UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FERNANDO KENJI KAMEI

Understanding and Supporting the Decision-making Whether to Use Grey Literature in Software Engineering Research

Recife

2022

FERNANDO KENJI KAMEI

Understanding and Supporting the Decision-making Whether to Use Grey Literature in Software Engineering Research

A Ph.D. Thesis presented to the Center for Informatics of the Federal University of Pernambuco in partial fulfillment of the requirements for the degree of Philosophy Doctor in Computer Science.

Concentration Area: Software Engineering

Advisor: Sérgio Castelo Branco Soares

Co-advisor: Gustavo Henrique Lima Pinto

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

K15u Kamei, Fernando Kenji
Understanding and supporting the decision-making whether to use grey literature in software engineering research / Fernando Kenji Kamei. – 2022.
275 f.: il., fig., tab.

Orientador: Sérgio Castelo Branco Soares.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
.
Inclui referências e apêndices.

1. Engenharia de software. 2. Literatura cinza. I. Soares, Sérgio Castelo Branco (orientador). II. Título.

005.1 CDD (23. ed.) UFPE - CCEN 2022-116

Fernando Kenji Kamei

"Understanding and Supporting the Decision-making Whether to Use Grey Literature in Software Engineering Research"

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Engenharia de Software e Linguagens de Programação.

Aprovado em: 22/03/2022.

_____

Orientador: Prof. Dr. Sérgio Castelo Branco Soares

BANCA EXAMINADORA

_____

Prof. Dr. André Luís de Medeiros Santos
Centro de Informática /UFPE

_____

Prof. Dr. Fabio Queda Bueno da Silva
Centro de Informática /UFPE

_____

Prof. Dr. Marcos Kalinowski
Departamento de Informática /PUC-RJ

_____

Prof. Dra. Tayana Uchôa Conte
Instituto de Computação/ UFAM

_____

Prof. Dra. Katia Romero Felizardo Scannavino
Departamento de Computação/ UTFPR

To my family, that is my base, my everything, without whom none of this would be possible.

ACKNOWLEDGEMENTS

Sorry the reader, but this part was written in Portuguese.

É uma frase clichê, mas sim, esta tese foi uma construção e conquista coletiva. Primeiramente agradeço a minha família, em especial aos meus pais, pois tudo o que consegui conquistar até aqui, foi fruto de muito amor e dedicação deles. Agradeço imensamente à minha querida e amada esposa (Mozão), por todo amor, suporte, e compreensão em todos os momentos de ansiedade, medo, estresse ao longo dessa jornada. Sou muito grato por ter você em minha vida.

Aos meus orientadores, Sérgio Soares e Gustavo Pinto, agradeço imensamente por todo o apoio, dedicação, e paciência, e compreensão nesta jornada. Obrigado pelas palavras de encorajamento e motivação, sempre mostrando que seria possível e ia dar certo. Obrigado por mostrarem o caminho da pesquisa em engenharia de software (aprendi muito com vocês), e por me acalmarem quando precisei.

Aos meus irmãos e irmã, por sempre estarem na torcida por mim. Às minhas sobrinhas Lilica e Isinha, vocês nem imaginam o quanto minha vida é mais feliz e leve com vocês. À minha família por parte da minha esposa, que estiveram sempre na torcida, em especial à minha sogra, João (cunhado) e Hélcia (concunhada). Aos meus afilhados, Mary, Maryel, e Nati, que em tão pouco trouxeram tanta luz e amor.

Aos meus amigos(as) e colegas pesquisadores(as), Igor Wiese, Waldemar Ferreira, Vilmar Nepomuceno, Márcio Ribeiro, Crescencio Lima, Ivanilton Polato, Carolline Pena, César França, e a Profa. Renata Souza, meus sinceros agradecimentos. Vocês foram imprescindíveis. Gratidão especial ao Igor Wiese, por sua amizade e por todas as conversas e conselhos nesta jornada. Que sorte tive de ter o apoio de vocês!

Agradeço também aos colegas da CINFO (IFAL) pelo apoio que recebi desde o início até o final do doutorado. Aos pesquisadores dos grupos ESEG e SPG (CIn-UFPE), Meme-oriented advising (UFPA), e EASY (UFAL), por todas as discussões, colaborações e aprendizados.

Por fim, agradeço aos revisores anônimos dos artigos que submeti. Aos membros do comitê de avaliação do IDoESE (2019) e do WTDSoft (2019), pois as críticas e discussões foram extremamente importantes para o rumo que esta tese tomou. Aos pesquisadores(as) que disponibilizaram o seu tempo participando das minhas pesquisas. E, aos membros da banca de defesa por todas as críticas, revisões, e sugestões para a melhoria da tese.

ABSTRACT

Recently, there has been growing interest in exploring Grey Literature (GL) in Software Engineering (SE) research. However, researchers are skeptical of its use due to credibility concerns. Still, the area lacks investigation to support its use in Secondary Studies (SSs). Therefore, in this thesis, we focus on improving GL's understanding and propose strategies to support its use in SSs. For this, we first interviewed 76 Brazilian SE researchers to understand their perceptions and experiences about GL. We identified several reasons to use and avoid, beyond the benefits and limitations of its use, and criteria for assessing GL credibility. Afterward, we investigated 34 Brazilian researchers to understand their experience assessing the credibility of the different GL types. Second, based on a tertiary study, we analyzed the use of GL in 126 SSs to present trends, concepts employed, GL types used, and methods used to select and perform quality assessment of GL. Our findings showed a recent increase in GL use. We did not identify a common definition of GL, and there are different interpretations of the GL types among the investigated studies. Interestingly, only about 1/4 of the LC fonts are no longer available. Third, we conducted another tertiary study, where we assessed how GL use contributed to nine Multivocal Literature Reviews. We noted that GL contributed to these studies, mainly providing recommendations and explaining topics. In addition, we perceived that some information would not have been presented in the studies if the studies did not consider GL. Fourth, we conducted two focus groups with ten SE researchers to assess Garousi's guidelines. The investigated researchers generally considered the guidelines helpful, although many issues have been noted. We provide recommendations to address these issues and improve guidelines. Finally, based on these four studies, we proposed two processes. One is a decision-making process to help SE researchers decide whether to include GL in a SS. The other process with recommendations to improve the conduction of SSs with GL. To conclude, GL showed to be an important source for SE research. However, there are challenges that researchers may face in using it. We endeavored to mitigate these challenges by providing criteria to assess GL credibility and recommendations for researchers to better deal with its use.

Keywords: grey literature; multivocal literature review; software engineering.

RESUMO

Recentemente, há um interesse crescente em explorar a Literatura Cinza (LC) na pesquisa de Engenharia de Software (ES). No entanto, existem pesquisadores céticos ao seu uso, devido às preocupações de credibilidade. Ainda, a área carece de investigações para apoiar o seu uso em Estudos Secundários (ESSs). Portanto, nesta tese, focamos em melhorar a compreensão sobre a LC e propomos estratégias para apoiar seu uso em ESSs. Para isso, primeiramente entrevistamos 76 pesquisadores brasileiros de ES para entender suas percepções e experiências sobre LC. Identificamos diversos motivos para usar, razões para evitar, benefícios e limitações de seu uso, e critérios para avaliar a credibilidade da LC. Em seguida, investigamos 34 pesquisadores brasileiros de ES, para entender como estes avaliam a credibilidade dos diferentes tipos de LC. Em segundo lugar, com base em um estudo terciário, analisamos 126 ESSs para apresentar tendências, conceitos empregados, tipos de LC e métodos usados para selecionar e realizar a avaliação da qualidade da LC. Nossas descobertas mostraram um recente crescimento no uso de LC. Não identificamos uma definição comum de LC, e há diferentes interpretações de seus tipos entre os estudos investigados. Curiosamente, apenas cerca de $1/4$ das fontes de LC não estão mais disponíveis. Terceiro, conduzimos outro estudo terciário, onde avaliamos nove Revisões Multivocais da Literatura. Identificamos que o uso da LC contribuiu com esses estudos, principalmente fornecendo recomendações e explicando tópicos. Além disso, encontramos informações que não seriam encontradas nos estudos se a LC não fosse considerada. Quarto, realizamos dois grupos focais envolvendo dez pesquisadores de ES para avaliar as diretrizes de Garousi. Em geral, as diretrizes são úteis, embora vários problemas tenham sido percebidos. Fornecemos recomendações para lidar com esses problemas e melhorar as diretrizes. Por fim, com base na condução desses estudos, propusemos dois processos. Um processo para apoiar a tomada de decisão dos pesquisadores para avaliar se devem incluir LC em um ESS. E outro com recomendações para melhorar a condução de ESSs com o uso de LC. Para concluir, a LC mostrou-se importante para a pesquisa de ES. No entanto, existem desafios que os pesquisadores podem enfrentar. Nos esforçamos para mitigar esses desafios, como, por exemplo, fornecendo critérios para avaliar sua credibilidade e recomendações para os pesquisadores lidarem melhor com o seu uso.

Palavras-chaves: literatura cinza; revisão multivocal da literatura; engenharia de software.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

EBSE	Evidence-Based Software Engineering
GL	Grey Literature
GLR	Grey Literature Review
MLR	Multivocal Literature Review
MS	Mapping Study
SE	Software Engineering
SLR	Systematic Literature Review
TL	Traditional Literature

CONTENTS

1 INTRODUCTION

Evidence-Based Software Engineering (EBSE) is an approach of Software Engineering (SE) research to help practitioners improving their decision-making practices in their work environment (KITCHENHAM; DYBÅ; JØRGENSEN, 2004). According to Kitchenham et al. (KITCHENHAM; DYBÅ; JØRGENSEN, 2004), the goal of EBSE is *"to provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision-making process regarding the development and maintenance of software."* As we can note, *evidence* is the basis for decision-making (WOHLIN, 2013) and for EBSE.

According to Schum (SCHUM, 2001), it is certainly true that the value of the evidence to build knowledge differs from one context to another. That is, evidence is context-dependent. In this regarding, Wohlin (WOHLIN, 2013) highlighted that something that may be perceived as relevant evidence in research, may not be as highly valued in another context. It is supported by the Marculescu et al.'s (MARCULESCU; JABBARI; MOLLÉRI, 2016) investigation, that perceived that academia focuses on collecting evidence and generalizing from evidence to build theories. On the other hand, industry focuses to find and apply knowledge that is applicable in their context and to provide solutions for those problems. Thus, it is common that researchers and practitioners feed from information of different sources or channels. While academics value formal information (e.g., conference or journal papers), practitioners tend to search for information using social media and communication channels (FIDEL; GREEN, 2004; YITZHAKI; HAMMERSHLAG, 2004; STOREY et al., 2017). The last are using these sources motivated mainly to exchange problems and acquire and share information more easily and quickly (STOREY et al., 2017). For this reason, over the last years, there is an increasing interest of SE researchers investigating its use. In SE, the data produced in these media are known as source of Grey Literature (GL) (RAINER, 2017).

Grey Literature was defined as *"Grey Literature is produced on all levels of government, academics, business, and industry in print and electronic formats, but which is not controlled by commercial publishers."* This definition is well-known as Luxembourg definition, proposed in Third International Conference on Grey Literature in 1997 (FARACE; SCHÖPFEL, 2010). According to Bonato (BONATO, 2018), GL could be interpreted in diverse manners that vary from area to area, despite this well-known definition. For Schöpfel and Prost (SCHöPFEL; PROST, 2020) the GL definition depends on the context and the people.

In SE research, *"GL can be defined as any material about SE that was not subject to a peer-reviewed process or nor-formally published"* (GAROUSI et al., 2020). As we can perceive, the concept of GL is quite broad. For this reason, Adams and colleagues (ADAMS et al., 2016), in the health area, introduced the idea of "Grey data" and "Grey information" to distinguish the different grey forms, including grey literature, grey information, and grey data. In SE, according to Williams and Rainer (WILLIAMS; RAINER, 2019), "grey data" means the content of tweets, blogs, and posts on Q&A websites, and by "grey information" is informally published or not published at all, e.g., meeting notes and emails. Other SE researchers share this vision, for instance, Garousi and colleagues (GAROUSI et al., 2020), and Zhang and colleagues (ZHANG et al., 2020).

The use and benefits of GL have been investigated in several areas of knowledge. For instance, in Medicine (PAEZ, 2017), Nutrition (ADAMS et al., 2016; GODIN et al., 2017), and Management (ADAMS; SMART; HUFF, 2017). As examples of these benefits, we can mention the gaining of significant knowledge obtained from practitioners in addition to academic articles, providing data not found within scientific and commercial literature (ADAMS; SMART; HUFF, 2017). Others are using GL to reduce the impact of publication bias, as studies with null or negative findings are less likely to be published in peer-reviewed journals (PAEZ, 2017) and also to address missing topics from conventional academic sources (GUL et al., 2020).

On the opposite side of GL is the Traditional Literature (TL), that is the main source used in Secondary Studies, mainly in terms of Systematic Reviews and Mapping Studies. The interest for these methods increased over the years. For instance, Zhang and Ali Babar (ZHANG; BABAR, 2013) found 148 SLRs published from January 2004 to December 2010. This growth shows the emerging character of the field. Nevertheless, the increasing interest in SLRs revealed important issues, for instance, the lack of connection between SE research and practice. In this regard, Da Silva and colleagues (SILVA et al., 2011) claimed that only a few SLRs provide guidelines to practitioners, and Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2016) observed that most results provided by SLRs were focused on academic needs, meaning the practitioner's voice is limited.

Focusing on investigating how to approximate the research from the practice, Williams investigated how to use blogs of SE practitioners (one type of GL) as a source of evidence for Secondary Studies (WILLIAMS, 2018). In this study, Williams claimed that the evidence retrieved from GL could create opportunities to mitigate SLR issues related to the lack of connection between the Secondary Studies and the current state of the SE practice. Garousi

and colleagues (GAROUSI; FELDERER; MÄNTYLÄ, 2016) also investigated those gaps, exploring the impact on the results if Secondary Studies did not use GL, showing that several pieces of information could be missed if we did not consider GL. Considering the use of GL in Secondary Studies, previous investigations perceived growth in its use over the last years (YASIN et al., 2020; ZHANG et al., 2020). More specifically, the studies that have the intention to search for GL, such as Multivocal Literature Reviews (MLR) and Grey Literature Reviews (GLR) (GALINDO NETO et al., 2019; ZHANG et al., 2020). Focusing on supporting the use of GL in Secondary Studies of SE, Garousi and colleagues (GAROUSI; FELDERER; MÄNTYLÄ, 2019) proposed a set of guidelines. Despite this growing interest, previous literature identified gaps, as discussed in the following section.

We highlight that in our conception, despite GL could contribute aggregating value to certain situations (as we described in this thesis), we value the importance of scientific epistemology to knowledge evolution. For this reason, we believe that the context of the research is important to analyzed to decide whether to use or avoid GL, as we cleared in this thesis.

1.1 PROBLEM STATEMENT AND MOTIVATION

Based on the previous literature, we guided this research to investigate the following problems:

Problem 1: *Despite the increase use of GL in SE research, as its use and investigations are recent, Zhang and colleagues (ZHANG et al., 2020) and Rainer and Williams (RAINER; WILLIAMS, 2018b) claimed for more investigations to clear the importance and improve the use of GL in SE research.*

Problem 2: *According to Flynn and Williams (FLYNN; WILLIAMS, 1998), 'Decision-making' is usually defined as the act of choosing between alternative courses of action. Aurum and Wohlin (AURUM; WOHLIN, 2002) pointed that for effective decision-making, it is crucial that decision-makers select the "best" course of action based on the information available at the time. In our context, there is a lack of support for the decision-making of SE researchers on whether to include GL in Secondary Studies, considering the researcher's knowledge about GL and its sources for the investigated topic and how this inclusion could impact the research.*

Problem 3: *Despite the SE research community's growing interest in GL, the main concern about its use is related to the lack of its credibility. Although studies have previously investigated criteria to assess GL credibility (WILLIAMS, 2019), they were focused on assessing blog posts content. In this regard, Williams (WILLIAMS, 2019) pointed out the importance to investigate credibility criteria that cover other GL types. For this reason, there is a lack of understanding of these criteria and how to assess the diversity of GL types.*

Problem 4: *Although previous studies showed that GL has been used in SE Secondary Studies (YASIN et al., 2020; ZHANG et al., 2020), and more specifically with the increase in studies based on the MLR (GALINDO NETO et al., 2019), there is a lack of investigations showing how GL properly contributed to MLR studies (we identified only the study of Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2016) in this regard).*

Problem 5: *Despite Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019) being one of the most used guidelines in SE research to conduct MLR and GLR, we did not find any in-depth investigations evaluating them, although, Kitchenham and colleagues (KITCHENHAM et al., 2008) claimed the importance of assessing the guidelines before widely being adopted. In this regard, we identified some studies that presented an experience report about processes, challenges, and recommendations to deal with a GLR (WEN et al., 2020; ZHANG et al., 2021; MELEGATI; GUERRA; WANG, 2021). However, they did not assess all guidelines and did not present a broad and in-depth view because they were based only on the viewpoints of authors' perspectives of a single study.*

1.2 OBJECTIVE

The general goal of this thesis is twofold. First, we aim to create a knowledge base about GL from the perspective of SE researchers and previous literature. Second, to design two decision-making processes to support SE researchers. One of these processes is to support the decision on whether to use GL, and the other is with recommendations to improve the conduction of Secondary Studies that intend to search and include GL. Table 1 summarize our research goal using a Goal Question Metric (GQM) template.

Table 1 – GQM related to our goal.

Analyze	Object under measurement
For the purpose of	*improving the understanding and use of GL in SE research.*
With respect to	*the perceptions to decide whether to use GL and strategies to support its use.*
From the point of view of	*the SE research community and the previous literature that investigated GL.*
In the context of	*the use of GL in SE research.*

Source: the author.

1.3 RESEARCH QUESTIONS

To achieve the stated goal, this thesis proposes to answer a General Research Question (GRQ).

GRQ: *What is the Role of Grey Literature in Software Engineering Research?*

This general research question intends to guide our investigation to understand better Grey Literature use in SE research, support SE researchers in the decision-making on whether to include GL, and support its use when SE researchers decide to use it in Secondary Studies. This question is important to provide an overview about GL in SE research.

In what follows, we describe and present the motivation to our Research Questions (RQs):

RQ1: *What Are the Perceptions of Software Engineering Researchers About Grey Literature?*

According to Zhang and colleagues (ZHANG et al., 2020), GL is a theme that needs more investigations to be mature and prove its importance to contribute to SE research. For this reason, in this question, we investigated SE researchers to understand: (i) What motivates to use GL or the reasons not to use it; (ii) What types of GL are mostly used; and (iii) The criteria employed to assess GL credibility; (iv) The potential benefits and challenges of its use. This investigation is important to the SE research community to improve the understanding of how researchers could take advantage of GL, using it in SE research with more credibility.

In our investigation, we used the meaning of *credibility* as the quality of being trusted and believed in[1], showing that it is a subjective concept often defined in terms of a list of criteria for a particular group, in our case, a group of SE researchers.

RQ2: *How do Secondary Studies Use Grey Literature?*

Even with GL's growth in interest in SE, there is a lack of understanding how GL is used in Secondary Studies and what researchers reported in their studies about its use. For this reason, in this research question we investigated: (i) The usage of GL in Secondary Studies of SE; (ii) How they are used; (iii) If the use of GL has increased in Secondary Studies over the years; (iv) What types of GL were most used, and who are their producers; (v) The methods employed to search, select, and perform a quality assessment on GL; (vi) The definitions employed about GL; and (vii) The motivations that SE researchers mentioned to use or avoid GL. This research question is essential to understanding the current landscape of GL in Secondary Studies.

RQ3: *How the Grey Literature Use Contributed to Multivocal Literature Review Studies?*

Recently, Zhang and colleagues (ZHANG et al., 2020) showed a growing increase in Secondary Studies using GL over the years in SE, especially the Multivocal Literature Reviews (MLR) studies. However, there is little evidence assessing to what extent GL sources contribute to the MLR studies' findings, despite its interest. Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2016) investigated what is gained when considering GL in an MLR study. However, several MLRs have been published since Garousi et al.'s study, and no other research has deeply investigated how GL affected the MLR studies' synthesis. This lack of understanding could make SE researchers skeptical about using GL and conducting an MLR study since the search for GL increases the effort compared with traditional Secondary Studies (RAULAMO-JURVANEN et al., 2017). Thus, answering this research question is important to understand how GL could contribute to MLR studies could increase researchers' confidence in using it.

RQ4: *What Are the Perceptions of Software Engineering Researchers About Garousi's Guidelines?*

[1] Oxford definition

Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019) proposed a set of guidelines for including GL and conducting an MLR in SE. It is the unique and well-known guideline used by SE researchers who intend to conduct GLR or MLR. Despite the importance of this guideline, there is no evidence assessing its use until now. For Kitchenham and colleagues (KITCHENHAM et al., 2008), it is essential to evaluate the guidelines as soon as possible, avoiding causing problems of poor quality reporting with the studies that followed it. For this reason, with this question, we focused on assessing Garousi's guidelines to identify the researcher's perceptions about its usefulness, issues, and points for improvement.

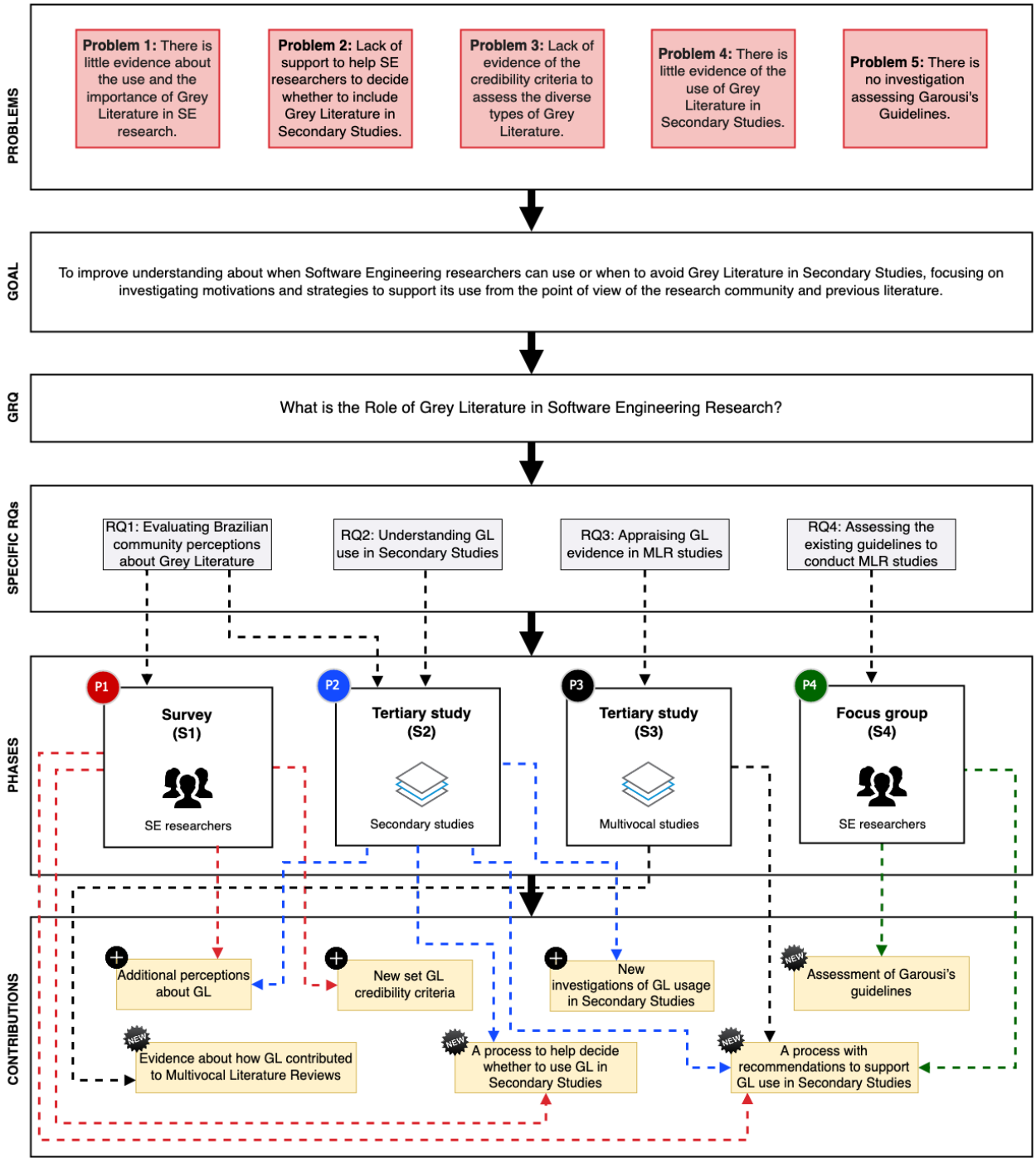
1.4 RESEARCH METHODS OVERVIEW

This thesis followed the process presented in Figure 1. The process started investigating *previous literature*, where we identified four **problems**, previously presented in Section 1.1 and summarized in this figure. After, we proposed our **objectives**, that in general, focused on improving the state-of-art to present better definitions when to use or avoid GL and strategies to support its use in Secondary Studies. To achieve this objective, we proposed a general research question (**GRQ**), that was answered with discussions based on the investigations of the entire thesis (see Chapter 7). To answer the GRQ, we proposed four specific research questions (**RQ1–RQ4**). Finally, to answer these questions, we conducted four phases in this thesis (**P1–P4**) by conducting various empirical studies, adopting a mixed-methods strategy. Each phase is related to specifics RQs.

In the ***first phase*** **(P1)**, presented in Chapter 3, we conducted a *survey* (**S1**) based on Linåker's guidelines (LINÅKER et al., 2015) with the objective to conduct explanatory research, focusing on answering the **RQ1**. We investigated Brazilian SE researchers to provide their viewpoints about GL use, mainly related to their benefits, challenges, motivations to use or avoid, the criteria used to assess the GL credibility, and the perceptions about the different GL types concerning the outlet control and source expertise of its production.

In the ***second phase*** **(P2)**, presented in Chapter 4, we performed a *tertiary study* (**S2**), following Kitchenham et al.'s guidelines (KITCHENHAM; BUDGEN; BRERETON, 2015) with the objective to perform an exploratory study and conduct a critical review of Secondary Studies, seeking to provide answers to **RQ2**. We searched for Secondary Studies in the top SE conferences and journals, using automatic and manual search strategies. In these studies, we investigated the usage of GL and what the SE researchers reported about the benefits and

Figure 1 – Overview of the process followed and the contributions of this thesis.



Source: the author.

The process presents the problems and the objective of the thesis's investigation, followed by the general research question (GRQ), which are answered by four specific research questions (RQ1–RQ4). These questions are answered along with the four phases (P1–P4, each one with a specific color). Above each phase, there is a description of the research method employed with the identification number for each investigation and, at the bottom, the focus of each phase. Below the phases, there are the main contributions of this thesis. Dotted lines are linking the phases with the contributions. By the signal "+" indicates that additional contributions were provided compared with previous studies and "New" indicates new contributions not previously explored.

challenges of using GL in the studies. In addition, we identified challenges in conducting our investigation and provided potential recommendations or solutions to deal with each one.

In the ***third phase*** **(P3)**, presented in Chapter 5, we conducted another *tertiary study* (**S3**), this time to perform explanatory research by investigating MLR studies that followed Garousi's guidelines (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019), focusing on assessing how the use of GL contributed to MLR studies, seeking to provide an answer to **RQ3**. By contributing, we mean understanding to what extent GL provides information that an MLR indeed uses to answer its research questions.

Finally, in the ***fourth phase*** **(P4)**, presented in Chapter 6, we performed two *focus groups* (**S4**) based on the procedures proposed by Kontio et al. (KONTIO; BRAGGE; LEHTOLA, 2008), to identify insightful perceptions of Garousi's guidelines from the researchers interactions, aiming to provide answers to **RQ4**. This investigation focused on understanding the perceptions of experienced SE researchers in conducting Secondary Studies and of a group of researchers that conducted MLR or GLR adopting Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019).

We summarized our main contributions in Figure 1 below the presented phases, as shown in each yellow box. The contributions represented with the signal "+" indicate that we provided additional contributions to what has been done in previous studies, and "NEW" indicates our contributions that are new to the literature. In general, the most significant contributions of the phases conducted were to serve as input to our proposed processes *to help SE researchers decide whether to include GL in Secondary Studies* and *support with recommendations to use GL with more credibility in Secondary Studies*. In the following section, we briefly discussed each contribution.

1.5 MAIN CONTRIBUTIONS

This work presents the following contributions:

- **Perceptions about GL.** We conducted the first investigation focusing on Brazilian SE researchers and the first investigating the motivations to use and reasons to avoid GL. We also identified additional benefits and challenges of GL use and confirmed some findings of previous studies. In addition, we identified to what purpose GL could be used or when it should be avoided in SE research.

- **Assessing GL credibility.** We explored how SE researchers assess the GL source's

credibility. We identified a set of criteria, confirming some findings of previous studies, and identified news one. Furthermore, we were the first exploring how the different GL types were perceived according to the outlet control and expertise levels;

- **Investigations of GL usage in Secondary Studies.** We investigated 446 Secondary Studies, identifying 126 that used GL. We provided an overview of GL use in these studies, for instance, showing that GL is not extensively used, and almost half of the GL used are unavailable;
- **Additional evidence of the contributions of GL to Multivocal Literature Reviews.** Exploring nine MLR studies, we perceived that several findings were exclusively retrieved from GL sources. The main contributions of GL were related to providing explanations and recommendations about a topic. We were the first classifying the type of contribution;
- **Perceptions and points of improvement of Garousi's guidelines.** We conducted the first broad assessment of Garousi et al.'s guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019). We identified that, in general, these guidelines are useful to support SE researchers, but we also identified some points that need a particular attention. Based on the researchers' opinions, our investigations, and previous literature, we provided a set of recommendations to deal with the identified issues;
- **A process to decide whether to use GL in Secondary Studies.** Based on our investigations, we designed the first process to help SE researchers to think about the benefits, challenges, motivations to use, and reasons to avoid and decide whether to use GL in their Secondary Studies.
- **A process to support GL use in Secondary Studies**. Based on the challenges and recommendations to better use GL in Secondary Studies identified in our investigations and in previous literature, we designed a process to support the conduction of Secondary Studies using GL and avoid the identified challenges.

1.6 OTHER CONTRIBUTIONS

In the following, we present the contributions that resulted in scientific publications and partnerships established related to the Ph.D.

Publications Related to this Thesis

In the following, we listed the works published/accepted related to this thesis.

1. (Paper 1) A full paper at the 34th Brazilian Symposium on Software Engineering (SBES 2020), reporting the results of the first survey of Phase 1 (KAMEI et al., 2020). This paper was one of the Best Paper Awards.

2. (Paper 2) A full paper at the Journal of Software Engineering Research and Development (JSERD), reporting the results of the second survey of Phase 1 (KAMEI et al., 2022). This paper extends Paper 1. This extended version provides the perceptions of SE researchers about the control and source expertise criteria of the different GL types.

3. (Paper 3) A full paper at the Information and Software Technology Journal (IST), reporting the Phase 2 results (KAMEI et al., 2021b).

4. (Paper 4) A full paper at 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2021), reporting the Phase 3 results (KAMEI et al., 2021a).

We also participated in two doctoral symposiums/workshops, receiving important feedback on this thesis.

1. Participation at the International Doctoral Symposium on Empirical Software Engineering (IDoESE 2019) (KAMEI, 2019) and at the Workshop on Theses and Dissertations of CBSoft (WTDSoft 2019) (KAMEI; SOARES; PINTO, 2019) proposing the goal and method of this research, as well as preliminary results we obtained with our tertiary study.

1.7 DOCUMENT STRUCTURE

The remainder of this work is organized as follows:

- **Chapter 2** presents the theoretical foundations and essential concepts of GL in SE.

- **Chapter 3** exposes the perceptions (motivations to use or avoid and the perceived benefits and challenges with their use) about GL and identifies criteria used by Brazilian SE researchers to assess GL credibility.

- **Chapter 4** provides a critical investigation about the usage of GL in Secondary Studies (e.g., GL types frequently used, the extent of GL included in the Secondary Studies, and GL definition employed).
- **Chapter 5** presents a critical review of how the use of GL contributed to MLR studies. Moreover, this chapter provides a process used to conduct this investigation and presents challenges identified by investigating GL contributions in MLR studies.
- **Chapter 6** provides an investigation with SE researchers, showing their perceptions of usefulness, issues, and points for improvement of Garousi's guidelines.
- **Chapter 7** discusses and relates our main findings to propose a process to help SE researchers to decide whether to include or avoid GL in their Secondary Studies. Moreover, we designed an experience-based process through the conduction of this thesis, with recommendations to support the conduction of Secondary Studies using GL. Finally, we compared some of our findings.
- **Chapter 8** presents an overview of the related works compared with ours.
- **Chapter 9** exposes our main conclusions and direction for future works.

2 BACKGROUND

This chapter presents the main concepts related to this work. Section 2.1 provides a brief description and the importance of empirical studies in Software Engineering (SE). Section 2.3 shows a theoretical foundation about the Grey Literature (GL) and its use in a diverse area of knowledge. Section 2.4 explains the use of GL in SE research, initially presenting an overview, followed by addressing studies (primary, secondary, and tertiary studies) that explored the GL content to answer their research questions. Finally, Section 2.5 presents a summary of this chapter.

2.1 EMPIRICAL SOFTWARE ENGINEERING

Empirical Software Engineering (ESE) is a field of science that emphasizes the use of empirical methods in Software Engineering to study, evaluate, improve, and propose new technologies, practices, processes, and tools to support SE through qualitative or quantitative analysis (BASILI, 1996 apud FELDERER; TRAVASSOS, 2020).

The empirical methods could be classified according to terms used by Sjøberg et al. (SJØBERG; DYBÅ; JØRGENSEN, 2007), *primary research*, and *secondary research*. Kitchenham et al. (KITCHENHAM; BUDGEN; BRERETON, 2015) added another term, *tertiary research*. In the following, we present their descriptions according to Kitchenham et al. (KITCHENHAM; BUDGEN; BRERETON, 2015):

- *Primary research:* a study in which we directly make measurements about objects of interest using empirical methods (e.g., experiment, case study, survey);
- *Secondary research:* a study that analyses a set of primary studies and usually seeks to aggregate the results from these to provide stronger forms of evidence about a particular phenomenon;
- *Tertiary research:* a secondary study that uses the outputs of Secondary Studies as its inputs.

As our investigations focused on Secondary Studies of SE, in the following, we briefly describe Evidence-Based Software Engineering.

2.2 EVIDENCE-BASED SOFTWARE ENGINEERING

Kitchenham et al. (KITCHENHAM; DYBÅ; JØRGENSEN, 2004), inspired by the Evidence-Based Medicine (EBM) (GUYATT; CAIRNS; CHURCHILL, 1992), proposed the term Evidence-Based Software Engineering (EBSE), aiming *"to provide the means by which current best evidence from research can be integrated with practical experience"*, in other words, to identify and appraise all relevant evidence to the problem or technology under consideration.

The EBSE has some Secondary Studies methods (GAROUSI; FELDERER; MÄNTYLÄ, 2019), but the most common are Systematic Literature Review and Mapping Study (SILVA et al., 2011). In the following, we describe the Secondary Studies types according to Kitchenham (KITCHENHAM; BUDGEN; BRERETON, 2015):

- Systematic Literature Review (SLR) (or Systematic Review);
- Mapping Study (MS) (or Systematic Mapping Study);
- Meta-Analysis (MA);
- Grey Literature Review (GLR);
- Multivocal Literature Review (MLR).

Over the years, it is noticeable the increase in the number of Secondary Studies published in Software Engineering. For instance, Kitchenham et al.'s work (KITCHENHAM et al., 2009) identified 20 studies through conducting the first tertiary study in SE. One year later, this number increased to 33 studies (KITCHENHAM et al., 2010). In 2011, Da Silva and colleagues (SILVA et al., 2011) identified 67 more new Secondary Studies, totaling 120 studies. In 2019, Galindo Neto and colleagues (GALINDO NETO et al., 2019) identified 12 studies based on Grey Literature Reviews and Multivocal Literature Reviews.

As this thesis focused on GL, we present its concepts, its use, and investigations in SE studies.

2.3 GREY LITERATURE

The term Grey Literature has some definitions, but the most widely used and accepted is the so-called Luxembourg definition (GAROUSI; FELDERER; MÄNTYLÄ, 2019), approved at

the Third International Conference on Grey Literature in 1997: *"that which is produced on all levels of government, academics, business, and industry in print and electronic formats, but which is not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body"*. In summary, the term "grey" literature is often used to refer to literature that is not obtainable through normal publishing channels, which was not subject to quality control mechanisms (peer-review) before a publication (PETTICREW; ROBERTS, 2006).

