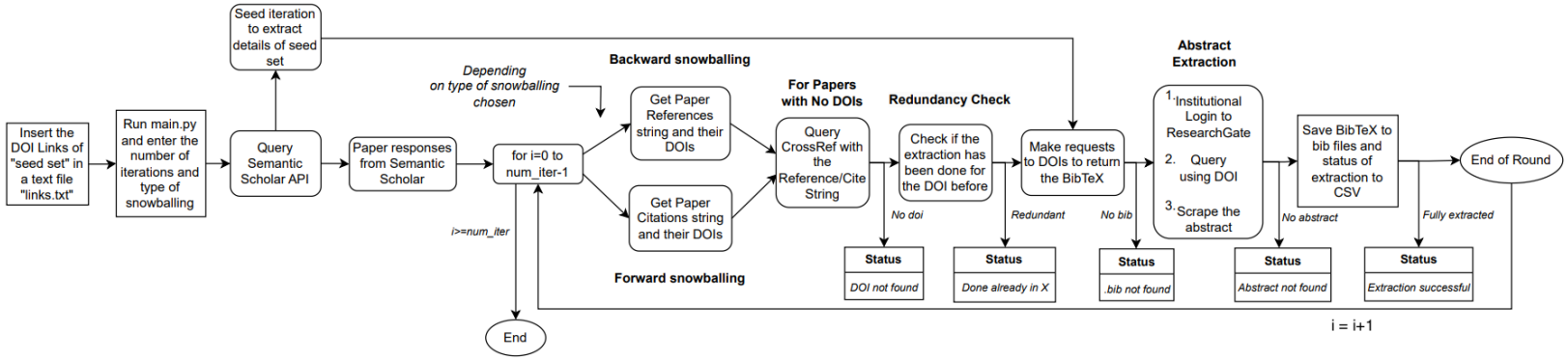


Figure 6.1 : Execution flow of the snowballing tool. ©Bianca Minetto Napoleão.

The abstract is usually missing in the response and hence we need to apply other methods to get its content. For the abstract extraction, we make use of ResearchGate (ResearchGate, 2023), which allows researchers and students free access to abstracts of scientific publications and preprints of research work. We employed web scraping to log in to ResearchGate with institutional credentials, and query for the studies using the DOI. As the page is rendered by the browser, the abstract is extracted and saved to the BibTeX of the corresponding study.

The results of the multiple iterations of the snowballing process are stored in one common CSV file and one common BibTeX file at the end of the runs. This applies to each type of snowballing. Therefore, if the user wishes to perform both backward and forward snowballing together, there will be 2 separate files (CSV and .bib) for each type of snowballing. The CSV file stores the reference strings, the corresponding DOIs, the status of the extraction, and the iteration number. The reference string is generated using the keys 'authors', 'title', 'venue', and 'year' returned in the metadata by Semantic Scholar (Semantic Scholar, 2023), which are joined to produce a string similar to the Chicago format of referencing. This is done because the full reference of a study is not returned by Semantic Scholar. The status of the extraction for a particular study uses the following implicit phrases: "Extraction successful", "DOI not found", ".bib file not found", "Abstract not found", and "Done already in X" where X is the iteration number.

Since we store the results of all the iterations in a common CSV file, the iteration number tells us in which iteration of snowballing a particular study was discovered. The BibTeXs of the studies are stored in a common .bib file. Each BibTeX is appended to the common BibTeX file after each extraction phase. Another feature of the tool helps us to obtain the BibTeX file for the "seed set" as well.

Regarding the selection of studies, we opted to consider in our evaluation the ML algorithm that has shown the most promising results in the selection of studies for SLRs as well as other ML algorithms known to perform well for text classification (Aggarwal & Zhai, 2012; Peterson, 2009). As mentioned in Chapter 2, Section 2.2.4, SVM is the supervised learning algorithm most adopted in SE and medicine that showed promising results on the selection of studies for SLRs. Besides, it was also investigated by Watanabe *et al.* (2020) in the context of SLR updates providing the best performance result in terms of recall and precision. The selected ML algorithms are: (Aggarwal & Zhai, 2012; Peterson, 2009): XGBoost (Chen *et al.*, 2015), Linear Support Vector Machines (LSVM) (Lilleberg *et al.*, 2015), Logistic Regression (Hosmer Jr *et al.*, 2013), and Multinomial Naïve Bayes (MNB) (Kibriya *et al.*, 2005). We chose them because all four algorithms have a regularization term, which plays an important role in combating overfitting and underfitting in unbalanced datasets by adding penalties to the loss function. The second-most important term is "class weight" which prevents the prejudice of the model towards the minority class, by assigning a higher weight to it. They are reciprocal of the class frequencies. We detail the tool's selection process and parametrization in Section 6.2.2 by providing a practical evaluation example.

6.2 SMALL-SCALE EVALUATION

To evaluate our prototype tool, we performed a small-scale evaluation (Wohlin & Rainer, 2022). Chapter 2 reports that a highly adopted form of assessment for text classification techniques for SLRs is experiments (also referred as case studies) considering data from published SLRs performed manually. According to the smell indicator proposed by Wohlin & Rainer (2022), the correct label for our evaluation is small-scale evaluation instead of experiment or case study. However, to guide and report our evaluation process we followed

the five main steps for case studies proposed by Runeson *et al.* (2012): design, preparation, collecting data, analysis, and reporting.

6.2.1 DESIGN

Our design consists in selecting a published SLR replication (Wohlin *et al.*, 2022) and its ongoing SLR update evaluation instrument to search for studies through snowballing iterations, and perform an initial selection of potentially relevant studies to be included in an SLR update.

To evaluate the prototype tool's capability of performing backward and forward snowballing, we opted to use the SLR replication study (Wohlin *et al.*, 2022) since it documents in its supplementary material[27] the results of each snowballing iteration performed manually by the authors. In summary, our goal with this is to illustrate the tool's potential to be used to perform both snowballing search types in an SLR conduction process when a "seed set" of studies is known by the authors (e.g. selected from database search (Wohlin *et al.*, 2022)).

Next, to evaluate the tool's capability of being employed in the SLR update context, the main goal of our study, we used the 45 manually selected studies by the SLR replication (Wohlin *et al.*, 2022) as a "seed set" to perform an iteration of forward snowballing and then apply the ML algorithms on a reliable and complete dataset from the ongoing SLR update of (Wohlin *et al.*, 2022). In this replication and the ongoing update, the inclusion and exclusion of new studies were conducted based on individual assessments and the consensus of three experienced SLR researchers, allowing us to have confidence in this data for building reliable training and testing sets.

[27] https://ars.els-cdn.com/content/image/1-s2.0-S0950584922000659-mmc2.pdf

6.2.2 PREPARATION AND DATA COLLECTION

To evaluate the tool's capabilities to perform backward and forward snowballing, we first prepared our "seed set" by obtaining the DOI of the studies mentioned as "seed set" (9 studies) in the supplementary data of the SLR replication (Wohlin *et al.*, 2022). Next, we replicated with our tool the same 5 iterations of backward and forward snowballing performed manually by the authors. Finally, we compared our tool's results with the manual execution.

Regarding the analysis of the search and selection for SLR updates, we conducted three steps. First, we search performing a single forward snowballing iteration using the 45 included studies by the SLR replication (Wohlin *et al.*, 2022) as the "seed set". Second, we train our ML algorithms with the "training set" containing both included and excluded studies of the SLR replication (Wohlin *et al.*, 2022). Finally, we perform the selection of studies using the trained algorithms on the results of the forward snowballing in the first step ("testing set").

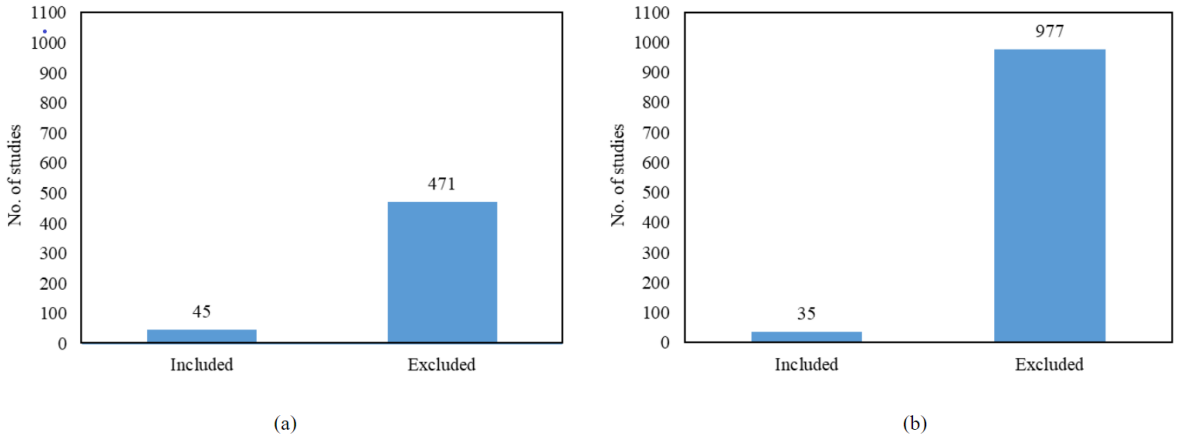


Figure 6.2 : The data distribution of (a) Training set and (b) Testing set. ©Bianca Minetto Napoleão.

The distribution of included and excluded studies in the training and the testing sets is shown in Figure 6.2. It is highly imbalanced with a minority of included studies. The

imbalance represents the real-world scenario where out of a large number of studies only a few are typically relevant to the focus of a particular research topic.

Regarding the training process, for *Training Data Collection*, as shown in Figure 6.3, the input consists of the included and excluded studies in BibTeX format (.bib) of the SLR replication (Wohlin *et al.*, 2022).

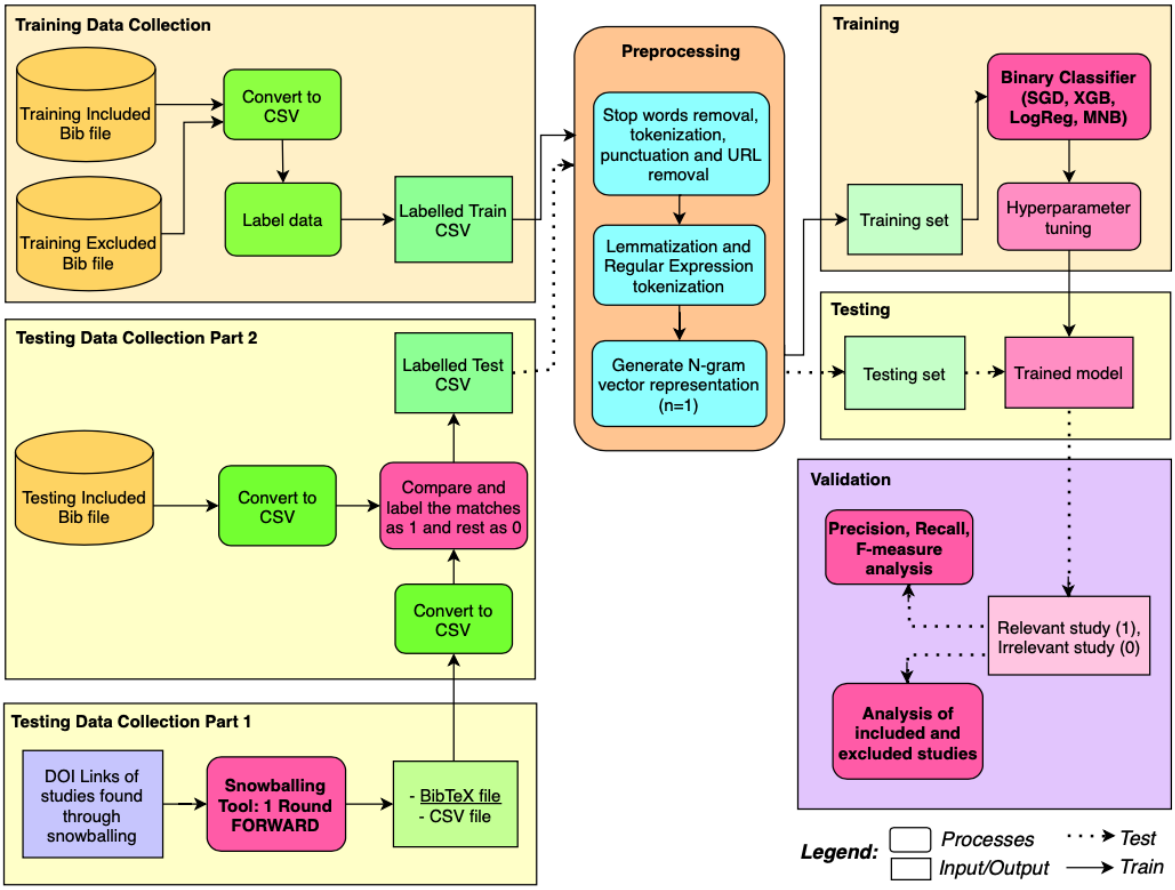


Figure 6.3 : Tool process to select studies for SLR Updates. ©Bianca Minetto Napoleão.