GL has been investigated in several areas of knowledge, such as management (ADAMS; SMART; HUFF, 2017), medicine (PAEZ, 2017), and nutrition (ADAMS et al., 2016; GODIN et al., 2017), aiming to gain benefits from GL content, such as to provide data not found within scientific and commercial literature, to reduce publication bias in which more positive findings tend to be published in Traditional Literature (TL), and to facilitate a more balanced view of using multiple sources of evidence (PAEZ, 2017). Unlike the SE area, the use and investigations of GL in other disciplines are not recent. There are many databases, repositories, and search engines (e.g., GreyNet[1], OpenGrey[2], and The Grey Literature Report[3]) in those areas.

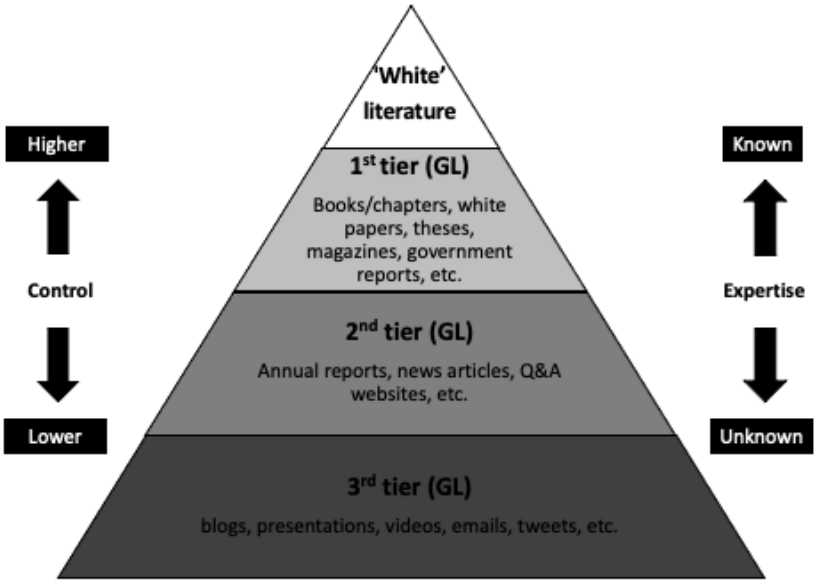
GL has a diversity of types that varies according to the discipline area (BONATO, 2018). This heterogeneity of GL is a specific problem that makes it less amenable to traditional forms of archiving, retrieval, analysis, synthesis, bibliographic data capture, data extraction, and integration (ADAMS; SMART; HUFF, 2017). Focusing on extending the concept of GL to a broader range of sources, researchers proposed the terms grey literature, grey information, and grey data to distinguish the different forms of grey (ADAMS et al., 2016). The term *"grey data"* describes user-generated web content, e.g., tweets and blogs (WILLIAMS; RAINER, 2019). On the other hand, *"grey information"* is information informally published or not published at all, e.g., meeting notes and emails (RAINER; WILLIAMS, 2018a). Nevertheless, SE literature hardly distinguishes these terms[4].

Aiming to classify the different GL types, Adams et al. (ADAMS; SMART; HUFF, 2017) proposed the "shades of GL" that classify GL into three tiers according to two dimensions: control and expertise. *Control* refers to the rigor that the source is produced (not related to the construct of the experimentation method), and *Expertise* to the producer's authority and knowledge. Both dimensions run between the extremes "unknown" and "known". The first tier considers GL with high expertise and high outlet control. The second tier considers

[1] www.greynet.org
[2] www.opengrey.eu
[3] www.greylit.org
[4] Similarly, in our work, we considered all forms of grey data and grey information as GL.

GL with moderate expertise and moderate control. Finally, the third tier considers GL with low expertise and low outlet control. In SE, Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019) classified and adapted these shades, as shown in Figure 2. This figure shows examples of GL sources used in SE to each tier. On the top of the pyramid is the "traditional literature" with scientific articles from conferences and journals. On the rest of the pyramid are what we called as three tiers of GL. The darker the color, the less moderated or edited the source in conformance with explicit and transparent knowledge creation criteria.

Figure 2 – The "shades" of Grey Literature for Software Engineering.



Source: adapted from Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019).

2.4 GREY LITERATURE IN SOFTWARE ENGINEERING

In the context of SE, interest in GL is recent and increased over recent years. In particular, due to the widespread presence of GL media used by SE professionals, including various communication channel types (STOREY et al., 2014). Rainer (RAINER, 2017) has pointed out that those media, which practitioners are naturally producing, could help researchers gain additional insights into the dynamics and challenges that occur during the software development process. It potentially explains why SE researchers have been paying attention to the potential of GL lately.

GL has been investigated in various studies types in SE. There are studies that feed on GL data, others investigated how researchers are using them, and others explored how to support SE researchers to use them.

As example of primary studies, there are that feed on GL, for instance, its relation to question and answer (Q&A) websites such as Stack Overflow (ZAHEDI; RAJAPAKSE; BABAR, 2020) and news aggregator websites, such as Reddit and Hacker News (ANICHE et al., 2018).

We also identified several Secondary Studies conducted using GL. For instance, we identified one SLR that investigated the use of Agile methods with CMMI and included technical reports as primary studies (SILVA et al., 2015), and a MS that explored knowledge related to software smells to identify challenges as well as opportunities and included books as primary study (SHARMA; SPINELLIS, 2018). We have other examples of Secondary Studies that were conducted with the intention to search for evidence on the content of GL. For example, the studies based on MLR (e.g., GAROUSI; FELDERER; MÄNTYLÄ (2016)) and GLR (e.g., RAULAMO-JURVANEN et al. (2017)). Beyond that, previous tertiary studies investigated GL in Secondary Studies (e.g., ZHANG et al. (2020) and YASIN et al. (2020)). These studies will be explained in the following sections.

In the following sections, we group the studies in terms of who explores how researchers are using GL to support their research (Section 2.4.1) and who shows how SE researchers are investigating GL (Section 2.4.2).

2.4.1 Using Grey Literature to Support Software Engineering Studies

SE researchers rely on several GL sources to help them answer their research questions. Examples include screencasts, videos on YouTube, Twitter posts, and Stack Overflow posts. This section briefly describes some studies that used some GL sources.

MacLeod and colleagues conducted two studies exploring the use of screencasts. The first study published in 2015 (MACLEOD; STOREY; BERGEN, 2015) investigated how and why developers create and share screencasts through YouTube. Motivations (e.g., learning, documenting their code, and self-promotion) and a diversity of goals and techniques for creating such screencasts (e.g., code demonstrations and describing code functionality in different ways) were identified. The second study published in 2017 (MACLEOD; BERGEN; STOREY, 2017) explored how screencasts are used for software documentation. The findings revealed video as a useful medium for developers communicating their program knowledge and building an online reputation.

Previous researchers investigated the use of Twitter in SE as an important communication channel for keeping up with new technologies and the fast-paced development land-

scape (STOREY et al., 2017; SINGER; FILHO; STOREY, 2014). Twitter was also associated with communicating issues, documentation, advertising blog posts to its community, and soliciting contributions from users (STOREY et al., 2014). In another example, Singer et al. (SINGER; FILHO; STOREY, 2014) investigated 271 developers who claimed that Twitter helps them keep up with the fast-paced development landscape.

SE researchers also investigated Questions and Answers (Q&A) websites. For instance, Zahedi et al. (ZAHEDI; RAJAPAKSE; BABAR, 2020) employed an empirical study to explore Continuous Software Engineering from the practitioners' perspective by analyzing 12,989 questions and answers from Stack Overflow. The findings presented trends (questions are becoming more specific to technologies and more difficult to attract answers) and the most challenging areas in this domain from the practitioners' perspective (variety of interpretations and understanding among researchers and practitioners).

2.4.2 Understanding How Software Engineering Researchers Are Investigating Grey Literature

This section presents studies that investigated the potential of GL for SE research. In the following, we briefly describe those studies in terms of primary, secondary, and tertiary studies.

Primary Studies

We explored three studies conducted by Williams and Rainer that addressed blog articles as GL sources in SE research. In the following, we briefly described these studies.

The first study (WILLIAMS; RAINER, 2017) examined criteria to evaluate blog articles to be used as a source of SE research evidence through two pilot studies (a systematic mapping study and preliminary analyses of blog posts). The findings showed criteria to select content (e.g., authentic, informative, and trusted) and assess blog article credibility (e.g., rigor, relevance, evidence-based, and quality of writing). Benefits (e.g., evidence timeliness and trends analysis) and drawbacks (content diversity) using blog content as evidence in SE research were also identified.

The second study (RAINER; WILLIAMS, 2018b) examined the opportunities for using blog as source of evidence in SE research, by performing informally review practitioners use of blogs, review the research literature, and present the findings of a survey. An overview of research

on this topic was presented, exploring potential benefits (e.g., trend analysis and practitioners insights evidence) and challenges (e.g., the variability of blog content and the non-established process for assessing the quality).

The third study (WILLIAMS; RAINER, 2019) focused on finding credibility criteria to assess blog posts by selecting 88 candidate credibility criteria from a previous Mapping Study (WILLIAMS; RAINER, 2017). Then, 43 SE researchers were surveyed to gather opinions on a blog post to assess those credibility criteria. A set of criteria were identified, for instance, the presence of reasoning, reporting empirical data, and reporting data collection methods.

Secondary Studies

We briefly explored studies that investigated the use of GL in Secondary Studies. First, we explored studies based on MLR, followed by studies based on GLR.

Garousi and colleagues (GAROUSI; FELDERER; MÄNTYLÄ, 2016) were one of the first to investigate the use of GL in Secondary Studies of SE, in addition to TL. The authors compared the results of two Secondary Studies. One considered GL sources, while the other did not. In this investigation, Garousi and colleagues highlighted the importance of GL to cover technical research questions. Afterwards, the same researchers continued their investigations on using GL in addition to traditional literature by proposing the use of Multivocal Literature Reviews (MLR) (GAROUSI; FELDERER; MÄNTYLÄ, 2019). Since these investigations of Garousi and colleagues, several MLR studies have been conducted, as shown in Galindo Neto and colleagues (GALINDO NETO et al., 2019).

Raulamo-Jurvanen and colleagues (RAULAMO-JURVANEN et al., 2017) conducted the first GLR we have known in SE to analyze how software practitioners address the practical problem of choosing the appropriate test automation tool. The resultant data was derived mainly from experiences based on opinions. Moreover, this research examined the credibility of GL sources based on the information available about the number of readers, number of sharing, number of comments, number of Google Hits for the title, and the number of backlinks, by analyzing the websites that mentioned the source (a reference comparable to a citation). We identified another GLR that investigated the pains and gains of using microservices (SOLDANI; TAMBURRI; VAN DEN HEUVEL, 2018). This study observed that, in traditional literature, academic research on the topic is still in its early stage even though companies are working day-by-day with microservices, as also witnessed by the considerable amount of GL on the topic.

Tertiary Studies

There are also tertiary studies that explored the use of GL in Secondary Studies (YASIN et al., 2020; GALINDO NETO et al., 2019; ZHANG et al., 2020). In the following, we briefly described them.

Yasin and colleagues (YASIN et al., 2020) investigated the evidence of GL use in Secondary Studies published until 2012. In this study, was mentioned that GL was found in 76% of the Secondary Studies, and that the level of GL evidence of Systematic Literature Review (SLR) synthesis discussion was around 9%. This work employed some GL definitions; for instance, it mentioned the Luxembourg definition. This study also mentioned that GL is always referred to as "fugitive literature" or a semi-published source. This study considered theses, conference proceedings, technical reports, official documents, company white papers, discussion boards, and blogs as GL types. We highlight that, in this thesis, we have a different interpretation of Yasin et al., as we did not consider conferences proceedings as a GL type, as in SE, in general, conference papers undergo a peer-review process (GAROUSI et al., 2020). Thus, in our interpretation for this study, the GL included did not reach 76% of the primary sources. In addition, we did not agree with the GL definition employed in this study, as we did not consider GL as a "semi-published source" because it is always a published source but, in general, not using conventional channels of academia.

Galindo Neto and colleagues (GALINDO NETO et al., 2019) conducted the first tertiary study that focused solely on MLR and GLR. This study investigated the motivations SE researchers of 12 Secondary Studies had to conduct MLR. Their findings showed that MLR use was in the early stages. In addition, they identified that the main motivations to conduct MLR were the lack of academic research on the topic, the amount of evidence available in GL, and when the research topic was emerging.

Most recently, Zhang and colleagues (ZHANG et al., 2020) conducted a tertiary study focusing on two perspectives: (i) To understand the definitions of GL and their types used by researchers, and (ii) To survey 35 SE researchers from the studies identified in their tertiary study, to understand the motivations and challenges to use GL, how GL was used in their studies, and how they search for it.

In Chapter 8, we discussed these three studies (YASIN et al., 2020; GALINDO NETO et al., 2019; ZHANG et al., 2020) in more detail.

2.5 SUMMARY

This chapter introduced important concepts of GL used through this work. We also presented the importance, use, and previous investigations about GL in SE research, focusing on primary studies in which GL was used as a source of evidence. Then, we explored studies in which researchers are investigating how SE researchers feed on GL.

3 PERCEPTIONS OF GREY LITERATURE IN SOFTWARE ENGINEERING RESEARCH

This chapter presents two investigations based on survey research that focused on understanding the perceptions of Software Engineering (SE) Brazilian researchers about Grey Literature (GL) and answer RQ1. In particular, our study is unique in its focus on the Brazilian SE research community, and it improves the understanding of how researchers could explore and take advantage of GL.

In the following, Section 3.1 presents our research questions. Section 3.2 describes the survey methodology and the procedures used to collect and analyze the data. Section 3.3 presents our findings. Section 3.4 provides the discussions about this research. Finally, Section 3.5 presents a summary of this research.

3.1 OVERVIEW

To achieve the stated goal, we explored the following research question:

RQ1: *What Are the Perceptions of Software Engineering Researchers About Grey Literature?*

The above research question was divided into specific questions. They are:

RQ1.1: *Why do Brazilian Software Engineering Researchers Use Grey Literature?*

Rationale: Recently, SE practitioners have relied on social media and communication channels to share and acquire knowledge (STOREY et al., 2017). On the one hand, we identified researchers trying to take advantage of its use in SE research. For instance, Rainer and Williams (RAINER; WILLIAMS, 2018b) explored the benefits and challenges of blog articles as evidence in SE research. On the other hand, concerns (e.g., lack of detail and lack of empirical methods) related to GL could make SE researchers skeptical about their credibility (RAINER; WILLIAMS, 2019b). In this broad question, we investigated (i) if Brazilian SE researchers were using GL and, if so, (ii) what motivated them to use, or if not, (iii) the reasons that led them not to use GL.

RQ1.2: *What Types of Grey Literature Are Used by Brazilian Software Engineering researchers?*

Rationale: Nowadays, GL is available in many forms, from traditional mediums such as blogs, and question & answer websites, to more dynamic mediums such as Slack and Telegram, to videos on YouTube, to interactive gaming discussions on Twitch. Each one of these forums offers researchers a rich spectrum of unstructured data, which could bring specific benefits and limitations. This research question sought to investigate what Brazilian SE researchers often use as GL sources. A better understanding of the GL source would be important to guide future research in this area.

RQ1.3: *What Are the Criteria Brazilian Software Engineering Researchers Employ to Assess Grey Literature Credibility?*

Rationale: GL is, by nature, non-peer-reviewed; that is, when writing a blog post, practitioners are free to share their thoughts without worrying too much about methodological concerns. This freedom, however, may come with a cost: GL may be inaccurate, lacking context or details, or may even be incorrect. For instance, Fischer and colleagues (FISCHER et al., 2017) analyzed 1.3 million Android applications, and 15.4% of them contained security-related code snippets from Stack Overflow. Out of this, 97.9% contain at least one insecure code snippet. Therefore, researchers should employ additional assessment levels to ensure the selected GL is appropriate for the study when using GL in research works. This question investigated the reliability criteria that Brazilian SE researchers consider.

RQ1.4: *What Benefits and Challenges Brazilian Software Engineering Researchers Perceive When Using Grey Literature?*

Rationale: According to Storey and colleagues (STOREY et al., 2014), the SE research community has increased its interest in GL since the widespread presence of SE professionals using social media and communication channels. For instance, exploring the Stack Overflow, Zahedi et al. (ZAHEDI; RAJAPAKSE; BABAR, 2020) identified trends and challenges in continuous SE that researchers could better explore. This question explored the understanding of the (i) benefits and (ii) challenges that researchers may face when resorting to GL. Answering this question is important to understand the potential benefits and challenges of using GL more broadly by researchers.

RQ1.5: *How do Software Engineering Researchers Prioritize a Set of Criteria to Assess Grey Literature Credibility?*

Rationale: In Survey 1, we provided a set of criteria used by Brazilian SE researchers to assess GL credibility. Previous literature (WILLIAMS; RAINER, 2019) also identified another set of criteria. This question focused on understanding the importance of those criteria to assess GL credibility.

RQ1.6: *What Is the Perception of Brazilian SE Researchers About the Different Types of Grey Literature According to the Perspective of Control and Credibility?*

Rationale: Due to the diverse nature of the GL types, some studies suggested that GL needs to be assessed differently (GAROUSI; FELDERER; MÄNTYLÄ, 2019). For this reason, Adams et al. (ADAMS; SMART; HUFF, 2017) classified its types according to the shades of grey. This classification is based on two dimensions: control and expertise. *Control* refers to the rigor that the source is produced (not related to the construct of the experimentation method). Expertise is the extent to which the knowledge and producer authority can be determined. Nevertheless, this understanding and classification are still confused. This research question sought to understand how Brazilian SE researchers commonly perceived GL types according to the dimensions of (i) Control and (ii) Expertise.

To answer these research questions, we opted to employ survey research because, according to Linåker and colleagues (LINÅKER et al., 2015), this method collects information from a group of people by sampling individuals from a large population. In our case, we conducted two surveys to collect perceptions from Brazilian Software Engineering researchers from the largest Brazilian software conference. Focusing on answering the RQ1.1–RQ.1.4, we investigated how 76 Brazilian SE researchers use GL to understand which criteria they employed to assess its credibility, as well as the benefits and challenges they perceived. Then, in a follow-up survey, focusing on answer RQ1.5–RQ1.6, we investigated how 34 Brazilian SE researchers that previously mentioned the use GL perceived the criteria to assess the different GL types according to the dimensions of Control and Expertise. We employed a mixed-method approach based on qualitative and quantitative methods to analyze and synthesize the data.

In summary, our main findings with our investigations are:

- We elucidated what are the main GL sources used by the Brazilian SE researchers;
- We identified several motivations to use (or to avoid) GL;
- We identified what are the main criteria employed by Brazilian SE researchers to assess GL credibility (GL source be provided by renowned authors, institutions, companies, or cited by a renowned source);
- We perceived that GL is not widely used as a reference in scientific studies;
- We identified different interpretations to assess GL types, showing the importance to consider own source in particular;
- We identified for most of the GL types a strong to very strong positive correlations (*p-value* $<= 0.05\%$) between the perceptions of the level of Control and Expertise;
- We did not find a significant correlation (*p-value* $<= 0.05\%$) between the perceptions of Control and Expertise to GL types when considering the respondent's profile, i.e., the respondent profile did not influence their answers;
- We perceived misunderstandings about whether a source type is considered a GL type or not, mainly related to the most classified sources as High Control and High Expertise.

By describing these findings, we expect to improve the understanding of the importance of GL and its use by SE researchers.

3.2 SURVEY RESEARCH

This work focused on SE researchers potentially interested in using GL in their works. We followed Linåker et al.'s guidelines (LINÅKER et al., 2015), aiming to use a survey methodology to collect information from a group of people by sampling individuals from a large population. Section 3.2.1 presents the procedures used to conduct Survey 1. Section 3.2.2 presents the procedures used for Survey 2. Finally, Section 3.2.3 provides the methods used to analyze both surveys.

For replication purposes, the data used in this chapter is available online at:
<https://doi.org/10.5281/zenodo.5164714>.

3.2.1 Survey 1: Overview and Perceptions of Grey Literature

In Survey 1, we aimed to gather a broad perception of GL used by Brazilian SE researchers, focusing on understanding the *motivations to use (or avoid)*, the *GL types used*, the *benefits* and *challenges*, and the *criteria used to assess its credibility*. In what follows, we present the procedures employed to conduct our survey.

3.2.1.1 Survey Design

Our survey focused to investigate the Brazilian SE researchers' perceptions regarding the benefits, challenges, motivations to use, reasons to avoid, and the credibility criteria use to assess GL sources. To do so, we conducted our survey with participants of the 10th Brazilian Conference on Software: Practice and Theory (CBSoft), the largest Brazilian software conference with many SE researchers' participating. It includes well-established and specialized satellite SE conferences in its domain. Our population comprehends SE researchers are potentially interested in using GL in their research. We chose our sample using non-probabilistic sampling by convenience (BALTES; RALPH, 2020).

Before sending the final survey version, an experienced researcher (Ph.D. SE researcher with more than 15 years of experience in research) reviewed our draft. We also conducted a **pilot study** by randomly selecting two participants and explicitly asking for their feedback. We received feedback suggesting changing the order and re-writing some questions to make them more understandable to the target population.

Furthermore, we obtained the contact of all participants (n=252), asking the conference's general chair whether s/he could share this information with us, which s/he gently provided.[1]

We used two approaches to invite the researchers to answer our questionnaire. First, we placed posters on the event's walls and tables with a brief description of the work and the link to the online survey. Second, we sent the actual survey to the 250 remaining participants of the event. In the invitation email, we briefly introduced ourselves, presented the research's purposes, highlighted that the invite was to the participant of the CBSoft, and the link to the online survey. We also mentioned that the participant was free to withdraw at any moment, and all information stored was confidential.

[1] In the period of this research, the Brazilian General Data Protection Law was not yet officially published.

The survey was open for responses from September 26th to October 11th, 2019. We received a total of 76 valid answers (30.4% response rate). We did not consider the pilot survey answers.

3.2.1.2 Survey Respondents

Among the survey respondents, 48.7% have a Ph.D., 31.6% have a Master's, 2.6% are graduate specialization, 14.5% have a Bachelor's degree, and 2.6% are undergraduates. Among them, 72.4% are men, and 27.6% are women. Table 3 presents the demographics' information about the respondents and their experience using GL or not. This table shows that most respondents with Ph.D. and Master's degrees answered that they were using GL.

Table 3 – Demographics information of the Survey 1 respondents.

Gender	Level of course	Used GL	Not used GL
Woman	Doctorate	5	5
Man	Doctorate	24	3
Woman	Master	4	2
Man	Master	15	3
Woman	Expert	1	1
Man	Expert	0	0
Woman	University graduate	0	2
Man	University graduate	2	7
Woman	Technical education	0	0
Man	Technical education	0	0
Woman	High school	1	0
Man	High school	1	0

Source: the author.

3.2.1.3 Survey Questions

Our survey had 11 questions (three were required, nine were open). We used different questions flow for those who used GL (did not answer question 10) from those who did not (answered only questions 1 to 4 and questions 10 and 11). Table 4 presents the questions covered in this survey. A complete version of Survey 1 questionnaire is available on Appendix A.

Table 4 – Questions covered in Survey 1.

#	Question	Type of question	Options of answers (for closed questions)	Required?	RQ
Q1	What is your e-mail?	Open	-	No	-
Q2	What is your gender?	Open	-	Yes	-
Q3	Please list the highest academic degree you have received.	Closed	High school, Technical education, University graduate, Expert, Master's degree, Doctorate.	Yes	-
Q4	Have you used grey literature? If you never used, go to question *Q10*.	Closed	Yes, No.	Yes	RQ1
Q5	What sources of grey literature did you use?	Open	-	No	RQ2
Q6	In which conditions *do you use* grey literature?	Open	-	No	RQ1
Q7	In which conditions do you do *not use* grey literature?	Open	-	No	RQ1
Q8	Could you list any *benefits* in using grey literature?	Open	-	No	RQ4
Q9	Could you list any *challenges* in using grey literature?	Open	-	No	RQ4
Q10	If you answered 'no' in question four, please state why did you *never use* or *avoid* use grey literature?	Open	-	No	RQ1
Q11	What would be a *reliable source* of grey literature for you?	Open	-	No	RQ3

Source: the author.

3.2.2 Survey 2: Credibility Criteria Considering the Different Types of Grey Literature

In this survey, we conducted a follow-up survey to collect perceptions only from the Brazilian SE researchers from Survey 1, who answered that they have previously used GL, focusing on

understanding the credibility perceptions of the different GL types concerning the dimensions of Control and Expertise. In what follows, we present the procedures employed to conduct our survey.

3.2.2.1 Survey Design

We invited by email once again the 53 researchers that participated in our Survey 1 and mentioned the use of GL. In this follow-up survey, we had the intention to understand their perceptions regarding the credibility criteria in deep for the different GL types.

We first drew our questionnaire and improved it through the conduction of three sequential steps: 1) A pilot study with five Ph.D. SE researchers; 2) Another SE researcher specialist assessed the questionnaire; and 3) Received feedback of a participant relating a problem in the first hours after opening the survey. For this reason, we closed the survey to stop receiving answers.

Then, we deleted all answers previously received and sent a new questionnaire version to the researchers. We opened the survey for answers from February 10th to March 4th, 2021. We received 34 valid answers (64.1% response rate). We did not consider the pilot survey answers.

3.2.2.2 Survey Respondents

In this survey, as we retrieved our sample from the previous one who answered that they had used GL, we did not ask the same questions (e.g., gender, academic degree). Instead, we collected information about their experience in SE research and using GL in scientific articles.

The respondents' profile of our survey was composed of 76.5% of professors or researchers and 23.5% of undergraduates. Regarding SE research experience, 55.9% of the respondents had more than ten years. Considering the experience using GL, 47% had conducted between 2 and 5 scientific studies using GL, although 26.5% were unable to answer.

3.2.2.3 Survey Questions

Our second survey had ten questions (six were required, and four were open). Table 8 presents the questions covered in this survey. We highlighted that, before question 4, we

Table 6 – The number of scientific articles produced using Grey Literature by Software Engineering researchers, analyzed in Survey 2.

Type of Occupation	Number of scientific articles using GL				
	I do not know	Only one	From 2 and 5	From 6 and 10	More than 10
Professor/Researcher	9	2	13	-	2
Student (M.Sc. or Ph.D.)	-	5	3	-	-

Source: the author.

included a video[2] to summarize and explain the "shades of GL" and the dimensions of the level of Control and Expertise. A complete version of the questionnaire for Survey 2 is available in Appendix B.

Table 8 – Questions covered in Survey 2.

#	Question	Type of question	Options of answers (for closed questions)	Required?	RQ
Q1	What is your occupation?	Closed	Professor/Researcher, Student (M.Sc. or Ph.D.), Other (open).	Yes	-
Q2	How many years of experience did you have conducting SE research?	Closed	Until 1 year, From 1 and 3 years, From 4 to 6 years, From 7 to 9 years, 10 years or more.	Yes	-
Q3	How many scientific studies have you conducted using GL as source of evidence?	Closed	I do not know, No one, Only one, From 2 and 5, From 6 and 10, More than 10.	Yes	-

[2] Video explaining the "shades of GL" (in Portuguese): https://youtu.be/hGMkVXIApR0

Q4	We are aware that the *level of Control* varies from source to source. For this reason, we ask you to consider your experience more frequent in relation to each source type in relation to the *Control* dimension of the production.	Closed	Source types: {adapted from Maro et al. (MARO; STEGHÄFER; STARON, 2018); Level of Control: I did not consider it as a GL type, Low Control, Moderate Control, High Control, No opinion.	Yes	RQ6
Q5	Please, explain what did you consider to classify each source type with the *Control criteria* presented in Question 5.	Open	-	No	RQ6
Q6	We are aware that the *level of Expertise* varies from source to source. For this reason, we ask you to consider your experience more frequent in relation to each source type in relation to the *Expertise* dimension of the production.	Closed	Source types: {adapted from Maro et al. (MARO; STEGHÄFER; STARON, 2018); Level of Expertise: I did not consider it as a GL type, Low Expertise, Moderate Expertise, High Expertise, No opinion.	Yes	RQ6

Q7	Please, explain what did you consider to classify each source type with the *Expertise criteria* presented in Question 7.	Open	-	Yes	RQ6
Q8	Considering a GL source with important information to your research, would you include a GL source if it is produced by/with.	Closed	Choices for Credibility criteria: Be produced by a renowned author, Be produced by a renowned institution, Be produced by a renowned company, Be cited by others renowned sources, Describe the methods of collection, Cites an academic reference, Cites a practitioner source, Presents information with rigor, Presents empirical data; Choices for answers: No opinion, No, Yes.	Yes	RQ5
Q9	Could you cite any additional potential aspect to assess the GL credibility source that was not mentioned before?	Open	-	No	RQ6

Q10	We are planning to conduct a future research about Quality Assessment in Grey Literature. Please, could you inform your mail to future contact?	Open	-	No	-

Source: the author.

3.2.3 Data Analysis and Synthesis

This section presents our approach used to analyze and synthesize the data. Section 3.2.3.1 presents our qualitative and Section 3.2.3.2 presents our quantitative approaches.

We used a *qualitative* approach when we were interested in questions about "what" and "how" and a *quantitative* analysis using descriptive statistics to discuss frequency and distribution and correlation analysis between the dimensions of Control and Expertise to each GL type. We describe these methods in the following.

3.2.3.1 Qualitative Analysis

We used a qualitative approach based on the thematic analysis technique (BRAUN; CLARKE, 2006). This process involved three SE researchers with previous qualitative research experience (one Ph.D. student (R1) and two Ph.D. professors (R2–R3)) for both surveys.

We performed an agreement analysis with the codes and categories generated by each researcher using the Kappa statistic (VIERA; GARRETT, 2005) to Survey 1. According to the Kappa reference table (VIERA; GARRETT, 2005), the Kappa value was 0.749, indicating a Substantial Agreement level. For Survey 2, we do not calculate Kappa due to the analysis process that occurred with the researchers working together.

Figure 18 presents a general overview of the process employed. In the following, we detailed the procedure used to analyze all answers (adapted from Pinto and colleagues (PINTO et al., 2019)) of both surveys, showing the differences employed in each survey research:

(a) *Familiarizing with data:* The process started with two independent researchers reading the answers of the survey respondents, as expressed in Figure 18-(a).

(b) *Initial coding:* Then, for Survey 1, two independent researchers (R1 and R2) individually analyzed and added codes. For Survey 2, the researchers analyzed, discussed, and coded together (R1 and R2, into a dotted box). We used a post-formed code, so we labeled portions of text that expressed the meaning of the excerpts without any previous pre-formed code. The initial codes are temporaries, since they still need refinement. We refined the emerged codes throughout all analyses. An example of coding is present in Figure 18-(b).

(c) *From codes to categories:* Here, we already had an initial list of codes. For Survey 1, two researchers individually conducted this process (R1 and R2). For Survey 2, this process occurred with two researchers working together (R1 and R2). This process begins to look for similar codes in the data. We grouped the codes with similar characteristics in broader categories. Eventually, we also had to refine the categories identified, comparing and re-analyzing them in parallel, using an approach similar to axial coding (SPENCER, 2009). Figure 18-(c) presents an example of this process.

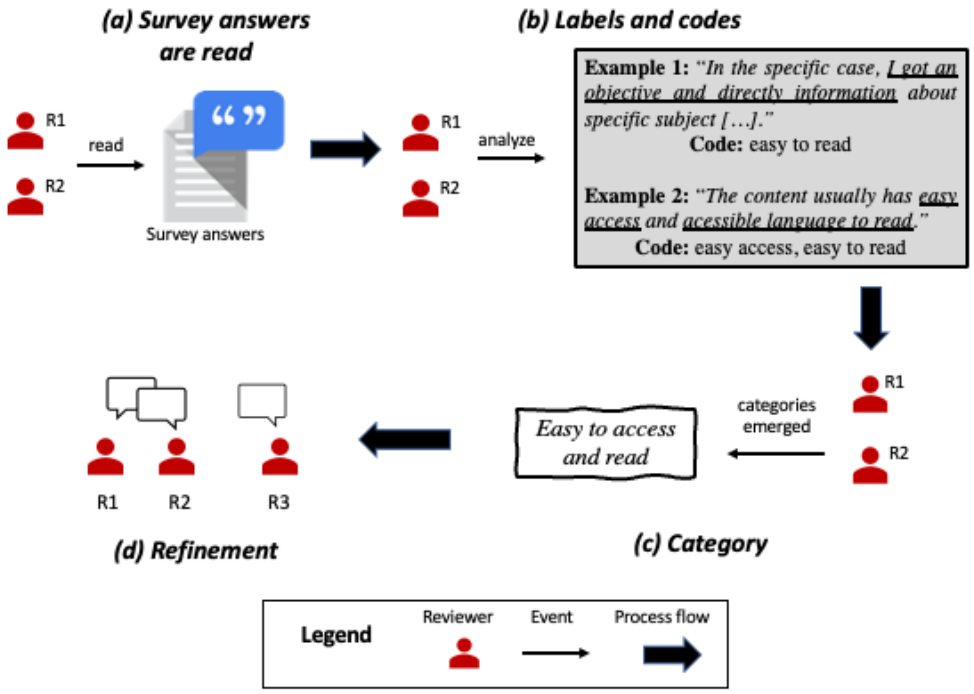
(d) *Categories refinement:* Here, we have a potential set of categories. For both surveys, in a consensus meeting between R1 and R2 (Figure 18-(d)), the categories were evaluated and solved the disagreements of interpretation for evidence that supported or refuted the categories found. We also renamed or regrouped some categories to describe the excerpts better there. In cases where disagreements remained, we invited a third researcher (a Ph.D. professor) to review and solve them for both surveys.

3.2.3.2 Quantitative Analysis

We based our quantitative investigation on three samples: (i) We used the answers from 76 SE researchers to answer RQ1.1; (ii) We used the answers from 53 researchers that mentioned using GL to answer RQ1.2, RQ1.3, and RQ1.4; and (iii) We used the answers from 34 researchers to answer RQ1.5 and RQ1.6.

For the descriptive statistics, we highlighted that one answer of a respondent could be related to more than one category found. In the investigations related between the GL types

Figure 3 – Example of a coding process used to analyze the questionnaire answers of Survey 1 and 2.



Source: the author.

and the dimensions of Control and Expertise, we present it into boxplots to show the differences of interpretations of each GL type.

We used Spearman's rank correlation coefficient to analyze the Control and Expertise perceptions for each GL type. Then, we transformed the answers related to the level of Control and Expertise (Low, Moderate, High) into non-linear scales: Low = 0, Moderate = 50, and High = 100.

For the quantitative data analysis, we used R language and Python. This last, with the support of Google Colab[3].

3.3 RESULTS

This section presents our main results organized concerning the research questions. Section 3.3.1 presents the motivations to use or reasons to avoid GL. Section 3.3.2 presents the GL types used. Section 3.3.3 exposes the criteria employed to assess GL credibility. Section 3.3.4 presents the perceived benefits and challenges when using GL by SE researchers. Section 3.3.5 shows how the criteria used to assess GL credibility are prioritized. Finally, Section 3.3.6 pro-

[3] https://colab.research.google.com

vides the perceptions about the different GL types according to the perspective of Control and Expertise.

To enable traceability, we include direct quotes from respondents along with the answer identified in open-ended questions, and we present the discovered codes slanted. We also presented the list of categories identified in tables, with the total number of occurrences of a given category in the column "#". An important observation is that some researchers may have reported more than one answer per question, which may happen to be grouped into different categories. Still, most of our questions are not required. Then, when summarizing the categories in tables, the overall results might not reach 100% of respondents.

3.3.1 RQ1.1: Why do Brazilian Software Engineering Researchers Use Grey Literature?

(i) Overall Use

Most of the respondents of our work (53/76 occurrences, 69.7%) are using GL for some purpose, which means they had previous experience using GL. This value was used to analyze all categories about motivations to use GL in the following. Moreover, to understand better why and how SE researchers are using GL, we asked questions that included the motivations to use GL or reasons to avoid it. We observed several driving motivations *to use GL*, as present in Table 9. We describe some of them in the following.

We highlight that one answer of researchers could be related to more than one category identified. This is worth to the next following categories of other RQs.

Table 9 – Motivations to use Grey Literature identified in Survey 1.

Motivation	#	%
To understand the problems	28	52.8%
To complement research findings	12	22.6%
To answer practical and technical questions	10	18.9%
To prepare classes	4	7.5%
To conduct government studies	1	1.9%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

(ii) Motivations to Use

To understand the problems **(28/53 occurrences; 52.8%)**. This category was mentioned by more than half of the respondents, which means when the researcher uses GL to understand or investigate a new topic that has no previous knowledge, or when s/he searches for something aiming to solve problems, or when they want to acquire specific information to deepen the knowledge. Regarding this category, some respondents have pointed out: *"I used initially to learn a topic that I don't have knowledge"*, *"In most cases, to understand how the problem happens in the society"*, and *"When I want to search for deep references and in a large amount about a specific theme"*.

To complement research findings **(12/53 occurrences; 22.6%)**. A researcher mentioned that used GL to complement a Mapping Study, as we quoted out: *"GL was used to complement a data of a Systematic Mapping"*. Another respondent raised using GL for a specific context, in which the peer-reviewed content is still scarce, as pointed out: *"I use it when I don't find many studies in a specific context, for instance, in the use of SE in the context of digital games there are process models that are not described in articles that are considered by game developers"*.

To answer practical and technical questions **(10/53 occurrences; 18.9%)**. This category was quite mentioned, mainly about understanding the state of the practice. In this sense, a respondent pointed out: *"([. . .] I use it when I have the perception that the theme has an origin on the industry and is on discussion or an increase of adoption in the industry"*.

To prepare classes **(4/53 occurrences; 7.5%)**. Few SE researchers mentioned the use of GL to support the material to prepare classes, as a respondent pointed out: *"(Use GL) When I'm searching for something for a class"*. It is common for SE researchers to be professors at universities in the investigated research community. For this reason, some researchers have used the GL to support them.

(iii) Reasons to Avoid/Never Use

Even though the perception of several motivations to use GL, 50.9% of SE researchers (27/53) ***avoid using*** GL as a reference or to reinforce some claims in **Scientific papers**, or any other type of scientific documents, such as thesis and SLR because they argued that GL

evidence is usually scarce of scientific value, that makes it is not often well-regarded research community. In this regard, a SE researcher mentioned: *"I try to avoid the use of in research papers and systematic reviews. Generally, the community belittles such references"*. Furthermore, we identified respondents that ***never used GL*** (23/76 occurrences; 30.3%), which means they did not have previous experience using GL in any research situations. This value was used to analyze all reasons for never using GL in the following. Of those 23, 15 answered our question to understand why never use GL. The summary of the findings for this question is presented in Table 10. We describe some of them in the following.

Table 10 – Reasons to avoid/never use Grey Literature identified in Survey 1.

Reason	#	%
Lack of reliability	6	26%
Lack of scientific value	3	13%
Lack of opportunity to use	3	13%
Others	3	13%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Lack of reliability **(6/23 occurrences; 26%)**. This category was the main motivation that our respondents mentioned not to use GL in their research. It is related to the lack of rigor in which GL content is written and published, putting into question the credibility of information presented due to the lack of quality control that makes it difficult to ensure their quality. Regarding this motivation, we present two quotes: *"Because grey literature has no scientific or commercial control, it can produce unreliable content with bias from the scientific point of view"* and *"GL is very open, without a deeper assessment of the material available"*.

Lack of scientific value **(3/23 occurrences; 13%)**. In this category, due to the lack of scientific value of GL by the scientific community, the respondents were afraid that GL use would weaken a research paper when submitted to the peer-review process, as a respondent cited: *"Formally I never used it because I believe that will not be considered by academia. [. . .] academia only accepts peer-reviewed references"*.

Lack of opportunity to use **(3/23 occurrences; 13%)**. This category was mentioned due to the nature of research employed, and GL is recent in the context of SE, as a respondent

mentioned: *"I never had an opportunity to use"* and the another mentioned: *"I met this type of review recently and have not yet had the opportunity to adopt it in my research"*.

***Others* (3/23 occurrences; 13%)**. Here we group other responses that we were unable to group. Among them: (i) One that was removed from the entire analysis, where a researcher mentioned that s/he had never used GL because s/he never heard about GL before, showing that s/he did not understand what the question asks for, and (ii) Another mentioned due to the lack of support for GL search. *"Because I don't know where to search for relevant content"*.

> **Summary of RQ1.1:** Brazilian SE researchers are using GL motivated mainly to understand new topics, find information about practical and technical questions, and complement research findings. However, some researchers affirmed avoiding using GL, particularly in scientific papers.

3.3.2 RQ1.2: What Types of Grey Literature Are Used by Brazilian Software Engineering researchers?

In this question, we investigated the GL source used. To answer this question, we used the responses of the 53 respondents that mentioned using GL. When analyzing these questions, we identified several sources used by SE researchers, as listed in Table 11. We highlight that 11 out of 53 (20.7%) respondents mentioned using a search engine (e.g., Google and Google Scholar) as a start point to find GL content. However, we did not consider Google a source of GL, although our survey respondents had considered it. In the following, we present some of our findings.

***Community websites* (16/53 occurrences; 30.2%)**. The most common source used was the community website, i.e., websites where users can interact with others, e.g., creating content, posting comments, and assessing the content. Some researchers mentioned the use of Stack Overflow and Quora as a GL source, as mentioned by a respondent: *"Communities that bring together a variety of developers profile, such as Stack Overflow"*.

***Blogs* (15/53 occurrences; 28.3%)**. The use of blogs as a source of GL was the second most common category identified. A respondent used blogs from renowned practitioners, as s/he pointed out: *"Sites or blogs by well-known authors in a particular area"*. Another respondent mentioned the content of blogs derived from companies that produce a diversity of material

and content for SE and software development in general: *"Blogs by SE firms (Netflix, Uber, Facebook engineering) [. . .]"*.

Technical experience/report/survey **(14/53 occurrences; 26.4%)**. Most of the respondents that mentioned this category used technical experience, reports, and surveys derived from industry, as a respondent pointed out: *"Usually websites of companies that provide technical reports, for instance, such as SEI, CMU, Jetbrains, among others"*. Another SE researcher mentioned that there are also technical reports derived from academic settings: *"Technical Reports published in national and international research groups, available on publications repositories"*.

Companies website **(8/53 occurrences; 15%)**. This category means the website of companies, e.g., Google, Facebook, and ThoughtWorks, that contains information regarding their technologies, methods, practices. Some respondents mentioned browsing these websites to find news about a specific technology to help decision-making. Regarding this category, a SE researcher pointed out: "*I have always used blogs and companies' websites to help me with decision-making to select a specific software or tool to use"*.

Others **(3/53 occurrences; 5.7%)**. This last category group responses that we could not group elsewhere, including government publications, open data portal, and class material.

Table 11 – Grey Literature sources used by Software Engineering researchers identified in Survey 1.

Source	#	%
Community website	16	30.2%
Blogs	15	28.3%
Technical experience/report/survey	14	26.4%
Companies website	8	15%
Preprints	5	9.4%
Books	5	9.4%
Whitepapers	4	7.5%
Data repository	4	7.5%
Videos	3	5.7%
Non-scientific magazines	3	5.7%
News	2	3.8%
Others	3	5.7%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Summary of RQ1.2: We identified several GL sources used by Brazilian SE researchers. The most common sources are community websites, blogs, technical experience/reports/-surveys, and companies' websites.

3.3.3 RQ1.3: What Are the Criteria Brazilian Software Engineering Researchers Employ to Assess Grey Literature Credibility?

With this research question, we explored the criteria of how the SE researchers assess GL credibility. We asked it using one open-ended question. In this research, we identified 16 cases of mention in a general way to the criteria of GL source need to be trustworthy. Table 12 summarizes the results, and some of them are described in the following.

Renowned authors **(15/53 occurrences; 28.3%)**. Some respondents mentioned that the GL content provided by renowned authors is an important criterion to assess its credibility. For instance, they assess the author's experience and reputation about the topic in the community, as some respondents cited Martin Fowler as an important software engineer with notorious knowledge. A respondent mentioned the importance of relying on a renowned author: *"One source that shows an author with in-depth knowledge about they are writing"*. Another respondent mentioned the importance of searching practitioners: *"popular blogs and websites from important people of the industry"*.

Renowned institutions **(14/53 occurrences; 26.4%)**. Similar to the above category, in this, we perceived that an important criterion of credibility is the use GL source produced by renowned institutions or research groups, as a respondent mentioned: *"Something [GL] that is produced by an institution with credibility on the topic"*. Regarding this criterion, another researcher pointed out: *"When one recognized institution is supporting (e.g., whether reviewing, following up) the work. For instance, the technical reports produced by SEI or by the Institute of Fraunhofer because their institutions follow a scientific rigor and are concerned with the production of the material"*. Still, the groups of research of an institution were mentioned: *"Repositories of research group publications with a history and reputation to conduct research on the topic"*.

Cited by others **(8/53 occurrences; 15%)**. This category was mentioned to express those respondents considered a trusted source, which was cited by others (studies or people). In

this sense, a respondent affirmed: *"The ResearchGate shows the citations and recommendations of works by other researchers, even though some of them were non peer-reviewed"*. Still, another researcher affirmed: *"A source of information attested by the community that used certain information"*. This last mention refers to the Stack Overflow, in which the users can comment, "up vote", and "down vote" the answers.

***Renowned companies* (7/53 occurrences; 13.2%)**. Some respondents considered a GL source as trusted when it was produced by renowned software industries or portals, as mentioned by a respondent: *"Sites or blogs of large software engineering companies (Netflix, Uber, Facebook)"*.

***Others* (9/53 occurrences; 17%)**. This last category group responses that we could not group elsewhere, including the criteria that were not related to the GL producer. Instead, these criteria were related to the *quality of GL information*, meaning that it is important to allow the reader to assess the accuracy and precision of their procedures. Another respondent pointed out that it is important that the *source is well written with all the important information for analysis*. Beyond that, another one pointed out that the GL needs to *present a certain scientific rigor (e.g., presenting references)*. Moreover, another respondent pointed out that an important criterion is to use a source that permits the user to *assess the content or answer* and *give feedback* to data posted by others.

Table 12 – Criteria to assess Grey Literature credibility identified in Survey 1.

Source	#	%
Renowned authors	15	28.3%
Renowned institutions	14	26.4%
Cited by others	8	15%
Renowned companies	7	13.2%
Others	9	17%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Summary of RQ1.3: The main criteria of credibility is about who produces the GL content, whether produced by a person, institution, or company since the source is renowned.

3.3.4 RQ1.4: What Benefits and Challenges Brazilian Software Engineering Researchers Perceive When Using Grey Literature?

Our last research question aimed to explore the perceived benefits and challenges (problems or difficulties) on the GL use by SE researchers. We asked them using two open-ended questions. The results regarding the benefits are presented in Table 13 and the challenges in Table 14. In the following, we present some discussions about our findings.

(i) Benefits

Easy to access and read **(16/53 occurrences; 30.2%)**. This category was the most common benefit perceived by the respondents, mainly because most GL sources are open access, are easily recovered by free search engines, and the contents are usually easy to read. Moreover, another respondent considered GL information to be written in a less formal language: *"Easy to access and is written in less formal language"*. Other respondent shares the same opinion: *"The content usually has easier access and a more accessible language"*.

Provides practical evidence **(13/53 occurrences; 24.5%)**. Respondents mentioned that GL evidence from the industry is important to understand the state of the practice. A respondent mentioned using GL to find practical information: *"To discover practical information and practices not reported in traditional literature"*. Another researcher shared the same opinion: *"Understanding how things happen in the industry [...]"*.

Helps in knowledge acquisition **(13/53 occurrences; 24.5%)**. Some respondents mentioned that using only the traditional literature limits the knowledge. For this reason, the GL use could widen the knowledge with different information, as a researcher mentioned: *"The industry experience reports brought different facets about the phenomenon they were studying"*. Another situation was pointed out by a respondent that read a researcher's blog: *"[...] more complete and detailed data about one scientific research than scientific articles of the same author"*.

Up to date information **(6/53 occurrences; 11.3%)**. Since it often takes a reasonable time to publish a scientific paper, some papers' content may become technically outdated shortly after publication. In this sense, as our respondents mentioned, GL is often more up-to-date regarding technical details. Regarding this situation, a respondent affirmed: *"[...]*

Additionally, newer technologies tend to appear faster in GL". Another one claimed: *"I have found very interesting (blog) articles about new topics"*.

Advances the state of the art/practice **(5/53 occurrences; 9.4%)**. Some respondents perceived the importance of GL to understand the industry better and conduct research aiming to find important practice gaps. A respondent affirmed: *"Understanding how things happen in the industry, and which technologies derived from academia are in use. The GL also reveals many gaps and opportunities to applied research and transfer of knowledge"*.

Covers different results from scientific studies **(3/53 occurrences; 5.7%)**. Some researchers revealed the GL importance in providing additional knowledge not yet available in the research area. Regarding this benefit, a respondent pointed out: *"Data and evidence (of GL) are different from peer-reviewed articles that do not always provide original data for replication and also by limiting the coverage and comprehensiveness of data available from non-GL sources"*.

Trend analysis **(1/53 studies; 1.9%).** One researcher mentioned the importance of GL to understand new technologies, as we pointed out: *"[. . .] Additionally, newer technologies tend to appear more quickly in GL"*.

Table 13 – Benefits of the Grey Literature use identified in Survey 1.

Benefit	#	%
Easy to access and read	16	30.2%
Provides practical evidence	13	24.5%
Helps in knowledge acquisition	13	24.5%
Up to date information	6	11.3%
Advances the state of the art/practice	5	9.4%
Covers different results from scientific studies	3	5.7%
Trend analysis	1	1.9%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

(ii) Challenges

Lack of reliability **(34/53 occurrences; 64.2%)**. This category was the main challenge perceived by the respondents. Some of them put in check the reliability of GL's content, as a researcher pointed out: *"The biggest challenge, in my opinion, represents the validation of what is being reported"*. Still, another pointed out: *"How to ensure the quality of information maybe is the big challenge to use GL"*.

Lack of scientific value **(15/53 occurrences; 28.3%)**. This was the second category most cited by the respondents. This category is closely related to the one mentioned before. Respondents cited this problem because they are not comfortable to the lack of GL recognition by scientific area or to use it as a reference in scientific work, as two respondents affirmed: *"It has not scientific rigor"*, and *"[. . .] The diversity of channels that they are published hinder the search, defy replicability [. . .]"*.

Difficult to search/find information **(6/53 occurrences; 11.3%)**. The diversity of sources to search for GL content was a challenge perceived for some respondents, as pointed out: *"The diversity of channels in which the content of GL are published hinder the search, defy replicability, and increase the effort to select content"*.

Non-structured information **(6/53 occurrences; 11.3%)**. Another challenge perceived is the GL lack of content structure. For instance, for some respondents, there is a lack of a writing pattern and a large variety of formats in which the sources are published. Regarding those challenges, a respondent mentioned: *"The lack of pattern to the structure and writing"*, and another complement: *"The variety of formats in which the sources (non-standard) are reported in GL also configured as another significant challenge"*.

Table 14 – Challenges on the Grey Literature use identified in Survey 1.

Challenge	#	%
Lack of reliability	34	64.2%
Lack of scientific value	15	28.3%
Difficult to search/find information	6	11.3%
Non-structured information	6	11.3%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Summary of RQ1.4: We identified several benefits of GL use. The most common was that the content of GL is easy to access and read, and it is important to knowledge acquisition, mainly for providing evidence derived from SE practitioners. The most cited challenges were lack of reliability and scientific value, making it difficult to be used in scientific research.

3.3.5 RQ1.5: How do Software Engineering Researchers Prioritize a Set of Criteria to Assess Grey Literature Credibility?

In our second survey, we asked 53 researchers to prioritize the importance of a set of criteria to assess GL credibility. These criteria were derived from our first investigation and identified in Williams and Rainer's study (WILLIAMS; RAINER, 2019). We received answers from 34 SE researchers. Table 15 presents the result of the ranking prioritization of credibility criteria, revealing that essential criteria perceived by SE researchers are: GL source be provided by a *Renowned author*, *Renowned institution*, or *Cited by a renowned source*.

We also investigated whether the SE researchers have any additional criteria to assess GL credibility not mentioned in the previous survey questions. By analyzing the answers, we did not find any new criterion that was not related to the criteria as earlier presented in Table 15. For instance, some researchers mentioned that the detailed description of the publication context is an important criterion. For this case, we considered that it is already contemplated in *Rigor in presenting information* criterion, previously mentioned by Williams and Rainer (WILLIAMS; RAINER, 2019). The author's experience with the topic was another criterion mentioned. We considered this criterion related to the *Renowned author's* criterion identified in our first survey.

Table 15 – Prioritized criteria to assess Grey Literature credibility investigated in Survey 2.

Criteria	#	%
Renowned authors	30	88.2%
Renowned institutions	30	88.2%
Cited by a renowned source	27	79.4%
Cites academic source*	26	76.5%
Present empirical data*	26	76.5%
Renowned companies	25	73.5%
Cites practitioner source*	16	47.1%
Rigor in presenting information	12	35.3%
Describe the methods of collection*	6	17.6%

Notes:
1) The column "#" shows the total number of occurrences of a given category.
2) The symbol "*" indicates a criterion proposed in previous study (WILLIAMS; RAINER, 2019).

Source: the author.

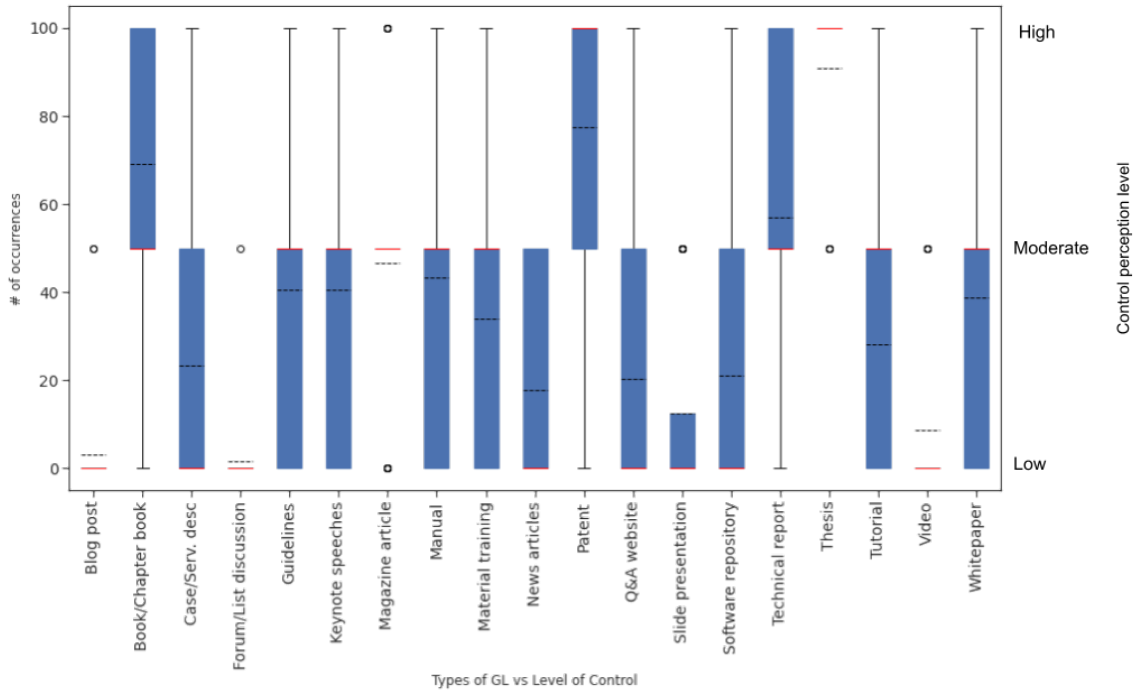
Summary of RQ1.5: We assessed the prioritization of credibility criteria identified in the Survey 1, in addition to those identified in previous study (WILLIAMS; RAINER, 2019). We identified that the most used criteria by SE researchers are when the GL is produced by a renowned source, cited by a renowned authority, cites an academic source, and presents empirical data.

3.3.6 RQ1.6: What Is the Perception of Brazilian SE Researchers About the Different Types of Grey Literature According to the Perspective of Control and Credibility?

Our last research question explored how the researchers perceived the different GL types concerns to the dimensions of Control and Expertise. These dimensions are used to classify the tiers of the "shades of GL". Each dimension could be evaluated into three levels (Low, Moderate, High). Figure 4 presents the results of classifications according to the level of Control, and Figure 5 shows the results of the level of Expertise.

Even though we are investigating different dimensions, interestingly, in some cases, the

Figure 4 – Classification of the Grey Literature types according to the Control level identified in Survey 2.



Each level of Control indicates: Low = 0; Moderate = 50; High = 100.

Source: the author.

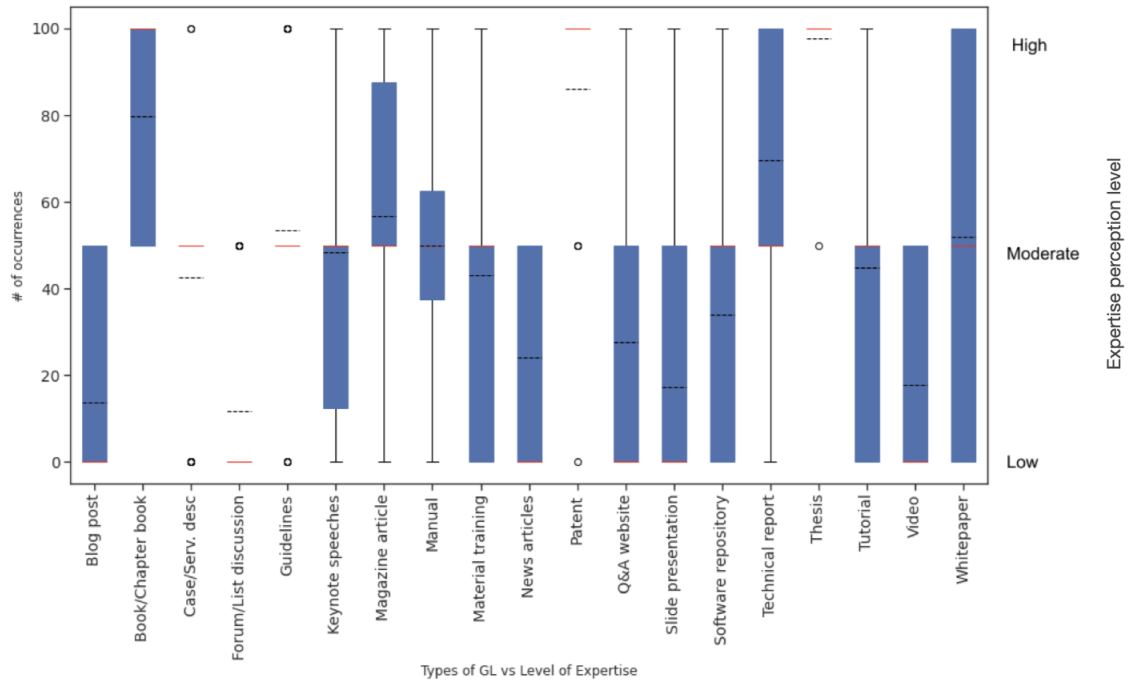
figures 4 and 5 presented similar behaviors. For instance, for some GL types (e.g., *blog posts*, *forums/list of discussions*), the Low level was predominantly in both dimensions. We also identified similarities concerning the other levels for both dimensions. For instance, some types (e.g., *materials training*, *news articles*, *software repositories*, and *tutorials*) run between Low (1st Quartile) to Moderate (2nd Quartile). Although, for a diversity of cases, the median behavior varied.

We also identified differences. For instance, considering the level of Control to *cases/services descriptions* and *guidelines*, the classifications run between Low (1st Quartile) to Moderate (2nd Quartile). In contrast, for the level of Expertise to these GL types, we identified outliers on the Low level (1st Quartile) and outliers on the High level (3rd Quartile).

Other classifications caught our attention. For instance, regarding the Control dimension, the opinions about the *magazine articles* are not equalized, as we identified some outliers in both extremes (Low and High). A similar classification we identified related to *guidelines* for the Expertise dimension.

In addition to classifying the levels (Low, Moderate, and High) of the dimensions (Control and Expertise), we offered the possibility to the researcher to choose the options of *"I did not consider it a GL type"* or *"I have no opinion"*. We included these options because even

Figure 5 – Classification of the Grey Literature types according to the Expertise level identified in Survey 2.



Each level of Expertise indicates: Low = 0; Moderate = 50; High = 100.

Source: the author.

previous studies (e.g., Maro et al. (MARO; STEGHÄFER; STARON, 2018)) presented the GL types for SE research; in our previous investigation (KAMEI et al., 2021b), we identified different interpretations, for instance, in which some types were not considered as GL. Table 16 shows the results of these classifications.

Comparing the findings presented in Table 16 with the information presented in figures 4 and 5, we perceived that most of GL types classified with High Expertise and High Control were also, many times, considered as not a GL type (e.g., *thesis*, *books/book chapters*, and *patents*). Moreover, we identified that *patents* are still unknown to several researchers.

(i) Rationale to employ classification of each dimension (Control and Expertise)

We asked why the researchers employed the classifications of each GL type according to the Control and Expertise. We identified four main reasons that are summarized in Table 17 and described in the following.

Table 16 – Grey Literature types in which Software Engineering researchers have no opinion regarding the level of Control and Expertise identified in Survey 2.

	Control	Expertise	
Type of source	No opinion	No opinion	✖ GL
Thesis	0	1	12
Patents	7	10	7
Books/Book chapters	2	1	6
Magazine articles	1	2	3
Case/Serv. desc	1	5	3
Manuals	1	3	3
Materials training	0	3	3
Software repositories	0	3	3
Blog posts	1	3	2
Forums / Lists	0	2	2
News articles	0	3	2
Slide presentations	0	6	2
Keynote speeches	0	2	2
Videos	3	4	2
Technical reports	3	2	2
Q&A websites	1	3	1
Guidelines	1	4	1
Tutorials	0	4	1
Whitepapers	2	5	1

Note: "✖GL" represents when SE researcher does not consider a source type as GL.

Source: the author.

Table 17 – Reasons to classify Grey Literature types according to the Control and Expertise levels identified in Survey 2.

Reasons	#	%
Rigor	23	67.6%
Producer reputation	14	41.2%
Own research experience	13	38.2%
Peer interaction	5	14.7%

Source: the author.

***Rigor* (23/34 occurrences; 67.6%)**. Researchers considered the rigor of each source's production, for instance, the degree of formality present. In this regard, one researcher pointed out: *"Technical reports, for instance, present systematic studies with high control (of produc-*

tion)". This category was also related to the Control dimension, as one researcher affirmed: *"I consider that credibility is directly related to the rigor of the publication/availability of an artifact"*.

Producer reputation **(14/34 occurrences; 41.2%)**. The producer's reputation was considered an essential criterion to assess Control and Expertise, as one researcher pointed out: *"The credibility relates to who is the author of the material and to the platform being conveyed*. Another one mentioned: *"Depending on the publisher, I can consider high (e.g., Elsevier) or low (e.g., autonomously published book) control. The same applies to news: the credibility of the source influences the level of control regarding stricter editorial control in favor of the integrity of the information"*.

Own researcher experience **(13/34 occurrences; 38.2%)**. The researcher considered the own experience using GL to employ the classification of the Control and Expertise levels. In this regard, one researcher pointed out: *"I thought of the examples for each type that I have used and classified them according to my experience in dealing with each material"*. Another one mentioned that: *"I considered what I have read about grey literature"*.

Peer interaction **(5/34 occurrences; 14.7%)**. Another criterion considered for assessing GL Control was the users' interactions in GL sources. In this regard, one researcher mentioned: *"Another point is that if I have a lot of people interacting and building the content (such as Q&A websites), I consider that it has a certain control in the final knowledge presented there"*. Another one pointed out: *"In general, I consider the control to be higher when there is a peer review in some way, as in the case of theses and Stack Overflow"*.

(ii) Correlation analysis between the level of the dimensions (Control and Expertise) and each Grey Literature type

We conducted our analysis using correlation statistics between the two variables (Control and Expertise) to each GL type using the Spearman coefficient. We interpreted the Spearman coefficient according to Dancey and Reidy study's (DANCEY; REIDY, 2004). To conduct this analysis, aiming to pair the samples, we removed the answers in which one respondent answered that *"I did not consider it a GL type"* or *"I have no opinion"* to at least one dimension to the same GL type.

Based on the results of Spearman's rank correlation presented in Table 18, we identified 13

GL types (13/19; 68.4%), with correlations that varies from **strong to very strong positive correlations (p-value <= 0.05% of significance).** It indicates that when the Control's level increases, the Expertise tends to increase.

Considering only the group of GL types that presented **less than 95% of significance**, we identified six types. Among these types, 4 out of 6 (*forums/list of discussions*, *cases/services descriptions*, *keynote speeches*, *materials training*) had **moderate correlations**. For the remaining two (*books/book chapters* and *magazine articles*), we identified the **negligible correlations**.

Table 18 – Correlation test between the level of Control and Expertise of Grey Literature types analyzed in Survey 2.

Type of Grey Literature	Spearman coefficient	*P*-value
Blog post	**.441***	.017
Book/Book chapter	.106	.607
Case/Soft. description	.341	.082
Forum/Discussion list	.337	.069
Guideline	**.518***	.004
Keynote speeches	.305	.101
Magazine article	.167	.377
Manual	**.620***	.000**
Material training	.308	.104
News articles	**.525***	.003
Patent	**.550***	.027
Q&A websites	**.656***	.000**
Slide presentation	**.593***	.001
Soft. Repository	**.652***	.000**
Technical report	**.527***	.005
Thesis	**.546***	.013
Tutorial	**.688***	.000**
Video	**.671***	.000**
Whitepaper	**.769***	.000**

Notes:
*Correlation is significant (strong) at the rho >= 0.4 and p-value <= 0.05 level;
***p-value* is not zero (we used three decimal places.

Source: the author.

(iii) Correlation analysis between the level of the dimensions (Control and Expertise) and the respondent profiles

After analyzing our data, a chi-square test of independence was conducted between the respondent profiles and their inclination to answer *"I did not consider it a GL type"* or *"I have no opinion"*. Therefore, we evaluated if the fact that the respondent is a professor or not has any influence in not considering as GL or not having an opinion. Table 19 presents our result.

Table 19 – Chi-square test between respondent profiles analyzed in Survey 2.

Type of GL	i	ii	iii
Blog post	.769	.526	.959
Book/Book chapter	.925	.959	.526
Case/Soft. description	.959	.959	.439
Forum/Discussion list	.959	.999	.959
Guideline	.526	.526	.579
Keynote speeches	.959	.999	.959
Magazine article	.959	.526	.959
Manual	.769	.526	.769
Material training	.959	.999	.769
News articles	.959	.999	.769
Patent	.883	.393	.726
Q&A websites	.769	.526	.959
Slide presentation	.959	.999	.925
Soft. Repository	.769	.999	.769
Technical report	.959	.769	.769
Thesis	.526	.999	.194
Tutorial	.526	.999	.579
Video	.959	.769	.579
Whitepaper	.959	959	.711

Notes:
Column "i" shows the opinion of *Not considered as GL*,
Column "ii" shows the number of researchers with *No opinion about Control*.
Column "iii" shows the number of researchers with *No opinion about Expertise*.
Source: the author.

As we can see in Table 19, we did not have found a statistically significant association ($p < 0.05$) between respondent profile and their inclination to have no opinion regarding the level of Control and Expertise, or did not consider a GL type. Therefore, based on our results, we

did not reject any null hypothesis, i.e., the respondent profile did not influence their answers, or our sample is not large enough to show this influence.

We performed another Chi-square statistical test to discover if the respondent profiles affect results to their opinion on Low, Moderate, or High level of Control and Expertise. For each factor (Control or Expertise) and GL (e.g., *blog posts*, *books/book chapters*), we populated a 2X3 contingency table composed of rows (i.e., respondent profile) and columns (i.e., their opinion as Low, Moderate, or High) variables. Table 20 presents the p-value from the chi-square statistical test for each contingency table.

Table 20 – Chi-square test between respondent profiles and (i) Expertise and (ii) Control levels analyzed in Survey 2.

Type of GL	Expertise	Control
Blog post	.785	.100
Book/Book chapter	.958	.722
Case/Soft. description	.632	.293
Forum/Discussion list	.720	.557
Guideline	.769	.853
Keynote speeches	.185	.853
Magazine article	.539	.692
Manual	.496	.069
Material training	.316	.690
News articles	**.049**	.205
Patent	.651	.905
Q&A websites	.567	.289
Slide presentation	.478	.157
Soft. Repository	.387	.261
Technical report	.848	.743
Thesis	.746	.844
Tutorial	.132	.707
Video	.755	.894
Whitepaper	.925	.752

Source: the author.

Table 20 shows the distribution of the p-values per comparison from each Chi-squared test of independence. As we can see, there is no evidence that different respondent profiles have different opinions. The only exception regards *news articles* credibility. The contingency table (see Table 21) summarizes the results from comparing answers from professors/students and

news articles credibility. We conclude that students think that *news articles* are more believable by analyzing this result.

Table 21 – Contingency table from respondent profiles and the levels of Expertise for "News articles" analyzed in Survey 2.

Respondent profile	Low	Moderate	High
Professors/researchers	7	1	0
Students	8	13	0

Source: the author.

Summary of RQ1.6: We identified similar behaviors when considering the same GL type concerning the two dimensions: Control and Expertise. Most GL types ran between the Low and Moderate levels in these dimensions. We also identified some differences, such as the median of answers for Control were at the Low level and a Moderate level for the Expertise dimension. The production rigor, the producer's reputation, researcher experience, and the permission of peer interaction are the criteria employed by the researchers to assess GL source. Moreover, we identified some misunderstandings to consider or not some data sources as GL, mainly related to *thesis*, *patents*, *magazine articles*, and *books/book chapters*. Considering the correlation analysis, we identified that it varied from strong to very strong between Control and Expertise dimensions for most GL types. Our investigation also shows a correlation analysis between the level of Control and Expertise for most GL types, showing that when one dimension increases, the other one tends to increase too. The same happens when the level decrease. Considering the researcher profile, we did not find evidence that different researcher's profiles have different opinions, except for the *news articles*.

3.4 DISCUSSION

This section discusses each research question, relating them with previous studies (Section 3.4.1). We provide a discussion about some findings unrelated to a specific research question (Section 3.4.2). We present some lessons learned with this research (Section 3.4.3). Finally, we discuss some threats to validity (Section 3.4.4).

3.4.1 Revisiting Findings

In this section, we discussed our findings to each RQ.

(RQ1.1) Motivations to use or reasons to avoid GL

(i) Even though our first investigation showed several motivations and benefits in using GL. Our second investigation shows that most researchers avoid its use as a reference in scientific papers.

(ii) We organized the motivations to use GL into five categories. Only the motivation *"to complement research findings"* was similar to previous works (RAINER; WILLIAMS, 2019b; GALINDO NETO et al., 2019; ZHANG et al., 2020).

(RQ1.2) Types of Grey Literature used

We did not find previous primary studies focusing on this research question. We identified tertiary studies that investigated the most GL types used in selected studies. For instance, Zhang and colleagues (ZHANG et al., 2020) identified that the most common GL types used in the list of selected secondary studies were (in order) *technical reports*, *blog posts*, *books/book chapters*, and *thesis*.

Considering the GL types used by Brazilian SE researchers, the most common are the *Q&A websites* (e.g., Stack Overflow), *blog posts* (e.g., SE firms, such as Netflix, Uber, Facebook), and *technical reports* (e.g., from SEI). Our investigation shows that most of these types are related to SE practice, mainly retrieved from renowned firms or research institutions.

(RQ1.3) Criteria used to assess Grey Literature credibility

We identified several criteria to assess the GL credibility, showing that most of them are related to the GL producer *be renowned (authors, institutions, and companies)*. These criteria caught our attention because we did not find any criterion mentioning to assess the GL content. However, the challenge of *Lack of reliability* identified is related to this, and previous work (WILLIAMS; RAINER, 2019) have investigated a set of criteria to assess GL content (e.g., *rigor in presenting information, presenting empirical data, describe the methods of data*

collection).

(RQ1.4) Benefits and Challenges using Grey Literature

We identified some contradictory findings between the benefits and challenges of GL use. They are part of the trade-off between traditional literature and GL nature. For instance, on the one hand, SE researchers mentioned that it is *Easy to access and read* the GL content. On the other hand, they said the *Difficult to search/find information*. Regarding the benefit, it is related to accessing the GL content without paywall restriction and to the informal language usually written. However, these benefits hinder the use of automatic data extraction.

We identified another trade-off, for instance, even though the perceived benefit of *Advances the state of the art/practice*, several researchers are avoiding the use of GL due to the challenges of *Lack of reliability* and *Lack of scientific value*. In part, those trade-offs are expected, showing the necessity for further investigations on how to improve the use of GL in SE research. For instance, as we have done in this research.

(RQ1.5) Prioritizing the Criteria to Assess Grey Literature Credibility

This investigation confirmed some findings of Survey 1 (KAMEI et al., 2020), showing that the most important credibility criteria are related to the GL source be produced by a *renowned source*. However, using the prioritization criteria, some of these findings contrasted partly because, in Survey 1 results, no criteria were related to assessing the GL content. At the same time, in Survey 2, several SE researchers considered important criteria of *Citing academic sources* and *Presenting empirical data*.