We converted the .bib file into CSV format by taking the 'Title' and 'Abstract' fields of the studies. We also labeled them with relevance 1 for included studies in the SLR replication and relevance 0 for excluded studies from the SLR replication (Wohlin *et al.*, 2022).

Then the labeled "training set" CSV is ready to be passed to the *Preprocessing* phase. The block *Training* shows the training process after the training set is fed to the Binary Classifiers LSVM, XGBoost, Logistic Regression and MNB, for training. After the initial training, hyperparameter tuning is done using the different parameters of the models to improve their performance on the minority class (here, 1). Hyperparameter tuning was done by rerunning the algorithms several times with different values to find the model parameters most suited to our goal: maximizing recall (finding most of the relevant studies) (Kitchenham *et al.*, 2015) and precision (reducing the load on reviewers to check irrelevant studies). Our goal is to get at least an acceptable trade-off between recall and precision according to the classification presented in (Dieste *et al.*, 2009).

The best hyperparameter configurations obtained during the training phase using Sklearn toolkit (Scikit-Learn, 2023) were as follows: the hyperparameter term "alpha" is set as 2 to perform strong regularization on the LSVM model, and both classes are given due importance by setting the "class_weight" term as 'balanced' which follows a weighted loss function. This linear model is then trained using Stochastic Gradient Descent (SGD) (Bottou, 2012) to optimize the loss function with a decreasing learning rate. A similar approach was followed for XGBoost and Logistic Regression. In XGBoost we set "gamma" (regularization term) as 20, the "scale_pos_weight" term as (number of articles in class 0)/(number of articles in class 1), and "sub_sampling" ratio term to 0.2 to prevent overfitting. In Logistic Regression, "C" (the regularization term) is set to 0.01 and the "class_weight" term is set to 'balanced'. For Multinomial Naïve Bayes the default parameters of Sklearn are used. In the first three models, strong regularization was done to maximize the recall and precision of the minority class by making the models more conservative and generalize better on testing data.

Concerning the testing process, *Testing Data Collection* is divided into two parts. First, *Testing Data Collection Part 1* implements one round of forward snowballing using the

snowballing tool on the update "seed set" (45 selected studies from the SLR replication). The output of the forward snowballing process is a CSV file keeping track of the study extraction and a BibTeX file which holds the bibliographical references of all the studies including their abstracts, in order to aid the authors in the selection of relevant studies. This task which is usually done manually, is here completely automated. Similarly as done for the training, the BibTeX file from the output of forward snowballing is converted to a CSV file taking the 'Title' and 'Abstract' fields of the studies. Second, in *Testing Data Collection Part 2* the BibTeX file of included papers is also converted to CSV and a comparison is done between the two CSVs to generate a unique labeled CSV file for testing, labeling the relevance of included studies as 1 and excluded ones as 0.

Thereafter, the labeled testing set CSV is passed through the *Preprocessing* phase. Finally, during *Testing* the trained model is used to predict inclusion or exclusion for the preprocessed testing data.

Preprocessing is done similarly for the training and test CSV files. First, the 'Title' and 'Abstract' columns are merged to form a single string for each study under a column named 'Merged', which now serves as the text for text classification. The preprocessing treats this text by removing stop-words using the Nltk library (NLTK Team, 2023) in Python (Merzouki, 2023), tokenizing the text, removing punctuation and URLs, performing lemmatization (Plisson *et al.*, 2004), and finally vectorizing using the Bag of Words count vectorizer (Wallach, 2006). The count vectorizer gives poor results when the n-gram range is increased as it predicts with a greater precision only for the majority class. In this sense, unigrams are used. The preprocessed text is then passed as input to the ML algorithm models for training.

Finally, for *Validation*, we record the number of included studies identified during the forward snowballing round. The labels of the test set allowed us to compute the performance

measures Precision, Recall, and F-measure based on the predictions obtained during the *Testing* phase. They are defined in Chapter 2 (Tables 2.6 and 2.7) and detailed as follows (Napoleão *et al.*, 2021a; Kitchenham *et al.*, 2015):

$$recall = \frac{number\ of\ relevant\ studies\ retrieved\ as\ relevant}{total\ number\ of\ relevant\ studies}$$

$$precision = \frac{number\ of\ relevant\ studies\ retrieved\ as\ relevant}{total\ number\ of\ studies\ retrieved\ as\ relevant}$$

$$F\text{-}measure = \text{harmonic mean of the precision and recall}$$

The results of our evaluation are presented in Section 6.2.3.

6.2.3 RESULTS - ANALYSIS OF OUR EVALUATION

The evaluation of our tool prototype for the reproduction of backward and forward snowballing for the SLR replication is presented in Table 6.1. In total, our tool was able to automatically identify 41 of the 43 (95.3%) studies manually identified by the authors when performing the same snowballing iterations (see Table 6.1 sum of column "Studies detected" + nine studies from the initial "seed set"). However, during iteration 2, a study identified manually during forward snowballing was identified by our tool during the backward snowballing execution (see lines highlighted with (*) in Table 6.1). We opted to conserve this result since it does not interfere with the final result of the snowballing tool in this automated execution scenario.

The two missing studies could not be identified because they do not have a DOI and the tool was not able to locate them through the implemented reference building (Chicago format). It is worth mentioning that the set of included studies of the SLR replication is composed of 45 studies in total, but two of them are not considered in this analysis because they were not

Table 6.1 : Results of the snowballing search replicated by our tool. ©Bianca Minetto Napoleão.

Iteration 1		
Snowballing type	**"seed set" of the iteration**	**Studies detected (%)**
Backward	9	12/12 (100%)
Forward	9	1/1 (100%)
Iteration 2		
Snowballing type	**"seed set" of the iteration**	**Studies detected (%)**
Backward	13 (1+12)	1+1* = 2/1 (100%)
Forward	13 (1+12)	12/14 (85.7%)
Iteration 3		
Snowballing type	**"seed set" of the iteration**	**Studies detected (%)**
Backward	14 (2*+12)	3/4 (75%)
Forward	14 (2*+12)	1/1 (100%)
Iteration 4		
Snowballing type	**"seed set" of the iteration**	**Studies detected (%)**
Backward	4 (3+1)	1/1 (100%)
Forward	4 (3+1)	0/0

retrieved by snowballing in the replication (authors' suggestions). In addition, our tool was able to execute automatically four iterations instead of five (manual) because the single study resulting from iteration 4 did not have a DOI available, which caused the snowballing process to stop. However, in this case, the final result is still the same (manual versus automated) since the manual process also stopped for not retrieving any other relevant study.

Regarding the evaluation of our tool applied in an SLR update scenario, the snowballing tool was able to retrieve 1012 unique studies in a single round of forward snowballing based on a "seed set" of 41 out of 45 studies included in the SLR replication. The 4 remaining studies not identified did not have DOIs besides two of them were identified with another search technique instead of snowballing. Consequently, we missed out on 28 citations (data from Google Scholar in February 2023). This leads to our final "seed set" being formed by the 41 studies and having 1012 unique citations to be analyzed in the selection phase by the ML algorithms.

Out of 35 studies contained in the "testing set - included", that are to be included in the SLR update, our search and selection tool identified 33 studies (94.3%) through the forward snowballing iteration, even without being able to include the fourno-DOI studies in the "seed set".

The precision, recall, and F-measure scores on the testing for the class of interest (included papers) are illustrated in Figure 6.4. It can be seen that the best-performing model in terms of recall is LSVM (74.3%), followed by XGBoost (63.6%), Logistic Regression (45.4%), and Multinomial Naïve Bayes (42.4%). In terms of precision, the algorithms had almost similar results ($\sim$15%) with the exception of XGBoost which had a lower precision (11.6%). Considering the fact that the F-measure combines the effect of both metrics, a high recall will give a low F-measure if the precision is low. The best F-measure value was observed with the LSVM model (24.9%). It is able to predict the highest number of studies belonging to the positive class 1 (included), correctly. The precision value can be explained by the high false positive value which is due to the strict regularization performed in LSVM, hence, a trade-off between precision and recall is remarked. In fact, Dieste *et al.* (2009) highlight that there will always be a trade-off between recall and precision because irrelevant studies are more likely to be returned by a search execution, the higher the recall is.

Out of 33 studies, 26 studies were identified by the LSVM model, 21 by XGBoost, 15 by Logistic Regression, and 14 by MNB. The number of studies to be excluded correctly that were identified was 820 by LSVM and XGBoost, 890 by Logistic Regression, and 900 by MNB. Logistic Regression and MNB give biased results for the majority class, hence the true negative values are much higher than LSVM, but they have lower true positive values, which is the number of included studies predicted correctly by the model.

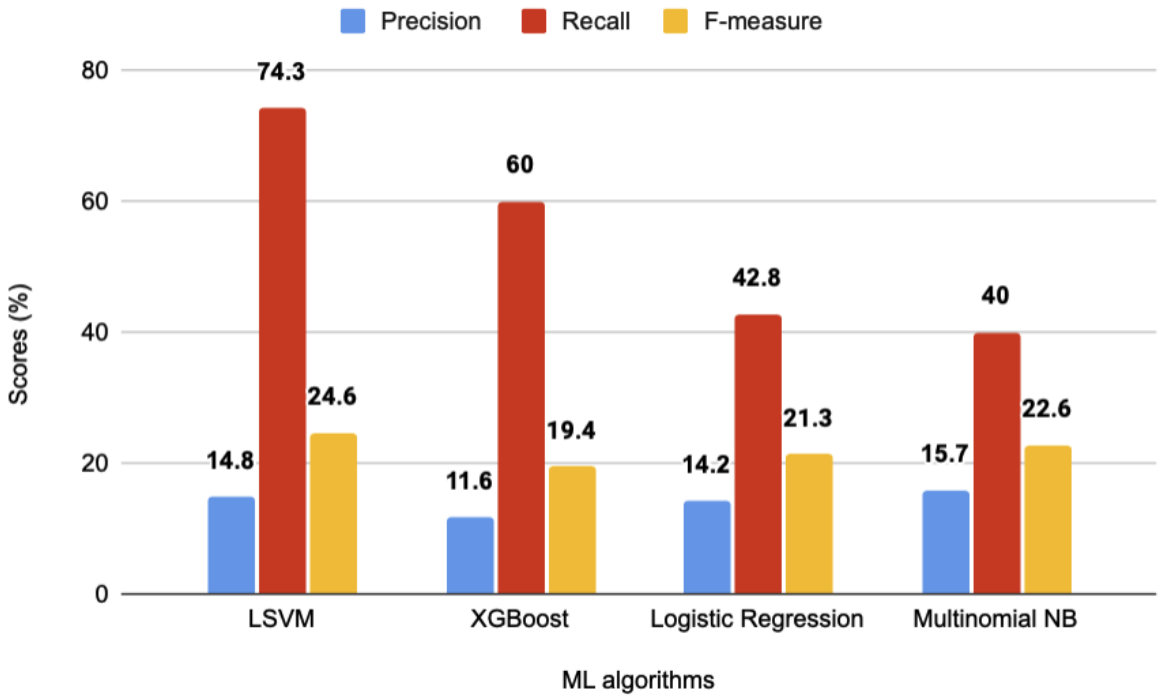


Figure 6.4 : The performance report of the ML models. ©Bianca Minetto Napoleão.