The criteria of citing academic sources, describing the collection methods, and presenting empirical data caught our attention due to the emphasis on applying scientific perspectives to assess GL sources. In our opinion, these criteria are difficult to be used, as we discuss in the following: 1) According to Wiiliams (WILLIAMS, 2018), online articles and blogs produced by SE practitioners rarely mentioned academic sources; 2) GL sources are produced mainly by practitioners (KAMEI et al., 2021b), and consultant/companies have different manners of expressing than academics one; and 3) Most of the GL sources do not present empirical data. Instead, they are primarily based on their opinions and belief (RAINER, 2017).

(RQ1.6) Types of Grey Literature vs. Dimensions of Control and Expertise

Some findings caught our attention because some GL types run between two and sometimes into three levels of the classification of the dimensions, showing that different interpretations may occur for the same type. Although, the correlation analysis showed a strong correlation between these interpretations for most of the GL types investigated. Considering the respondent's profiles, different from what we expected, our statistical analysis based on the Chi-square test showed that different respondent profiles shared similar opinions about each source type being considered a GL or not and concerning the level of control and expertise.

The criteria used by SE researchers to classify these dimensions are mostly related to the rigor of source, researcher experience, and the interaction permitted for the user to deal with each GL type. Although some of them considered it is challenging to classify considering only the source type, without a real example to be deeply assessed, as one researcher pointed out: *"[. . .] the credibility will depend on who produced that content"*. Moreover, we perceived that sources (e.g., *technical reports*, *books/book chapters*, *thesis*) produced by companies and institutions mainly were considered with Moderate to a High level of Control and Expertise. In contrast, the sources commonly produced by SE practitioners (e.g., *forums/list of discussions*, *blog posts*, *videos*) have a Low level of Control and Expertise. These findings caught our attention because, in RQ1.2 results, the most used GL sources runs between Low to Moderate level. It appears that the benefits and the motivations to use GL outweigh the Low level of Control and Expertise presented in these sources.

With these findings, we reinforce the claim of Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019) that it is complicated to assess the dimensions of Control and Expertise alone. Although they could bring us one direction, other essential criteria include identifying GL's producer and content. For this reason, we advocate that SE researchers use the concept of the "shades of GL" (that considered that GL has a diversity of types that differs according to the three-tiers - low, moderate, high – according to level of control and expertise. It was better explained in Chapter 2) to classify and assess a GL source because it recognizes the different perspectives of the nature of GL, although future investigations to set a limit between tiers of the shades are essential. Beyond that, we claimed the importance of employing objective criteria to assess GL sources and better permit the GL classification according to the shades. Although, as our findings showed, it could be essential to propose intermediate shades between each tier.

3.4.2 Other discussions

In this section, we discussed some findings and important discussions unrelated to a specific research question. First, we discussed the relations among the researcher's perceptions' of GL. Second, describe the relationship between the credibility criteria and the dimensions of credibility investigated. Lastly, discuss our findings of the perceptions of the different GL types.

Perceptions of Grey Literature

We identified relations between the perceptions of GL, as shown in Figure 6. For instance, some *motivations to use* GL related to some *benefits* identified (slashed line) and some *reasons to avoid* GL with some *challenges* by GL use (dotted line). In what follows, we discussed some of them.

Regarding the *motivation to use* "To complement research findings" is related to the *benefit* of use GL to provide "Covers different results from scientific studies" as some respondents informed that the inclusion of GL could provide evidence not explored or identified in the research area. Another one is "To answer practical and technical questions" related to the benefit of "Practical evidence", which was not perceived using only traditional literature.

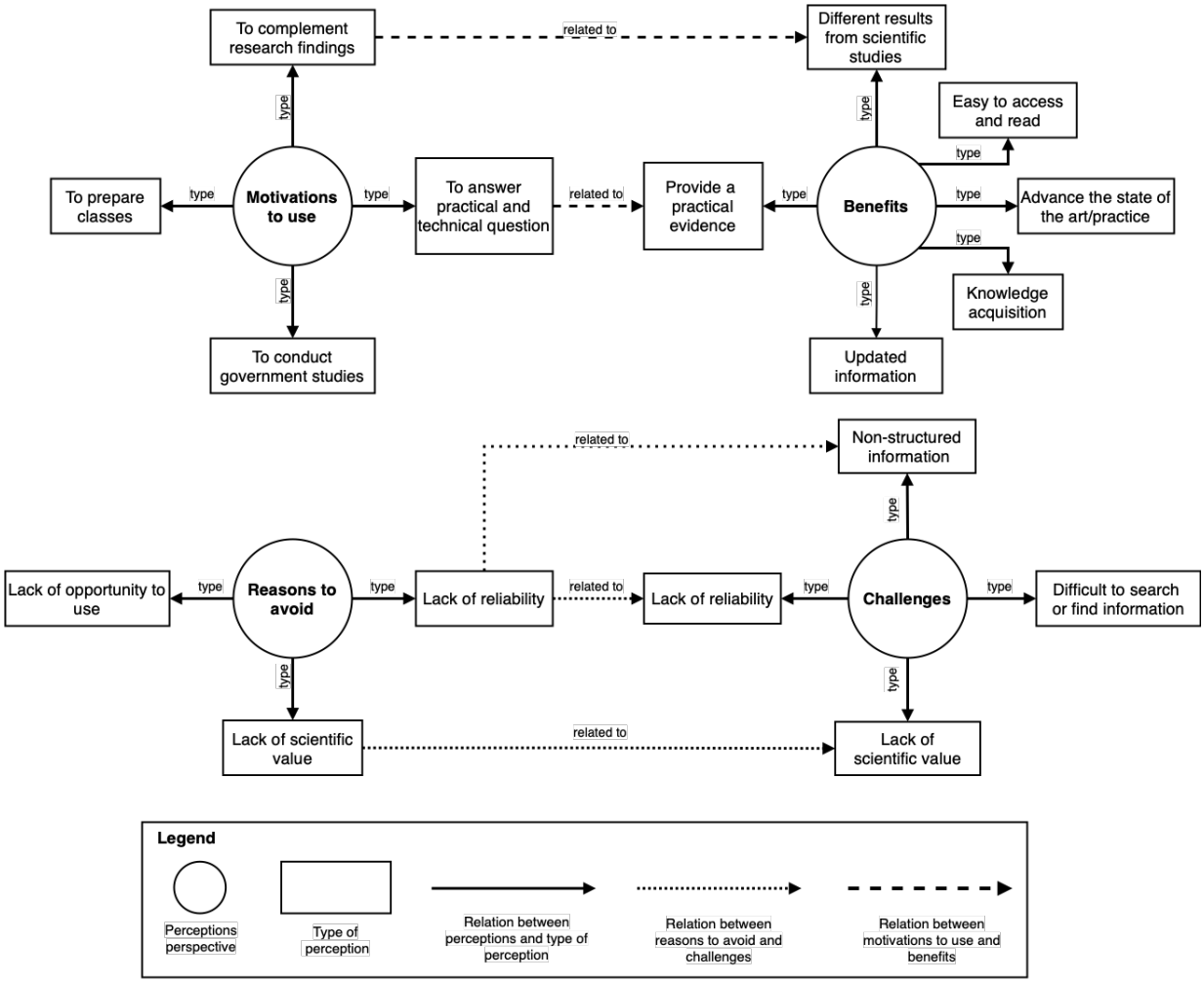
The *reasons to avoid* GL and the *challenges* identified are almost the same. Except for the "Lack of reliability" that hinders the replicability of the search for GL. It could be motivated due to the "Non-structured information" of a GL source.

Credibility criteria vs. Dimensions of Control and Expertise

The most important criteria identified to assess GL credibility are related to the "Producer reputation" and the "Rigor" presented in the GL source. The first is related to the source be produced by a renowned author, institution, or companies. The second with how the information is presented, for instance, if it describes the methods used to collect the data. Figure 7 presented these criteria.

We also identified some relations between the credibility criteria with some reasons to classify the Control and Expertise dimensions, as shown in Figure 7. The Control (slashed line) is related to the "peer interaction", "producer reputation", and the "rigor". The Expertise (dotted line), their relations are the same as the Control dimension, including the "researcher

Figure 6 – Relationships identified between the Motivations to Use Grey Literature with Benefits and the Reasons to avoid with the Challenges, analyzed in Survey 2.



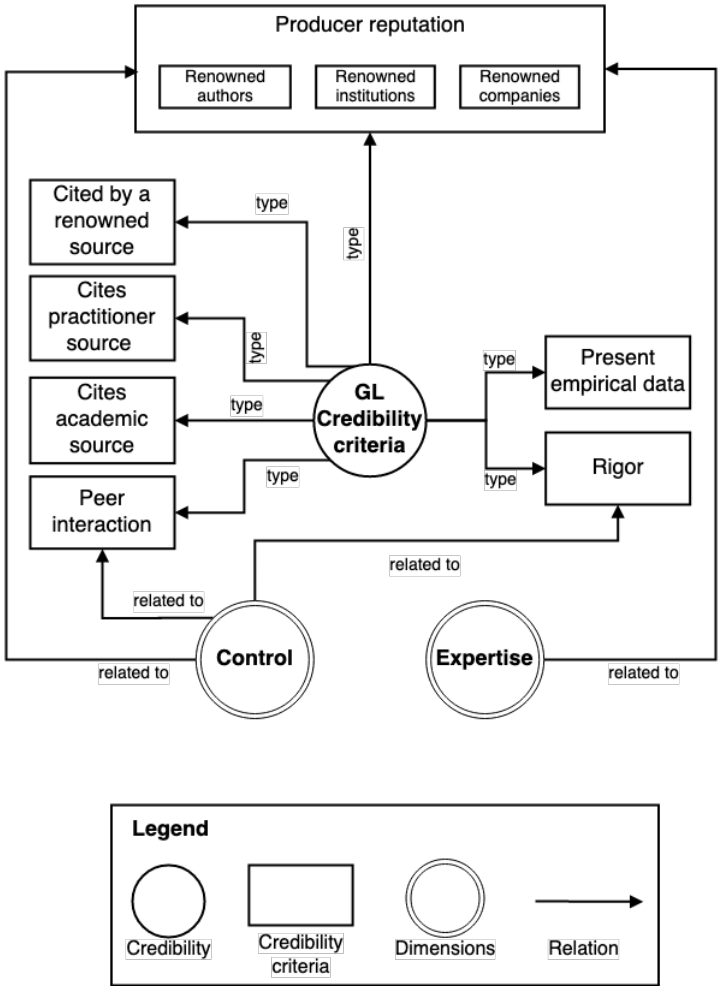
Source: the author.

experience". This last is related to the researcher's experience using GL to assess its credibility.

Grey Literature types interpretation

In our second investigation, we identified some misunderstanding in interpreting GL types (see Table 16), even though those types were recognized as GL in some previous SE works (e.g., Maro et al. (MARO; STEGHÄFER; STARON, 2018), Zhang and colleagues (ZHANG et al., 2020)). In the following, we present the most common types that were not considered GL: *thesis* (11/34 occurrences), *patents* (6/34 occurrences), *books/book chapters* (6/34 occurrences), and *magazine articles* (3/34 occurrences). In this regard, for instance, one researcher pointed out: *"I understand that thesis and dissertations are not Grey because external researchers formally assess them"*.

Figure 7 – Relationships between the Grey Literature Credibility criteria with the Dimensions of Control and Expertise identified in Survey 2.



Source: the author.

We also identified in previous studies some contradictions in interpreting a source type as a GL type or not. For instance, while Hosseinzadeh and colleagues (HOSSEINZADEH et al., 2018) considered *books/book chapters* as a GL type, Berg et al.'s study (BERG et al., 2018) did not. We identified another conflict, for instance, while Galindo Neto et al.'s study (GALINDO NETO et al., 2019) considered *thesis* a peer-reviewed source, Rodríguez-Pérez et al. (RODRíGUEZ-PéREZ; ROBLES; GONZáLEZ-BARAHONA, 2018) classified them as GL types. These misunderstandings were also identified in our Tertiary Study 1 (KAMEI et al., 2021b).

In our opinion, these misunderstandings reflect on each source's classification regarding Control and Expertise. For instance, for most researchers, *books/book chapters*, *technical reports*, *thesis*, and *patents* were not considered a GL type and related them to a High level of Control and Expertise (figures 4 and 5). It shows that the peer-reviewed process and grey literature boundary are unclear when considering only the source type.

3.4.3 Lessons Learned

With these investigations, we showed how GL could contribute to SE research. However, some advice is important to its use could be improved.

For **SE researchers,** we highlight the importance to pay attention when search, select, and use grey literature in SE research, focusing on: 1) As there are several GL types source, it is important that researchers explore the GL sources before using it on their research, focusing on understand, in general, what information could be retrieved in each source type, how each one could benefit the research, and how to retrieve information from them. This last is due to the issues about the difficulty to search in GL sources that may interfere with the search process.; 2) The researchers need to be aware of a set of credibility criteria that could be used to assess GL sources. For instance, by selecting data produced by renowned sources (e.g., authors, institutions, companies) and understanding how each credibility criteria could better fit each type of GL; 3) Another criterion to improve GL credibility could be used, considering the various interpretations for GL assessment related to the Control and Expertise aspects; and 4) Understand how to improve the search for GL using a systematic approach with methods and techniques to better deal with the content, improving their reliability.

For **Practitioners**, our findings show the importance of the content provided by them for the research community. However, for this information be consumed by researchers and to create a relevant impact on academia, we include some advice for practitioners: 1) Substantiate the data presented in an accessible language and with detail of information, e.g., explaining the context and making the used data available; 2) Adopting some quality criteria to improve the credibility of their content, e.g., use a checklist to verify if the information is well described; and 3) Adopting a pattern to provide information makes easier to retrieve information using an automatic approach. We understand that it is difficult to apply them since the gap exists between the industry and academia (CARTAXO et al., 2016).

3.4.4 Threats to Validity

We organized our threats to validity according to the classification proposed by Kitchenham and Pfleeger (KITCHENHAM; PFLEEGER, 2008). Although, only two types of validity fit to our research: content validity and construct validity.

Content Validity. We took some care to design and improve our questionnaires before sending them to the respondents in our surveys. For Survey 1, we invited an experienced researcher (Ph.D. SE researcher with more than 15 years of experience in research) to review our draft questionnaire. We also conducted a pilot study by randomly selecting two participants and asking for their feedback. We received feedback suggesting changing the order and rewriting some questions to make them more understandable to the target population. For Survey 2, we improved our questionnaire through three approaches: we conducted a pilot study with five experienced Ph.D. SE researchers; invited a SE researcher specialist in the survey method; and collected feedback from a participant.

Construct Validity. Even though our efforts to improve our questionnaire, we identified one potential threat in our research, specifically on the questions that we asked for the participant to classify each source type concerning the Control and Expertise dimensions. We mitigate this, informing the researchers that we know that Control and Expertise vary from source to source, and asked them to consider the most frequent experience for each data source. However, three researchers reported that assessing these GL types' dimensions was difficult without considering the content and the producer. This difficulty may have introduced some bias.

3.5 SUMMARY

This chapter described our investigation with 76 Brazilian SE researchers based on two surveys. Using qualitative and quantitative approaches, our findings reveal that most the researchers are motivated to use GL, however, with some restrictions because there are some challenges that researchers may face when they use GL. We identified a diversity of benefits, but also challenges were identified that need the attention of the SE research community. In addition, we identified that due to the diverse forms of GL, SE researchers assessed them in different forms, considering the level of control and expertise of their content and who produced the source.

By describing our findings, we expect to provide additional evidence to understand the motivations better to use GL and what criteria SE researchers could use to assess its credibility.

We inform that the results presented in this chapter were published in the 4th Brazilian Symposium on Software Engineering (SBES 2020) (KAMEI et al., 2020) and accepted to the

Journal of Software Engineering Research and Development (JSERD) (KAMEI et al., 2022).

4 CRITICAL REVIEW OF GREY LITERATURE IN SECONDARY STUDIES

This chapter presents a tertiary study to contribute to the literature, by compiling additional evidence on the *use*, *motivations*, *benefits*, and *challenges* of using Grey Literature (GL) in Secondary Studies. To achieve this aim, we investigated 446 Secondary Studies, of which we conducted an in-depth analysis of the GL use in 126 studies and answer the RQ2. Our work is unique because we explored the criteria employed in Secondary Studies to select studies, the coverage of GL to support answers to research questions, the motivations of studies in their decision to use or to avoid GL, and the definitions of GL used in different circumstances.

In the following, Section 4.1 presents our research questions. Section 4.2 provides the procedures used to conduct, collect, extract, analyze, and synthesize the data. Section 4.3 presents our findings. Section 4.4 presents the discussions about this research. Finally, Section 4.5 presents a summary of this research.

4.1 OVERVIEW

To achieve the stated goal, we explored the following research question:

RQ2: *How do Secondary Studies Use Grey Literature?*

The above research question was divided in six more specific questions, they are:

RQ2.1: *What Definitions of Grey Literature Are Employed in Secondary Studies?*

Rationale: Even exists a common definition for GL, called as Luxembourg definition, Farace and Schöpfel (FARACE; SCHÖPFEL, 2010) affirmed that its definition needs to be refined and/or redefined by a way of an accurate analysis of new means of access and distribution. According to Bonato (BONATO, 2018), GL definition varies from area to area. In SE research, as GL is a recent topic, the work of Zhang and colleagues (ZHANG et al., 2020) did not find a common definition for GL used by researchers that might influence its use. For Bonato (BONATO, 2018), it is important to define it to make it easier to search and assess a GL source. This research question investigated the way GL is defined by SE researchers. Answering this question is essential to improve the state of the art of GL in SE.

RQ2.2: *How Is Grey Literature Used in Secondary Studies?*

Rationale: When conducting Secondary Studies, researchers often employ inclusion criteria to filter out non-peer-reviewed works. This decision is motivated by the fact that peer-reviewed work is considered more reliable than not peer-reviewed work (SALAH; PAIGE; CAIRNS, 2014). However, more recently, a growing number of researchers are arguing that not peer-reviewed work could also be incorporated into scientific studies (GAROUSI; FELDERER; MÄNTYLÄ, 2016). Some studies even go further and consider not peer-reviewed works as the exclusive data source (SOLDANI; TAMBURRI; VAN DEN HEUVEL, 2018). In this research question, we sought to understand if Secondary Studies are using GL and, if so, how. More precisely, we aim to investigate: (i) Frequency of studies using GL; (ii) Frequency of GL use between the types of Secondary Studies that used it; and (iii) Frequency to which GL is used to support answers to research questions.

RQ2.3: *How Is Grey Literature Searched, Selected, and Has the Quality Assessed in Secondary Studies?*

Rationale: Using or searching for GL is not a trivial task. Some reasons for this lie in the diversity of its sources. GL differs in the type of structure and content provided by each source. This could even make it difficult to search for or find information in GL (KAMEI et al., 2020). In this research question, we sought to investigate the methods employed to (i) search, (ii) select, and (iii) perform a GL quality assessment. A better understanding of these procedures would be essential to guide future research in this area.

RQ2.4: *What Types of Grey Literature Are the Most Frequently Used in Secondary Studies?*

Rationale: GL is available in many forms, varying from traditional mediums, such as books and technical reports, to more dynamic mediums, such as forums and Question & Answer websites. These mediums offer researchers a rich spectrum of unstructured data, which brings specific benefits and limitations. However, since these mediums are often hosted on specific platforms (e.g., from official websites to personal blogs), it is not clear if (and how long) this information will be accessible. In this research question, we sought to investigate more concretely: (i) The different GL types used; (ii) Who are the producers; and (iii) The availability of the GL. A

better understanding of the GL source would be important to guide future research in this area.

RQ2.5: *What Motivates Researchers to Use/Avoid Grey Literature?*

Rationale: The process of conducting a literature review is far from trivial. It is important to plan according to the research focus. In this question, we want to understand if researchers are properly justifying GL's use in their research. More concretely, we aim to investigate the: (i) Motivations to use and (ii) Reasons to avoid GL. We know aware that this information is sometimes not explicitly described in the research papers, but rather implied in the context. Answering this question is important for SE researchers who do not know how GL could improve their research.

RQ2.6: *How do Researchers Perceive the Use of Grey Literature?*

Rationale: Some researchers advocate for GL because of some known benefits, including practical orientation and the appeal to complement formal literature. In this research question, we aimed to complement the state-of-art with the (i) perceived benefits of using GL, and wait to uncover eventually (ii) challenges related to its usage. Answers to this question could support SE researchers in understanding potential benefits that could be unlocked (and the challenges behind them).

To answer these research questions, we performed a tertiary study focusing on synthesizing GL use in Secondary Studies, using automatic and manual searches to identify potential studies published between 2011 and 2018. Our search process identified 20,181 studies, from which we identified 446 Secondary Studies that fulfilled our eligibility criteria. We noticed that 126 out of 446 Secondary Studies used or searched for GL. We used these two sets (446 and 126 Secondary Studies) for a more in-depth analysis, using qualitative and quantitative approaches.

In summary, our main findings in this research are:

- We perceived that GL is not extensively used in Secondary Studies of SE (126/446), although we noticed a growth over the years;
- We identified that at least 75% of the studies (95/126) used GL to support answers to at least one research question, even GL represents less than 21% of all 446 Secondary

Studies;

- We perceived that the understanding about GL types were sometimes controversial elements among the studies;
- We identified that almost 50% of the GL reported in investigated studies (n=126) are now unavailable;
- We showed few studies (14/126) employing specific criteria to search for and additional criteria for assessing the GL quality (7/126);
- We identified that consultants and companies were the ones that most produced the GL sources analyzed. Although, we perceived an increasing amount of content produced by practitioners over the years;
- We elucidate diverse challenges in dealing with GL that SE researchers may have to face. Nevertheless, we provided a potential list of ways to address that could help SE researchers to deal with each one.

By describing these findings and a list of challenges with potential ways to deal with them for SE researchers, we expect to help others better conduct Secondary Studies using GL to take advantage of SE practice.

4.2 TERTIARY STUDY

In this tertiary study, we followed the guidelines of Kitchenham et al. (KITCHENHAM; BUDGEN; BRERETON, 2015). This research started at the end of 2018 and finished at middle of 2020. In what follows, we present the procedures employed to search for the studies and the steps taken for selection, data extraction, and data analysis.

For replication purposes, the data used in this chapter is available online at: <https://doi.org/10.5281/zenodo.6780520>.

4.2.1 Research team

A team of seven researchers conducted this research. Two are Ph.D. students, and five are full-time lecturers. Furthermore, two advisors acted as reviewers. These researchers are affiliated

with the Federal University of Pernambuco, Federal Institute of Alagoas, Federal University of Technology of Paraná, Federal Institute of Bahia, Federal Institute of Pernambuco, University Catholic Pernambuco, Federal University of Alagoas, Federal Institute of Pernambuco, and the Federal University of Pará.

4.2.2 Search Strategy

We began the search procedures for Secondary Studies at the beginning of 2019, using a combination of three approaches: (i) a selection of previous tertiary studies, (ii) automatic searches in digital libraries, and (iii) a manual analysis of a small selection of premier SE journals and conference proceedings.

Since we were not interested in specific studies (e.g., a tertiary study about pair programming), we started by searching for tertiary studies that focused on general aspects of SE. We selected the following tertiary studies: Kitchenham and colleagues (KITCHENHAM et al., 2009), Kitchenham and colleagues (KITCHENHAM et al., 2010), Da Silva and colleagues (SILVA et al., 2011), Cruzes and Dybå (CRUZES; DYBÅ, 2011b), and Galindo Neto and colleagues (GALINDO NETO et al., 2019). These studies covered Secondary Studies published until 2010, excluding Galindo Neto et al.'s study (GALINDO NETO et al., 2019), which covered only MLR and GLR studies until 2019. We complement the search with an automatic and a manual search to find studies published between 2011 and 2018. We opted for this range because the years before 2011 were covered by the previous tertiary studies, and chose 2018 as the end because we started our investigation at the beginning of 2019.

We used the most relevant SE digital libraries for the automatic search, as recommended by Kitchenham et al. (KITCHENHAM; BUDGEN; BRERETON, 2015), namely ACM Digital Library, IEEE Xplore, ScienceDirect, and Scopus.

For the manual search, we chose the most prestigious SE journals and conferences related to this research topic, namely:

- Journals: *ACM Transactions on Software Engineering Methodology (TOSEM), IEEE Transactions on Software Engineering (TSE), Empirical Software Engineering Journal (EMSE), Information and Software Technology (IST)*, and *Journal of Systems and Software (JSS);*
- Conferences: *International Conference on Software Engineering (ICSE), Empirical Soft-*

ware Engineering and Measurement (ESEM), and *Evaluation and Assessment in Software Engineering (EASE)*.

4.2.2.1 Search String

We took advantage of the same search string adopted by Da Silva and colleagues (SILVA et al., 2011) to conduct our search. In our case, we adapted this search string to cover additional terms, such as Grey Literature Reviews and Multivocal Reviews. The updated search string is as follows:

> *("software engineering") AND ("review of studies" OR "structured review" OR "systematic review" OR "grey review" OR "grey literature" OR "gray review" OR "gray literature" OR "multivocal literature" OR "multi-vocal literature" OR "literature review" OR "literature analysis" OR "in-depth survey" OR "literature survey" OR "meta analysis" OR "past studies" OR "subject matter expert" OR "analysis of research" OR "empirical body of knowledge" OR "overview of existing research" OR "body of published research" OR "evidence-based" OR "evidence based" OR "study synthesis" OR "study aggregation")*

The Scopus library has a limited length on its search field. In this case, we had to break down the search string into 24 sub-queries. After we retrieved all studies from the search procedures, we organized them all onto a shared spreadsheet, which we used to conduct the next methodological steps.

4.2.3 Selection Criteria

When manually investigating the retrieved papers, we focused on selecting Secondary Studies. We paid particular attention to the following kinds of Secondary Studies:

- Systematic Literature Review (SLR) (or Systematic Review);
- Mapping Study (MS) (or Systematic Mapping Study);
- Meta-Analysis (MA);
- Grey Literature Review (GLR);

- Multivocal Literature Review (MLR).

For each candidate paper, we applied a set of **exclusion** criteria. Table 22 describes each exclusion criterion. We excluded any candidate study that complies with at least one exclusion criterion. The only exception is the criterion EC1, which was not applied to the studies retrieved from previous tertiary studies.

Table 22 – List of exclusion criteria used in Tertiary Study 1.

#	Description
EC1	The study was published before 2011 or after 2018.
EC2	The study was duplicated.
EC3	The study was not written in English.
EC4	The study was not a full paper (e.g., position papers, abstracts, posters, etc.).
EC5	The study was not peer-reviewed (e.g., editorials, summaries, letters, keynotes, slides, etc.).
EC6	The study did not report a Secondary Study (i.e., SLR, MS, MLR, GLR, or MA).
EC7	The study was not related to Software Engineering (e.g., Information Systems and Computer Science).
EC8	The venue in which the Secondary Study was published did not have a minimum h5-index (20 for conferences and 25 for journals).

Source: the author.

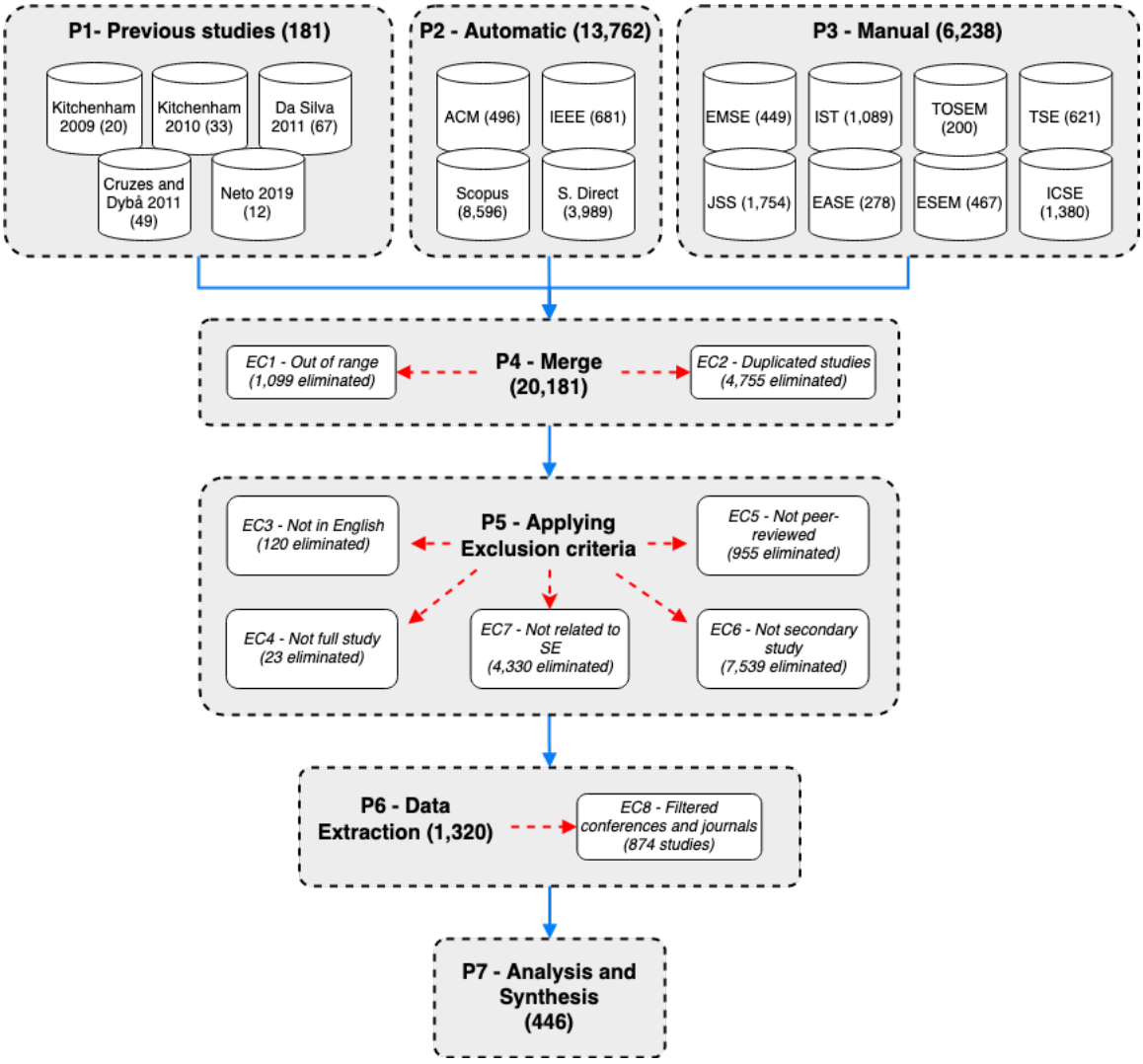
4.2.4 Study Selection

The procedure of study selection was conducted in seven sub-phases, as depicted in Figure 8. There is a number indicating each phase (**P1–P7**).

At phase **P1**, we selected a total of 181 Secondary Studies cited in the tertiary studies. We identified these studies at Kitchenham and colleagues (KITCHENHAM et al., 2009) (20 studies), Kitchenham and colleagues (KITCHENHAM et al., 2010) (33 studies), Da Silva and colleagues (SILVA et al., 2011) (67 studies), Cruzes and Dybå (CRUZES; DYBÅ, 2011b) (49 studies), and Galindo Neto and colleagues (GALINDO NETO et al., 2019) (12 studies).

At phase **P2**, we selected a total of 13,762 studies using the automated search. We identified these studies at ACM Digital Library (496), IEEE Xplore (681), Science Direct (3,989), and Scopus (8,596).

Figure 8 – Process of selecting studies in each phase of the Tertiary Study 1.



Source: the author.

At phase **P3**, we identified 6,238 studies using the manual search. These studies were identified at EMSE (449), IST (1,089), TOSEM (200), TSE (621), JSS (1,754), EASE (278), ESEM (467), and ICSE (1,380). The phases P1–P3 retrieved a total of 20,181 potential studies. Each potential study retrieved received a singular identification (ID).

Next, at phase **P4**, we sorted the Secondary Studies by title and organized them on a spreadsheet. We applied the EC1 and EC2 to remove the studies out of the range of our investigation and the studies with the same bibliographical information (i.e., title, abstract, and author(s)). For EC2 criterion, we employed the following next steps: (i) We compared paper titles; (ii) For papers with the same title, we looked at the abstracts and; if they are different, we considered the complete study as recommended by Kitchenham and Charters (KITCHENHAM; CHARTERS, 2007); if they are the same, we exclude one of them, if the publication years

are different, we excluded the least recent study. We removed 5,854 studies, 1,099 studies published before 2011 or after 2018 (EC1), and 4,755 instances of duplicated studies (EC2), respectively. At the end of this phase, 14,327 studies remained.

At phase **P5**, we read the studies thoroughly and applied the exclusion criteria (EC3–EC7) to all the 14,327 potentially relevant studies.

Determining whether a Secondary Study fits in the inclusion/exclusion criteria is a subjective task. To reduce the subjectivity and gain an alignment of understanding among the researchers involved in this work, we applied a pilot of the criteria to a small set of the selected studies. Since it would be too time-consuming to apply the exclusion criteria in pairs for all selected works, we applied those criteria in pairs in a random sample of 21% (=3,030) of the total of studies. Six authors participated in this review process (the first author paired with the additional co-authors). Each author applied the criteria individually, and, in the case of disagreements, we discussed them in conflict resolution meetings. In the case that no agreement was achieved, a third author joined the discussion. To evaluate the agreement level, we performed an agreement analysis using the Kappa scale (LANDIS; KOCH, 1977). Kappa scores are generally interpreted as slight (≥ 0 and ≤ 0.20), fair (≥ 0.21 and ≤ 0.40), moderate (≥ 0.41 and ≤ 0.60), substantial (≥ 0.61 and ≤ 0.80), and almost perfect (≥ 0.81 and ≤ 1.00). The Kappa value was *0.571*, which means a moderate agreement level. The remaining 79% (=11,297) of studies were applied individually.

Then there was the elimination of 13,007 studies based on the following criteria: 120 studies not written in English (EC3); 23 studies not reaching the status of a full paper (EC4); 995 not peer-reviewed studies (EC5); 7,539 studies that did not report a Secondary Study (EC6); and 4,330 studies not related to SE (EC7). At the end of this process, 1,320 Secondary Studies remained.

At phase **P6**, we applied the EC8 criterion to filter the studies from the top conferences and journals, there are remaining **446 Secondary Studies** to be analyzed. We used this criterion to select studies published in venues with a high potential impact on academic research and industry with international coverage, and eliminate studies published in predatory venues. Then, we extracted data (see Section 4.2.5) of all the information needed from those studies to answer our research questions.

Finally, at phase **P7**, we conducted the data analysis and synthesis (see Section 4.2.6) employing a qualitative and quantitative approach.

We separated and analyzed the selected studies (n=446) into two samples: **126 Secondary**

Studies using GL (see Appendix C) – the characteristics of these studies are in Appendix D – and **320 Secondary Studies that did not use GL** (see Appendix E). In Section 4.2.6, we explain how each sample was used in our analysis.

4.2.5 Data Extraction

We extract data relevant to answer each RQ. For RQ2.2, we extracted from each study the general information proposed by Da Silva and colleagues (SILVA et al., 2011). For the remaining RQs, we extracted similar data to that reported by the study of Zhang and colleagues (ZHANG et al., 2020). Some data extracted from each study includes (but is not limited to): (i) names of authors, (ii) year of publication, (iii) authors' institution, (iv) institutions' country, (v) quantity of included studies, and (vi) motivations to use or reasons to avoid GL. In addition, considering each study that included GL, we extracted the following information: (i) GL type, (ii) whether the GL data is still available online.

4.2.6 Data Analysis and Synthesis

We employed a mixed-method approach based on both qualitative and quantitative methods to analyze data. We used a *qualitative* approach when we were interested in questions about "what" and "how". To complement this qualitative analysis, we used descriptive statistics to discuss frequency and distribution.

4.2.6.1 Qualitative Analysis

Our qualitative approach followed a thematic analysis technique (CRUZES; DYBÅ, 2011a). Figure 18 presents a general overview of this approach. We describe it in greater detail in these next points (adapted from Pinto and colleagues (PINTO et al., 2019)):

(a) *Familiarizing ourselves with data:* Each researcher, involved in the data analysis, read (and re-read) every Secondary Study, as expressed in Figure 18-(a).

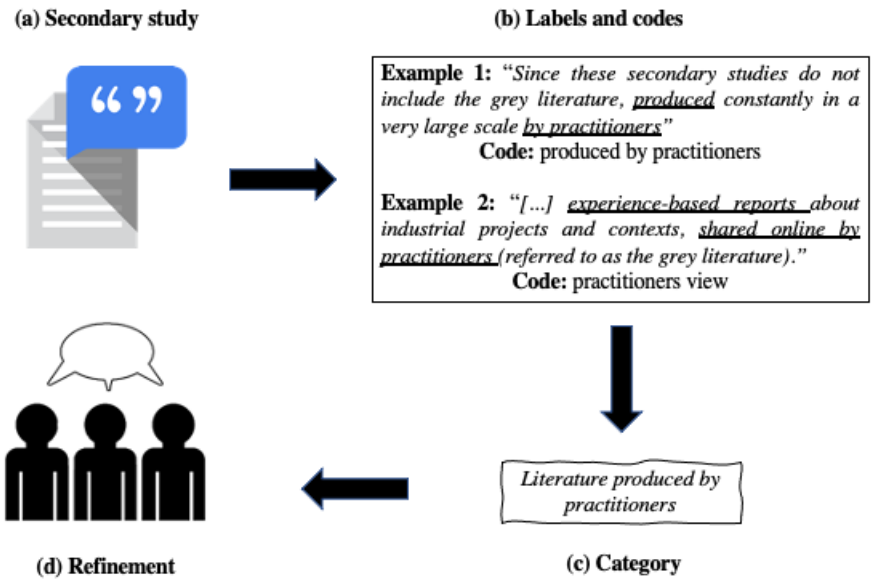
(b) *Initial coding:* In this step, each researcher individually added codes. We used a post-formed code. We labeled portions of text without any pre-formed code. Labels express the meaning of excerpts from the answer that represented appropriate actions or perceptions.

The initial codes were temporary, since they needed refinement. All identified codes were refined through a collective analysis. Figure 18-(b) presents an example of this analysis. Considering this example, there are two examples of portions of text extracted in which two codes were generated, *produced by practitioners* and *practitioners' views*. However, these codes were classified with a single code since they have the same meaning.

(c) *From codes to categories:* Here, we already had an initial list of codes. A single researcher looked for similar codes in data. Codes with similar characteristics were grouped into broader categories. Eventually, we also had to refine the categories found, comparing and re-analyzing them in parallel. Figure 18-(c) presents an example of this process. This example exhibits how the category "definitions of Grey Literature" emerged.

(d) *Categories refinement:* Having finished the previous step, we gathered a set of candidate categories. In this step (Figure 18-(d)), we involved three researchers: two to evaluate all categories, and a third researcher to resolve any disagreements (if needed). Besides aiming for accurate results, we renamed and regrouped some categories. The third researcher, once again, was invited to review and provide comments on those categories. In the cases of any doubt, we resolved them through conflict resolution meetings.

Figure 9 – Example of the coding process used to analyze the studies of Tertiary Study 1.



Source: the author.

4.2.6.2 Quantitative Analysis

We used two samples to our quantitative approach: (i) 446 Secondary Studies used to answer RQ2.1, RQ2.2, and RQ2.5 (*precisely the reasons to avoid its use*), and (ii) 126 studies that used GL answered the questions RQ2.3, RQ2.4, RQ2.5 (*specifically the motivations to use*), and RQ2.6.

We highlight that one Secondary Study could be related to more than one category found for quantitative analysis. Moreover, we calculated the percentage of answers based on the total population investigated (126 or 446). For example, to address RQ2.1, we considered all 446 studies, and we found 150 answers which were reported in 116 studies. Thus, we calculate to this category the value of 150/446.

4.3 RESULTS

In this section, we present our main results organized concerning the research questions. Section 4.3.1 presents the definitions of GL identified in the studies. Section 4.3.2 presents an overview of the usage of GL. Section 4.3.3 shows how SE researchers searched, selected, and conduct the quality assessment of GL in Secondary Studies. Section 4.3.4 presents the types and the producers of GL, and the availability of a GL source. Section 4.3.5 presents the motivations and the reasons to avoid GL reported in the Secondary Studies. Finally, Section 4.3.6 exposes the benefits and challenges presented in the studies by the GL use.

4.3.1 RQ2.1: What Definitions of Grey Literature Are Employed in Secondary Studies?

From the 446 selected Secondary Studies investigated, we identified 150 ones (33.6%) presenting some GL definitions. Among them, only 34 studies used general terms such as "grey or gray". On the other hand, however, 116 studies did not use any clear definition; instead, they used GL characteristics to express its definition. We identified the following categories (some studies were classified into more than one category).

Expressed by the types of Grey Literature **(108/446 studies; 24.2%)**. Some studies expressed GL regarding its types. For example, the study of Do Carmo Machado et al. (SS50)

pointed out that: *"Gray literature herein includes technical reports and book chapters"*. Similarly, Irshad et al. (SS42) informed that: *"New search terms were collected by browsing through Grey Literature (technical reports, non-peer-reviewed articles, and webpages)"*.

***Non peer-reviewed* (75/446 studies; 16.8%)**. This category groups "non-peer-reviewed" documents. For example, Tripathi et al. (SS34) excluded papers that did not pass through a peer-review process, as quoted here: *"[...] with its results, researchers and practitioners can consider both viewpoints that were non-peer-reviewed (gray literature) [...]"*. Li et al. (RQSS30) mentioned that *"A publication that has not undergone a peer-review is considered informal and not included"*, and is therefore GL.

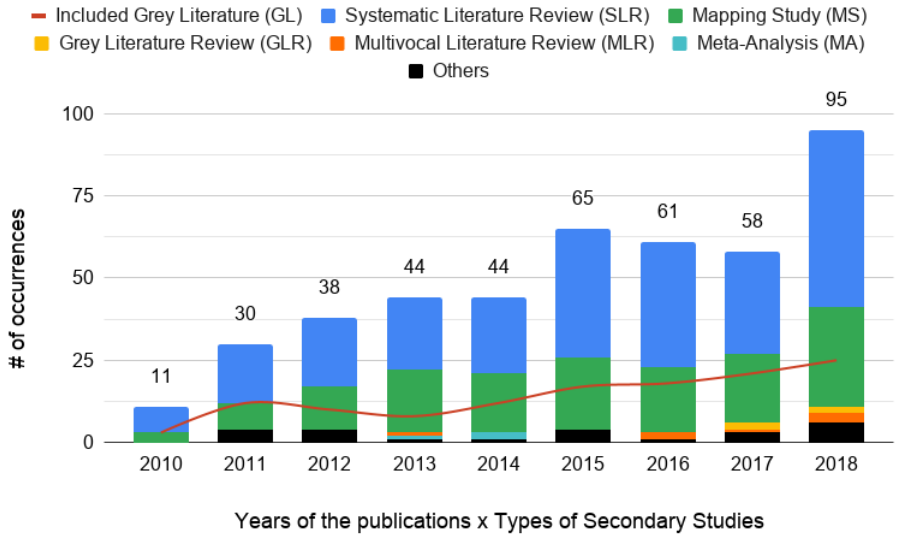
***Literature produced by practitioners* (40/446 studies; 9%)**. This category mentioned GL as literature published in the industry but not in academic settings. For example, Garousi et al. (SS98) pointed out that: *"These Secondary Studies do not include the grey literature produced constantly in a very large scale by practitioners"*. Raulamo-Jurvanen et al. (SS47) complements that: *"[...] experience reports about industrial projects and contexts, shared online by practitioners (referred to as grey literature)"*.

***Non-published literature* (24/446 studies; 5.4%)**. This category refers to studies not published, such as work in progress or non-indexed works. For example, Nadal et al. (SS9) mentioned it as we quoted: *"[...] five non-indexed works considered grey literature were additionally added to the list"*. In contrast, there was an exclusion criterion for the work of Mohabbati et al. (SS70) as we pointed out: *"[...] excluding 'gray publications' such as short papers, works in progress, unpublished, or non-verified literature"*.

***Others* (13/446 studies; 2.9%)**. We group here studies that employed other definitions, including the mapping study of Sharma and Spinellis (SS13) that mentioned GL as a secondary source, as quoted: *"We did not limit ourselves only to the primary studies. We included secondary sources of information and articles as and when we spotted them while studying primary studies"*. The research of Sharafi et al. (RQSS29) attributed GL to a lack of trust: *"Papers in 'grey' literature, which are not published by trusted, well-known publishers"*.

> **Summary of RQ2.1:** We perceived that most of the investigated studies did not use the term grey/gray literature. Instead, they refer to GL in terms of its types or characteristics (literature produced by practitioners and non-published literature).

Figure 10 – Distribution of each type of Secondary Study with a line trend of Grey Literature uses over the years investigated in Tertiary Study 1.



Source: the author.

4.3.2 RQ2.2: How Is Grey Literature Used in Secondary Studies?

(i) Usage of Grey Literature in Secondary Studies

Figure 10 presents the temporal distribution of 446 Secondary Studies, showing an increase in studies published over the years, along with the development of a trend line (represented by the red line) of studies using GL.

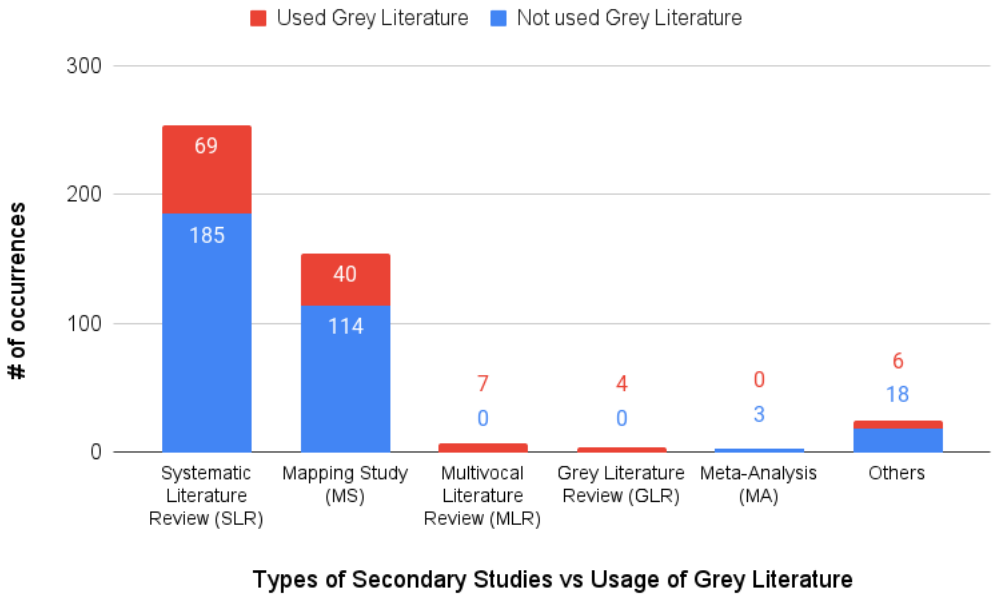
Overall, from the 446 selected Secondary Studies, we identified a subset of **126 studies that used GL (28.2%)**. To understand how our selected studies are using GL, we start by investigating the Secondary Studies methodological section, which often provides information about the search and selection procedures. As a shortcoming, we noticed that some studies did not mention such information, despite the lack of studies mentioning any criteria for using (or not using) GL. Our manual analysis found GL references in the list of selected studies.

(ii) Grey Literature vs. Types of Secondary Studies

Figure 11 shows a different scenario. It breaks down GL usage in terms of Secondary Study (i.e., SLR or MLR). As we could see, GL was not widely used in Secondary Studies.

Table 23 shows the distribution (%) of GL between included studies according to the

Figure 11 – Distribution of the studies using and not using Grey Literature, grouped by the Grey Literature types Secondary Studies identified in Tertiary Study 1.



Source: the author.

Secondary Study type, showing that for most of the Secondary Studies, there was the inclusion of up to 10% of GL studies. The MLR studies caught our attention because, for most of them, the GL studies represented over 50% of the studies included, despite we expected a different result from the other Secondary Study types, as MLR studies have a clear intention to search for GL sources.

Table 23 – Intervals of distribution of Grey Literature included over the total of studies included by each type of Secondary Study identified in Tertiary Study 1.

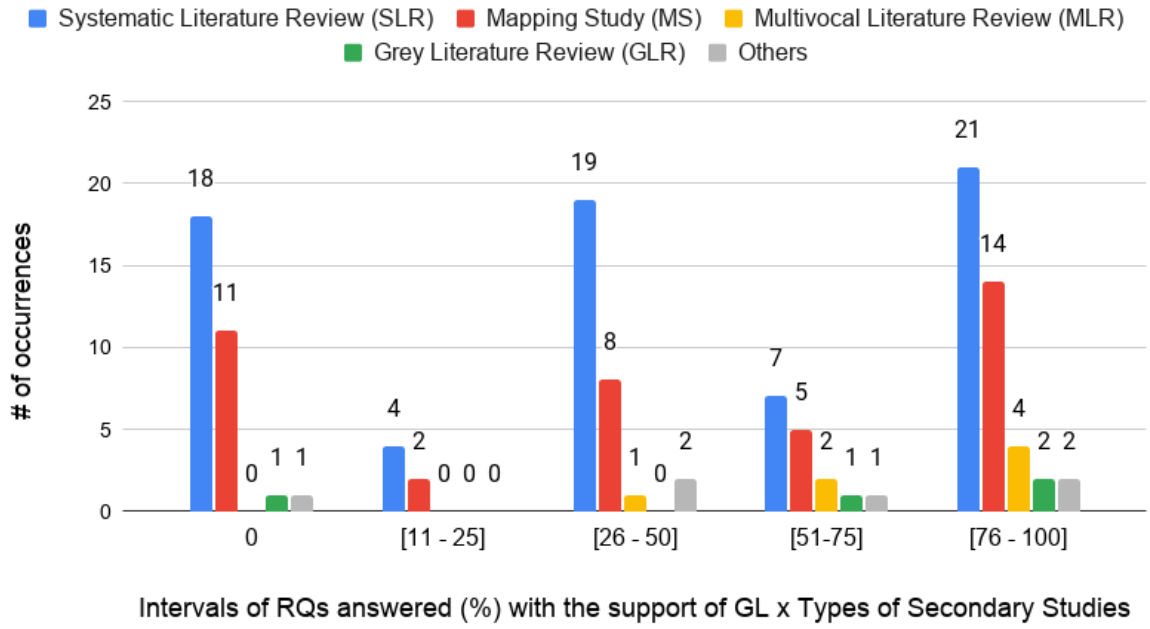
Type of Secondary Study	Interval of GL (%) among the primary studies				
	<=10%	11-25%	26-50%	51-75%	76-100%
Systematic Literature Review (SLR)	48	16	3	1	1
Mapping Study (MS)	31	4	4	1	0
Multivocal Literature Review (MLR)	0	0	3	3	1
Grey Literature Review (MLR)	0	0	0	0	4

Source: the author.

(iii) Usage of Grey Literature To Support Research Questions

We also investigated to what extent GL is used to support the answers to the research questions distributed between the types of Secondary Studies.

Figure 12 – Coverage (%) of research questions answered with the support of Grey Literature distributed for each Secondary Study type identified in Tertiary Study 1.



Source: the author.

Figure 12 shows the percentage of RQs answered by each type of Secondary Study using at least one example of GL. Our analysis identified **95 studies (95/126 studies; 75.4%) using GL to support answers to at least one research question**. However, by the individual analysis of each study, we perceived less than half of the studies using GL to answer more than half of the RQs. We perceived two interesting results: (i) We identified 31 studies (24.6%) that included GL but did not use them to answer any of the RQs, and (ii) RQs of six MLR studies (4.8%) were not answered with the support of GL.

We classified all the RQs explored in the 126 Secondary Studies. At total, 266 RQs were classified according to classification of Easterbrook et al.'s study (EASTERBROOK et al., 2008). We identified that almost 95% of the RQs answered using GL are Exploratory questions (existence, description and classification, descriptive-comparative).

Summary of RQ2.2: We perceived that Secondary Studies do not widely use GL, although there is an increase in its use over the years. We also identified that 75% of studies that included GL, used GL sources to answer at least one of their research questions (almost 95% are exploratory questions).

4.3.3 RQ2.3: How Is Grey Literature Searched, Selected, and Has the Quality Assessed in Secondary Studies?

In this section, we investigate the procedures used to collect, select, and analyze data. Firstly, we started by investigating if the study followed any guidelines to support its research. We found that Kitchenham and Charters' guidelines (KITCHENHAM; CHARTERS, 2007) were most used (67 studies) among Secondary Studies, followed by Wohlin's guidelines used to conduct snowballing (WOHLIN, 2014). We did not find a specific guideline to conduct GLR.

(i) Search Methods

We discovered that most of Secondary Studies (62.7%) used multiple search methods (automatic, manual, and snowballing) to search for studies. Table 24 describes the procedures employed, with the total number of occurrences of a given procedure in column "#". We found most Secondary Studies using *academic search engines* (e.g., ACM Digital Library, IEEE Xplore, Scopus, Science Direct, and SpringerLink) for data collection because most of them are studies using traditional SLR and MS focused on academic findings. The use of *Google Scholar* was the second most used search engine, followed by *Google*. All GLRs and MLRs have used *Google* as a primary search engine.

Table 24 – Procedures used to search for Grey Literature in Tertiary Study 1.

Source	#	%
Academic search engines*	102	80.9%
Google Scholar*	47	37.3%
Google*	28	22.2%
Microsoft Academic Research*	3	2.4%
Manual search	31	24.6%
Specialized databases	14	11.1%
Snowballing	50	39.7%

Note 1: "*" means automatic search.
Note 2: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Based on our sample analysis, from 10 GLR and MLR studies that intend to search for

GL studies, only two of them used different sources from the traditional Secondary Studies to search for GL. These findings show that **most of the studies did not apply any particular strategy to search for GL** (only 14 studies used specialized databases, e.g., Agile Alliance, YouTube, and Stack Overflow). Zhang and colleagues (ZHANG et al., 2020) also found this lack of a specific strategy to search for GL.

(ii) Selection Criteria

We found 20 studies did report their intention to include only peer-reviewed studies. On the one hand, 13 out of 20 studies described exceptions to the inclusion criteria, most of them related to specific GL types, for example, technical reports (e.g., (SS3)) and books (e.g., (SS28)). In the following sections, we present the groups of inclusion and exclusion criteria that emerged from our analysis.

- **Inclusion.** The most common inclusion criterion was to include *Specific GL types* found in 45 studies. For example, Fernandéz et al. (SS61) include Ph.D. theses and technical reports. Tiwari and Gupta (SS5) study mentioned that: *"The technical/experience reports, whitepapers, and books' chapters were searched by reviewing the references of the selected papers"*. Other studies (4 studies) used the criterion to include *industrial publication*, as quoted in the work of Lewis and Lago (SS17): *"A study that is in the form of a published scientific paper or industrial publication"*. Other inclusion criteria were also used but to a lesser extent. For example, Garousi et al. (SS112) specifically included GL from the *Seminal source*, as mentioned here: *"[. . .] one non-peer-reviewed technical report published by the Software Engineering Institute (SEI), which is considered a highly credible institute for software engineering research"*. Another criterion was to consider the *authors' credibility* to include specific studies, as quoted: *"Publishing companies or websites suggested by experts"*.

- **Exclusion.** The most common exclusion criterion was to exclude *Specific GL types* found in 33 studies as the inclusion criteria. For example, Pedreira et al. (SS27) searched for GL. However, personal blogs or web pages were excluded from the search. Mahdavi-Hezavehi et al. (SS19) also excluded specific GL types, as quoted: *"(Exclusion criteria) The study is an editorial, position paper, abstract, keynote, opinion, tutorial summary, panel discussion, or technical report"*. We found other exclusion criteria but to a lesser extent,

including reports with a *Lack of details* (e.g., from Quora, Slideshare, or LinkedIn), and *Web resources without keywords from the search string* (e.g., if one of the keywords was missing from the resource, it was automatically discarded). Another restriction was to exclude *websites without text* because they considered it hard to analyze them, as pointed out in the study of Tripathi et al. (SS34): *" If the webpage is only videos, audio, or images without text, it should be excluded"*.

(iii) Quality Assessment

We discovered that **only seven studies employed specific criteria to assess the quality of the GL.** Among those studies, we found three MLRs, one GLR, and three of the other types. The GLR conducted by Soldani et al. (SS113) employed specific criteria to assess the GL combined with inclusion/exclusion criteria to filter the studies. These criteria were grouped into four groups: *practical experience* measured in years of experience in the subject, *industrial case* that reported previous experience on the subject, *heterogeneity* of the results, and the *implementation quantity* that refers to the detail in which the results were discussed.

Tom et al. (SS67) informed that, due to the diverse nature of an MLR, it was necessary to consider the particularity of each type of GL. They assessed the studies in terms of the position and certainty of the source, clarity, detail, consistency, and plausibility. In the study of Garousi et al. (SS90), specific criteria were used to assess GL quality which covered the following aspects: authority, accuracy, coverage, objectivity, date, and significance.

Summary of RQ2.3: We did not find any specific guidelines to conduct a GLR, and most of the studies, even the MLR or GLR studies, used Kitchenham's or Petersen's guidelines. Moreover, few studies used specific criteria to search for GL or used specific criteria to assess its quality.

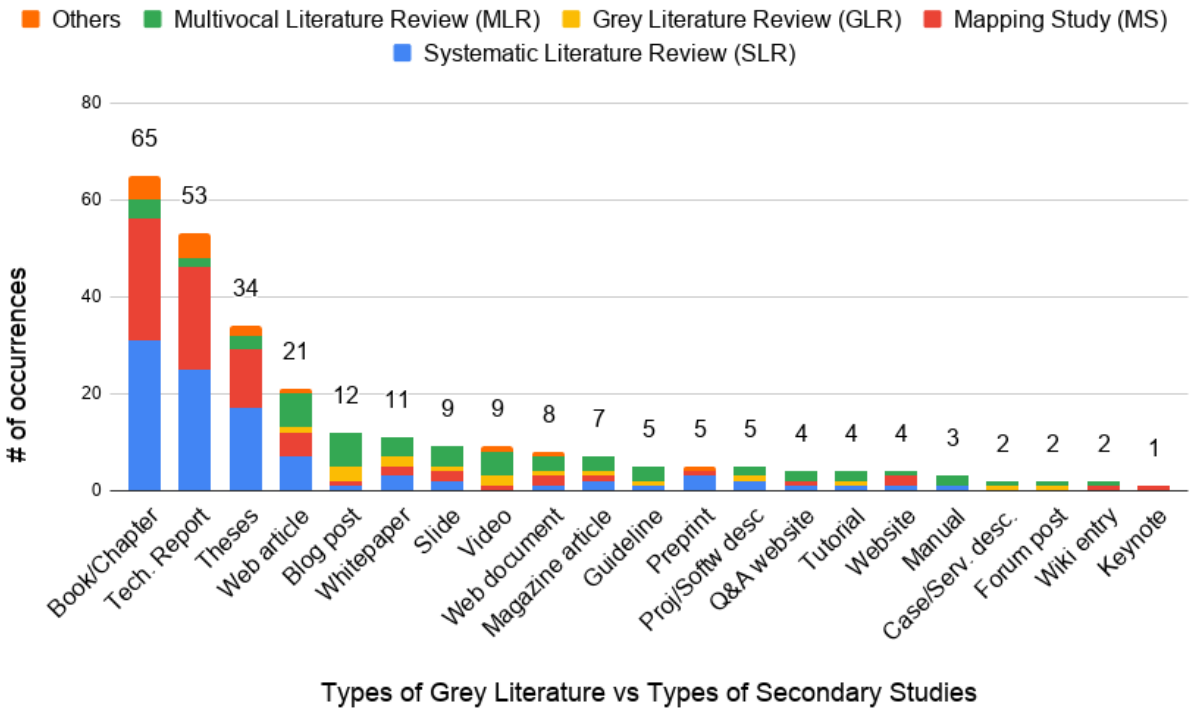
4.3.4 RQ2.4: What Types of Grey Literature Are the Most Frequently Used in Secondary Studies?

In this section, we present our investigation about the GL types and their producers identified among the studies. We also assess the GL availability.

We perceived that most of the studies (54%) (68/126 studies) did not classify the GL included. We used the classification of those that made it available. Since they were not classified, we had to classify them according to our interpretation. For instance, by reading the reference meta-information or accessing the link when it is available. However, there were some cases where we could not perform this assessment process (e.g., on a website/link no longer available, discontinued blog post, vague references). We classified these examples of GL as "Unknown".

The Secondary Studies mentioned a total of 1,314 examples of GL included. However, from this amount, when investigating the list of references of those studies, we retrieved only 1,273 GL studies (41 were missing). Moreover, from this list, we removed 25 peer-reviewed studies erroneously classified as GL. At the final, 1,246 GL studies remaining, distributed into 21 types.

Figure 13 – Distribution of the Grey Literature types used between the Secondary Studies identified in Tertiary Study 1.



Explanation: here it is evident, that 65 studies used books/book chapters; 53 studies used technical reports.

Source: the author.

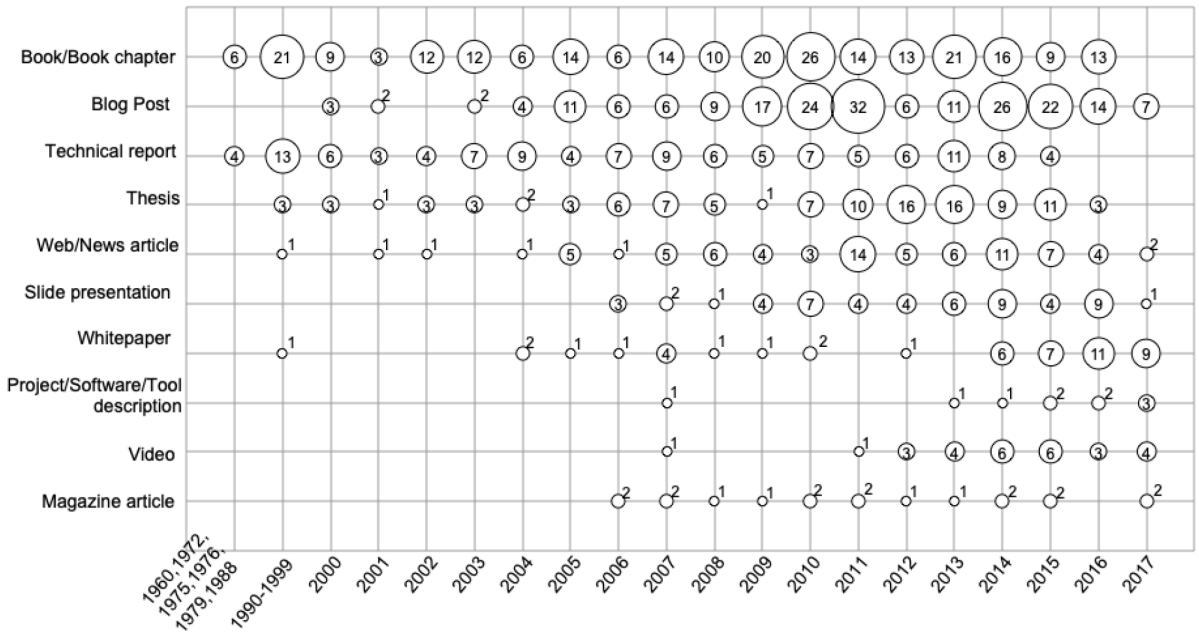
(i) Types of Grey Literature in Secondary Studies

Firstly, Figure 19 shows the GL types most commonly used distributed between the types of Secondary Studies. The figure shows that *books/book chapters* were the most common type of GL found among the Secondary Studies, followed by *technical reports*, *theses*, *web articles*, *blog posts*, and *whitepapers*.

Secondly, we analyzed all the 21 GL types found in the Secondary Studies and classified them according to the "shades" of grey proposed by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019) (see Figure 2). We found four GL types on the first tier, nine on the second tier, and eight on the third tier, showing that most GL types used have a medium level of control and expertise.

Blog posts (third tier) were most commonly found in the MLR and GLR studies. We also found many web/news articles, whitepapers, and descriptions of projects, software, and tools (second tier).

Figure 14 – Distribution of Grey Literature types found in the Secondary Study, distributed over the years of publication, identified in Tertiary Study 1.



Source: the author.

Figure 14 shows the evolution of the ten GL types most used over the years because the other types become inexpressive. Still, due to the small amount of GL arising in these years, in the first column, we combined the GL types published between 1960 and 1988. Further still, in the second column, we combined the GL published from 1990 to 1999. Overall, we found a

growth in GL's use since 2009, which was mainly driven by blog posts, theses, and web/news articles. The use of books/book chapters did not change significantly over the years. On the other hand, Secondary Studies are lately adopting whitepapers, videos, and descriptions of projects, software, and tools. In particular, blog posts start to be frequently published from 2000, and their inclusion as GL steadily increased over the years till the point it became one of the most used GL documents.

(ii) Grey Literature Producers

One of the criteria proposed by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019) to conduct the quality assessment in GL is to check the reputation of the authors and/or publishing organization. However, the information on who produces each example of GL has not always available. In these cases, we used our interpretation to fill this gap (e.g., accessing the links). To classify different producers of GL, we followed Maro et al. as a reference (MARO; STEGHÄFER; STARON, 2018).

We analyzed all the 1,246 GL to identify: (i) Who the producer was and (ii) Which GL types each producer was related to. Our first analysis is related to the data present in Table 25, which shows the total number of occurrences of contents produced by each type of GL producer in the column "#". Three types of producers (*Consultants / Companies, Academia,* and *Practitioners*) caught our attention, responsible for producing almost 80% of the GL included. Other important information is that for 13.2% of the GL studies, it was not possible to determine its producer.

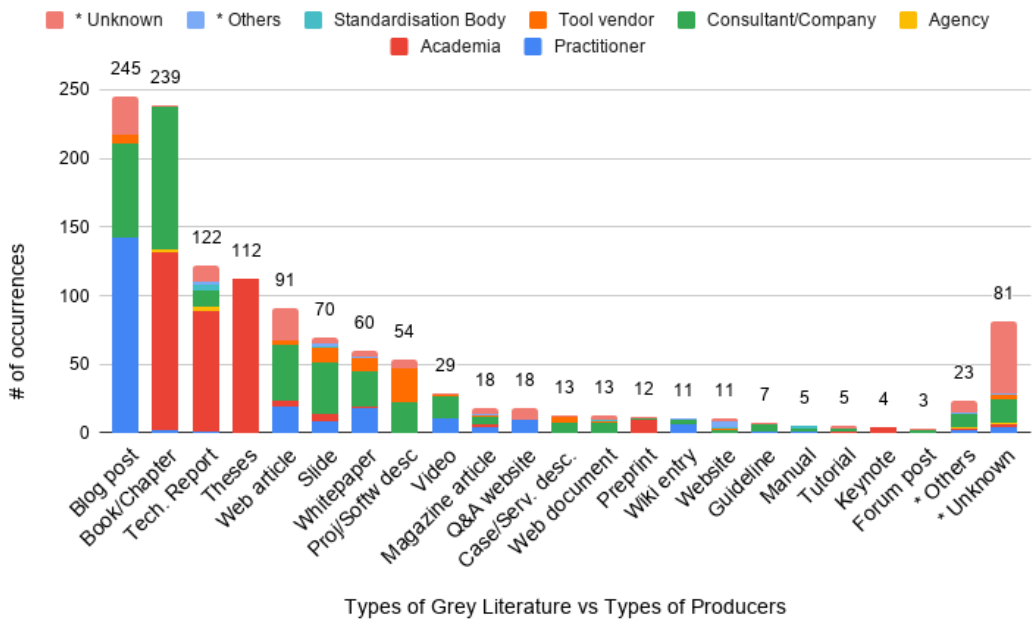
Table 25 – Grey Literature producers identified in Tertiary Study 1.

Producer	#	%
Consultant/Company	391	31%
Academia	361	28.6%
Practitioner	230	18.2%
Tool vendor	67	5.3%
Standardization Body	10	0.8%
Agency	7	0.6%
* Others	14	1.1%
* Unknown	166	13.2%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Figure 15 – Distribution of each Grey Literature type between the diverse types of producers, identified in Tertiary Study 1.



Source: the author.

Our second analysis is related to Figure 15, which shows the relationship between the GL types and producers types. We noted that producers from *Consultants / Companies* were the ones that most produced GL content, as one could expect, related to diverse GL types, but the most common were *Books/Book chapters*, *Blog posts*, and *Web articles*. Moreover, what caught our attention is that *Consultants / Companies* were the ones that produced the

most content of *Web articles*, *Slides*, and *Videos*. The second was *Academia*, mainly related to *Theses*, *Technical reports*, and *Books/Book chapters*. The third was the *Practitioners*, where the production was most related to *Blog posts*, *Web articles*, and *Whitepapers*.

(iii) Availability of the Grey Literature

Kitchenham et al. (KITCHENHAM; BUDGEN; BRERETON, 2015) argued on the importance of traceability in a Secondary Study so that the data would be available to others in the future. For this reason, we investigated all the 1,246 GL references included in the Secondary Studies (n=126) to verify: (i) If the URL of each GL was informed; (ii) If each URL reported is still working, that means we did not have any problem accessing it, and it was directed to the correct link of the source; and (iii) If were reported any access problems with the URLs.

Our findings show that 24.8% of GL sources (309/1,246) only presented the reference title but *did not report a URL*. From the remaining that *URLs were informed*, we found 23.7% (295/1,246) that exhibited *access problems* (e.g., server not found, page not found), leaving 51.5% (642/1,246) of GL references *still available*.

We also investigated the producer of each GL source, showing that among the sources in which URLs were not informed, more than half was produced by *Academia*, and among the URLs that returned some access problem, more than 30% of the items were from *Consultant / Company*.

Summary of RQ2.4: The most commonly found GL types among the studies were books/book chapters, technical reports, and theses, produced mainly by academia and consultants/companies. The practitioners produced most of the blog posts. Our investigation for GL references showed that, unfortunately, a quarter of the GL URLs were not reported, a little more than half of the GL URLs are still working, and for 23.7% of GL, we found some impediments to access.

4.3.5 RQ2.5: What Motivates Researchers to Use/Avoid Grey Literature?

In this question, we used two samples for our analysis. For the motivations *to use* GL, we investigated only the 126 selected studies; for the reasons to avoid, we investigated all the 446 Secondary Studies. We maintained the studies that did not use GL in the analysis because

some reported the reasons to avoid GL. Some studies, though, did not explain. We did not consider them in this research question.

(i) Motivations to Use Grey Literature

We noted that the discussion in favor of using GL is often provided in the search process or in the Secondary Study's inclusion criteria. For the GLR and MLR studies, it was common to find some of this information in the introduction as well. We found 35 out of 126 Secondary Studies that presented motivations to use GL. After grouping the answers, we presented the categories found in Table 26 and then discuss them.

Table 26 – Motivations to use Grey Literature identified in Tertiary Study 1.

Motivation	#	%
To identify more studies	16	12.7%
To incorporate practitioners' point of view	10	7.9%
To reduce publication bias	5	4%
Others	4	3.2%
No motivation was given	91	72.2%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

To identify more studies **(16/126 studies; 12.7%)**. This motivation is the most common among the Secondary Studies. Tüzün et al. (SS26) state that they: *"searched for company journals, gray literature, conference proceedings, and the internet, which led us to new papers that we could not identify in our regular search"*. The study of Carrizo et al. (SS52) reviewed books and Ph.D. theses to identify more studies. Besides, Garousi and Mäntylä (SS126) investigated the ideal moment to automate, and what to automate in software testing and decided to include GL because *"the academic studies on the topic were rare"*.

To incorporate the practitioners' point of view **(10/126 studies; 7.9%)**. This category was the second most common motivation among the studies. Tripathi et al. (SS34) state that: *"we need to incorporate the practitioners' point of view, which is shared through internet channels in the form of webpages and whitepapers"*. Other studies (SS47, SS112, SS126) included GL because they wanted to hear the practitioners' "voice".

To reduce publication bias **(5/126 studies; 4%)**. Some studies were motivated to include GL to reduce publication bias. As mentioned by Patel and Hierons (SS6), *"including grey literature is an important step to combatting publication bias"*. The study of Rizvi et al. (SS8) mentioned that *"We should point out that grey literature, such as the organization of whitepapers and lessons learned, were reviewed manually to address bias in paper selection"*.

Others **(4/126 studies; 3.2%)**. We group here the studies that employed other motivations. For instance, Soldani et al. (SS113) pointed out that companies are working day-by-day on the design, development, and operation of research interest, as also witnessed by the high number of GL items on the topic. Another motivation was the use of a *trustworthy source* such as Chen and Babar's study (SS109), which decided to manually search the technical reports from the Software Engineering Institute (SEI) because they considered it a trustworthy source, as pointed out here: *"SEI's series of technical reports is the main channel of grey literature in the research area"*.

(ii) Reasons to Avoid Grey Literature

We also analyzed the reasons for the studies to *avoid using* GL. Distinct from the motivation to use it, in this question, we also looked at the 446 studies, both the studies that *used* and did *not use* GL. It was most common to find information in the studies' inclusion/exclusion criteria section and the section on the threats to validity. We found 28 out of 446 Secondary Studies that presented reasons to avoid GL. After grouping the answers, we present the categories found in Table 27 and then discuss them.

Table 27 – Reasons to avoid Grey Literature identified in Tertiary Study 1.