6.2.4 REPORTING - OBSERVATIONS FROM THE EVALUATION

According to the analysis of the performance measures, the LSVM model showed the best result among the evaluated models. Moreover, following the search strategies scale proposed by Dieste *et al.* (2009), the recall and precision range resulted from the LSVM model showed to be acceptable (recall 72-80% and precision 15-25%) meaning a "good enough strategy".

6.3 DISCUSSION

The motivation of this study was to investigate automated alternatives to support both activities of searching and selecting studies during the execution of the CSLR process (SLR update scenario). Due to the effort required for these activities, authors often end up not giving

enough attention to keeping SLRs up to date. While we present emerging results, they already shed light on some meaningful automation limitations and possibilities.

With respect to searching for new studies, we described solution options used when building our snowballing tool to face automation challenges (e.g., querying CrossRef to find DOIs, complementing BibTex data with web scrapping). Besides providing an example of such options, our results indicate that snowballing-based search strategies can be fully automated with minor losses, at least for classical white literature (the only identified limitation was related to papers without DOI), reducing the effort of laborious manual snowballing iterations (*cf.* Table 6.1).

Regarding the selection of studies in SLR updates using ML algorithms trained based on papers included and excluded in the original SLR, the recall and precision obtained by these algorithms still represent only a "good enough strategy", according to the scale proposed by (Dieste *et al.*, 2009). The full automation of the selection of studies including full-text analysis can be questionable and it is not the goal of our study. However, we demonstrated that the support of automation would allow to conservatively save some manual selection effort. In fact, if we take our best model (LSVM), even if we use a threshold that results in a max recall scenario of 97% for the classifier (one false negative) and only 8% of precision, the SLR update author would have to manually analyze only 396 papers instead of 1012. *I.e.*, for our investigation, in a scenario of conservatively minimizing the risk of missing papers to be included, it would still be possible to reduce the number of papers to be manually analyzed by more than 2.5 times.

Overall, we believe that our results provide preliminary indications that strengthen the belief that automated approaches could significantly help reduce the CSLR update time and effort.

6.4 THREATS TO VALIDITY

Construct Validity. With respect to searching for studies using snowballing, the adoption of Semantic Scholar has not been formally evaluated by researchers in the context of SLR updates (Wohlin *et al.*, 2020), as is the case for Google Scholar. However, we noticed that, in our case, the results showed to be relevant for the study and comparable, with less noise. Also, Google Scholar does not allow the use of an API to perform searches. For study selection, even with the CSLR guidelines and Kitchenham *et al.* (2015) suggesting keyword analysis of studies for selection along with titles and abstracts, we chose not to consider keywords in our automated analysis because many of the studies under evaluation did not have keywords available, which could affect the overall reliability of the results. For this reason and others Watanabe *et al.* (2020) also chose not to use keywords in their analysis for the initial selection of studies for SLRs. Our evaluation results might have been affected by the choice of ML algorithms. Other ML algorithms could have been explored in our study. Large Language Models (LLMs) could also be explored to select studies. However, due to the small size of the training set, the model ran out of data to train and suffered from overfitting as reported in (Alchokr *et al.*, 2022). LLM is still an emerging topic and requires further investigation and validation to be applied to the selection of studies context for SLRs (see Section 6.5.2). Both alternatives can be considered as part of future work.

External Validity. The dataset used in our analysis might not represent the diversity of SLR Updates in SE. Similar analyses could have been conducted based on other SLRs to improve the generalizability of our results. However, replicating our emerging results on other SLRs to strengthen external validity would require significant effort. Furthermore, it is challenging to acquire a reliable and detailed SLR dataset for SLRs that could potentially need an update.

Reliability. One limitation of our study is associated with the dataset used in our experiment and the possibility of sample bias. For the snowballing analysis, we used a dataset of an SLR replication that involved experienced SLR researchers following strict guidelines for searching and selecting evidence (Wohlin *et al.*, 2022). The data used for the SLR Update analysis was acquired from the same authors who performed the SLR replication, also through a rigorous analysis process. Also, to improve the reliability of our results, the tool prototype and the small-scale evaluation datasets are openly available (Napoleão *et al.*, 2023b).

6.5 FUTURE DIRECTIONS ON CSLR AUTOMATION

The previous sections of this chapter present a prototype alternative to automate the activities of execution of forward snowballing and perform the initial selection of potentially relevant studies by analyzing the title and abstract of retrieved studies.

In fact, as illustrated in Chapter 5, Figure 5.6, several activities of the CSLR process were designed to rely on automation to facilitate the process execution. These activities demand a repository structure and pipeline/workflow to be executed. Therefore, in Section 6.5.1 we present a brief discussion about the implementation of a repository for SLRs in SE.

In the following, we present an outline of CSLR activities that do not rely on automation but can benefit from it.

Decision point: Does the SLR have a published update or replication? – Usually, a replication or upgrade of an SLR cites the original SLR. In this way, this activity can benefit from automating the execution of the snowballing forward to detect the citations of the SLR under investigation, facilitating the search work by the researcher.

Activity: Obtain the SLR protocol information – Protocol information is described in the study and/or in any referenced supplemental data document. This activity would benefit from automation related to data extraction from studies. An alternative we suggest to investigate is the use of LLM-based applications such as ChatPDF[28] in which the user inserts a .pdf file and asks questions about the file's contents. The application has an API that offers up to 500 queries and 5000 .pdf pages per month for free. For more queries and .pdf pages, there is a monetary cost to the user.

Decision point: Does the SLR pass the 3PDF? – The steps of the 3PDF rely on human interpretation. However, two steps in particular can benefit from automation support. These steps are described in Chapter 5, Section 5.2.6. They are: *Step 1.b Has the SLR had good access or use?* – automation of the detection of citations of the study and publication year to calculate automatically the coefficient proposed in (Octaviano *et al.*, 2022); and *Step 2.b Are there any new studies or information?* – The results obtained from the selection based on title, abstract and keywords can provide an indication of new relevant studies.

Activity: Contact the SLR authors (alerts) – A script for sending automatic emails or alerts (in the repository) can be developed to contact authors when necessary.

Activity: Update (if necessary) and execute the SLR protocol update – In summary, this activity consists of actually performing the SLR update including the protocol review and update as well its execution by performing additional searches (if applicable), selecting the potential relevant evidence identified by applying the IC/EC in their full-text, extracting data and synthesizing the evidence to answer the RQs. Automation alternatives applied in the SLR context for these tasks can also be explored in the SLR update context. Felizardo & Carver (2020) present an overview of existing strategies to automate the mentioned SLR tasks.

[28] https://www.chatpdf.com

Finally, in light of the current surge in popularity of LLMs, in Section 6.5.2 we present an overview of the state-of-the-art on the application of LLMs in SE SLRs. Considering that the CSLR process is grounded in the SLR process, the possible applications presented to provide automated support for SLRs can be extended to the respective activities of the CSLR process.

6.5.1 SE SLR REPOSITORY

As exemplified in Chapter 5 - Section 5.3.2, medicine benefits from several SLR repositories. A dedicated repository can be seen as the foundation of the CSLR process automation since its goal is to collect SE SLRs' information and data in a unique place. As a consequence, a repository dedicated to SLRs opens possibilities for the integration of automation to support the execution of activities, execution flow and management of the CSLR process. Besides, an SE SLR unified repository could offer many benefits to the SE community, such as:

- **facilitating the identification of potential outdated SLRs:** a dedicated and open-access platform to maintain the results from the execution of the CSLR process;
- **facilitating the full update of outdated SLRs:** reuse of protocol information can potentially save time and effort by building upon an existing protocol;
- **avoiding unnecessary duplicated SLRs:** find out if a review has already been carried out, in order to avoid wasting time, effort and resources answering research questions that have already been answered and are still up-to-date;
- **facilitating methodological decisions:** protocol definition based on other protocols can support researchers in making well-informed methodological decisions; and

- **searching evidence for tertiary studies:** look for secondary studies (SLRs and SMs) to answer proposed research questions of tertiary studies in a single, centralized and open location streamlining the evidence-gathering for tertiary studies conduction and update.

Mertz *et al.* (2018) introduced a web repository of literature reviews in Computer Science called LRDB[29]. They considered as the initial population of their database literature reviews on the SE field. The repository counts with an initial population of 71 SLRs retrieved from a search on ACM DL. After over five years of the repository online (July 2023), the repository has only 113 SLRs. According to (Mendes *et al.*, 2020), between 2004 and May 2016 there are more than 430 SLRs published in SE. A more recent study (Napoleão *et al.*, 2021) relates 1000 SLRs and SMs published in SE between 2004 and February 2020. Therefore, the repository is not maintained by the SE community neither the authors nor the research group that proposed the initiative.

Unlike the repository proposition of Mertz *et al.* (2018) that focuses on gathering computer science SLRs through insertion and search of SLRs by the community, our recommendation focuses on establishing a repository of SLRs in SE that is capable to support using automation alternatives both the activities and the execution flow of the CSLR process. Considering that the CSLR process comprises a continuous and systematic surveillance and analysis of potential new relevant evidence for published SLRs aiming to contribute to keeping SLRs up to date, we envision the creation of a repository that can integrate a CI platform that allows the creation and execution of workflow or pipeline of the CSLR process. Examples in the software development area are Jenkins (Jenkins, 2023) and GitHub Actions (GitHub Docs, 2023). In fact, Jenkins' declarative pipelines are similar to GitHub workflow files. While Jenk-

[29] http://prosoft.inf.ufrgs.br/lrdb/Home/Index

ins uses stages to execute a group of steps, GitHub Actions uses jobs to group one or several steps. The choice between GitHub Actions and Jenkins depends on your specific requirements, familiarity with the tools, and the project's hosting platform. GitHub Actions might be more appealing for projects hosted on GitHub while Jenkins remains a strong contender for its flexibility and vast plugin ecosystem.

In the following, we outline a high-level overview of a CI pipeline/workflow for the CSLR process based on CI/CD process (Duvall *et al.*, 2007; Humble & Farley, 2010).

1. **Create and set up a repository:** The first step consists in having a version-control repository to allow the version management of SLR and its replications and updates. The next step consists in setting up the repository by making sure that the SLRs' information and data are organized within the repository.

2. **Version control initialization:** This step consists of the initialization of the CSLR process: the input of an SLR in the repository and labeling it as the original version.

3. **Define of the pipeline/workflow:** Before defining the pipeline/workflow steps which are the steps of the CSLR process with the exception of the two CSLR process activities that are manual (Update and execute the SLR protocol, and Report/publish the SLR update and the SLR update and make its artifacts openly available), it is essential to choose and set up a CI tool to build the SLR project whenever changes are pushed in the repository.

4. **Write automation scripts:** In this step, we will develop the scripts to automate the CSLR process activities. In the CSLR process (see Figure 5.6) all activities that rely on automation dependent on automation might be developed. Scripts that fully or partially automate activities that do not fully depend on automation can be developed to support

and seed up the process execution, for example, an adapted version of the automation presented in Section 6.1 to support the forward snowballing activity.

5. **Configure the CI tool and the pipeline/workflow stages:** Configure the pipeline/workflow to listen for changes in the repository and trigger the CSLR activities whenever necessary. For example, wherever an automated forward snowballing search is performed and new potential evidence is detected, this new evidence must be pushed to the repository project folder of the SLR under investigation.

6. **Artifact and results storage:** This step consists in define where (a storage system such as Jenkins artifacts) and how (suitable format for the automated activities) store the artifacts generated during the pipeline execution, such as the SLR versions, protocol information, retrieved studies from snowballing forward, potential relevant studies selected, etc.

7. **Testing and Validation:** The goal of this step is to include tests and validation activities into the pipeline to ensure the accuracy and validity of the results generated during the SLR update process. However, as the last two activities of the CSLR process that constitute the update of the SLR and the publication of the results are manual activities, it is difficult to verify the accuracy and validity of the results generated during the execution of the pipeline. Once the new version of the SLR is published, it will be possible to verify this information and improve the pipeline and automation scripts.