Reason	#	%
Lack of Quality	23	5.1%
Hard to identify Grey Literature	3	0.7%
Others	2	0.4%
No reason was given	418	93.7%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Lack of Quality **(23/446 studies; 5.1%)**. The most common motivation to avoid GL

concerns the lack of quality that may affect the validity of the results. This category consists of a set of subcategories related to the nature of lack of peer-review process, some validity constraints, and the reliability of this type of source. The lack of peer-review processes was the most common subcategory found to avoid GL as Alkhanak et al. (RQSS02) pointed out to explain the motivation to exclude GL: *"[. . .] there could be a threat associated with Grey literature, that no peer-reviewed process might have been adopted"*. Another motivation that GL might negatively affect the validity because it is not externally valid as pointed out in Vilela et al. (RQSS22): *"[. . .] the external validity depends on the identified literature: if the identified literature is not externally valid, neither is the synthesis of its content [. . .]"*.

***Hard to identify Grey Literature* (3/446 studies; 0.7%)**. Some authors eliminated GL because they considered it hard to identify GL in their research interest (e.g., RQSS80). Moreover, Lane and Richardson (RQSS14) excluded GL *to make the SLR more straight-forward and repeatable but at the cost of potentially excluding valuable studies*.

***Others* (2/446 studies; 0.4%)**. We group here the studies that employed other motivations. For instance, Arvanitou et al. (RQSS11) eliminated GL because it considered that they had included the majority of "good quality" studies published in the selected venues, and increasing the number of studies would seriously threaten the feasibility of their investigation. A study by Alkhanak et al. (RQSS02) mentioned two reasons to exclude GL: lack of technical details and the lack of quality.

> **Summary of RQ2.5:** According to our findings, it was possible to categorize the studies' motivation to use and reasons to avoid GL. On the one hand, studies were motivated to use GL to identify more studies, incorporate practitioners' points of view, and reduce their studies' publication bias. On the other hand, they avoided using GL because of lack of quality (e.g., non peer-review and validity constraints).

4.3.6 RQ2.6: How do Researchers Perceive the Use of Grey Literature?

In this question, we explored the 126 Secondary Studies that used GL to investigate both the *benefits* and *challenges*. Although, not all the included Secondary Studies have data that could be used to answer this research question.

(i) Benefits

We analyzed the benefits mentioned in the studies (n=126) concerning the use of GL. We found only 13 Secondary Studies that present answers to this question. In the following, we present categories of the benefits found.

Table 28 – Benefits of Grey Literature use identified in Tertiary Study 1.

Reason	#	%
Provides practical evidence	13	10.3%
Helps in knowledge acquisition	9	7.1%
Makes academic studies more interesting	6	4.8%
Covers different results from scientific studies	3	2.4%
Easy to access and read	1	0.8%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Provides practical evidence **(13/126 studies; 10.3%).** This category indicates that GL brings practical experience. Usually, SLRs are based on academic papers. However, in some areas, this information is not enough to assess a topic of interest. For example, Raulamo-Jurvanen et al. (SS47) highlight: *"[...] the importance of grey literature for topics where the "voice of practice" is broad (and more active than academic literature)"*. In the same way, Soldani et al. (SS113) point out: *"[...] grey literature studies can be valuable to shed light on yet uncharted areas of software engineering research, especially when such areas are seeing massive industrial adoption"*.

Helps in knowledge acquisition **(9/126 studies; 7.1%).** According to this category, the traditional literature is not enough to cover some topics completely, missing important knowledge on specific research areas. For this reason, some studies stressed the importance of GL to fill this gap. For instance, Garousi et al. (SS112) informed us that: *"[...] if we were to exclude the grey sources from the pool, we would simply miss a major pile of experience and knowledge from practicing test engineers on the topic"*. Another study by Garousi et al. (SS98) compares MLRs with SLRs and says: *"We believe that conducting an MLR in the area of TMA/TPI will be more useful compared to an SLR since there is a large body of knowledge"*.

Makes academic studies more interesting **(6/126 studies; 4.8%).** Making academic work more interesting for practitioners is what reveals this category. For example, Garousi and Mäntylä's study SS126) reports: *"[. . .] grey literature in SLR studies is insightful, and thus the authors recommend including it when the topic has a low number of academic studies but high practitioner interest"*. Furthermore, Raulamo-Jurvanen et al.'s study (SS47) adds: *"Grey literature seems to have its place in SE, not only in serving the practitioners but also in providing an interesting aspect into academic studies"*.

Covers different results from scientific studies **(3/126 studies; 2.4%).** This category indicates that GL brings results uncovered by scientific literature to the discussion, as the study of Tripathi et al. (SS34) points out: *"Since the research on software startups, especially in the area of RE, is still in the nascent stage, we need to incorporate the practitioners' point of view, which is shared through the internet channel in the form of webpages and whitepapers"*. Moreover, Soldani et al. (SS113) complements: *"[. . .] with grey literature studies can be valuable to shed light on yet uncharted areas of software engineering research, especially when such areas are seeing massive industrial adoption"*.

Easy to access and read **(1/126 studies; 0.8%).** According to this category, GL provides for ease of access, as Raulamo-Jurvanen et al. (SS47) point out: *"[. . .] grey literature is freely and easily available for the public"*.

(ii) Challenges

We analyzed the challenges mentioned in the studies into the use of GL (n=126). We found only 14 Secondary Studies that present answers to this question. In the following, we present the categories of challenges found.

Table 29 – Challenges of Grey Literature use identified in Tertiary Study 1.

Reason	#	%
Non-structured information	8	6.3%
Epistemological problem	5	4%
Time-consuming	4	3.2%
Difficulty in measuring quality	2	1.6%
Others	2	1.6%

Note: The column "#" shows the total number of occurrences of a given category.

Source: the author.

Non-structured information **(8/126 studies; 6.3%).** This category brings evidence about how GL can pose great difficulties for assessment due to its lack of format and undisciplined characteristics in its structure and writing. For example, the study of Williams (SS122) reports: *"blog articles and much grey literature differ in that they are varied in their structure and the formality of their language"*. Soldani et al. (SS113) complements the position: *"This is mainly because grey literature lacks a unique format acknowledged across all sources and available data"*.

Epistemological problem **(5/126 studies; 4%).** According to this category, an issue in GL is experts' opinions without the support of empirical evidence about the subject. This discloses an experience of personal opinion without providing a reliable source. It was reported in the work of Garousi and Mäntylä (SS126): *"[. . .] it may be that some practitioners could simply be repeating the ideas/opinions that they had heard from other practitioners. Thus, as the source of knowledge was not typically revealed in GL, we are faced with an epistemological problem. We do not know how we know what we know"*. In the same way, Garousi et al. (SS112) pointed out: *"We found that sources of evidence in grey literature were often opinion or experience based rather than relying on systematic data collection and analysis as done in scientific papers"*.

Time-consuming **(4/126 studies; 3.2%).** This category indicates that the amount of literature included by GL sources demands an effort not foreseen in a common SLR. Raulamo-Jurvanen et al. (SS47) pointed out: *"Screening of grey literature sources can be a time-consuming process since usually, there is no applicable abstract or summary available"*. Also, Garousi and Küçük's (SS90) state: *"Since there is a vast grey literature as well as a large body*

of research studies in this domain, it is not practical for practitioners and researchers to locate and synthesize such a large literature".

***Difficulty in measuring quality* (2/126 studies; 1.6%).** This category arises from the difficulty of assessing GL quality. For example, Garousi and Küçük (SS90) pointed out: *"[. . .] we discovered that it is very difficult to uniquely measure the quality of grey literature when conducting a systematic, controllable and replicable secondary study"*. Garousi and Mäntylä (SS126) complemented: *"This suggests that requirements placed on formal publishing actually increase the amount of empirical evidence in software engineering"*.

***Others* (2/126 studies; 1.6%).** Here we group studies that adopted other challenges. According to Anjum and Budgen (RQSS92): *"The verification of references proved a troublesome process as more references were taken from the grey literature, often being provided on websites. As a consequence, many were either not available or had changed their URLs"*. This is an issue also seen during our study. Turner et al. (SS124) mentioned the difficulty in searching or finding information on GL, as pointed out here: *"The identification of 'grey literature', however, may be more problematic due to the digital libraries and search engines used and the lack of available benchmarks to use for validation"*.

Summary of RQ2.6: Our investigation found several benefits of GL uses. The most common was that GL provides practical evidence and helps in knowledge acquisition. The studies did not widely mention challenges, but the most common was the lack of structured information. This makes it difficult to retrieve information from a GL source. Another challenge was an epistemological problem concerning the lack of GL reliability.

4.4 DISCUSSION

This section discusses each research question, relating them with previous studies (Section 4.4.1). We present a discussion about some challenges identified based on the Secondary Studies investigated and our experience-based with this research (Section 4.4.2). Finally, we discuss some threats to validity (Section 4.4.3).

4.4.1 Revisiting Findings

(RQ2.1) Definitions of Grey Literature

As the investigations of GL in SE research are recent, we believe that our findings of GL definitions are compatible with an area still under development and that needs more in-depth studies. For instance, most investigated studies did not explicitly use the term grey or gray, making it difficult to find GL's common definition. This finding was also identified by the tertiary study of Zhang and colleagues (ZHANG et al., 2020). The same situation was found in Schöpfel and Prost's study (SCHöPFEL; PROST, 2020), leaving the reader to guess what GL is and what it is not. We highlighted that Schöpfel and Prost's (SCHöPFEL; PROST, 2020) findings supported some of our definitions. For instance, "unpublished works" and "non peer-reviewed studie".

We agree with Zhang and colleagues (ZHANG et al., 2020) that misunderstandings may influence the use of GL in SE without a common definition, as we discussed in RQ2.4. Recently, Garousi and colleagues (GAROUSI et al., 2020) proposed a definition of GL in SE research that states *"Grey Literature in SE can be defined as any material about SE that is not formally peer-reviewed nor formally published"*. We believe that future SE works will benefit from this definition.

(RQ2.2) Use of Grey Literature in Secondary Studies

(i) Our investigation showed an increase in Secondary Studies published over the years, together with the increase in studies using GL. Even though GL has not been used extensively in Secondary Studies (126/446; 28.2%), it stands out as an important source of evidence, as shown in our findings for RQ2.5 and RQ2.6, and in previous studies (GAROUSI; FELDERER; MÄNTYLÄ, 2016; GALINDO NETO et al., 2019; ZHANG et al., 2020), that showed a diversity of benefits by using GL.

(ii) We also investigated 1,246 GL data identified among the 126 Secondary Studies included, representing $<$21% of the 446 included studies. Nevertheless, a considerable amount, taking into consideration that GL use is recent in SE research. Zhang and colleagues (ZHANG et al., 2020) also found a similar ratio (22%), while Yasin and colleagues (YASIN et al., 2020) found a different picture: 76% of the Secondary Studies included GL. This difference occurred

because Yasin et al.'s study (YASIN et al., 2020) considered conference proceedings and workshop papers as GL types. In this regard, Garousi and colleagues (GAROUSI et al., 2020) pointed out that, in some disciplines, conferences accept all submitted papers with no peer-review. However, SE conferences usually have established peer-review processes. For this reason, we did not consider conference proceedings and workshop papers as GL sources, differently of Yasin and colleagues inter interpretation.

(iii) Moreover, from the 126 studies, 75.4% of them used GL to support an answer to their research questions, showing the importance of GL evidence to contributing with the findings of Secondary Studies. In this regard, our findings did not corroborate with previous studies. For instance, Zhang and colleagues (ZHANG et al., 2020) found that only 25% of studies used GL to evaluate their conclusions, while Yasin et al.'s work (YASIN et al., 2020) mentioned that only 9.2% of the GL was used to support the findings. We found a different interpretation of GL *usage* between the three studies. While our study interprets this usage by analyzing each research question, Zhang et al. wanted to determine if GL was used to evaluate the conclusions.

(RQ2.3) Process of Search, Select, and Quality Assessment of Grey Literature

(i) We observed that the majority of studies had used academic search engines to perform their search. What caught our attention was that of all the GLR and MLR studies, which are naturally inclined to seek GL, used Google's general search engine and, and only 10% of them used a more specific source (e.g., YouTube, Stack Overflow, Blogs, and Twitter) or used different forms to search for GL. For the last case, two studies specifically caught our attention. The first one, conducted by Soldani et al. (SS113), specifically searched on GL using Google, Bing, DuckDuckGo, Yahoo!, and Webopedia. The second one, the work conducted by Williams (SS122), used reasoning makers to search for rigorous blog articles. This finding agrees with Zhang and colleagues (ZHANG et al., 2020), who found that researchers are not adopting specific strategies to search for GL. We emphasize that future research needs to focus on using relevant and specific GL sources to the investigated topic, avoiding retrieving a large amount of sometimes irrelevant data using Google. It was perceived as a challenge in the study Raulamo-Jurvanen et al.'s study (SS47), that pointed out: *"Screening of grey literature sources can be a time-consuming process"*.

(ii) Some investigated studies used inclusion criteria to specific GL types, even they have

mentioned to include only peer-reviewed studies. This conflict could result from the difference in interpretation of the GL types, as discussed in the RQ2.4 overview. Another criterion was to include seminal sources (e.g., those provided by SEI). In concern to the exclusion criteria, most of them were also related to GL types (e.g., blogs, personal web pages, and videos).

(iii) Interestingly, only seven studies (8.1%, considering the Secondary Studies without Mapping Studies) employed criteria to assess GL. Al-Baik and Miller (SS58) mentioned that there is a lack of guidelines to assess GL. Unfortunately, these seven studies did not employ the existing quality assessment criteria proposed by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019), which is currently the state-of-the-art method for assessing GL.

We consider the lack of quality assessment approaches for GL as a problem because the nature of GL is different from peer-reviewed studies. Loading criteria for only one side might compromise the evaluation of the other. However, another problem was raised: it is difficult to use a single type of assessment because of the different forms of GL, as pointed out in the study of Tom et al. (SS67).

(RQ2.4) Types of Grey Literature in Secondary Studies

(i) Our research identified 21 GL types used in Secondary Studies. The most common among the studies were books/book chapters, technical reports, and theses. The study of Zhang and colleagues (ZHANG et al., 2020) found similar characteristics in terms of GL's most common types, but our study differs in terms of the proportions found. For example, Zhang et al. mentioned that technical reports are present in almost 66% of the studies (we found 42%), blog posts in 22% (we found 9%), books/book chapters in 22% (we found 54%), and theses in 17% (we found 26%).

In our investigation, we found difficulty in interpreting GL types. For example, studies of Irshad et al. (SS42) and Turner et al. (SS124) considered technical reports as GL, different from Tahir et al. (SS21). Therefore, we agree with Bonato (BONATO, 2018) that a lack of GL definition is the cause of the difficulty in its interpretation.

(ii) In our analysis of GL producers, we perceived an increase in the importance of GL produced by *Consultants / Companies* and *Practitioners* over the years. They represent, respectively, 31% and 18.2% of the content found. *Academia* had 28.6%. Our findings did not corroborate with Yasin et al.'s work (YASIN et al., 2020), who placed Academia first (38.3%). As we pointed in the RQ2.2 Overview, this difference occurred because the Yasin et al.'s in-

terpretation differs from ours, as we did not consider conference proceedings and workshop papers as GL.

(iii) We found many URLs used to reference GL were no longer available, which reduces GL's value to other researchers and the credibility of the study that cited it.

(RQ2.5) Reasons to Use or to Avoid Grey Literature

(i) Approximately 30% of the studies clearly state their motivations to use GL. We organized these studies into four categories. We identified that these categories were similar to the categories found by Zhang and colleagues (ZHANG et al., 2020), namely the following: to seek more related research, to understand the views of the practitioner's community, and to avoid publication bias. However, differently from Zhang et al., we identified that only 9.7% of the 446 studies described the motivation to avoid GL. We organized these studies into three categories.

Two motivations to use GL caught our attention: (i) "To incorporate the practitioners' point of view" because, for many years, researchers have been calling for the importance of incorporating industry evidence for SE research (e.g., in (KITCHENHAM; DYBÅ; JØRGENSEN, 2004)). However, only 8.7% of studies mentioned this motivation, and (ii), the category "To reduce publication bias" has been discussed through several areas of knowledge (e.g., in SE (GAROUSI; FELDERER; MÄNTYLÄ, 2019; ZHANG et al., 2020), Medicine (PAEZ, 2017), and Nutrition (ADAMS et al., 2016)). It was found to some degree in the Secondary Studies investigated.

(ii) We found few studies presenting their reasons for avoiding GL use. The main cited reason was the "Lack of Quality" of the studies, usually related to the lack of formal peer-review processes for publication. This reason was also found in survey research with Brazilian SE researchers (KAMEI et al., 2020).

We also found that SE researchers investigating a well-established research field tend to avoid the use of GL because of the availability of numerous peer-reviewed papers, as pointed out in the study of Vallon et al. (RQSS110). On the other hand, the lack of studies on a new research topic motivates GL use. For instance, Garousi and Küçük (SS90) noted that academic research concerning microservices was still at an early stage. However, companies were working daily on the design, development, and operation of the field, resulting in a considerable GL on the topic.

Moreover, we identified two trade-offs between the motivations to use and the reasons to avoid GL: (i) some researchers claimed that the motivation to use GL is the possibility to

include practitioner experience, as seen in (SS9, SS4). However, others tend to avoid its use because they were worried about the study reliability (RQSS19, RQSS22), and (ii), there was the motivation to use GL to reduce publication bias (SS6). However, again, concerns related to GL studies quality did some studies to question its credibility. To deal with those trade-offs, we recommended the set of criteria to assess the GL credibility found in our previous study (KAMEI et al., 2020) by selecting GL sources retrieved from renowned producers or cited by others.

(RQ2.6) Benefits and Challenges

(i) Only a few studies (15%) reported benefits in the use of GL. We found five categories in which the studies were placed. Comparing our findings of the benefits with previous studies, we have: (i) the tertiary study conducted by Zhang and colleagues (ZHANG et al., 2020) found four categories similar to ours, namely: to seek more relevant research, to avoid publication bias, to understand the views of the practitioner's community, and to explore uncharted research areas. Zhang and colleagues (ZHANG et al., 2020) pointed out one more category not identified in our work: to compare different perspectives between researchers and practitioners, and (ii) the review presented by Rainer and Williams (RAINER; WILLIAMS, 2019b) also corroborate with all of our five categories.

(ii) Also investigating the challenges in GL use, we identified five categories. As for benefits, only a few studies (11.1%) made clear the challenges of using GL. Comparing our findings with previous studies, we have: (i) four categories, out of the five, have similar categories to Zhang and colleagues (ZHANG et al., 2020), namely: noise in GL, paucity in ways of obtaining reliable GL, difficulty in quality assessment, and uncertain availability of GL. Zhang and colleagues (ZHANG et al., 2020) present one more category: Differences in the understanding of GL definition, and (ii) all of our identified categories have similar categories to Rainer and Williams (RAINER; WILLIAMS, 2019b). These authors perceived one more category: the lack of a mechanism to control the contents' variability.

We perceived some findings of the benefits and challenges to be contradictory. They are part of the trade-off between traditional literature and GL. For instance, on the one hand, GL helps in knowledge acquisition and practical evidence. On the other hand, the epistemological problem related to lack of reliability arises. In part, these trade-offs were expected, but they also show the need for further investigation on improving the use of the content provided and to better deal with it. For this reason, in the next section, we proposed some recommendations

to deal with this problem.

4.4.2 Challenges for Dealing with Grey Literature

In this section, we present some challenges identified based on the Secondary Studies investigated and our experience based on this research. First, we describe the challenge. In the following, we present potential ways to address or some existent proposals on how to deal with them, as we describe in the following.

Challenge 1: Lack of Grey Literature definition and misunderstanding about its types. Our investigation for RQ2.1 found little agreement exists about GL definition, corroborating with the study of Zhang and colleagues (ZHANG et al., 2020). Instead, we observed that most of the studies did not explicitly mention "GL". In 2020, Garousi and colleagues (GAROUSI et al., 2020) proposed a definition for GL in SE. Thus, as the formal concept of GL is recent, it is not yet widespread. We suggest that this lack of agreement on the unique definition for GL introduces a bias. Accordingly, different sources can be interpreted differently to be or not classified as a GL type. For instance, while Galindo Neto et al. (SS54) considered Ph.D. and master theses as a peer-reviewed source, the study of Rodríguez-Pérez et al. (SS83) did not consider this to be so. The same conflict was found in interpreting books/book chapters as GL types (e.g., SS56) or not (e.g., SS97).

Potential way(s) to address: We considered it essential to clarify what GL is about and the GL types included (or excluded) to make clear decisions employed, avoiding using only its characteristics. We also recommend using Garousi's definitions (GAROUSI et al., 2020) that stated, *"Grey literature in SE can be defined as any material about SE that is not formally peer-reviewed nor formally published"*. As GL includes many different types, we advocate using the concept of "shades of grey" in SE to classify GL material, as proposed by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019), to avoid misunderstanding about its types.

Challenge 2: Lack of search efforts for Grey Literature in specific data sources. Our investigation for RQ2.3 observed most of the studies used Google (search or scholar) as a primary source to search for studies. To search for GL, Bonato (BONATO, 2018) emphasized the importance of using specialized data sources because they are reproducible, and these

sources provide a means to identify Deep Web content, while Google may not identify more than 16% of the content available. In the SE area, several specialized data sources provide important GL content that could be useful for researchers (e.g., blogs, Q&A websites, and videos).

Potential way(s) to address: As GL in SE can be published in different ways, it is essential to understand the sources that could provide valuable information to the research and understand the viability to use the data available because each source provides different characteristics. This advice was also partially recommended by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019). Another issue to be considered is how to find relevant and rigorous GL in a considerable amount of information that could be retrieved if deemed the use of search engines (e.g., Google). Rainer and Williams (RAINER; WILLIAMS, 2019a) proposed using heuristics to improve GL searches' relevance and rigor to address this challenge. These recommendations avoid retrieving a vast amount of irrelevant data.

Challenge 3: Lack of specific quality assessment criteria for Grey Literature and its particular types. By analyzing RQ2.3, we noticed that most of the Secondary Studies that included GL (126) did not employ specific criteria for quality assessment, even among the studies that explicitly search for GL. Despite that, some previous studies perceived the difference in GL studies' nature compared to traditional literature (GAROUSI; FELDERER; MÄNTYLÄ, 2019), suggesting that these studies need to be evaluated in different ways. Moreover, Tom et al. (SS67) reported that due to the heterogeneity of the studies included investigated in an MLR, it was necessary to consider each type of GL's specific nature. This claim is also supported by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019), although it is not restricted to studies that included GL. Kitchenham (KITCHENHAM; CHARTERS, 2007) also drew attention to quality assessment instruments that meet the different types of studies.

Potential way(s) to address: When looking for GL, SE researchers should define a set of quality assessment criteria appropriate to assess these studies, in particular, by observing if the requirements are adequate for GL types retrieved in the search. Although some previous studies have already defined some quality criteria assessment for GL (e.g., (GAROUSI; FELDERER; MÄNTYLÄ, 2019; SOLDANI; TAMBURRI; VAN DEN HEUVEL, 2018)), we believed that more effort and attention by the SE research community is needed.

Challenge 4: Lack of Grey Literature classification. Our analysis for RQ2.4 found only

46% of Secondary Studies classifying the GL studies they have used. This lack of classification increased our effort to interpret the GL used (e.g., accessing the online link available), which could also introduce additional interpretation bias. This problem hinders a comprehensive understanding of the GL types used. For example, to understand better which GL types are commonly investigated.

Potential way(s) to address: As occurs with scientific papers included in Secondary Studies, they are usually classified by the publication channel (e.g., as a journal, conference, or workshop paper); we highlight the importance of classifying the GL with their types for the reader to understand what types were used and to guide future research that may want to investigate specific GL types.

Challenge 5: Grey Literature availability. Our analysis for RQ2.4 investigated 1,246 GL included in investigated Secondary Studies (n=126). We found that 24.8% did not provide the GL URLs, and almost half of the URLs informed were not working. Farace and Schöpfel (FARACE; SCHÖPFEL, 2010) also recognized this problem with GL availability. It happened because some websites were broken, or the URLs had changed. This challenge hinders the appraisal of the evidence retrieved and limits the Secondary Studies' replicability that used GL.

Potential way(s) to address: For this challenge, we perceived two possible ways to deal with this challenge. The first one is storing all data searched and collected in an external database (preference for Open Access) for later consultation, such as archiving data on preserved archives such as Zenodo and Figshare, as recommended by Mendez and colleagues (MENDEZ et al., 2020). Preserving the GL data in those websites makes it possible to cite those source, as each research artifact receive a Digital Object Identifier (DOI). Although websites that could significantly mitigate this problem exist, we believe we need a more robust culture to widen searches or promote permanent GL. The second one is trying to minimize this challenge using web archiving initiatives (e.g., Internet Archive <https://archive.org>) that preserves information published on the web or digitized from printed publications (COSTA; GOMES; SILVA, 2017). For example, accessing one GL URL <http://weblogs.asp.net/astopford/archive/2010/07/19/technical-debt.aspx> is returned that the page was not found. Nevertheless, by using the Internet Archive, we could find the web content.

Challenge 6: Lack of reliability/credibility. Our investigation for RQ2.5 and RQ2.6 found that even with the perceived benefits and motivations to use GL, several researchers avoid using it due to the lack of reliability or credibility (RQSS19, RQSS02, and RQSS22). This trade-off between the benefits of "hearing the practitioner's voice" and "lack of reliability or credibility" were expected, in part, but they also show the need for further investigation on how to improve the selection of content provided in GL and to better deal with it.

Potential way(s) to address: One possible way to deal with this challenge is selecting GL sources based on the 1st and 2nd tier of the "shades" of GL, aiming to retrieve evidence from sources produced by authors with high or moderate expertise and with high or moderate outlet control/credibility of the content production. Another possibility is the researchers employed a set of criteria to assess the GL credibility as discussed by previous studies. For example, Kamei and colleagues (KAMEI et al., 2020) investigated the importance of selecting sources from renowned authors, institutions, companies, or a renowned producer cited that, and Williams and Rainer (WILLIAMS; RAINER, 2017) proposed another set of criteria claiming that the GL source needs to be rigorous, relevant, well written, and experience-based. Thus, from these possibilities, SE researchers could take a decision whether GL is suitable to use or not.

4.4.3 Threats to Validity

We organized our threats according to the classification schema proposed by Ampatzoglou and colleagues (AMPATZOGLOU et al., 2020): Study Selection Validity, Data Validity, and Research Validity.

Study Selection Validity. Focusing on adequacy to identify relevant publications, we endeavored to mitigate study selection validity by first using a search string used and tested in a prior study (SILVA et al., 2011). Second, we performed a broad automatic search using the most relevant SE digital libraries, as recommended by Kitchenham et al. (KITCHENHAM; BUDGEN; BRERETON, 2015). Third, we selected secondary studies identified in previous tertiary studies (KITCHENHAM et al., 2009; KITCHENHAM et al., 2010; SILVA et al., 2011; CRUZES; DYBÅ, 2011b; GALINDO NETO et al., 2019). Fourth, we opted to select studies published in premier SE conferences (EASE, ESEM, and ICSE) and journals (EMSE, IST, TOSEM, TSE, and JSS), focusing on studies with a high potential impact on academic research and industry with in-

ternational coverage, and avoid predatory publications. We understand that this last decision may have introduced a bias to under-represent or over-represent the use of GL.

In the study selection applying exclusion criteria, we performed pilot studies and used a paired process and discussions between the involved researchers and invoked a third researcher to solve any disagreement. The Kappa value was calculated (0.571), which means a moderate agreement level. One threat related to the selection process is that we selected only studies wrote in English that can lead to the omission of important studies written in other languages.

Data Validity. Although we conducted a comprehensive search process, it is possible that some studies were missed, and as we avoided selecting non-English studies and those published in non-premier SE conferences and journals, it could introduce a bias in the data analyzed. For this reason, it is not possible to generalize the results.

Considering the data extracted, we chose them based on our research questions. However, as the process used a personal interpretation, it could not be very objective. We mitigate this process by involving two researchers and conducting pilot extractions to assess the understanding of the data extraction definition and agreement between the researchers. In addition, we used predefined fields in a spreadsheet with the data that should be extracted. We highlighted that the involved researchers refined this spreadsheet during the process.

At least two researchers analyzed all the data, and a third researcher analyzed the categories that emerged from the thematic analysis.

Research Validity. Focusing on mitigate the research validity, during the process of drawing our protocol, before adequately starting the conduction of the tertiary study, the protocol was revised by experienced SE researchers. In addition, in most phases of this research, we endeavored to minimize those by using a paired approach with a constant discussion between the researchers and invoking a third researcher to revise the entire analysis. We also compared our findings with previous works, showing that many findings corroborates with these works, although we identified news and controversial findings.

Focusing to permit the replicability, we made available all the process and data extracted and analyzed used in this Tertiary Study, focusing on permitting the future analysis of replication of the process.

4.5 SUMMARY

This chapter presented a tertiary study showing an overview of GL use in Secondary Studies. We identified a total of 446 Secondary Studies, within which we investigated 126 (28.2%) for a more comprehensive understanding.

We found a lack of GL definition among the studies and different interpretations of a GL type. We believe that the use of "shades" of grey could be promoted and help solve this challenge.

Our findings have several implications for SE research. We highlighted the importance of GL to Secondary Studies, and we presented several benefits and motivations to use it. We also found some challenges and reasons to avoid GL, showing that future investigations are necessary. Moreover, we discovered the need for specific guidelines to search, select, and assess, considering GL types' plurality. Researchers should also consider developing methods to improve GL's availability, allowing their data to be preserved and accessed by others in the future. Those guidelines could be important to the SE research to take better advantage of using GL.

By describing our findings and a list of challenges with the potential ways to address them, we expect to help others to use GL in SE research. To conclude, in this investigation, GL shows as an important source of evidence for Secondary Studies but needs more maturity for researchers' broad acceptance.

We inform that the investigation presented in this chapter was published in the Information and Software Technology Journal (IST) (KAMEI et al., 2021b).

5 CONTRIBUTIONS OF GREY LITERATURE IN MULTIVOCAL LITERATURE REVIEWS STUDIES

This chapter presents a tertiary study that investigated nine Multivocal Literature Review (MLR) studies that followed Garousi's guidelines (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019) to understand how the use of Grey Literature (GL) contributed to MLR studies and answer the RQ3. By *contributing*, we mean understanding to what extent the GL is providing evidence that is used by an MLR to answer its research question.

In the following, Section 5.1 presents our research questions. Section 5.2 presents the research methods employed. Section 5.3 presents our findings. Section 5.4 presents the discussions about this research. Finally, Section 5.5 presents a summary of this research.

5.1 OVERVIEW

To achieve the stated goal, we explored the following research question:

RQ3: *How the Grey Literature Use Contributed to Multivocal Literature Review Studies?*

The above research question was divided in three more specific questions, they are:

RQ3.1: *How Commonplace Is to Employ Grey Literature in Multivocal Literature Review Studies?*

Rationale: Recently, it was perceived an increase in Secondary Studies using GL, mainly related to MLR or GLR studies (GALINDO NETO et al., 2019). Despite this interest, there is limited knowledge about how GL is used in those studies. In this question, we aimed to present an overview and a comparison over the findings of the investigated studies, on the perspective of TL and GL contributions.

RQ3.2: *To What Extent Grey Literature Contributes With the Findings of Multivocal Literature Review Studies?*

Rationale: In a previous study conducted by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2016) was conducted an investigation about the type of GL contribution in MLR studies. In this question, we aimed to provide a profound investigation by classifying each type of contribution to understanding how each GL properly contributed to each MLR study.

RQ3.3: *What Types of Grey Literature Sources Are Most Commonly Observed in Multivocal Literature Review Studies?*

Rationale: As GL could be produced in several forms, as shown in previous investigations (e.g., (YASIN et al., 2020; ZHANG et al., 2020)), our intention in this question was to explore the GL types and its producers and correlate them with the type of contribution identified. This question is important to guide future investigations to understand better what each GL type could contribute to their investigation.

To answer these questions, we employed a tertiary study to find potential MLR studies. In the end, we qualitatively explored nine MLR studies.

In summary, our main findings in this research are:

- We identified that several findings of MLR studies were exclusively retrieved from GL sources. For instance, we perceived that some RQs from two MLR studies (MLR4, MLR6) were answered using only GL;
- We perceived that MLRs are benefiting from GL mainly to provide *explanation* about a topic (e.g., explaining how DevOps could help in operations process and manage risks of companies (MLR2)) and to *classify* the findings (e.g., when classifying libraries, architectural style, and architectural guidelines about Android apps (MLR3)). Moreover, contributions providing recommendations (e.g., a recommendation of the use of dependency injection approach to fix the heavy of the setup of test smells (MLR6)) are presented in 66.6% of the MLR studies;
- We identified several GL types used among the MLR studies. The most common types were blog posts, web articles, books and book chapters, and technical reports. These GLs were produced mainly by SE practitioners, consultants and companies, and tool vendors.

By describing these findings, we expect to: (i) provide a further understanding of how the use of GL contributed to MLR studies, (ii) provide a structured process to help SE researchers

that intend to investigate how GL is used in MLR studies, (iii) present an overview of the types and producers of GL sources used in MLR studies, and (iv) help others to conduct MLR studies by describing potential challenges faced in conducting this analysis, along with potential ways to deal each one.

5.2 TERTIARY STUDY

We conducted a tertiary study to identify MLR studies published in the SE literature to investigate to what extent GL contributed to multivocal studies and answer RQ3. This research followed the most well-known guideline to conduct secondary studies in SE, produced by Kitchenham et al. (KITCHENHAM; CHARTERS, 2007). In what follows, we present the procedures employed to conduct this tertiary study.

For replication purposes, the data used in this chapter is available online at: <http://doi.org/10.5281/zenodo.5090736>.

5.2.1 Search Strategy

In this investigation, we restricted our investigation to MLR studies that strictly followed Garousi's guidelines (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019). We took this decision because these are the main and most recent guidelines in SE research to conduct MLR studies. Although the most recent Garousi's guidelines were published (in peer-review format) in 2019 (GAROUSI; FELDERER; MÄNTYLÄ, 2019), an earlier version of it (published in 2017 as a preprint (GAROUSI; FELDERER; MäNTYLä, 2017)); this is why we considered both of them in our research.

We started our research at the beginning of 2020. For this reason, we decided to limit our scope to studies published since 2017 (the first publication of Garousi's guidelines (GAROUSI; FELDERER; MäNTYLä, 2017)) until the end of 2019. We started by using the Google Scholar search engine to find works that cited Garousi's studies published (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019).

5.2.2 Selection Criteria

When manually investigating the 60 potential studies, we focused on selecting only MLR studies. For each candidate study, we applied a set of exclusion criteria described in Table 30. We excluded any candidate study that complies with at least one exclusion criterion. At the end of this process, we were left with nine MLR studies.

Table 30 – List of exclusion criteria used in Tertiary Study 2.

#	Description
EC1	The study was published before 2017 or after 2019.
EC2	The study was duplicated.
EC3	The study was not written in English.
EC4	The study was not related to Software Engineering.
EC5	The study was not a full paper (e.g., a position paper).
EC6	The study did not report an MLR study.
EC7	The study did not follow Garousi's guidelines (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019).

Source: the author.

5.2.3 Study Selection

We conducted this research in five phases, as detailed in Figure 16. There is a number indicating each phase (**P1–P5**).

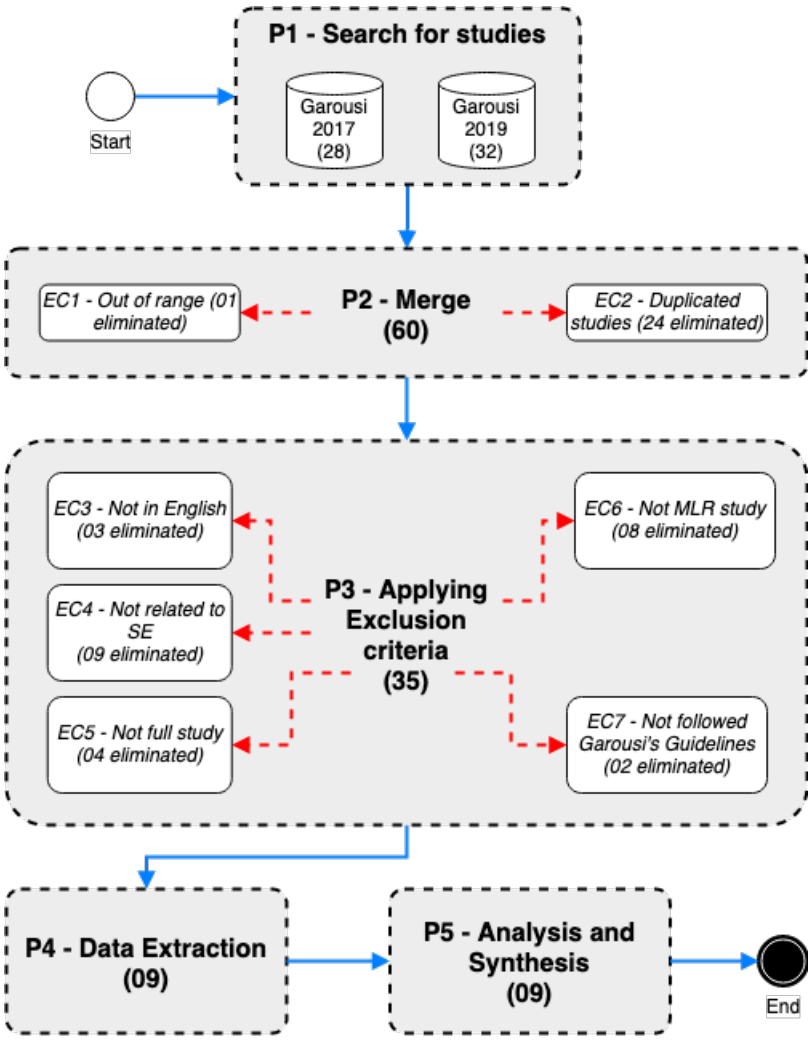
At phase **P1**, we selected a total of 60 potential studies. From these, 28 cited the first version of the guideline, published in technical report format (GAROUSI; FELDERER; MäNTYLä, 2017), and 32 mentioned the final version of the MLR guidelines for SE (GAROUSI; FELDERER; MÄNTYLÄ, 2019).