8. **Notifications and Reporting:** The goal of this step is to configure notifications and alerts to inform relevant stakeholders about the pipeline execution status and issues that arise during its execution. In the CSLR context, this activity is represented by the activity “Flag the SLR (repository) and contact the SLR authors”. In this case, the SLR authors that signed up to receive alerts and updates about an SLR will receive a status

update. Regarding errors, in the first moment, we see them be informed to the repository development and contributors team in order to perform adjustments to the pipeline or automation scripts as necessary.

9. **Documentation:** The documentation of the pipeline/workflow consists of the CSLR process and guidelines, script usage, and any configuration required.

10. **Iteration and improvements:** Since the CSLR process is an iterative process, based on feedback from the SE community and evolving needs, improvements of the CSLR pipeline could be done.

In summary, the creation of a dedicated SLR repository that integrates the CSLR process pipeline/workflow can offer several benefits to the SE community. Despite the easier access to SLRs public information and their data, the automation of the CSLR process allows systematic automation of the CSLR activities and execution flow (pipeline/workflow) bringing other advantages such as (i) reproducibility - by defining the process pipeline/workflow, researchers can precisely reproduce the same steps and analyses in future reviews or updates, (ii) efficiency and speed - the automation provided can reduce manual efforts and accelerating the review process. Researchers can focus on more complex tasks, data interpretation and synthesis, rather than repetitive tasks. (iii) traceability and transparency - documentation of an SLR including protocol information and outputs of scripts are centralized and organized, ensuring clarity and transparency throughout the update process. It is important to highlight the necessity for a comprehensive investigation into the repository functionalities, choice of platforms and tools, and script implementation concerning the specific requirements of the CSLR process. Moreover, rigorous evaluation through practical application is essential to draw further conclusive insights and recommendations.

6.5.2 TOWARDS THE APPLICATION OF LARGE LANGUAGES MODELS IN SE SLRS

LLMs such as GPT (Generative Pre-trained Transformer) 3 and 4 designed by OpenAI[30]; and BERT (Bidirectional Encoder Representations from Transformers) designed by Google[31] have gained a recent rise in popularity due their capability of answering questions in a natural-understandable way (OpenAI, 2023; Google, 2023).

In the SLR context, the main input to the SLR process is text from primary studies. Researchers have explored traditional ML techniques combined with NLP to provide support to several SLR activities (Felizardo & Carver, 2020). In Chapter 2 of this thesis, Section 2.2, there are several examples of traditional ML and NLP techniques applied to support the activities of search and selection of studies for SLRs. The main difference between LLMs and traditional ML algorithms combined with NLP techniques is that ML algorithms often rely on labeled or annotated datasets specific to the SLR task (Ray, 2019). These datasets are typically created manually by domain experts. On the other hand, LLMs are pre-trained on large general-domain unannotated datasets, and their language understanding capabilities can be fine-tuned with smaller domain-specific datasets (e.g. specific domain requirements of an SLR). LLMs can predict the next word in a sentence or fill in missing words, allowing them to learn the statistical patterns and structure of the language (Zoph *et al.*, 2022).

The conduction and update of an SLR is a time-consuming and labor-intensive task. In addition, in SE there are no well-established tools widely used in practice due to limitations such as the availability of the tools, lack of documentation and/or difficulty in their use (Felizardo & Carver, 2020; Napoleão *et al.*, 2021a). Moreover, activities of the SLR and

[30] https://openai.com/blog/

[31] https://cloud.google.com/ai-platform/training/docs/algorithms/bert-start

CSLR such as data extraction and analysis do not count with automation support for being complex and demanding human interpretation.

In order to obtain an overview of current research that addresses the application of LLMs to support SLR and underline potential application in the CSLR process automation, we performed searches on Google Scholar using the terms “Systematic Review”, “Large Language Models”, “ChatGPT”, “BERT”. The search was executed in June 2023. Because it is a recent topic, we found only three studies that discuss the applicability of LLMs in SLR activities: (Qureshi *et al.*, 2023), (Alchokr *et al.*, 2022) and (Wang *et al.*, 2023). The most recent one, (Qureshi *et al.*, 2023) is not published in a peer-review venue yet.

In Qureshi *et al.* (2023), the authors tested ChatGPT during a webinar hosted by PICO Portal[32] developers to explore ChatGPT’s capacity and get feedback on its outputs. The goal was to determine if ChatGPT could be used to assist in the planning of an SLR, refining research questions or supporting by drafting the search or analysis methods. Data-specific tasks such as data extraction were not addressed. Table 6.2 summarizes the findings regarding the performance of ChatGPT on the execution of SLR activities.

The authors also performed a minor test with GPT-4 (a more recent version released in March 2023). They only observed a mild improvement in the synthesis and summary of the three abstracts investigated. Another important point highlighted by the authors is the non-deterministic characteristic of the answers generated by ChatGPT: the answer will not be the same when the same question is asked multiple times. This fact could be an interference regarding the reproducibility aspect of an SLR (Qureshi *et al.*, 2023).

Regarding referencing capabilities, the authors mentioned ChatGPT’s inability to perform searches and perform real literature retrieval (Qureshi *et al.*, 2023). A recently released

[32] https://picoportal.org

Table 6.2 : Summary of ChatGPT performance on SLR tasks execution. Adapted from Qureshi *et al.* (2023).

SLR Task	Summary of the results
Formulating a structured SLR question	ChatGPT's output suggested an appropriate interpretation and contextualization of the presented results providing a starting point for refinement.
Creating eligibility criteria and screening titles	The proposed criteria and selected articles could be a useful starting point, but refinement may be needed depending on the complexity of the question.
Generating PubMed search strategy	A ChatGPT-generated search strategy could be helpful for those lacking access to an informationist, but the proposed strategy had multiple issues, including the fabrication of controlled vocabulary, requiring search strategy construction experience for troubleshooting.
Producing code for meta-analysis	ChatGPT was able to generate code for conducting a meta-analysis in Python and R. However, coding errors were present, necessitating corrections by a knowledgeable user.
Synthesis and summary of multiple studies	ChatGPT has the potential to assist in the initial stages of picking relevant information from abstracts and creating a summary. However, errors were observed, indicating that the technology is not yet fully prepared for this task.
Referencing	When asked for references, the model's response could not be verified. Many times ChatGPT created references that do not exist. Moreover, when asked to perform a search in bibliographic databases, ChatGPT responds that it is unable to perform any real literature retrieval. In fact, LLMs are not developed to look through literature to find real sources but to build a response using predictions.

web-based application called Consensus[33] (beta version) promising to answer questions based on scientific research. It blends NLP, ML, blockchain and LLM (GPT-4) to analyze scientific web content (openly available). It offers users a free limited version and paid plan options. However, we did not evaluate the applicability of it in the SLR context nor its effectiveness and accuracy of results generated by the application.

[33] https://consensus.app/search/

In summary, the authors conclude the study by reflecting on their concerns about the use of LLMs in research and expressing their uncertainty and hesitancy. They add that it is essential that the users that attempt to use ChatGPT must have expertise on the research topic under investigation to be able to verify and correct errors. Lastly, the authors acknowledge the potential of ChatGPT and other LLMs to be integrated into the SLR process, but their current capabilities are still insufficient to be confident and reliable in their use in any way (Qureshi *et al.*, 2023).

The study of Alchokr *et al.* (2022) explored a deep-learning-based contextualized embedding clustering technique employing two transformer-based language models BERT (Kenton & Toutanova, 2019) and S-BERT (Reimers & Gurevych, 2019) to perform the initial selection of studies for SLRs. More specifically, these models are employed to derive contextualized embeddings from the title and abstract of studies at various levels, such as word, sentence, and paragraph. Additionally, a weightage scheme is incorporated to prioritize studies closely related to the SLR search string. Finally, the study-level embeddings are clustered using the k-means algorithm to identify groups of similar documents. To evaluate their proposed technique, the authors compared the generated models' resulting clusters with the results of two SLRs manually conducted.

The results of the small-scale experiment performed by the authors show that clustering on contextualized embeddings obtained via language models (BERT and S-BERT) outperforms their traditional baseline model (TF-IDF). Regarding the model settings experimented, S-BERT-paragraph represents the best-performing model setting in terms of optimizing the required parameters such as correctly identifying primary studies, the number of additional studies identified as part of the relevant cluster and the execution time of the experiment. Moreover, the results outline that the weightage schemes are inconsistent, but they were

evaluated with only two SLR datasets. An extension with a larger dataset is needed to underline a conclusion (Alchokr *et al.*, 2022).

Although the study presents a preliminary investigation, the use of natural-language-based deep-learning architectures such as LLMs to automate the initial selection of studies is promising. The authors suggest as future work (i) extend the experiment with more SLRs from different SE research domains; (ii) perform an experiment using the SLRs' full-text instead of just title, abstract and keywords; and (iii) explore other LLMs such as GPT[34] and XLNet[35] and even LLMs trained specifically on scientific text such as SciBert (Beltagy *et al.*, 2019) to improve the presented results (Alchokr *et al.*, 2022).

Lastly, in Wang *et al.* (2023), a pre-print study not published until June 2023, investigated the use of ChatGPT to formulate and refine search strings (Boolean queries) for SLRs. The authors experimented with an extensive set of prompts on over 100 different SLR topics. In summary, the boolean queries generated by ChatGPT obtained higher precision but lower recall values when compared to manual string formulation. During the use of guided prompts (i.e. following instructions of the conceptual or objective procedures), the effectiveness of the results improved. However, the study presents a major limitation: each time a prompt is executed a different boolean query is generated. Thus, in the context of SLRs where reproducibility is a key aspect, ChatGPT cannot be used yet. Overall, the study demonstrated the potential of ChatGPT to generate search strings for SLR. On the one hand, the authors are still not certain of the use of ChatGPT to generate any SLR search string. On the other hand, they assert that the topic investigated is promising and an exciting foundation for future research.

[34] https://openai.com/product#made-for-developers

[35] https://huggingface.co/docs/transformers/model_doc/xlnet

6.6 CHAPTER FINAL REMARKS

In this chapter, we presented and investigated an automation solution proposal to support searching for new evidence and selecting evidence for SLR updates as well as we discussed future directions on CSLR automation addressing the creation of a dedicated SLR repository and a brief literature analysis on the application of LLMs to support the SLR process.

We built a tool prototype and described it in detail. Based on a small-scale evaluation, we discuss automation limitations and perspectives for the SLR update context.

Concerning the search for evidence, preliminary results of our investigation indicate that, while there are challenges faced when automating snowballing-based search strategies (e.g., to automatically gather DOIs for papers, to automatically complement BibTeX information of identified papers), these strategies can be fully automated with minor losses. This can be particularly helpful for updating SLRs, given that forward snowballing has been recommended for this context (Felizardo *et al.*, 2016; Wohlin *et al.*, 2020). Furthermore, applying automated snowballing iterations could also be employed to reduce the effort of applying SLR search strategies in general (Wohlin *et al.*, 2022).

We also investigated the selection of studies in SLR updates using ML algorithms trained based on papers included and excluded in the original SLR. While improvements can surely be obtained, emerging results obtained by our prototype tool are promising and already considered acceptable by literature (Dieste *et al.*, 2009). In our small-scale evaluation, it was also possible to observe that using our best-obtained ML model (Linear SVM optimized using the SGD algorithm) conservatively minimizing the risk of missing papers during the SLR update, it would still be possible to reduce the number of papers to be manually analyzed in about 2.5 times.

We argue that an SLR repository will provide the SE community easier access to SLRs and their data; time-savings to conduct and update an SLR; and improve collaboration by providing a central location to find specialists in a specific research area. Available protocols can also serve as a guide to the writing of other protocols, especially by novice researchers; to provide a useful tool for SE practitioners who need to make evidence-based decisions; and to promote integration between SE academia-industry. However, further clarification on the repository functionalities regarding the CSLR process requirements, user interface, and implementation details would be valuable for a comprehensive evaluation.

Regarding the adoption of LLMs in the SLR context including their application in the CSLR process, the existing literature corroborates on the need to further investigation to understand the current limitations and capacity of LLMs in the context of supporting SLR and CSLR activities such as formulation of RQs and search strings (SLR context), selection of studies, synthesis and summary of multiple studies, etc. Interesting results were observed during all the initial investigations (Qureshi *et al.*, 2023; Alchokr *et al.*, 2022; Wang *et al.*, 2023), but the authors emphasized that the investigated models are not yet ready to be used in practice with confidence by the SLR community.

We envision that automated approaches could significantly help to reduce the SLR update effort and time spent. Investigations in this direction should be encouraged and undertaken to help the community keep SLRs up to date at the pace of the rapid increase of new evidence.

CONCLUSION

The motivation behind the definition of the CSLR concept and process is to provide a systematic process and guidelines to help mitigate the intermittent SLR update problem. The CSLR contributes to avoiding missing new potential relevant research in evidence syntheses or decision-making. Actions to keep SLRs updated are of great importance to the SLR research field (Nepomuceno & Soares, 2019). Next, we summarize the thesis contributions, report limitations and future work.

SUMMARY OF CONTRIBUTIONS

The CSLR process and guidelines proposed (Chapter 3), evaluated (Chapters 4 and 5) and improved (Chapter 5) unifies several pieces of knowledge on SE SLR updates that have been investigated separately, integrating them with DevOps metaphors and open science concepts. As a consequence, we observed benefits such as (i) facilitating the identification if an SLR has been updated or not; (ii) assisting in the identification (search and selection) of potentially relevant evidence; (iii) promoting the sharing of potentially relevant evidence available in open repositories that are freely accessible by the SE community; (iv) supporting the decision on the need to update an SLR; and (v) supporting SLR authors throughout the update process.

In a more general view, the CSLR process and guidelines could also be an instrument to direct the SE community on research subjects that have been investigated. If an SLR is often cited, it proves that the subject of study is constantly evolving. This fact leads to questions such as: Does the SLR remain relevant? Is the SLR up to date? Does it need to be updated? As observed in the participative case study presented in Chapter 4, new research trends can be identified, leading to other research directions (questions) on a research subject. This

advantage was even pointed out by a SE SLR expert during the CSLR process and guidelines expert evaluation presented in Chapter 5.

Besides, the CSLR process and guidelines were effective during our evaluations and opened avenues for automating its activities. As can be observed in Chapter 6, the results obtained by our proposed prototype tool for searching and selecting studies are promising and demonstrated the potential to reduce by at least 2.5 times the effort potentially reflecting on the researchers' time spent during the search and selection of studies to update SLRs in SE.

The three RG contemplated in this thesis demonstrated positive results bringing contributions to the SE field. They are three-fold. First, a systematic, well-defined, and dual-validated process to update SLRs continuously in SE. Second, validated and improved guidelines for the CSLR process that describe details and examples on how to update SLRs in SE continuously. Third, a validated prototype tool addressing the search and selection of studies (triggers activities) of the CSLR process as well as an overview of research directions on CSLR automation.

LIMITATIONS AND FUTURE WORK

The limitations and mitigation actions of the research presented in this thesis are detailed at the end of each chapter respectively. Therefore, next, we summarize the main general limitations and provide research directions for future work.

One possible limitation is related to the sample size of the validations performed in this thesis. The participative case study presented in Chapter 4 used as a sample a SE SLR that was strategically chosen because it contemplated a series of requirements that allowed a complete evaluation of all stages of the CSLR process (see Section 4.1.1). The expert evaluation performed in Chapter 5 featured 6 SE SLR experts from different research groups

who thoroughly evaluated the CSLR process and guidelines, providing valuable feedback. Finally, the small-scale evaluation presented in Chapter 6 used a single SLR as an instrument due to the difficulty of obtaining reliable and detailed SLR datasets for SLRs that could potentially need an update. However, in all the validations performed, we carefully followed well-established guidelines and we had the collaboration of two renowned researchers who have been investigating the SE SLR process area for over ten years. Despite this, the CSLR process and guidelines could benefit from further practical evaluation.

Another limitation is the possibility of automating the CSLR process as a whole. In this thesis, we focus our automation analysis on the two trigger activities of the CSLR process, search and selection of studies (see Chapters 2 and 6). We also present some research directions on CSLR automation in Section 6.5. More specifically, Section 6.5.1 outlines a high-level overview of a pipeline/workflow for the CSLR process based on CI/CD process components. Furthermore, as explored in this thesis, adopting NLP and ML solutions is key for developing automation solutions to support several CSLR process activities. In Section 6.5.2, we provided a brief literature analysis on the application of LLMs to support activities of the SLR (and CSLR) process. The presented directions on automation were not validated, but they provide a starting point for further investigations.

The same way that the LSR was designed and integrated into medicine based on the needs of the area (Cochrane Reviews), the CSLR process and guidelines followed the same idea: we developed and validated them considering the SE context. Therefore, further investigations would be necessary to guarantee the applicability of the CSLR process and guidelines to other areas rather than SE.

Lastly, it was not possible to assess the acceptance and practical adoption of the CSLR process and guidelines by the SE community. The first publication that introduces the CSLR

concept and process was published in late September 2022 in a conference, and a journal paper with the CSLR guidelines is still under revision. However, the value of the CSLR process and guidelines was demonstrated during the validations performed in Chapters 4 and 5. More specifically, in Section 5.3.2 we reported observations on the relevance recognition of the CSLR value by SE researchers.

Several interesting research directions emerged during the research process of this thesis. They are mentioned afterward. It is worth mentioning that the first three directions are directly related to the limitations mentioned above.

– **Further practical evaluation of the CSLR process and guidelines:** We suggest extending the case study presented in Chapter 4 with a larger sample of SE SLRs and conducting a controlled experiment with a focus group to observe the live practical application of the CSLR process and guidelines.

– Development of a dedicated SLR repository in SE with the integration of the CSLR pipeline/workflow: We see this as the base step for the CSLR process automation due to the possibility of unifying SE SLR data in a unique place beside the opening of possibilities for the integration of automation support to the management and execution of CSLR process activities and execution flow. We strongly encourage research efforts on the design and implementation of a dedicated SE SLR repository as well as the integration of a pipeline/workflow of the CSLR process. Section 6.5.1 provides an initial high-level overview of an integration pipeline/workflow proposition for the CSLR process.

– Exploration of LLMs in the CSLR context: We encourage the investigation of modern ML algorithms including LLMs for the development of automated solutions for the CSLR process.

– Elaboration of an SLR update template: We suggest an elaboration of a protocol template proposition to support SE researchers that opted to update an SE SLR. This template should summarize the CSLR guidelines providing a ready-use template to support the protocol elaboration and especially the results reporting of SLR updates in SE.

REFERENCES

Aggarwal, C. C. & Zhai, C. (2012). A Survey of Text Classification Algorithms. In C. C. Aggarwal & C. Zhai (dir.), *Mining Text Data* 163–222. Springer.

Al-Zubidy, A. & Carver, J. C. (2019). Identification and prioritization of SLR search tool requirements: an SLR and a survey. *Empirical Software Engineering*, *24*(1), 139–169.

Al-Zubidy, A., Carver, J. C., Hale, D. P. & Hassler, E. E. (2017). Vision for SLR tooling infrastructure: Prioritizing value-added requirements. *Information and Software Technology*, *91*, 72 – 81.

Alabool, H., Kamil, A., Arshad, N. & Alarabiat, D. (2018). Cloud service evaluation method-based Multi-Criteria Decision-Making: A systematic literature review. *Journal of Systems and Software*, *139*, 161–188.

Alchokr, R., Borkar, M., Thotadarya, S., Saake, G. & Leich, T. (2022). Supporting Systematic Literature Reviews Using Deep-Learning-Based Language Models. In *IEEE/ACM 1st International Workshop on Natural Language-Based Software Engineering (NLBSE)*, 67–74. IEEE Computer Society.

Almeida, H., Meurs, M.-J., Kosseim, L. & Tsang, A. (2016). Data Sampling and Supervised Learning for HIV Literature Screening. *IEEE Transactions on NanoBioscience*, 354–361.

Ameller, D., Farré, C., Franch, X. & Rufián, G. (2016). A Survey on Software Release Planning Models. In P. Abrahamsson, A. Jedlitschka, A. Nguyen-Duc, M. Felderer, S. Amasaki, & T. Mikkonen (dir.). *Product-Focused Software Process Improvement - 17th International Conference, PROFES 2016, Trondheim, Norway, November 22-24, 2016, Proceedings*, Vol. 10027 of *Lecture Notes in Computer Science*, 48–65.

Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M. & Chatzigeorgiou, A. (2019). Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and Software Technology*, *106*, 201 – 230.

Ananiadou, S., Rea, B., Okazaki, N., Procter, R. & Thomas, J. (2009). Supporting Systematic Reviews Using Text Mining. *Social Science Computer Review - SOC SCI COMPUT REV*, 509–523.

Aphinyanaphongs, Y., Tsamardinos, I., Statnikov, A., Hardin, D. & Aliferis, C. (2005). Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association : JAMIA*, 207–16.

Babar, M. A. & Zhang, H. (2009). Systematic Literature Reviews in Software Engineering: Preliminary Results from Interviews with Researchers. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 346–355.

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A., Ananiadou, S., Liao, J. & Macleod, M. (2019). Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 1–12.

Barn, B. S., Raimondi, F., Athappian, L. & Clark, T. (2014). Slrtool: A Tool to Support Collaborative Systematic Literature Reviews. In S. Hammoudi, L. A. Maciaszek, & J. Cordeiro (dir.). *Proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS)*, 440–447. SciTePress.

Barros-Justo, J., Benitti, F. & Molleri, J. (2021, 11). Risks and risk mitigation in global software development: An update. *Journal of Software: Evolution and Process*, p. e2370.

Baskerville, R. L. (1997). Distinguishing action research from participative case studies. *Journal of Systems and Information Technology*, *1*, 24–43.

Bass, L., Weber, I. & Zhu, L. (2015). *DevOps: A Software Architect's Perspective* (1st). Addison-Wesley Professional.

Bekhuis, T. & Demner-Fushman, D. (2010). Towards Automating the Initial Screening Phase of a Systematic Review. *160*, 146–150.

Bekhuis, T. & Demner-Fushman, D. (2012). Screening Nonrandomized Studies for Medical Systematic Reviews: A Comparative Study of Classifiers. *Artificial Intelligence Medicine*, *55*, 197–207.

Bekhuis, T., Tseytlin, E., Mitchell, K. J. & Demner-Fushman, D. (2014). Feature Engineering and a Proposed Decision-Support System for Systematic Reviewers of Medical

Evidence. *PLOS ONE*, 1–10.

Beltagy, I., Lo, K. & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*.

Bezerra, R., da Silva, F., Santana, A., Magalhaes, C. & Santos, R. (2015). Replication of Empirical Studies in Software Engineering: An Update of a Systematic Mapping Study. In *9^{th} International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–4.

Bottou, L. (2012). Stochastic Gradient Descent Tricks. *7700*, 421–436.

Bowes, D., Hall, T. & Beecham, S. (2012). SLuRp: A Tool to Help Large Complex Systematic Literature Reviews Deliver Valid and Rigorous Results. In *Proceedings of the 2nd International Workshop on Evidential Assessment of Software Technologies*, EAST '12, p. 33–36., New York, NY, USA. Association for Computing Machinery.

Boyle, E., Hainey, T., Connolly, T., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C. & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education*, *94*, 178–192.

Brooker, J., Synnot, A., McDonald, S., Elliott, J., Turner, T. & et al. (2019). Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. 1–60.

Cartaxo, B., Pinto, G. & Soares, S. (2020). Rapid Reviews in Software Engineering. In M. Felderer & G. H. Travassos (dir.). *Contemporary Empirical Methods in Software Engineering* (357–384). Springer.

Cartaxo, B., Pinto, G., Vieira, E. & Soares, S. (2016). Evidence Briefings: Towards a Medium to Transfer Knowledge from Systematic Reviews to Practitioners. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 57:1–57:10., New York, NY, USA. Association for Computing Machinery.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T. *et al.* (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1–4.

Cohen, A., Ambert, K. & McDonagh, M. (2009). Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. *Journal of the American Medical Informatics Association : JAMIA*, 690–704.

Cohen, A., Ambert, K. & McDonagh, M. (2010). A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. *AMIA – Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 121–125.