At phase **P2**, we sorted the potential studies by title and organized them on a spreadsheet. We applied EC1 and EC2 to remove the studies out of the range of our investigation and the studies with the same bibliographical information (i.e., title, abstract, and author(s)). For EC2, we employed the following steps: (1) We compared paper titles; (2) For studies with the same title, we looked at the abstracts and verified if they differed. We considered the complete

study as recommended by Kitchenham and Charters (KITCHENHAM; CHARTERS, 2007); if they are the same, we exclude one of them. If the publication years are different, we excluded the oldest study. We removed 25 studies, one published after 2019 (EC1), and 24 instances of duplicated studies (EC2), respectively. At the end of this phase, 35 studies remained.

At phase **P3**, we read the studies thoroughly and applied the criteria EC3–EC7 in 35 potentially relevant studies. As the criteria employed to select studies were simple, only one researcher applied them alone. We removed 24 studies based on the following criteria: three studies are not written in English (EC3); nine studies are not related to SE (EC4); four studies are not full papers (EC5); six studies did not report an MLR (EC6); and two studies were eliminated because they did not follow the Garousi's studies (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019) to conduct their research (EC7). This way, at the end of this phase, **nine MLR studies remained**. The complete references of the included

Figure 16 – Process of selecting studies in each phase of the Tertiary Study 2.



Source: the author.

studies are presented in **Appendix F**.

At phases **P4–P5**, we applied the data extraction, analysis, and synthesis following the process depicted in Figure 17. These phases are fully described in Section 5.2.4.

5.2.4 Data Extraction and Analysis

Due to the lack of a process to help SE researchers to investigate how the use of GL contributed to MLR studies, we designed a process based on our experience. Three researchers refined this process to conduct phases P4 and P5. We conducted this process in pairs, and all researchers who participated in this research revised the emerged categories and classifications.

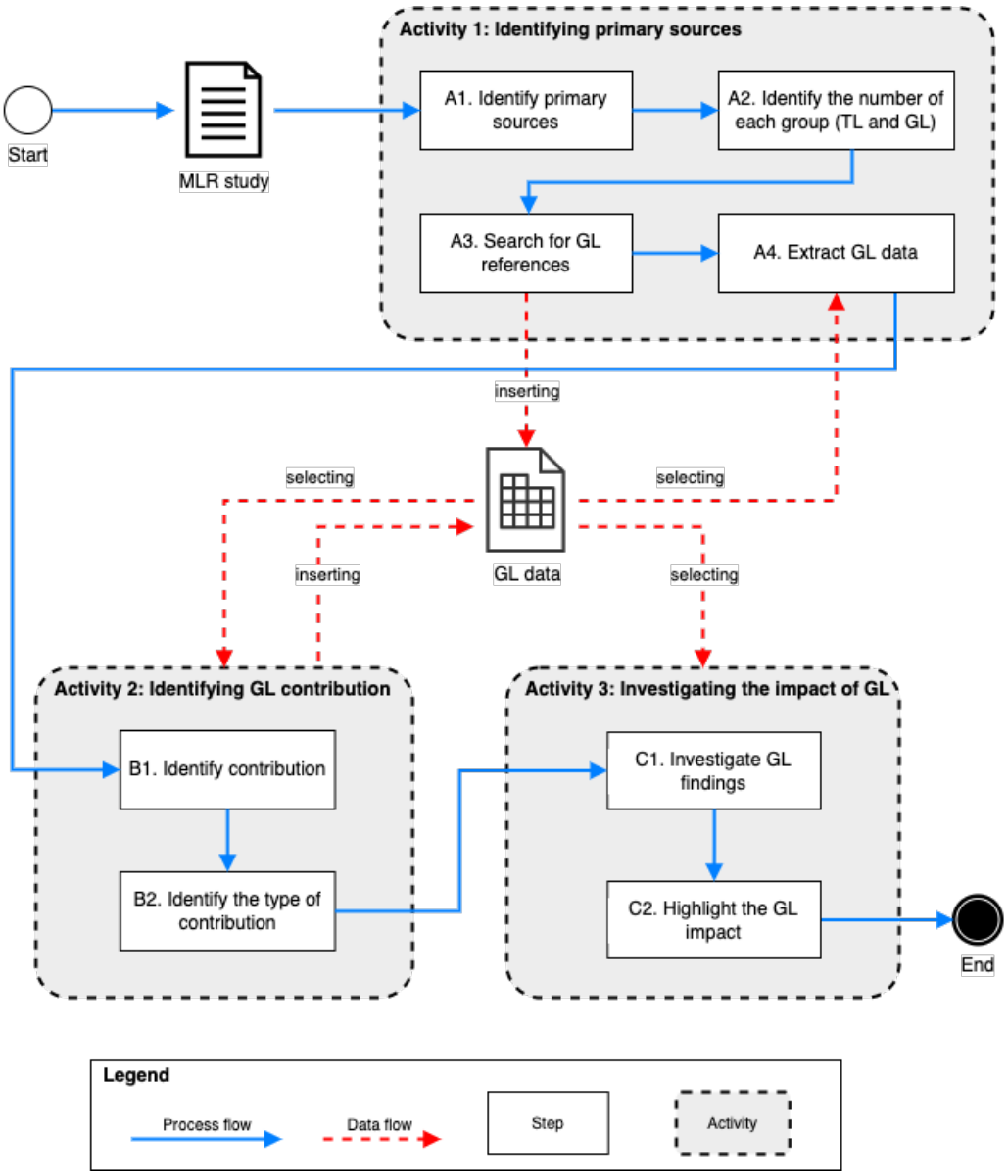
Our process started by investigating an MLR study distributed in three activities with their respective steps, as shown in Figure 17. In what follows, we describe our process.

Activity 1: Identifying primary sources

The first activity aims to identify the primary sources[1] included in an MLR study through four steps. The first step (*Step A1*) identifies the number of primary sources included in the MLR study. Then, we count the occurrences of each group: GL and TL (*Step A2*). These numbers are important for two moments: (i) to calculate the amount (%) of GL included (total of GL included / total of included studies), and (ii) to search for GL references in the studies. The following step is to find the reference of each GL included (*Step A3*), and add all data collected in A3 in a spreadsheet. The list of references for GL is usually found in the appendix, tables, or external files available. The final step (*Step A4*) consisted in selecting and extracting all data available of each GL, to permit traceability between the data extracted and the primary sources (as recommended by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019)). In our research, we collected data such as (but is not limited to): (i) the names of authors, (ii) the year of publication, (iii) total number of included studies, (iv) the total number of GL sources included, and (v) the guideline followed. In addition, considering each study that included GL, we also extracted: (i) the GL type, (ii) the evidence used from GL, (iii) the type of contribution, and (iv) the type of producer.

[1] In related studies, the term "primary sources" in GL is used as equivalent to the term "primary studies" in traditional literature (GAROUSI; FELDERER; MÄNTYLÄ, 2019)

Figure 17 – Process of identifying how Grey Literature use contributed to Multivocal Literature Review studies in Tertiary Study 2.



Source: the author.

Activity 2: Identifying the Grey Literature contribution

The second activity consists of selecting the GL data saved to identify how its use contributed to the MLR study. Then, inserting in the spreadsheet all the portions of GL used as evidence.

We used the following approach to identify these contributions *(Step B1)*: (i) after identifying the GL sources, we searched for any mention/discussion of each GL in the manuscript. We noticed it is common to find this information in tables, graphics, or as a citation during the manuscript; (ii) once we identified the contribution, we extracted the citation or the arti-

fact name used to highlight where the contribution occurred; (iii) we employed a qualitative analysis to classify the contribution of the use of each GL *(Step B2*) according to its type. We used the GL types' classification introduced by Maro et al. (MARO; STEGHÄFER; STARON, 2018); and (iv) we investigated the relation of the GL types and the contributions identified.

In the following, we present in greater detail the qualitative analysis process used in Activity 2 (Figure 18), based on the thematic analysis technique (BRAUN; CLARKE, 2006) and adapted from Kamei and colleagues (KAMEI et al., 2020):

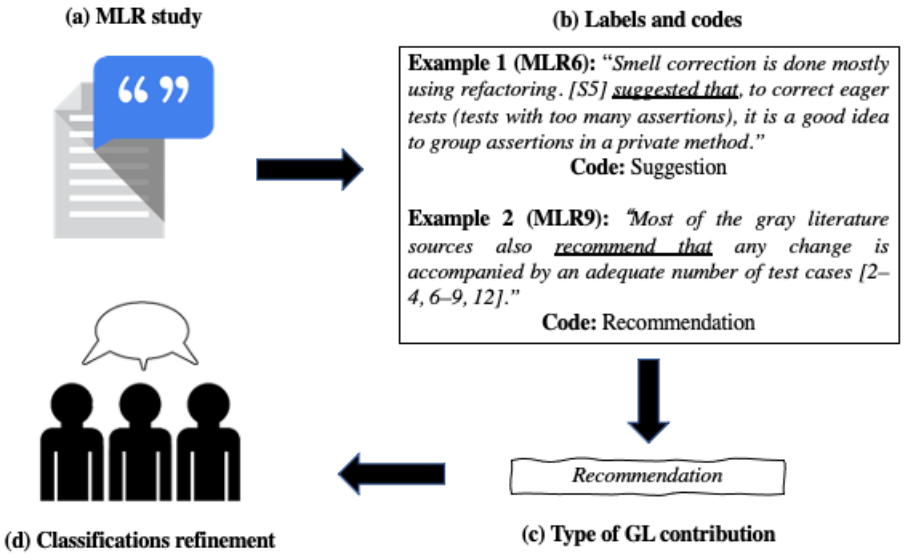
(a) *Familiarizing ourselves with data:* Each researcher involved in the data analysis becomes aware of which part of the MLR study the GL sources were referenced, as expressed in Figure 18-(a).

(b) *Initial coding:* In this step, each researcher individually added pre-formed codes. Our process of allocating data to pre-identified themes of contributions is based on the list of contribution facets proposed by Garousi and Küçük (GAROUSI; KüçüK, 2018) (e.g., recommendation, experience report, tool, solution proposal, opinion, empirical study, categorizing (or classification)). During the initial coding, we perceived categories not identified by Garousi and Küçük (GAROUSI; KüçüK, 2018). Thus, we extended the original contribution facets to add these categories. We briefly define each one as follows: *Programming* used to evidence programming-related techniques; *Concept Definition*, used for sources that present a concept or a definition of meaning; *Explanation*, used for evidence that provides an explanation or information about a topic; *Recommendation*, used for evidence that contributed by providing any recommendation to solve or support a problem or challenge. Figure 18-(b) presents an example of this analysis, where two portions from the texts were extracted and coded: Suggestion and Recommendation. Labels express the meaning of excerpts from the quote that represented appropriate types of contributions.

(c) *Classifying contributions by GL use:* Here, we already had an initial list of codes. A single researcher looked for similar codes in data. Codes with similar characteristics were grouped into broader categories. Eventually, we also had to refine the categories identified, comparing and re-analyzing them in parallel. Figure 18-(c) presents an example of this process. This example exhibits how the category "Recommendation" emerged.

(d) *Classifications refinement:* In this step (Figure 18-(d)), we involved two researchers (a Ph.D. student and a Professor) in evaluating all classifications and a third researcher (a Professor) to solve any disagreements (if needed). In the cases of any doubt, we solved them through conflict resolution meetings.

Figure 18 – Example of classification process used to analyze the contributions by Grey Literature use in Tertiary Study 2.



Source: the author.

Activity 3: Investigating the impact of GL

This activity consisted of investigating how GL usage contributed to MLR study. It started by selecting the data of GL stored to investigated GL findings *(Step C1)* and to understand how these findings contributed to the MLR study *(Step C2)*. The goal is to assess quantitatively and qualitatively these contributions. For instance, we presented the difference in the proportion of included studies and the number of studies related to a particular finding in terms of quantitative analysis. In qualitative aspects, we compared GL findings with TL findings, focusing on understanding if any finding was observed solely because of GL.

5.3 RESULTS

In this section, we present our main results organized concerning the research questions. Section 5.3.1 presents an overview of how the use of GL contributed to each MLR study. Section 5.3.2 provides our classification for the contributions identified and correlating them with the GL types and their producers. Finally, Section 5.3.3 shows the GL types and producers identified among the investigated MLR studies.

Table 31 – Characteristics of investigated studies in Tertiary Study 2.

ID	Total (#)	Total (%)	RQ	XRQ
MLR1	15	32.6%	2/2	0/2
MLR2	7	43.7%	3/3	0/3
MLR3	32	72.7%	3/3	0/3
MLR4	10	90.9%	3/3	2/3
MLR5	120	72.3%	8/9	0/9
MLR6	160	47.2%	1/3	1/3
MLR7	5	21.7%	1/3	0/3
MLR8	151	66.5%	1/1	0/1
MLR9	21	48.8%	1/2	0/2

Note 1: "Total (#)" means the total amount of GL as the primary source.
Note 2: "Total (%)" means the proportion of GL as primary source.
Note 3: "RQ" means the number of research questions answered with GL.
Note 4: "XRQ" means the number of research questions *exclusively* answered with GL.

Source: the author.

5.3.1 RQ3.1: How Commonplace Is to Employ Grey Literature in Multivocal Literature Review Studies?

An overview of the nine MLR studies is presented in Table 31, showing several interesting observations. First, the second column (Total (%)) shows that in the study (MLR4), GL accounted for more than 90% of primary sources overall. In three studies (MLR3, MLR5, MLR8), GL accounted for between 51–75% of the selected studies. Only one MLR study

(MLR8), GL was identified in less than 25% of included sources. This finding suggests that MLRs are taking serious advantage of GL. Second, in the third column (RQ), we depict how many GL sources were used to answer the research questions posed by the MLRs. We noticed that all studies used GL to answer at least one research question. The MLRs (MLR1, MLR2, MLR3, MLR4, MLR8), in particular, used GL as their basis to answer all research questions. When looking closer (last column, XRQ), we also observed two studies (MLR4, MLR6) that have some RQs that were exclusively answered using GL, for instance. Next, we assess what evidence was found in GL.

Garousi et al. (MLR1) conducted an MLR to provide a more "holistic" view about SE research's relevance. The study included 46 primary sources, 31 from TL (67.4%) and 15 from GL (32.6%). Although the number of TL studies was higher than GL sources, the evidence retrieved from GL was used to support most of the findings. The authors identified that the root causes of low relevance of SE research (e.g., simplistic view about SE in practice, wrong research problems identification, issues with research mindset) were observed in multiple sources (GL and TL), concluding that the community members share similar opinions on the debate.

Plant's study (MLR2) performed an MLR to investigate which risks types companies using DevOps are generally exposed to and proposed a framework that helps companies control their process and manage risks. The study identified 24 risk types. From these, nine were exclusively identified in GL sources (e.g., Automated change controls and thresholds, Automated production deployment, Static code analysis), eight were exclusively identified in TL sources, and seven were found in both groups (GL and TL). In particular, if the study did not consider GL sources, the MRL would not have discussions about *automated security tests* and *monitoring and logging*, which comes largely from GL.

Verdecchia (MLR3) investigated (through an MLR and interviews with SE practitioners) how developers architect their Android apps, what architectural patterns these apps rely on, and their potential impact on quality. The study identified 15 libraries and nine architectural patterns considered when developing Android apps. Considering only the libraries, 13 were found exclusively in GL (e.g., JUnit, Mockito, Mosby), and only two of them were found through the interviews. From the architectural patterns identified, 7/9 (77.8%) were exclusively found in GL (e.g., Clean, Hexagonal, Viper). Beyond that, 212 architectural practices were extracted and synthesized into 42 architectural guidelines. From these guidelines, 38/42 (90.5%) were retrieved from GL. According to the study, four main themes emerge from the

guidelines retrieved exclusively in GL. Regarding the quality requirements considered while architecting Android apps, seven (7/24; 29.1%) of them were exclusively retrieved from GL (e.g., Scalability, Interoperability, Maintainability). In particular, the scalability attribute was *exclusively* found in GL sources. On the other hand, 11 groups of quality requirements were exclusively found in TL sources.

Bhandari and Colomo-Palacios (MLR4) conducted an MLR to investigate holacracy, a practice to radically shift from the conventional ladder to a more decentralized organizational structure. This MLR investigated holacracy in software development teams, its features, benefits, and challenges. This study investigated three research questions: RQ1 covered the definitions of holacracy and was answered using only GL sources. RQ2 investigated the characteristics of holacracy, which were identified: roles, circles of small groups, and meetings. Circles and meetings, in particular, were derived only from GL sources, and the roles were identified in both GL and TL. Finally, RQ3 was answered using only GL sources, explored the benefits (e.g., increased product transparency, better decisions, fast improvement) and challenges (e.g., implementation difficulty, undefined job roles cause employee outflow) using holacracy.

Garousi and Küçük (MLR5) performed an MLR to summarize what is known about smells in test code. The authors highlighted that "most test smells and problems in this area are 'observed' by practitioners who are actively developing test scripts and are communicating by them via the GL (e.g., blog posts and industry conference talks)". In this study, GL sources represent 72 out of 81 (88.9%) primary sources that presented new smells names and types. For solution proposals, 72.4% of the sources were GL.

Maro et al. (MLR6) conducted an MLR to explore traceability challenges and solutions in the automotive software development domain. The study identified 22 challenges of software traceability (e.g., Lack of knowledge and understanding of traceability, Difficulty defining an information model for traceability, Unclear traceability process) distributed in seven groups of factors (e.g., Human, Uses of Traceability, Knowledge of Traceability). In this investigation, although the challenges identified in GL and TL were similar, the study mentioned that the solutions presented in GL were richer than TL due to the diversity of producers.

Freire et al. (MLR7) performed an MLR to evaluate integration platforms, specialized software tools with integration solutions, which aim to direct a set of applications to promote compatibility among their data and new features regarding the performance of their runtime systems. This study selected nine open-source integration platforms, of which two were exclusively found in GL sources (Petals and ServiceMix), five were found both GL and TL (e.g.,

Guaraná, Fuse, Mule), and two exclusively found in TL sources (Camel and Spring Integration).

Saltan and Smolander (MLR8) investigated a total of 13 SaaS pricing frameworks: seven retrieved from TL (e.g., Pricing process framework, Cloud solution pricing framework) and six from GL (e.g., Customer-centric value-based pricing framework, Pricing process framework, PWC pricing management framework). These frameworks cover the three SaaS pricing aspects (Pricing strategy, Pricing tactics, Pricing operations). Considering the pricing aspects observed, if the study did not include GL, no evidence of Pricing Operations would exist.

The study of Ram and Sawant (MLR9) focused on gaining a sound foundation about what aspects of a code change reviewers focus on, conducted two investigations: an MLR study and the other one using interviews. The study identified ten themes that constitute an excellent code change (e.g., Change description, Change scope, Code style). Two themes were identified only in TL sources (Nature of the change, Subsystem hotness). No theme was exclusively composed of GL, although in some themes, GL counts as the main source providing evidence (e.g., Change description, Commit history).

> **Summary of RQ3.1:** All the investigated MLR studies showed that GL contributed with several findings of these studies. Some of those contributions were exclusively found in GL sources, and some others confirmed the findings identified in TL. This highlighted the importance of GL to contribute to MLR studies.

5.3.2 RQ3.2: To What Extent Grey Literature Contributes With the Findings of Multivocal Literature Review Studies?

In this section, we present the results of our investigation of the 384 GL sources found in nine MLR studies, resulting in 326 contributions identified and classified. We also explored these contributions by analyzing their relation with each MLR study.

To improve the comprehension of the results and enable traceability, we include direct quotes extracted from the MLR studies representing the GL use in the study.

In the following, we describe each type of contribution.

Recommendation **(7/9 studies; 77.8%)**. GL evidence was found by providing recommendations to deal with something (e.g., project, tool) or some problems (e.g., lack of proper visualization and reporting tools for software traceability in the automotive domain). In Garousi

and Küçük (MLR5), the authors cited a blog post that suggested using dependency injection as an approach to fix one test smell. Maro et al. (MLR6) mentioned a service description presenting a recommendation to use a centralized data storage where all artifacts are stored and therefore accessible by the staff in different locations. This would solve the challenge of complexity added by distributed software development: *"[. . .] having tool support such as an integrated tool platform where all development activities are done, or a structured way of defining artifacts also helps to solve this challenge"*.

Explanation **(7/9 studies; 77.8%)**. This category (with the highest number of contributions) indicates that authors used GL to explain some topics explored in seven MLR studies. An example for this category, the study of Garousi et al. (MLR1) mentioned a blog post: *"Software research is biased toward huge projects and, thus, small to medium size projects may not benefit from most SE papers"*. The study of Plant (MLR2) used a whitepaper to explain how DevOps could manage risks in software companies: *"[. . .] Due to the increased speed, quality, and agility which DevOps brings about if implemented correctly, implementing DevOps processes can contribute significantly to achieving these objectives"*. In the study of Bhandari and Colomo-Palacios (MLR5), GL sources were used to characterize holacracy in software development teams. For instance, the information present in a blog post: *"In holacracy, instead of job titles, there is a strong focus on the roles that people take on within the organization. Every task or project assigned to an employee must be within the accountabilities of his or her role"*.

Classification **(6/9 studies; 66.7%)**. This category was also commonly observed, indicating that GL helped to classify the findings (e.g., types of concepts, tools, SE practices) of the MLR studies. Verdecchia et al. (MLR3) used 32 GL primary sources and 12 TL primary studies to classify the libraries, architectural style, and architectural guidelines found about Android apps. As an example, the study of Verdecchia et al. (MLR3) used GL evidence to classify 38 architectural practices found into four themes: general Android architecture, MVP, MVVM, and Clean Architecture. In Garousi and Küçük (MLR5), a GL based in a bachelor thesis was used to categorize test smells, as follows: *"[. . .] categorized 53 different test smells on several dimensions, e.g., test automation, determinism, correct use of assertions, and reliability"*.

Another example was the study of Ram et al. (MLR9) that used GL to classify the findings of what constitutes a good code change. This study used evidence from GL to classify eight themes (e.g., change description, change scope, code quality, code style).

Solution proposal **(5/9 studies; 55.5%)**. In this category, the use of GL contributed to proving solutions proposals to some problems or challenges faced. An example for this category, the study of Maro et al. (MLR6) identified some solutions proposals for software traceability in the automotive domain, in a presentation of one company, as we quoted: *"Two solutions have been suggested. One is to have tools that support the different disciplines with collaboration features such as chats, forums, and notifications. [...] Second is having a defined process on how the teams should collaborate [...]"*. The study of Plant (MLR2) used a book to show how they implemented their DevOps process: *"In order to ensure quality and information security, Muñoz and Díaz implemented phases from the OWASP Software Assurance Maturity Model (SAMM) [...]. The OWASP SAMM covers the phases of governance, construction, verification, and operations and therefore spans the complete DevOps life cycle [...]"*.

Opinion **(5/9 studies; 55.5%)**. This category was identified using opinions included in some GL sources. We employed the same meaning of Garousi and Küçük (GAROUSI; KüçüK, 2018) for "opinion" contributions, in which GL sources characterizing to emit "opinion". In this regard, an opinion about Android architecture based on a discussion from a blog post was used in Verdecchia et al.'s (MLR3): *"No. Do not retain the presenter I don't like this solution mainly because I think that presenter is not something we should persist, it is not a data class, to be clear"*. Another example was presented in Garousi et al. (MLR1) that used the content of a video presentation in a conference panel as evidence. A professor in the panel emitted an opinion about the root causes of low relevance of SE research, focusing on requirements engineering in the SE area: *"In my view, too often, research justified as satisfying the needs of the industry begins with a wrong or simplified understanding of industry's problems"*.

Concept Definition **(3/9 studies; 33.3%)**. GL was used to present some concepts and definitions in MLR studies. For instance, in Bhandari and Colomo-Palacios (MLR5), a web article presented the definition of holacracy, as followed: *"The literature defined holacracy in software development teams as a way of decentralized management and organizational governance where authority and decision-making are delivered throughout autonomous and self-organizing teams (circles)"*. Another use of this contribution was identified in Garousi's study (MLR1), in which a slide presentation defined the "impact" in SE research as *"How do your actions [in research] change the world?"*.

Experience report **(3/9 studies; 33.3%)**. To characterize the evidence found in experience-based studies, we employed the same approach of Garousi and Küçük (GAROUSI; KüçüK, 2018):

"Experience studies were those who had explicitly used the term "experience" in their title or discussions without conducting an empirical study". In this regard, the study of Verdecchia et al. (MLR3) used a guideline that provided diverse experience reports on how to test each code module (e.g., User interface and interactions, Webservice, Testing Artifacts). The study of Küçük (MLR5) used evidence from a blog post about unit testing that provided: *"a practitioner shared her experience of moving away from assertion-centric unit testing and fixing smells such as eager tests"*.

Others **(3/9 studies; 33.3%)**. Here we group the studies that the use of GL contributed with *"tools"*, *"code programming"*, and *"empirical evidence"*. In this regard, Plant (MLR2) presented a discussion from a *whitepaper* about the use of containers such as Docker in DevOps, as we quoted: *"They are therefore very resource efficient. However, configurations in Docker containers cannot be changed since containers cannot be updated. Updated software or configuration, therefore, requires a new image build"*. The study of Maro et al. (MLR6) used a *book* that explored test smells, as follows: *"[GL] explored a set of 'pitfalls' (smells) for JUnit and an Apache-based test framework named Cactus"*. The last example is about empirical study based in a blog post, present in Garousi and Küçük (MLR5), in which were explored open-source projects to investigate test redundancy, as we follow: *"[. . .] [GL] reported a study on more than 50 test suites from 10 popular open-source projects and found that higher amounts of test redundancy are linked to higher amounts of bugs"*.

Summary of RQ3.2: We identified that GL contributed in several manners to the MLR studies. Although the majority of these contributions were providing recommendations and explaining some topics or were used classifying the findings of the study. Other contributions were providing solution proposals, opinions, concept definition, and experience reports.

5.3.3 RQ3.3: What Types of Grey Literature Sources Are Most Commonly Observed in Multivocal Literature Review Studies?

In our investigation, we explored: (i) the use of each GL type in MLR studies and the relation between these types with the contribution identified by GL use, and (ii) the GL types and the types of producers identified.

For a better comprehension of Table 32, we informed: one GL type could be related to none, one or more of a type of contribution, and one study could be classified into none (blank), one, or in more than one type of contribution.

(i) Grey Literature vs. Contributions

We classified the 384 GL sources identified in MLR studies according to 19 GL types. Figure 19 shows the distribution of this classification from two perspectives. The first one (blue bar) presents the number of GL sources for each GL type. The second one (red bar) shows the number of MLR studies in which each GL type was found. The GL types identified were related to the type of contribution identified, as shown in Table 32.

Considering GL sources, *Blog posts* were the most common GL type found among the MLR studies (118 occurrences), used in six MLR studies (MLR1, MLR3, MLR5, MLR6, MLR7, and MLR9). Regarding the contributions related to its use, the most common was providing recommendations and opinions.

Slide presentations were the second type most commonly found in the studies (45 occurrences), used in four MLR studies (MLR1, MLR5, MLR6, and MLR8). Its use was commonly used providing recommendations and solution proposals.

Project or software descriptions were the third most found type (42 occurrences), although this type was used in only one study (MLR7). Its use provided the following contributions: solution proposals and recommendations.

Whitepapers were another type commonly found (25 occurrences), used in four MLR studies (MLR3, MLR4, MLR6, and MLR8). The main contributions related to this use were providing explanations, recommendations, and opinions.

(ii) Grey Literature Producers

We also investigated the producers of all 384 GL sources to identify who was the producer and to which GL types he/she was related. Figure 20 shows the results of these investigations.

Our first analysis shows that GL sources were produced mainly by SE *Practitioners* (130/384 GL sources; 31.9%), followed by *Consultants or Companies* and *Tool vendors*, each one representing, respectively, 21.3% (87/384 GL sources) and 21.1% (86/384 GL sources).

Our second analysis showed the relationship between GL types and producer types. Three

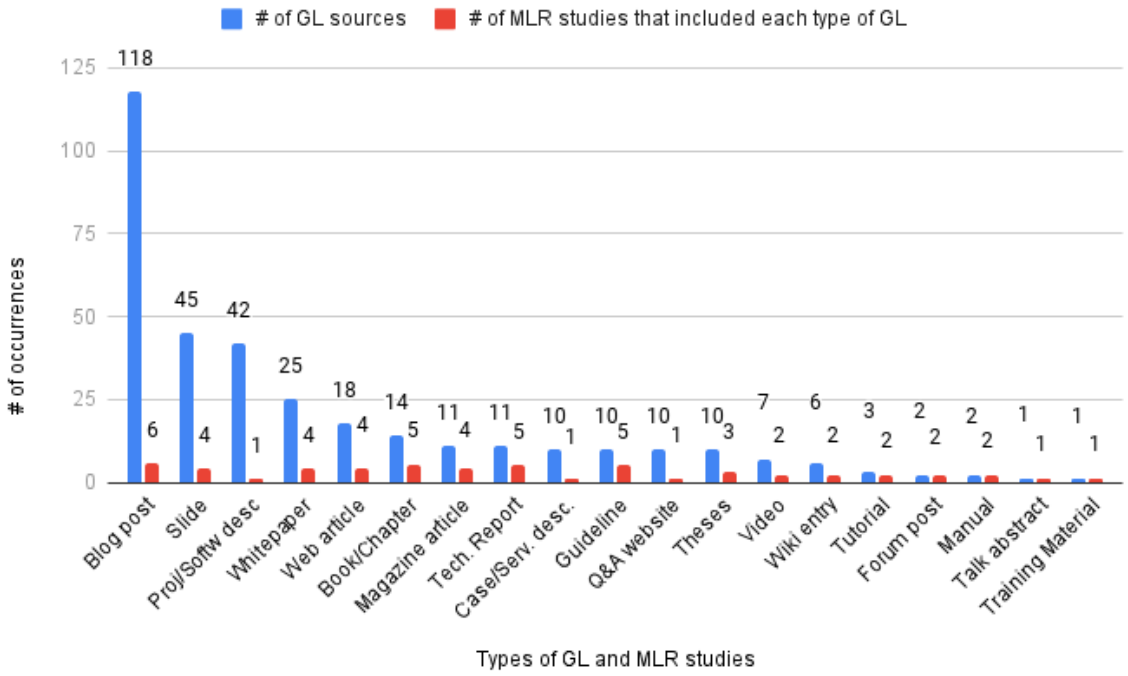
Table 32 – Grey Literature types vs. Contribution types in Tertiary Study 2.

Type of GL	Type of contribution									
	REC	EXPLA	CLA	SOP	OPN	DEF	EXP	TOOLS	PRO	EMP
Blog post	55	19	12	2	34	2	2	5		2
Book/Chapter	7	3	4	2	1			1	1	
Case/Serv. desc.	1			2						
Guideline	17	5	5		1		1			
Magazine article		1	2		1					
Q&A website	1	2								
Slide	9	5	3	6	2	1				
Proj/Softw desc	2			4						
Talk abstract	1									
Tech. Report	5	13	5	1	4					
Theses		3	2				2			
Video	4	8			1			1	3	
Web article	2	3		3		1				
Whitepaper	4	6	2	2	4	2		1		
Wiki entry	1	3								
* Unknown	1	3	1	1						
* Others	1	2	1	2	5	1				

CLA = Classification
EXP = Experience
SOP = Solution Proposal
PRO = Programming
EXPLA = Explanation
REC = Recommendation
DEF = Concept Definition
OPN = Opinion
TOOLS = Tools
EMP = Empirical Study

Source: the author.

Figure 19 – The amount of the Grey Literature found distributed by its types and the number of Multivocal Literature Review studies in which each Grey Literature type was used in Tertiary Study 2.

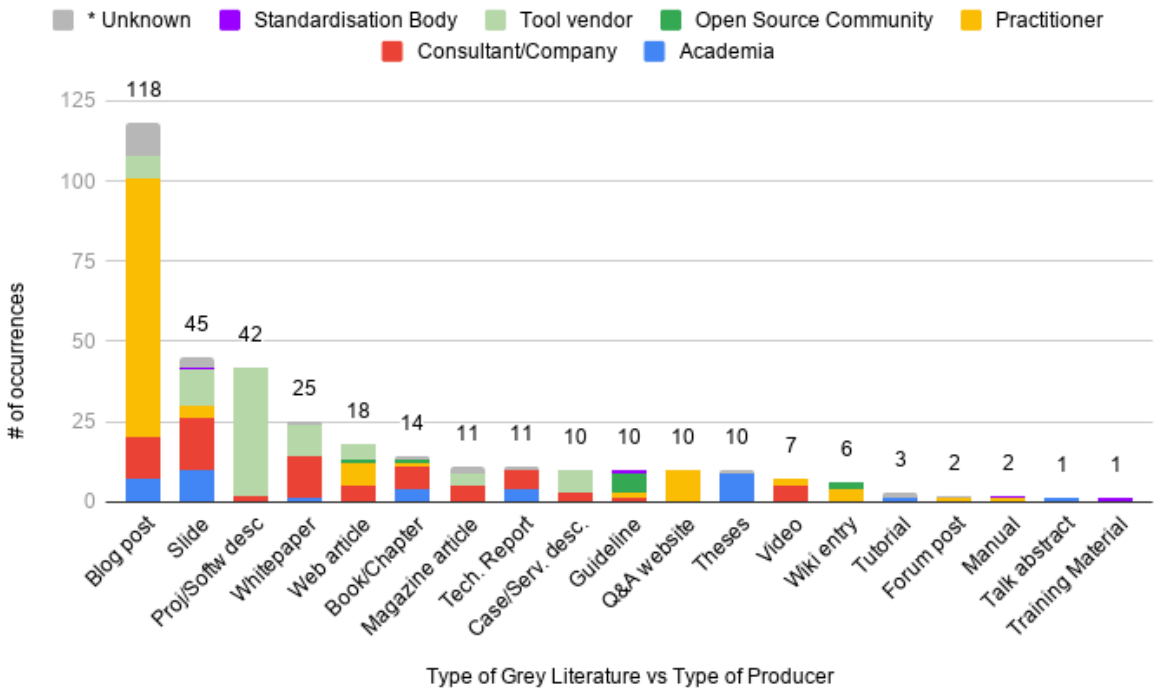


Source: the author.

types of producers (Practitioners, Consultant or Companies, and Tool vendors) caught our attention because they were responsible for almost 75% of the GL primary sources identified. We noted that *Consultants and Companies* contributed to more GL types. Their major contributions occurred with *slides* and *whitepapers*. *Practitioners* were the second one with more contributions in different GL types. The highlights of their contributions were mainly with *blog posts*, *web articles*, and *Q&A websites*. Finally, *tool vendors* were the ones that most produced *descriptions of projects or software* included in the MLR studies.

Summary of RQ3.3: The most common GL types identified among the studies were blog posts, slide presentations, project and software descriptions, and whitepapers. We also identified that most of the GL sources were produced by SE practitioners, consultants or companies, and tool vendors. Considering the GL type and the type of recommendation, we identified that blog posts were more related to providing recommendations and opinions. Guidelines were most related to recommendations, explanations, and classifications.

Figure 20 – Distribution of each Grey Literature type identified among the Multivocal Literature Review studies investigated, according to the producers types identified in Tertiary Study 2.



Source: the author.

5.4 DISCUSSION

This section discusses each research question, relating them to previous studies (Section 5.4.1). We present a discussion about some challenges that we faced to investigate MLR studies and conduct this research (Section 5.4.2). Finally, we discuss some threats to validity (Section 5.4.3).