Cohen, A., Ambert, K. & McDonagh, M. (2011). Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association: JAMIA*, *18*, 104–05.

Cohen, K., Johnson, H., Verspoor, K., Roeder, C. & Hunter, L. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, *11*, 492.

CrossRef (2023). *Crossref Metadata Search*. https://search.crossref.org/references. Online; accessed 19 March 2023.

Cruzes, D. & Dybå, T. (2010). Synthesizing evidence in software engineering research. In *International Symposium on Empirical Software Engineering and Measurement (ESEM*, 1–10. Association for Computing Machinery.

Cruzes, D. S. & Dyba, T. (2011). Recommended Steps for Thematic Synthesis in Software Engineering. In *2011 International Symposium on Empirical Software Engineering and Measurement*, 275–284. IEEE Computer Society.

da Silva, F., Santos, A., Soares, S., França, A. & Monteiro, C. (2010). Six Years of Systematic Literature Reviews in Software Engineering: an Extended Tertiary Study. In *International Conference on Software Engineering (ICSE)*, 1–10. IEEE Computer Society.

Dantas, E., Perkusich, M., Dilorenzo, E., Santos, D. F., Almeida, H. & Perkusich, A. (2018).

Effort estimation in agile software development: an updated review. *International Journal of Software Engineering and Knowledge Engineering*, *28*(11n12), 1811–1831.

de A. Cabral, J. T. H., Oliveira, A. L. I. & da Silva, F. Q. B. (2023). Ensemble Effort Estimation: An updated and extended systematic literature review. *Journal of Systems and Software*, *195*, 111542.

de Aguiar Beninca, R., Huzita, E. H. M., Cardoza, E., Leal, G. C. L., Balancieri, R. & Massago, Y. (2015). Knowledge Management Practices in GSD - A Systematic Literature Review Update. In *17th International Conference on Enterprise Information Systems (ICEIS)*, 365–373. SciTePress.

de Oliveira, L. (2015). *Architectural design of service-oriented robotic systems*. (Ph.d. thesis). São Carlos, Brazil / Vannes, France.

Dieste, O., Griman, A. & Juristo, N. (2009). Developing search strategies for detecting relevant experiments. *Empirical Software Engineering*, *14*(5), 513–539.

Dieste, O., López, M. & Ramos, F. (2008). Formalizing a Systematic Review Updating Process. In *6th Int. Conference on Software Engineering Research, Management and Applications (SERA)*, 143–150. IEEE Computer Society.

Dieste, O., López, M. & Ramos, F. (2008). Obtaining Well-Founded Practices about Elicitation Techniques by Means of an Update of a Previous Systematic Review. In *20th International Conference on Software Engineering & Knowledge Engineering (SEKE)*, 769–772. Knowledge Systems Institute Graduate School.

dos Santos, V., Iwazaki, A. Y., Felizardo, K. R., de Souza, E. F. & Nakagawa, E. Y. (2021). Towards Sustainability of Systematic Literature Reviews. In *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, New York, NY, USA. Association for Computing Machinery.

Duvall, P., Matyas, S. & Glover, A. (2007). *Continuous Integration: Improving Software Quality and Reducing Risk* (first). Addison-Wesley Professional.

Elliott, J., Synnot, A., Turner, T., Simmonds, M., Akl, E., Mcdonald, S., Salanti, G.,

Meerpohl, J., MacLehose, H., Hilton, J., Shemilt, I. & Thomas, J. (2017). Living systematic review 1: Introduction - the Why, What, When and How. *Journal of Clinical Epidemiology*, 23–30.

Fabbri, S., Felizardo, K., Ferrari, F., Hernandes, E., Octaviano, F., Nakagawa, E. & Maldonado, J. (2013). Externalising tacit knowledge of the systematic review process. *IET Software*, *7*(6), 298–307.

Fabbri, S., Silva, C., Hernandes, E., Octaviano, F., Di Thommazo, A. & Belgamo, A. (2016). Improvements in the StArt Tool to Better Support the Systematic Review Process. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 21:1–21:5., New York, NY, USA. Association for Computing Machinery.

Felizardo, K., Andery, G., Paulovich, F., Minghim, R. & Maldonado, J. (2012). A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology*, *54*(10), 1079–1091.

Felizardo, K., Mendes, E., Kalinowski, M., Souza, E. F. & Vijaykumar, N. (2016). Using Forward Snowballing to update Systematic Reviews in Software Engineering. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 53:1–53:6. Association for Computing Machinery.

Felizardo, K., Nakwgawa, E., MacDonell, S. & Maldonado, J. (2014). A Visual Analysis Approach to Update Systematic Reviews. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 1–10. Association for Computing Machinery.

Felizardo, K., Salleh, N., Martins, R., Mendes, E., MacDonell, S. & Maldonado, J. (2011). Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews. In *5th International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–10. Association for Computing Machinery.

Felizardo, K. R. & Carver, J. C. (2020). Automating Systematic Literature Review. In M. Felderer & G. H. Travassos (dir.). *Contemporary Empirical Methods in Software Engineering* (327–355). Springer International Publishing.

Felizardo, K. R., da Silva, A. Y. I., de Souza, E. F., Vijaykumar, N. L. & Nakagawa, E. Y.

(2018). Evaluating Strategies for Forward Snowballing Application to Support Secondary Studies Updates: Emergent Results. In *Proceedings of the XXXII Brazilian Symposium on Software Engineering (SBES)*, p. 184–189., New York, NY, USA. Association for Computing Machinery.

Felizardo, K. R., de Souza, E. F., Malacrida, T., Napoleão, B. M., Petrillo, F., Hallé, S., Vijaykumar, N. L. & Nakagawa, E. Y. (2020). Knowledge Management for Promoting Update of Systematic Literature Reviews: An Experience Report. In *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 471–478. IEEE Computer Society.

Felizardo, K. R., Éica Ferreira de Souza, Napoleão, B. M., Vijaykumar, N. L. & Baldassarre, M. T. (2020). Secondary studies in the academic context: A systematic mapping and survey. *Journal of Systems and Software*, *170*, 110734.

Feng, L., Chiam, Y., Abdullah, E. & Obaidellah, U. (2017). Using Suffix Tree Clustering Method to Support the Planning Phase of Systematic Literature Review. *Malaysian Journal of Computer Science*, 311–332.

Ferrari, F. & Maldonado, J. (2008). Experimenting with a Multi-Iteration Systematic Review in Software Engineering. In *5th Experimental Software Engineering Latin America Workshop (ESELAW)*, 1–10. ICMC/USP.

Franca, A., Gouveia, T., Santos, P., Santana, C. & da Silva, F. (2011). Motivation in software engineering: A systematic review update. In *15th Int. Conference on Evaluation and Assessment in Software Engineering (EASE'11)*, 154–163.

França, A. C. C., Gouveia, T. B., Santos, P. C., Santana, C. A. & da Silva, F. Q. (2011). Motivation in software engineering: A systematic review update. In *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE)*, 154–163. IET - The Institute of Engineering and Technology/IEEE Xplore.

Frunza, O., Inkpen, D. & Matwin, S. (2010). Building Systematic Reviews Using Automatic Text Classification Techniques. In *23rd International Conference on Computational Linguistics (COLING*, Vol. 2, 303–311. Chinese Information Processing Society of China.

Frunza, O., Inkpen, D., Matwin, S., Klement, W. & OaBlenis, P. (2011). Exploiting the

systematic review protocol for classification of medical abstracts. *Artificial Intelligence Medicine*, *51*(1), 17–25.

Fu, C., Zhang, H., Huang, X., Zhou, X. & Li, Z. (2019). A Review of Meta-Ethnographies in Software Engineering. In *Proceedings of the 23rd International Conference on Evaluation and Assessment in Software Engineering (EASE)*, p. 68–77. Association for Computing Machinery.

Garcés, L., Oquendo, F. & Nakagawa, E. (2020). Assessment of Reference Architectures and Reference Models for Ambient Assisted Living Systems: Results of a Systematic Literature Review. *International Journal of E-Health and Medical Communications*, *11*(1), 17–36.

Garcés, L., Felizardo, K., Oliveira, L. & Nakagawa, E. (2017). An Experience Report on Update of Systematic Literature Reviews. In *The 29th International Conference on Software Engineering and Knowledge Engineering (SEKE)*, 91–96. Research Inc. and Knowledge Systems Institute Graduate School.

García Adeva, J., Pikatza Atxa, J., Ubeda Carrillo, M. & Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, *41*(4, Part 1), 1498–1508.

Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Schünemann, H., Akl, E., Beyene, J., Chang, S., Churchill, R., Dearness, K., Guyatt, G., Lefebvre, C., Liles, B., Marshall, R., García, L., Mavergames, C., Nasser, M., Qaseem, A., Sampson, M. & Wilson, E. (2016). When and how to update systematic reviews: Consensus and checklist. *British Medical Journal*, *354*, 1–10.

Garousi, V., Felderer, M. & Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology*, *106*, 101–121.

Gates, A., Gates, M., Sebastianski, M., Guitard, S., Elliott, S. & Hartling, L. (2020). The semi-automation of title and abstract screening: A retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. *BMC Medical Research Methodology*, 2–12.

Gates, A., Guitard, S., Pillay, J., Elliott, S., Dyson, M., Newton, A. & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews*, p. 278.

Ghafari, M., Mortaza, S. & Touraj, E. (2012). A Federated Search Approach to Facilitate Systematic Literature Review in Software Engineering. *International Journal of Software Engineering & Applications*, *3*, 13–24.

GitHub Docs (2023). *Understanding GitHub Actions*. https://docs.github.com/en/actions/learn-github-actions/understanding-github-actions. Online; accessed 2 July 2023.

González-Toral, S., Freire, R., Gualán, R. & Saquicela, V. (2019). A ranking-based approach for supporting the initial selection of primary studies in a Systematic Literature Review. In *XLV Latin American Computing Conference (CLEI)*, 1–10. IEEE Computer Society.

Google (2023). *Getting started with the built-in BERT algorithm*. https://cloud.google.com/ai-platform/training/docs/algorithms/bert-start. Online; accessed 5 June 2023.

Götz, S. (2006). An effective general purpose approach for automated biomedical document classification. In *AMIA Annual Symposium proceeding*, 161–5. American Medical Informatics Association.

Götz, S. (2018). Supporting Systematic Literature Reviews in Computer Science: The Systematic Literature Review Toolkit. MODELS '18, p. 22–26. Association for Computing Machinery.

Guo, M., Zhang, C. & Wang, F. (2017). What is the Further Evidence about UML? - A Systematic Literature Review. In *2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW)*, 106–113. IEEE Computer Society.

Hamel, C., Thavorn, K., Wells, G. & Hutton, B. (2020). An evaluation of DistillerSR machine learning-based prioritization tool for title/abstract screening - impact on reviewer-relevant outcomes. *BMC Medical Research Methodology*, *20*, 1–14.

Hannes, K. & Lockwood, C. (2011). *Synthesizing qualitative research: choosing the right*

approach. John Wiley & Sons.

Hassler, E., Carver, J., Kraft, N. & Hale, D. (2014). Outcomes of a Community Workshop to Identify and Rank Barriers to the Systematic Literature Review Process. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 1–10. Association for Computing Machinery.

Hassler, E. E., Hale, D. P. & Hale, J. E. (2018). A comparison of automated training-by-example selection algorithms for Evidence Based Software Engineering. *Information and Software Technology*, *98*, 59–73.

Higgins, J. & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration.

Higgins, J., Green, S. & Scholten, R. (2008). Maintaining Reviews: Updates, Amendments and Feedback. In *Cochrane Handbook for Systematic Reviews of Interventions* (31–49). John Wiley & Sons, Ltda.

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. & Welch, V. (2019). Cochrane Handbook for Systematic Reviews of Interventions. *Cochrane Handbook for Systematic Reviews of Interventions*.

Hinderks, A., Mayo, F. J. D., Thomaschewski, J. & Escalona, M. J. (2020). An SLR-Tool: Search Process in Practice: A Tool to Conduct and Manage Systematic Literature Review (SLR). In *ICSE Companion*, p. 81–84. Association for Computing Machinery.

Hoisl, B. & Sobernig, S. (2016). Open-source development tools for domain-specific modeling: Results from a systematic literature review. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 5001–5010. IEEE Computer Society.

Hosmer Jr, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression*, Vol. 398. John Wiley & Sons.

Howard, B., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M., Holmgren, S., Pelch, K., Walker, V., Rooney, A., Macleod, M., Shah, R. & Thayer, K. (2016). SWIFT-Review: A text-mining workbench for systematic review. *Systematic Reviews*, 1–16.

Humble, J. & Farley, D. (2010). *Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation* (1st). Addison-Wesley Professional.

Hummel, M. (2014). State-of-the-Art: A Systematic Literature Review on Agile Information Systems Development. In *47th Hawaii International Conference on System Sciences (HICSS)*, 4712–4721. IEEE Computer Society.

Imtiaz, S., Bano, M., Ikram, N. & Niazi, M. (2013). A Tertiary Study: Experiences of Conducting Systematic Literature Reviews in Software Engineering. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 177–182. Association for Computing Machinery.

Jenkins (2023). *Jenkins User Documentation*. https://www.jenkins.io/doc. Online; accessed 5 November 2022.

Jiang, S., Zhang, H., Gao, C., Shao, D. & Rong, G. (2015). Process simulation for software engineering education. In *Proceedings of the 2015 International Conference on Software and System Process*, 147–156.

Joachims, T. (1999). Transductive Inference for Text Classification Using Support Vector Machines. In *ICML*, p. 200–209. Morgan Kaufmann Publishers Inc.

Kenton, J. D. M.-W. C. & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1, p. 2.

Kibriya, A. M., Frank, E., Pfahringer, B. & Holmes, G. (2005). Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, 488–499. Springer.

Kim, S. & Choi, J. (2012). Improving the Performance of Text Categorization Models used for the Selection of High Quality Articles. *Healthcare informatics research*, 18–28.

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews* publication no. TR/SE-0401 (Keele) - 0400011T.1 (NICTA). Software Engineering Group - Department of Computer Science - Keele University and Empirical SE - National ICT Australia Ltd.

Kitchenham, B., Brereton, P., Turner, M., Niazi, M., Linkman, S., Pretorius, R. & Budgen, D. (2010). Refining the systematic literature review process–two participant-observer case studies. *Empirical Software Engineering*, *15*(6), 618–653.

Kitchenham, B., Budgen, D. & Brereton, P. (2015). *Evidence-Based Software Engineering and Systematic Reviews*. Chapman & Hall/CRC Innovations in Software Engineering and Software Development Series. Chapman & Hall/CRC.

Kitchenham, B. & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering* publication no. EBSE 2007-001. Keele University and Durham University, UK.

Kitchenham, B., Mendes, E. & Travassos, G. H. (2006). A Systematic Review of Cross- vs. within- Company Cost Estimation Studies. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE)*. Association for Computing Machinery.

Kitchenham, B. A., Madeyski, L. & Budgen, D. (2023). SEGRESS: Software Engineering Guidelines for REporting Secondary Studies. *IEEE Transactions on Software Engineering*, *49*(3), 1273–1298.

Kitchenham, B. A., Mendes, E. & Travassos, G. H. (2007). Cross versus Within-Company Cost Estimation Studies: A Systematic Review. *IEEE Transactions on Software Engineering*, *33*(5), 316–329.

Kontonatsios, G., Brockmeier, A. J., Przybyła, P., McNaught, J., Mu, T., Goulermas, J. Y. & Ananiadou, S. (2017). A semi-supervised approach using label propagation to support citation screening. *Journal of Biomedical Informatics*, *72*, 67–76.

Kontonatsios, G., Spencer, S., Matthew, P. & Korkontzelos, I. (2020). Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, *6*, 100030.

Kuhrmann, M., Fernández, D. M. & Daneva, M. (2017, 6). On the Pragmatic Design of Literature Studies in Software Engineering: An Experience-Based Guideline. *Empirical Soft. Eng.*, p. 2852–2891.

Leyh, C. & Sander, P. (2011). Critical success factors for ERP system implementation projects: An update of literature reviews. *Enterprise Systems. Strategic, Organizational, and Technological Dimensions*, 45–67.

Likert, R. (2010). A technique for the measurement of attitudes. *Archives of Psychologyy*, *22*(140), 1–55.

Lilleberg, J., Zhu, Y. & Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In *14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 136–140. IEEE Computer Society.

Linneberg, M. S. & Korsgaard, S. (2019). Coding qualitative data: a synthesis guiding the novice. *Qualitative Research Journal*, *19*, 259–270.

Liu, J., Timsina, P. & El-Gayar, O. (2018). A Comparative Analysis of Semi-Supervised Learning: The Case of Article Selection for Medical Systematic Reviews. *Information Systems Frontiers*, *20*(2), 195–207.

Malheiros, V., Hohn, E., Pinho, R., Mendonca, M. & Maldonado, J. (2007). A visual text mining approach for systematic reviews. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 245–254. Association for Computing Machinery.

Manikas, K. (2016). Revisiting software ecosystems research: A longitudinal literature study. *Journal of Systems and Software*, *117*, 84–103.

Marcos-Pablos, S. & García-Peñalvo, F. (2020). Information retrieval methodology for aiding scientific database search. *Software Computing*, 5551–5560.

Marshall, C. & Brereton, P. (2013). Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 296–299. Association for Computing Machinery.

Marshall, C., Brereton, P. & Kitchenham, B. (2015). Tools to Support Systematic Reviews in Software Engineering: A Cross-domain Survey Using Semi-structured Interviews. In

International Conference on Evaluation and Assessment in Software Engineering (EASE), 26:1–26:6. Association for Computing Machinery.

Marshall, C., Kitchenham, B. & Brereton, P. (2018). Tool Features to Support Systematic Reviews in Software Engineering – A Cross Domain Study. *e-Informatica Software Engineering Journal*, *12*(1), 79–115.

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O. & O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association : JAMIA*, 446–53.

Mendes, E., Felizardo, K. R., Wohlin, C. & Kalinowski, M. (2019). Search Strategy to Update Systematic Literature Reviews in Software Engineering. In *45th Euromicro Conference on Software Engineering and Advanced Applications SEAA*, 355–362. IEEE Computer Society.

Mendes, E., Kalinowski, M., Martins, D., Ferrucci, F. & Sarro, F. (2014). Cross- vs. Within-company Cost Estimation Studies Revisited: An Extended Systematic Review. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 12:1–12:10. Association for Computing Machinery.

Mendes, E., Wohlin, C., Felizardo, K. & Kalinowski, M. (2020). When to update systematic literature reviews in software engineering. *Journal of Systems and Software*, *167*, 110607.

Mendez, D., Graziotin, D., Wagner, S. & Seibold, H. (2020). Open Science in Software Engineering. In M. Felderer & G. H. Travassos (dir.). *Contemporary Empirical Methods in Software Engineering* (477–501). Cham: Springer International Publishing.

Mergel, G. D., Silveira, M. S. & da Silva, T. S. (2015). A Method to Support Search String Building in Systematic Literature Reviews through Visual Text Mining. In *SAC*, SAC '15, p. 1594–1601., New York, NY, USA. Association for Computing Machinery.

Mertz, J., Corrêa, E. S., Gomes, W. & Nunes, I. (2018). *LRDB: a Database of Literature Reviews in Computer Science*. http://prosoft.inf.ufrgs.br/lrdb/Home/Index. Online; accessed 2 july 2023.

Merzouki, A. (2023). *What's new in python 3.8*. https://docs.python.org/3/whatsnew/3.8.html. Online; accessed 19 March 2023.

Meyer, M. (2014). Continuous Integration and Its Tools. *IEEE Software*, *31*(3), 14–16.

Miwa, M., Thomas, J., Mara-Eves, A. O. & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, *51*, 242–253.

Moher, D., Tsertsvadze, A., Tricco, A., Eccles, M., Grimshaw, J., Sampson, M. & Barrowman, N. (2008). When and how to update systematic reviews. *Cochrane database of systematic reviews (Online)*, *1*, 1–10.

Molléri, J. S. & Benitti, F. B. V. (2015). SESRA: A Web-Based Automated Tool to Support the Systematic Literature Review Process. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. Association for Computing Machinery.

Mourao, E., Kalinowski, M., Murta, L., Mendes, E. & Wohlin, C. (2017). Investiganting the use of a hybrid search strategy for systematic review. In *11st International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 193–198. Association for Computing Machinery.

Mourão, E., Pimentel, J. F., Murta, L., Kalinowski, M., Mendes, E. & Wohlin, C. (2020). On the performance of hybrid search strategies for systematic literature reviews in software engineering. *Information and Software Technology*, *123*, 106294.

Nair, S., De La Vara, J. L., Sabetzadeh, M. & Briand, L. (2014). An extended systematic literature review on provision of evidence for safety certification. *Information and Software Technology*, *56*(7), 689–717.

Napoleão, B., Felizardo, K., Souza, E. & Vijaykumar, N. (2017). Practical similarities and differences between systematic literature reviews and systematic mappings: a tertiary study. In *The 29th International Conference on Software Engineering and Knowledge Engineering (SEKE)*, 85–90. Research Inc. and Knowledge Systems Institute Graduate School.

Napoleão, B. M., Felizardo, K. R., Souza, E. F. d., Petrillo, F., Hallé, S., Vijaykumar, N. L. & Nakagawa, E. Y. (2021). Establishing a Search String to Detect Secondary Studies in Software Engineering. In *47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 9–16. IEEE Computer Society.

Napoleão, B. M., felizardo, K. R., Petrillo, F., Hallé, S. & Kalinowski, M. (2023). *Supplementary material - Guidelines to Continous Systematic Review in Software Engineering*. https://doi.org/10.5281/zenodo.7686514. Online.

Napoleão, B. M., Petrillo, F. & Hallé, S. (2021). Automated Support for Searching and Selecting Evidence in Software Engineering: A Cross-domain Systematic Mapping. In *47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*.

Napoleão, B. M., Petrillo, F. & Hallé, S. (2021). *Supplementary material - A cross-domain SM on automated support for searching and selecting evidence for SLRs in SE*. https://doi.org/10.5281/zenodo.4719161. Online.

Napoleão, B. M., Petrillo, F., Hallé, S. & Kalinowski, M. (2022). *Supplementary material - Towards Continuous Systematic Literature Review in Software Engineering*. https://doi.org/10.5281/zenodo.6503143. Online.

Napoleão, B. M., Petrillo, F., Hallé, S. & Kalinowski, M. (2022). Towards Continuous Systematic Literature Review in Software Engineering. In *48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 467–474. IEEE Computer Society.

Napoleão, B. M., Sarkar, R., Hallé, S., Petrillo, F. & Kalinowski, M. (2023). *Tool prototype and dataset - Emerging Results on Automated Support for Searching and Selecting Evidence for Systematic Literature Review Updates*. https://doi.org/10.5281/zenodo.7888955. Online.

Nepomuceno, V. & Soares, S. (2018). Maintaining Systematic Literature Reviews: Benefits and Drawbacks. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 47:1–47:4. Association for Computing Machinery.

Nepomuceno, V. & Soares, S. (2019). On the need to update systematic literature reviews. *Information and Software Technology*, *109*, 40–42.

Nepomuceno, V. & Soares, S. (2020). Avoiding Plagiarism in Systematic Literature Reviews: An Update Concern. In *Proceedings of the 14th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 32:1–32:6. Association for Computing Machinery.

NLTK Team (2023). *Natural Language Toolkit*. https://pypi.org/project/nltk. Online; accessed 19 March 2023.

Noblit, G. & Hare, R. (1988). *Meta-Ethnography*. Newbury Park, Calif.: Sage Publications Inc.

Nonaka, I. & Takeuchi, H. (1997). *The knowledge-creating company* (1). Oxford University Press.

Octaviano, F., Felizardo, K., Maldonado, J. & Fabbri, S. (2014). Semi-automatic selection of primary studies in systematic literature reviews: is it reasonable? *Empirical Software Engineering*, 1898–1917.

Octaviano, F., Felizardo, K. R., Fabbri, S. C. P. F., Napoleão, B. M., Petrillo, F. & Hallé, S. (2022). SCAS-AI: A Strategy to Semi-Automate the Initial Selection Task in Systematic Literature Reviews. In *48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 483–490. IEEE Computer Society.

O'Leary, D. & Studer, R. (2001). Knowledge Management: an Interdisciplinary Approach. *IEEE Intelligent Systems*, *16*(1), 24–25.

Olorisade, B. K., Brereton, P. & Andras, P. (2017). Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist. *Journal of Biomedical Informatics*, *73*, 1–13.

Olorisade, B. K., Brereton, P. & Andras, P. (2019). The use of bibliography enriched features for automatic citation screening. *Journal of Biomedical Informatics*, *94*.

Olorisade, B. K., de Quincey, E., Brereton, P. & Andras, P. (2016). A Critical Analysis of Studies That Address the Use of Text Mining for Citation Screening in Systematic Reviews. In *Proceedings of the 20th International Conference on Evaluation and Assessment in*

Software Engineering (EASE), 14:1–14:11. Association for Computing Machinery.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, *4*, 1–25.

OpenAI (2023). *ChatGPT: optimizing language models for dialogue*. https://openai.com/blog/chatgpt. Online; accessed 5 June 2023.

Ouhbi, B., Kamoune, M., Frikh, B., Zemmouri, E. M. & Behja, H. (2016). A Hybrid Feature Selection Rule Measure and Its Application to Systematic Review. iiWAS '16, p. 106–114., New York, NY, USA. Association for Computing Machinery.

Paula, A. C. M. d. & Carneiro, G. d. F. d. (2016). A systematic literature review on cloud computing adoption and migration. In *International Conference on Evaluation of Novel Approaches to Software Engineering*, 222–243. Springer.

Petersen, K., Feldt, R., Mujtaba, S. & Mattsson, M. (2008). Systematic Mapping Studies in Software Engineering. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 1–10. Association for Computing Machinery.

Petersen, K., Vakkalanka, S. & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: an update. *Information and Software Technology*, *64*(1), 1–18.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Pfleeger, S. L. (1995). Experimental design and analysis in software engineering. *Annals of Software Engineering*, *1*(1), 219–253.

Picard, R. R. & Cook, R. D. (1984). Cross-Validation of Regression Models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Pizzoleto, A. V., Ferrari, F. C., Offutt, J., Fernandes, L. & Ribeiro, M. (2019). A systematic literature review of techniques and metrics to reduce the cost of mutation testing. *Journal*

of Systems and Software, *157*, 110388.

Plisson, J., Lavrac, N., Mladenic, D. *et al.* (2004). A rule based approach to word lemmatization. In *Proceedings of IS*, Vol. 3, 83–86.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K. & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews*. A product from the ESRC Methods Programme.

Popoff, E., Besada, M., Jansen, J., Cope, S. & Kanters, S. (2020). Aligning text mining and machine learning algorithms with best practices for study selection in systematic literature reviews. *Systematic Reviews*, *9*, 1–12.

Przybyła, P., Brockmeier, A., Kontonatsios, G., Le Pogam, M.-A., McNaught, J., Elm, E., Nolan, K. & Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, 470–488.

Python Team (2023). *Urllib package documentation*. https://docs.python.org/3/library/urllib.html. Online; accessed 19 March 2023.

Qureshi, R., Shaughnessy, D., Gill, K., Robinson, K., Li, T. & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic reviews*, p. 72.

Rathbone, J., Hoffmann, T. & Glasziou, P. (2015). Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic reviews*, *4*, 80.

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. In *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39. IEEE Computer Society.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

ResearchGate (2023). https://www.researchgate.net. Online; accessed 19 March 2023.

Riaz, M. (2012). Maintainability prediction of relational database-driven applications: a systematic review. In *16th International Conference on Evaluation & Assessment in Software Engineering (EASE)*, p. 263–272. IET - The Institute of Engineering and Technology/IEEE Xplore.

Rico, S., Ali, N., Engström, E. & Höst, M. (2020). *Guidelines for conducting interactive rapid reviews in software engineering – from a focus on technology transfer to knowledge exchange*. Lund University.

Rizzo, G., Vetro, A., Ardito, L., Torchiano, M. & Troncy, R. (2017). Semantic Enrichment for Recommendation of Primary Studies in a Systematic Literature Review. *Digital Scholarship in the Humanities*, *32*, 195–208.

Ros, R., Bjarnason, E. & Runeson, P. (2017). A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, p. 118–127. Association for Computing Machinery.

Runeson, P., Host, M. & Rainer, A. (2012). *Case Study Research in Software Engineering: Guidelines and Examples*. Wiley.

Saha, T. K., Ouzzani, M., Hammady, H. M. & Elmagarmid, A. K. (2016). A large scale study of SVM based methods for abstract screening in systematic reviews. *CoRR*, *abs/1610.00192*.

Saldana, J. (2012). *The Coding Manual for Qualitative Researchers*. English short title catalogue Eighteenth Century collection. SAGE Publications.

Scikit-Learn (2023). *Scikit-Learn Documentation*. https://scikit-learn.org/stable. Online; accessed 19 March 2023.

Sellak, H., Ouhbi, B. & Frikh, B. (2015). Using Rule-Based Classifiers in Systematic Reviews: A Semantic Class Association Rules Approach. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services,*

(iiWAS), 43:1–43:5. Association for Computing Machinery.

Semantic Scholar (2023). *Semantic scholar - academic graph API*. https://api.semanticscholar.org/api-docs. Online; accessed 19 March 2023.

Shekelle, P., Newberry, S., Wu, H. K., Suttorp, M. J., Motala, A., Lim, Y. W., Balk, E. M., Chung, M., Yu, W. W., Lee, J., Gaylor, J. M., Moher, D., Ansari, M. T., Skidmore, R. & Garritty, C. M. (2011). *Identifying Signals for Updating Systematic Reviews: A Comparison of Two Methods*. Agency for Healthcare Research and Quality (US).

Shojania, K., Sampson, M., Ansari, M., Ji, J., Doucette, S. & Moher, D. (2007). How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis. *Annals of internal medicine*, *147*, 224–33.

Silva, F., Cruz, S., Gouveia, T. & Capretz, L. (2013). Using Meta-ethnography to Synthesize Research: A Worked Example of the Relations between Personality and Software Team Processes. In *7th International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 153–162. Association for Computing Machinery.

Silva, G., Santos Neto, P., Santos Moura, R., l. C. Araújo, Cury da Costa Castro, O. & Ibiapina, I. (2019). An Approach to Support the Selection of Relevant Studies in Systematic Review and Systematic Mappings. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 824–829. IEEE Computer Society.

Simmonds, M., Elliott, J. H., Synnot, A. & Turner, T. (2022). Living Systematic Reviews. In E. Evangelou & A. A. Veroniki (dir.). *Meta-Research: Methods and Protocols* (121–134). Springer.

Souza, F. C., Santos, A., Andrade, S., Durelli, R., Durelli, V. & Oliveira, R. (2017). Automating Search Strings for Secondary Studies. *Information Technology - New Generations*, *558*, 839–848.

Stol, K.-J. & Fitzgerald, B. (2015). A Holistic Overview of Software Engineering Research Strategies. In *3rd IEEE/ACM International Workshop on Conducting Empirical Studies in Industry (CESI)*, p. 47–54. IEEE Computer Society.

Sulayman, M. & Mendes, E. (2011). An extended systematic review of software process Improvement in small and medium web companies. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, Vol. 2011, 134–143. IET - The Institute of Engineering and Technology/IEEE Xplore.

Timsina, P., Liu, J. & El-Gayar, O. (2015). Advanced analytics for the automation of medical systematic reviews. *Information Systems Frontiers*, *18*, 237–252.

Timsina, P., Liu, J., El-Gayar, O. & Shang, Y. (2016). Using Semi-Supervised Learning for the Creation of Medical Systematic Review: An Exploratory Analysis. In *49th Hawaii International Conference on System Sciences (HICSS)*, 1195–1203. IEEE Computer Society.

Tomassetti, F., Rizzo, G., Vetro, A., Ardito, L., Torchiano, M. & Morisio, M. (2011). Linked data approach for selection process automation in systematic reviews. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 31–35. Association for Computing Machinery.

Tsafnat, G., Glasziou, P., Karystianis, G. & Coiera, E. (2018). Automated screening of research studies for systematic reviews using study characteristics. *Systematic Reviews*, *7*, 1–9.

Vallon, R., da Silva Estácio, B. J., Prikladnicki, R. & Grechenig, T. (2018). Systematic literature review on agile practices in global software development. *Information and Software Technology*, *96*, 161–180.

Vallon, R., da Silva Estácio, B. J., Prikladnicki, R. & Grechenig, T. (2018). Systematic literature review on agile practices in global software development. *Information and Software Technology*, *96*, 161–180.

Wagner, S., Méndez, D., Felderer, M., Graziotin, D. & Kalinowski, M. (2020). Challenges in Survey Research. In M. Felderer & G. H. Travassos (dir.), *Contemporary Empirical Methods in Software Engineering* 93–125. Springer.

Wallace, B., Trikalinos, T., Lau, J., Brodley, C. & Schmid, C. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, *11*, 55.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 977–984. Association for Computing Machinery.

Wang, S., Scells, H., Koopman, B. & Zuccon, G. (2023). Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? *CoRR*, *abs/2302.03495*.

Watanabe, W. M., Felizardo, K. R., Candido, A., de Souza, E. F., ao Ede de Campos Neto, J. & Vijaykumar, N. L. (2020). Reducing efforts of software engineering systematic literature reviews updates using text classification. *Information and Software Technology*, *128*, 106395.

Wohlin, C. (2014). A Snowballing Procedure for Systematic Literature Studies and a Replication. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 321–330.

Wohlin, C. (2016). Second-generation Systematic Literature Studies Using Snowballing. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 15:1–15:6. Association for Computing Machinery.

Wohlin, C., Kalinowski, M., Romero Felizardo, K. & Mendes, E. (2022). Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology*, *147*, 106908.

Wohlin, C., Mendes, E., Felizardo, K. R. & Kalinowski, M. (2020). Guidelines for the search strategy to update systematic literature reviews in software engineering. *Information and Software Technology*, *127*, 106366.

Wohlin, C. & Rainer, A. (2022). Is it a case study? A critical analysis and guidance. *Journal of Systems and Software*, *192*, 111395.

Xiong, Z., Liu, T., Tse, G., Gong, M., Gladding, P., Smaill, B., Stiles, M., Gillis, A. & Zhao, J. (2018). A Machine Learning Aided Systematic Review and Meta-Analysis of the Relative Risk of Atrial Fibrillation in Patients With Diabetes Mellitus. *Frontiers in Physiology*, *9*, 835.

Yu, Z., Kraft, N. & Menzies, T. (2018). Finding better active learners for faster literature reviews. *Empirical Software Engineering*, *23*, 3161–3186.

Yu, Z. & Menzies, T. (2018). FAST2: an Intelligent Assistant for Finding Relevant Papers. *Expert Systems with Applications*, *120*, 57–71.

Zack, M. & Serino, M. (2000). *Knowledge, GroupWare, and the Internet*. Knowledge Reader Series. Taylor & Francis Group.

Zhang, L., Tian, J.-H., Jiang, J., Liu, Y., Pu, M.-Y. & Yue, T. (2018). Empirical Research in Software Engineering — A Literature Survey. *Journal of Computer Science and Technology*, *33*, 876–899.

Zhou, Y., Zhang, H., Huang, X., Yang, S., Babar, M. A. & Tang, H. (2015). Quality Assessment of Systematic Reviews in Software Engineering: A Tertiary Study. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 1–14. Association for Computing Machinery.

Zoph, B., Raffel, C., Schuurmans, D., Yogatama, D., Zhou, D., Metzler, D., Chi, E. H., Wei, J., Dean, J., Fedus, L. B., Bosma, M. P., Vinyals, O., Liang, P., Borgeaud, S., Hashimoto, T. B. & Tay, Y. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*, 1–30.

APPENDIX A

ETHICS CERTIFICATION

This thesis has undergone ethical certification. The certificate number is 2021-731.