5.4.1 Revisiting Findings

(RQ3.1) Use of Grey Literature in Multivocal Literature Reviews Studies

Observing the number of primary sources included in MLRs, GL has a significant contribution, as most of the studies included more than 40% of GL among the primary sources. Although GL sources had low inclusions rates in some studies, as in Garousi's study (MLR1). In our opinion, it reflected the research topic investigated, which was focused on the SE research area. This finding is consistent with our investigation of Study 2 (Chapter 4), where

the MLRs found, GL commonly accounts between 26-75% of the primary sources included.

(RQ3.2) Contributions of Grey Literature in Multivocal Literature Reviews Studies

In our investigation, our findings show that beyond the GL evidence supported some findings of TL sources, its use contributed with exclusive evidence that would not exist if GL were not investigated. It shows the importance of GL to address topics that are missing from TL sources (GUL et al., 2020).

Considering the study of Verdecchia et al. (MLR3), if they did not consider GL, no library, architectural standards, and guidelines presented on Android apps would exist since all these findings were identified only in GL and through interviews with Android SE practitioners. Moreover, in some studies (MLR4, MLR6), GL was the only type of source that had answers to some RQs (see Table 31). Thus, it shows the importance of GL evidence in contributing to the synthesis of MLR studies. Although in Garousi's study (MLR1) the inclusion of GL did not add anything different from what was found in TL. For this last study, we believe this happened because all GL included were produced in *Academia* by professors or researchers.

Furthermore, GL evidence is perceived as a benefit for several studies, for including different perspectives of TL and the practice of SE, as we identified in our investigation of Study 1 (Chapter 3) and in previous studies (RAINER; WILLIAMS, 2019b; ZHANG et al., 2020). This allows reducing the bias to the propensity for only studies reporting positive findings to be published, increase reviews' comprehensiveness and timeliness, and foster a holistic view of available evidence (PAEZ, 2017).

Our investigation shows that GL has essential contributions to MLR studies by providing helpful content with recommendations, explanations, and several other contributions, although the empirical evidence is scarce. We identified that the evidence provided in the investigated MLR studies was mainly produced by SE practitioners, consultants or companies, and tool vendors. Our findings corroborate with Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2016) as we found contributions related to practical solutions proposals, recommendations, opinions, or guidelines.

(RQ3.3) Types of Grey Literature in Multivocal Literature Reviews Studies

We identified 19 GL types used between the MLR studies investigated. The most common types were blog posts, web articles, and book chapters, produced mainly for SE Practitioners, Consultants or Companies, and Tool vendors. These findings show that studies using GL took advantage of evidence retrieved from the practice of SE. When we considered the only MLR studies, these findings are similar to our results presented in Study 2 (Chapter 4). Nevertheless, it highlighted the differences between MLR and GLR with other types of Secondary Studies, such as SLR and MS, in which, in the last, the most common types were books/chapter books, technical reports, and theses. The same situation we perceived when analyzing the GL producers, where for MLR and GLR studies, it was common to identify practitioners and consultants as they are among those who produce the most GL sources.

Previous studies investigated the GL types of sources used, but not their producers. For instance, Zhang et al.'s work (ZHANG et al., 2020) investigated Secondary Studies and identified that the most common GL types used were technical reports, blog posts, books, and theses. Another tertiary study conducted by Yasin and colleagues (YASIN et al., 2020) investigated a different time-span (studies published until 2012) of our research. Our results were quite different because Yasin and colleagues' study considered conference papers as a GL type. It is not straightforward to compare our findings related to the GL types included specifically in MLR studies, as studies by Zhang et al. and Yasin and colleagues conducted a broad investigation for any Secondary Studies, and they did not separate their findings by the type of Secondary Study. Only the study of Galindo Neto and colleagues (GALINDO NETO et al., 2019) that also investigated MLR studies, but did not mention the amount of use for each GL type. Instead, they only mentioned that MLR studies included videos, whitepapers, books, web articles, magazine articles, blog posts, and technical reports.

5.4.2 Challenges Investigating Grey Literature Contributions in MLR Studies

In this section, we describe some challenges we faced to investigate the GL in MLR studies, with a possible way(s) to address each one.

Challenge 1: Difficulty to identify the GL sources included in MLR studies. In some studies, it was a time-consuming activity to identify GL contributions since some of them had

hundreds of primary sources and some others did not classify the primary sources (e.g., MLR5) or did not present their references (e.g., (SALTAN, 2019; ECK et al., 2019)).

Potential way(s) to address: We recommended that SE researchers intending to conduct MLR studies to classify all primary studies/sources (TL or GL) for the first challenge. Moreover, we also recommended that GL be classified (e.g., blog post, book, theses). These recommendations are helpful for a more comprehensive understanding of GL use and to guide future researchers that may want to explore a specific GL source.

Challenge 2: Lack of GL information. The second challenge is related to the lack of information about the GL. For instance, some essential pieces of information (e.g., the title of the source, URL, last accessed, name of the author(s), type of GL, type of producer) were not available for several GL sources in MLRs studies (SALTAN, 2019; ECK et al., 2019). This challenge precludes a better understanding of each GL source and answers our research questions. For this reason, we removed these studies (SALTAN, 2019; ECK et al., 2019) from our analysis, although they presented some important information about GL in their studies. For instance, Saltan (SALTAN, 2019) investigates challenges about flaky tests, mentioning the high number of relevant GL sources identified compared with TL sources, which shows that flaky test understanding is still scarce.

Potential way(s) to address: To address this challenge, we recommended to the researchers include all information available from GL sources. This information may be essential for the reader to understand better the GL source used and guide future research to a profound investigation of GL sources.

Challenge 3: Difficulty to identify and classify GL contributions. The third challenge relates to identifying and classifying contributions by GL use, which is a consequence of the first two challenges. For instance, it was not possible to conduct a profound investigation of the GL sources in two MLR studies (SALTAN, 2019; ECK et al., 2019). Moreover, we perceived that the studies often did not highlight the differences between the findings from GL and TL.

Potential way(s) to address: One possible way to address the third challenge is following the Garousi et al.'s guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019) which recommended that the data extraction be conducted separated by the different types of source (GL and TL) and a balanced synthesis using sources with varying levels of rigor. In our opinion, another possibility is the synthesis highlight the differences between GL and TL, aiming to the reader understand how each type of primary source contributed to the study and the relevance of each piece of evidence presented. Moreover, we designed a process focused on helping SE researchers identify GL contributions in MLR studies.

5.4.3 Threats to Validity

We organized our threats according to the classification schema proposed by Ampatzoglou and colleagues (AMPATZOGLOU et al., 2020): Study Selection Validity, Data Validity, and Research Validity.

Study Selection Validity. In our research, as we intended to investigate MLR studies, we opted to select only the MLR studies that followed Garousi's guidelines. We are aware that this decision might have introduced a bias in our findings, limiting the number of MLR studies investigated and, consequently, the discussions' scope about the contribution and GL types identified. In addition, another threat is related to the selection process, as we selected only studies wrote in English that can lead to the omission of important MLR studies that followed Garousi's guidelines in other languages.

To identify the potentially relevant studies, we opted to use only Google Scholar (GS). This could hinder to identify more potential studies. Although, we took this decision because the investigation of Yasin et al.'s study (YASIN et al., 2020) showed that the use of GS is a sufficient search engine to identify primary sources, when compared the use with other academic search engines (e.g., ACM, ScienceDirect, IEEE Xplore, Springer Link). In addition, GS had the feature to search for studies that cited one specific source. In our case, Garousi's guidelines (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019).

Data Validity. Although Yasin et al.'s study (YASIN et al., 2020) mentioned the high cover of GS to identify primary sources, it is possible that any potential study was omitted and as we limited our analysis to MLR studies that followed Garousi's guidelines, it is not possible to generalize the results.

Considering our investigation, we extracted data based on the proposed research question. However, it was difficult to identify the GL contributions in MLR studies because, mostly, these contributions were not explicitly mentioned. Thus, the qualitative data extracted, in most cases, were related to our interpretation. Mainly related when we classified those contributions of GL use. Aiming to mitigate this threat, we designed and followed a paired process during this research and a third researcher revised the derived categories.

In addition, another threat is that we did not assess the quality of the GL sources included in those studies. Thus, we can not say that the exclusive information identified in GL sources improved the synthesis of the evidence of the studies.

Research Validity. We mitigate the research validity by adopting a paired process during the research analysis, and a third researcher revised the derived categories. In addition, to better analyze the GL contributions, we draw a process. Two experienced researchers revised this process.

Focusing to permit the replicability and future analysis by other researchers, we made available all the data extracted and analyzed in this research.

5.5 SUMMARY

This chapter presented a tertiary study, providing an investigation of nine MLR studies to show what evidence we would miss if we did not consider GL. To conduct this research, we designed a process to improve the identification of GL contribution in MLR studies.

We identified that several exclusive information was found retrieved only from GL sources. In addition, we identified that much evidence retrieved in TL was also supported by GL. Most of the GL sources were produced by SE practitioners, consultants, companies.

Considering how GL contributed to MLR studies, we identified that most of them were classifying information, providing an explanation of some topic, and providing recommendations.

By describing our findings, we expect to contribute to the state of the art in this topic, providing additional evidence on how GL contributes to Secondary Studies.

We inform that the investigation presented in this chapter was published in the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2021) (KAMEI et al., 2021a).

6 ASSESSING GUIDELINES FOR INCLUDING GREY LITERATURE IN SECONDARY STUDIES

This chapter presents two focus group research conducted to understand the perceptions of Software Engineering (SE) researchers regarding the usefulness, issues, and points for improvement of Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019), used to conduct Secondary Studies that used Grey Literature (GL); then answering RQ4. This work is unique because it is the first that assessed Garousi's guidelines. For Kitchenham and colleagues (KITCHENHAM et al., 2008), guidelines need to be evaluated before being widely adopted because if the guidelines are themselves flawed, they could cause the problem of poor quality reporting. Our investigation focused on the authors who have previous experience conducting secondary studies and authors of MLR and GLR studies that adopted Garousi's guidelines.

In the following, Section 6.1 presents our motivation and research questions. Section 6.2 describes the focus group methodology with the procedures and strategies used to conduct, collect, and analyze the data. Section 6.3 presents our findings. Section 6.4 provides the discussions about this research. Finally, Section 6.5 describes a summary of this research.

6.1 OVERVIEW

To achieve the stated goal, we explored the following research question:

RQ4: *What Are the Perceptions of Software Engineering Researchers About Garousi's Guidelines?*

We employed focus group research to have a broader, richer, and more collaborative perception of Garousi's guidelines to answer this research question.

In summary, our main findings in this research are:

- We elucidated that, in general, according to the investigated researchers, Garousi's guidelines are useful to support the conduction of MLR and GLR studies, although we identified some issues;
- We identified a set of reasons why SE researchers considered Garousi's guidelines useful;

- We elucidated a set of reasons why SE researchers perceived some issues in Garousi's guidelines as what they considered not useful, problems/challenges, or points for improvement;
- We provided a set of recommendations, based on the researcher's opinions and previous investigations, to deal with some issues that could improve the support to SE researchers better conduct their MLR or GLR studies.

By describing these findings, we expect to: (i) provide a discussion of useful guidelines to conduct MLR and GLR studies; (ii) provide an understanding of potential issues related to problems/challenges, and points of improvement of the assessed guidelines; and (iii) provide a discussion with potential ways to deal with the issues identified.

6.2 FOCUS GROUP

According to Morgan (MORGAN, 1996), the focus group method is a research technique used to collect data through group interaction on a topic determined by the researcher. In this research, we conducted the focus group based on the process proposed by Kontio et al. (KONTIO; BRAGGE; LEHTOLA, 2008), as shown in Figure 21. In this figure, our focus group process was divided into five phases: *Planning*, *Designing*, *Conducting*, and *Data Extraction*, and *Data Analysis and Synthesis*. In the following, we describe each phase.

For replication purposes, the data used in this chapter is available online at: <https://doi.org/10.5281/zenodo.6320920>.

6.2.1 Focus Group Planning

We began our research by defining the research problem, research questions, and objectives, as we previously presented in Section 6.1. In the following, we present how we composed our teamwork.

6.2.1.1 Teamwork

In the conduction of the focus groups, three researchers were directly involved: one vesting the role of the **moderator-researcher** and two alternating as **observers**.

The **moderator-researcher** was responsible for facilitating the interaction and the discussions between the participants, following a predefined questioning structure. Moreover, this researcher was responsible for permitting the integration and exposure of opinions and making interventions when the discussion has some conflict that would not be solved. The author of this thesis played this role.

The **observer** was responsible for taking notes of the verbal and non-verbal events. Beyond that, the observer can make interventions to clarify some talking and helping the moderator with the discussions. This role was alternated between two Ph.D. professors.

6.2.2 Designing Focus Group

In this section, we describe the process adopted to conduct the focus groups. In the following, we present how we selected the participants, the focus group strategies, and the group settings.

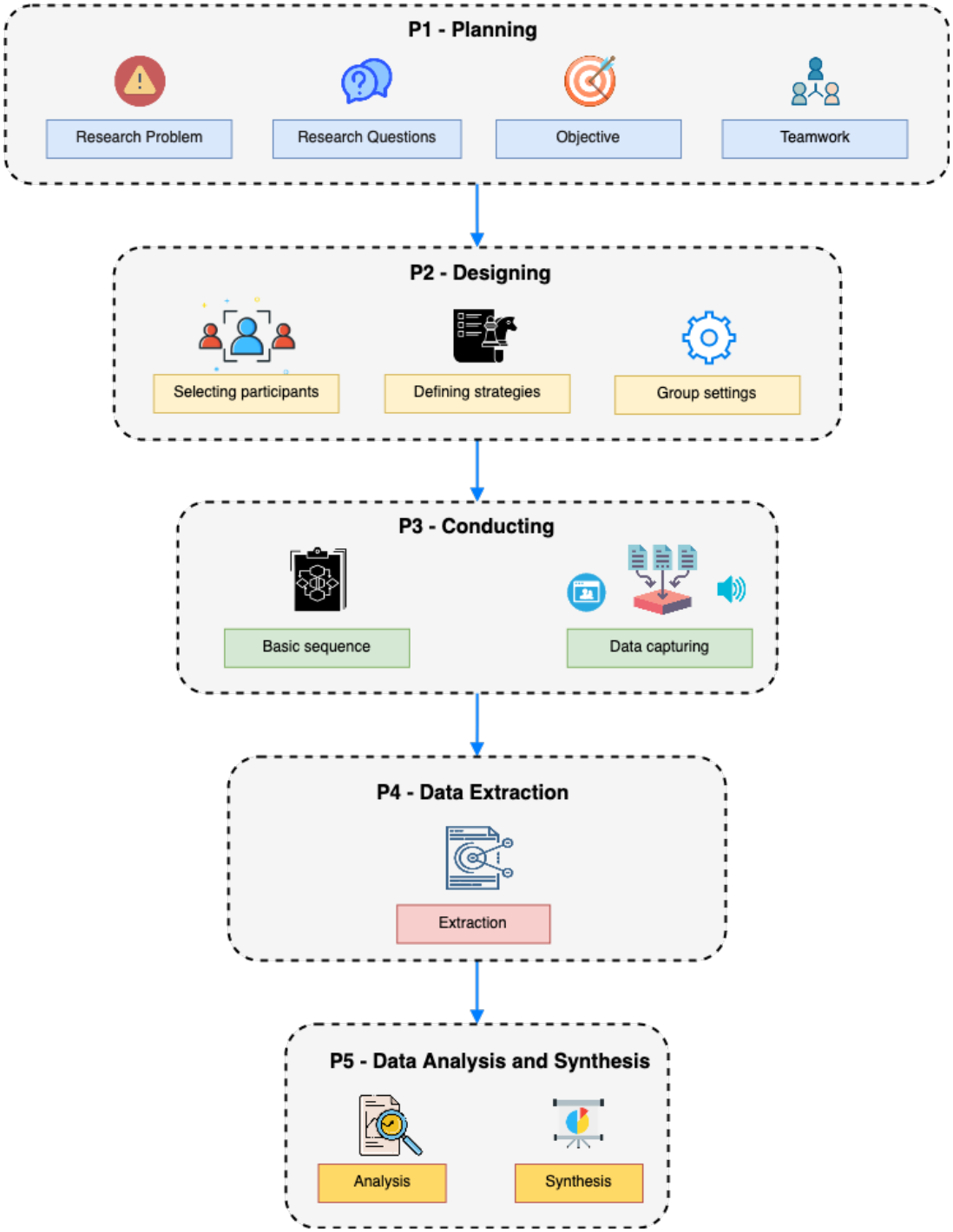
6.2.2.1 Selecting Participants

França and colleagues (FRANÇA et al., 2015) recommended that it is important to ensure the participants have the appropriate knowledge to participate in a focus group session. For this reason, we separated the participants into two groups: one with members of a research group, that focus on organize and publish empirical investigations to aid institutions and professionals in their decision-making regarding the evaluation of technologies, and the other with SE researchers with previous experience using Garousi's guidelines.

At first, we conducted a pilot study to validate our focus group plan, with the intention to discard all the data collected from this group. However, as few changes were required to conduct the second group and the approach used in the first did not affect or contaminate the data collected, we decided to keep the pilot study data in our investigation. This decision was based on Holloway's argument (HOLLOWAY, 1997), in which qualitative approaches, such as focus groups, pilot studies are not necessary because it is expected that the first analysis of the focus groups may help improve the later ones.

We called the pilot focus group as **first group (Focus Group 1 - FG1)**. To select their participants, we used a convenience sample (BALTES; RALPH, 2020) because the chosen participants were available or otherwise easy to study. They were composed of the SE researchers

Figure 21 – Overview of the Focus Group process.



Source: the author.

from the same research group as this author, with previous experience conducting secondary studies. However, most of them have no prior experience using Garousi's guidelines.

To select the participants of the **second group (Focus Group 2 - FG2)**, we employed a set of steps as we describe in the following:

1. We searched on Google Scholar for the studies that cited any version of Garousi's guide-

lines (GAROUSI; FELDERER; MäNTYLä, 2017; GAROUSI; FELDERER; MÄNTYLÄ, 2019);

2. We selected only studies based on Multivocal Literature Review or Grey Literature Review that clearly stated that they followed Garousi's guidelines;
3. We identified the authors of the selected studies;
4. We identified the number of studies that each author has participated in;
5. We invited all the authors by email, asking for the availability of participation in our focus group.

6.2.2.2 Defining strategies

We evaluated Garousi's guidelines from a theoretical perspective (KITCHENHAM et al., 2008) using the approach of "Perspective-based Reading" to identify useful points, problems, challenges, and areas for improvement, taking the viewpoints of the *researchers who read the guidelines* and the *authors of MLR or GLR studies that followed Garousi's guidelines*.

We organized the discussions similar to França et al.'s (FRANÇA et al., 2015) proposal, separating them into two steps. In step 1, the researchers were encouraged to argue the importance of each proposed guidelines' usefulness by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019). In step 2, to discuss the perceived issues related to what they considered not useful, problems/challenges, or points for improvement associated with Garousi's guidelines. In both phases, we encourage the researchers to interact and discuss, always asking if one person agrees with the other person's opinion.

We informed the researchers that the discussions and perceptions about each guideline would consider the entire section and the box related to each guideline. For instance, in the discussions about Guideline 10, the researchers should consider all the "Section 5.2.2. Source selection process" of Garousi's study. (GAROUSI; FELDERER; MÄNTYLÄ, 2019).

We highlighted that we did not assess guidelines 12 and 13 because, in our perception, they are general guidelines, not focused on the specific context for Grey Literature sources. Then, in the first focus group (FG1), we explored the remaining 12 guidelines. However, in general, FG1 researchers considered guidelines 1 (only one researcher perceived a useful point), 4, and 5 unnecessary. For these reasons, we improved our focus group strategy and removed these guidelines, remaining the guidelines 2, 3, 6–11, and 14 to be assessed by the second focus

group (FG2). Still, focusing on aggregating the guidelines by the topic explored (same section of Garousi's guidelines), we joint guidelines 1 and 2 and 6 and 7 to discuss them together.

6.2.2.3 Group Settings

To compose the groups, we tried to consider the expertise of the researchers conducting an MLR or GLR study using Garousi's guidelines, focusing on mixing researchers from different experiences to discuss the same topic in the same room to help them criticize each guideline, as recommended by França and colleagues (FRANÇA et al., 2015). However, due to lack of compatibility of the schedule of the researchers invited, it was not possible to consider the expertise. Thus, we grouped the researchers into two different focus groups sessions (FG1 and FG2), according to their availability and schedule.

Due to the diversity of time zones of the invited researchers, according to Kontion et al. (KONTIO; BRAGGE; LEHTOLA, 2008), it is important to focus on better identifying the time to schedule the focus group meeting. For this reason, we used Doodle[1] to support us on this.

6.2.3 Conducting the focus group sessions

In this section, we describe the basic sequence of how we conducted the focus groups. Then, we explained how the data of the focus group sessions were captured.

6.2.3.1 Basic Sequence

Each focus group begins with the moderator-researcher explaining how the focus group would work and describing Garousi's guidelines under evaluation. To each guideline investigated, we asked each researcher to think and write in the post-it card what they considered useful. Then, we encourage the discussion of each viewpoint, explaining whether they agree with other opinions or not, describing why, and if they have additional comments. Finally, we asked the researchers to arrange similar viewpoints cards and, in consensus, to prioritize them. The exact process occurred to investigate the issues of each guideline.

[1] www.doodle.com

6.2.3.2 Data Capturing

Each focus group session occurred using a Google Meet to record the capture the video, audio, and keyboard interaction. We used Miro[2], web-based software that permits interaction among the researchers, for instance, to write their opinions and vote on other researchers' views. Figure 22 presents an example of the use of Miro software by the participants of a focus group session. In this figure, it is possible to see the result of the researcher's interactions with the discussions about guidelines 9 and 10. To identify each researcher's perception, we attributed different cards (post-its) colors. In addition, we asked to identify the card with their names.

Figure 22 – Screenshot captured from a Focus Group session using Miro tool.

Source: the author.

6.2.4 Data Extraction

We performed the data extraction of the data collected from the focus group sessions conducted.

To help us with the data extraction, we used an online version of Microsoft Word (Office 365) that has the feature to transcribe the audio. Manually, corrections were essential to do in each transcription.

We extracted the data to a spreadsheet to support us in analyzing the audio transcribed. In the following, we present the data extracted: (i) who is the speaker, (ii) guideline discussed, (iii)

[2] www.miro.com

data extracted of the discussion, (iv) text of post-its, (v) usefulness mentioned, and (vi) issues mentioned. We separated the discussions related to what is not useful, problems/challenges, or points of improvement for the issues. In addition, we extracted the discussion about some specific points (e.g., if someone agrees or disagrees with another researcher's opinion).

6.2.5 Data Analysis and Synthesis

We employed a qualitative approach to analyze and synthesize the data collected from the answers using the Miro tool (video recorded), transcriptions of the audio discussions among the focus group participants, and the notes captured from the observers during each session.

One researcher performed the data analysis, although the final classifications of what was considered useful and issues with their sub-categories was revised and discussed by two other researchers. We inform that, for the issues, we had predefined sub-categories (not useful, problems/challenges, and points for improvement). The first we proposed, and the last two we retrieved from the study of Babar and Zhang (BABAR; ZHANG, 2009).

Our qualitative approach followed a thematic analysis technique (CRUZES; DYBÅ, 2011b), involving three researchers in this process. Figure 23 presents a general overview of this approach. We describe it in greater detail in these next points (adapted from Pinto and colleagues (PINTO et al., 2019)):

(a) *Familiarizing ourselves with data:* The process started with one researcher (R1) reading the focus groups' discussions transcriptions and listening and watching the video recorded from the Miro interaction, as expressed in Figure 23-(a).

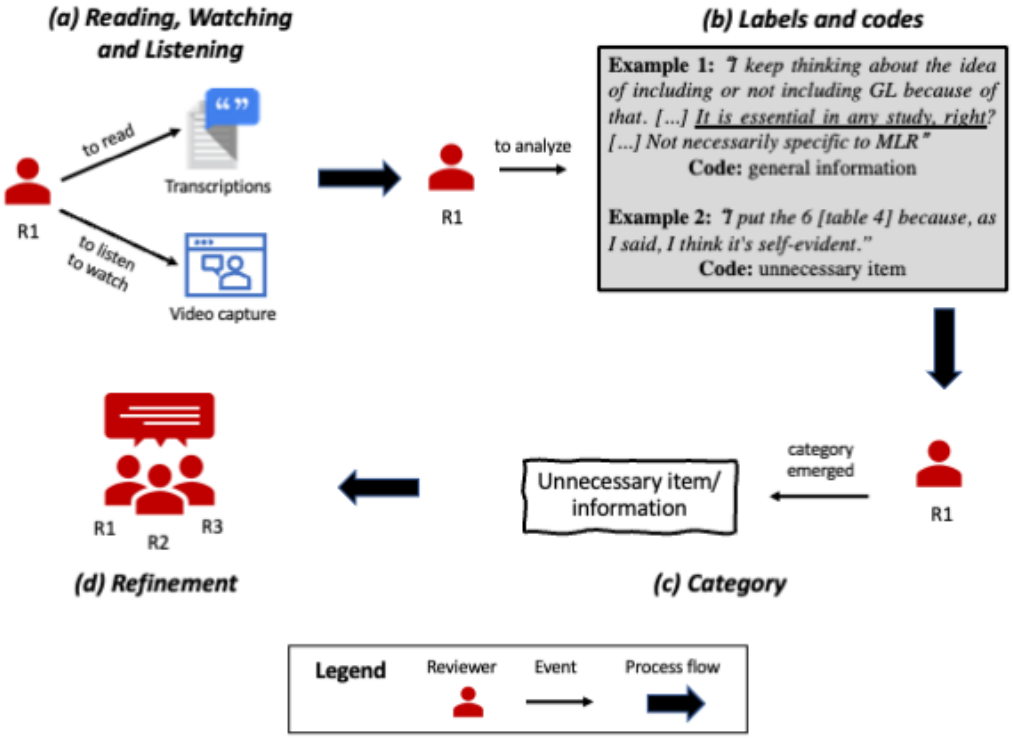
(b) *Initial coding:* In this step, researcher R1 individually added codes. We labeled portions of text without any pre-formed code, both to what researchers considered useful, the perceived issues, and points of improvement. Labels express the meaning of excerpts from the answer that represented appropriate actions or perceptions. The initial codes were temporary since they needed refinement.

(c) *From codes to categories:* Here, we already had an initial list of codes. Researcher R1 looked for similar codes in data. Codes with similar characteristics were grouped into broader categories. Eventually, we also had to refine the categories identified, comparing

and re-analyzing them in parallel. Figure 23-(c) presents an example of this process. This example exhibits how the category "Unnecessary item/information" emerged.

(d) *Classifications refinement:* In this step (Figure 23-(d)), we involved two researchers (R2 and R3) in evaluating all classifications. In the cases of any doubt, we solved them through discussions.

Figure 23 – Example of the coding process used to analyze the data collected in the Focus Groups sessions.



Source: the author.

6.3 RESULTS

This section presents our main results organized concerning the guidelines investigated. Section 6.3.1 gives an overview of the focus groups sessions conducted. Section 6.3.2 shows the findings related to each guideline, first presenting what were identified as useful, then the perceived issues.

6.3.1 Overview

Our search for MLR and GLR studies returned 211 potential studies. We removed studies that were not focused on SE and studies that did not follow Garousi's guidelines—in the end, resulting in 70 studies. Among these studies, we identified 182 authors.

The **first focus group (FG1)** occurred at the end of November 2021, involving five researchers. It was spent approximately three hours and was conducted in Portuguese. The **second focus group (FG2)** occurred at the beginning of January 2022, also involving five researchers. FG2 took approximately two hours and was conducted in English.

The participant's profiles were diverse. Table 33 describes **FG1** participants and Table 35 **FG2**'s. The majority of participants (70%) were men. The researchers were from different countries (Brazil, Mexico, Netherlands, and United States). Sixty percent of the participants had a Ph.D. degree. Considering the experience with a secondary study, all researchers had previous experience conducting SLR or MS, and seven also had experience conducting MLR or GLR.

Table 33 – Demographics information of the Focus Group 1 (FG1) participants'.

ID	Gender	Occupation	Country	Experience in SLR or MS?	Experience in MLR or GLR?*
FG1-01	Woman	Ph.D. student	Brazil	yes	no
FG1-02	Man	Ph.D. student	Brazil	yes	no
FG1-03	Man	Ph.D. professor	Brazil	yes	yes
FG1-04	Man	Ph.D. professor	Brazil	yes	yes
FG1-05	Woman	Ph.D. student	Brazil	yes	no

Note: (*) means the experience in conducting MLR or GLR studies following Garousi's guidelines.

Source: the author.

6.3.2 RQ4: What Are the Perceptions of Software Engineering Researchers About Garousi's Guidelines?

This section presents what SE researchers considered useful and a set of perceived issues for each guideline. To improve the comprehension of the results and enable traceability, we included

Table 35 – Demographics information of the Focus Group 2 (FG2) participants'.

ID	Gender	Occupation	Country	Experience in SLR or MS?	Experience in MLR or GLR?*
FG2-01	Man	Post doctor	Netherlands	yes	yes
FG2-02	Man	Post doctor	Netherlands	yes	yes
FG2-03	Man	Ph.D. professor	Mexico	yes	yes
FG2-04	Woman	Ph.D. student	United States	yes	yes
FG2-05	Man	Ph.D. professor	Netherlands	yes	yes

Note: (*) means the experience in conducting MLR or GLR studies following Garousi's guidelines.

Source: the author.

direct quotes extracted from the transcriptions and post-it cards, using the ID of the researchers presented in tables 33 and 35, to identify who is responsible for each argument/discussion.

An overview of the assessment of Garousi's guidelines, considering what researchers perceived as ***useful*** is presented in Table 37 and for the ***issues***, in Table 38.

In the following, we first briefly describe each guideline, what the researchers considered useful, and some identified issues.

Guideline 1 - Structure of the protocol

Guideline 1 is part of the Phase of Planning the Review. In this phase, Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019) explained the existing process for Secondary Studies (e.g., Kitchenham and Charters (KITCHENHAM; CHARTERS, 2007)) and proposed a new one called "A typical process for MLR studies".

In the following, we present our findings regarding what researchers considered useful and perceived issues related to Guideline 1.

Useful

When we asked opinions about Guideline 1 for FG1, only researcher *FG1-05* considered this guideline to be *important for beginners researchers* in secondary studies and *help understand an MLR process*. While the others considered it an obvious guideline, as we discussed in the

Table 37 – Overview of the usefulness of Garousi's guidelines perceived in the focus groups.

ID researcher	Guidelines											
	1	2	3	4	5	6	7	8	9	10	11	14
FG1-01		**x**	**x**			**x**	**x**	**x**	**x**	**x**	**x**	**x**
FG1-02		**x**	**x**			**x**	**x**	**x**	**x**		**x**	**x**
FG1-03		**x**	**x**			**x**	**x**	**x**		**x**	**x**	**x**
FG1-04		**x**	**x**			**x**	**x**	**x**			**x**	**x**
FG1-05	**x**	**x**	**x**			**x**	**x**				**x**	**x**
FG2-01	!	**x**	**x**	!	!	**x**	**x**	**x**		**x**	**x**	**x**
FG2-02	!	**x**	**x**	!	!			**x**		**x**	**x**	**x**
FG2-03	!			!	!			**x**	**x**	**x**	**x**	**x**
FG2-04	!	**x**	**x**	!	!			**x**			**x**	**x**
FG2-05	!	**x**	**x**	!	!	**x**	**x**	**x**			**x**	**x**
#Total	1	9	9	-	-	7	7	9	3	5	10	10

x = Guideline was considered useful.
! = Guideline was not assessed on Focus Group 2 (FG2).
Notes:
1) Guidelines 2 and 3, and 6 and 7 were assessed together;
2) Guidelines 12 and 13 were not included because they were not assessed.

Source: the author.

following.

Issues

Almost all FG1 researchers (4 out of 5) considered Guideline 1 **not useful** because they considered it, *"Similar guideline/information presented in previous studies"*. In this regard, *FG1-02* pointed out that: *"In my opinion, it would be more useful if it refers to Kitchenham's guidelines"*. Two other researchers in this group also had the same opinion. For instance, *FG1-04* pointed out that *"[...] I agree that this guideline is unnecessary since there are already others, which, if they are not the same, are very similar, which could be reused."* and *FG1-03* mentioned *"I have the same opinion as [FG1-04]. The entire paper could be simplified. For this case, in particular for Guideline 1, it was only to point to existing guidelines and show only specific parts of an MLR study"*.

We highlighted that, regarding the *FG1-05*, although we included their opinion pointed out

Table 38 – Overview of the issues of Garousi's guidelines perceived in the focus groups.

ID researcher	Guidelines											
	1	2	3	4	5	6	7	8	9	10	11	14
FG1-01	x	x	x	x	x	x	x	x	x	x	x	x
FG1-02	x	x	x	x	x	x	x	x	x	x	x	x
FG1-03	x	x	x	x	x	x	x	x	x	x	x	
FG1-04	x	x	x	x	x	x	x	x	x	x	x	x
FG1-05	x	x	x		x	x	x		x	x	x	x
FG2-01	!	x	x	!	!	x	x	x	x	x	x	x
FG2-02	!	x	x	!	!	x	x	x	x	x		
FG2-03	!	x	x	!	!	x	x	x	x	x	x	x
FG2-04	!	x	x	!	!	x	x	x	x	x	x	x
FG2-05	!	x	x	!	!	x	x	x	x	x	x	x
#Total	5	10	10	4	5	10	10	9	10	10	9	8

x = Were identified issue(s) in the Guideline.

! = Guideline was not assessed on Focus Group 2 (FG2).

Notes:

1) Guidelines 2 and 3, and 6 and 7 were assessed together;

2) Guidelines 12 and 13 were not included because they were not assessed.

Source: the author.

the usefulness of the guideline, she mentioned that this guideline is not useful for experienced researchers.

Guidelines 2 and 3 - The decision whether to include a GL in a review

Guidelines 2 and 3 are about raising (motivating) the need for an MLR, focusing on helping SE researchers to ensure that the usefulness of an MLR is necessary.

Guideline 3 brings a table, called *Table 4*. In this chapter, when we mention *Table 4*, we refer to a table of Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019), with seven questions used to help SE researchers decide whether to include a GL in Secondary Studies. To better comprehension of the reader, we included the questions of this table in Figure 24.

In the following, we present our findings regarding what researchers considered useful and the perceived issues related to guidelines 2 and 3.

Figure 24 – Questions presented in Guideline 3 (Table 4) of Garousi's guidelines to support decision whether to include Grey Literature in a review.

Table 4
Questions to decide whether to include the GL in software engineering reviews.

#	Question	Possible answers
1	Is the subject "complex" and not solvable by considering only the formal literature?	Yes/No
2	Is there a lack of volume or quality of evidence, or a lack of consensus of outcome measurement in the formal literature?	Yes/No
3	Is the contextual information important to the subject under study?	Yes/No
4	Is it the goal to validate or corroborate scientific outcomes with practical experiences?	Yes/No
5	Is it the goal to challenge assumptions or falsify results from practice using academic research or vice versa?	Yes/No
6	Would a synthesis of insights and evidence from the industrial and academic community be useful to one or even both communities?	Yes/No
7	Is there a large volume of practitioner sources indicating high practitioner interest in a topic?	Yes/No

Note: One or more "yes" responses suggest inclusion of GL.

Source: adapted of Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019).

Useful

When we analyzed the discussions for guidelines 2 and 3, we identified a set of reasons "why" the researchers considered this guideline useful (see Table 40), showing that these guidelines were considered one of the most useful guidelines mentioned by nine researchers. As we presented in the following, the main reason for this perception was related to some questions brought by *Table 4* in Garousi's guidelines.

The most useful point identified was that these guidelines helped *deciding whether to include* a GL in the study. This reason was most found when the researchers discussed Question 1 of *Table 4* in Garousi's guidelines. According to researcher *FG1-01*, this item helped to make this filter broader. Researcher *FG1-03* agreed with researcher *FG1-01* and said: *"I put the same too. I also consider that this item, especially item 1, brings an important thought that we need to know whether we will include a grey literature"*.

For *FG2-04*, *Table 4* in Garousi's guidelines is also important to argue why you have included GL for the article's reviewers, as we pointed out: *"[. . .] the reviewers ask you questions like, why do you think that it's useful to forward in the approach? Why do you choose to do this grey literature? So to answer that kind of question I guess following this guideline could be helpful because, as you said, you have a checklist to talk about, and I guess, in that case, these questions are useful [. . .]"*.

The second most useful point, mentioned in both focus groups, is that questions 1, 4, and 6 in *Table 4* (Garousi's guidelines) give importance to *bring practical experiences*, as researcher *FG1-04* wrote in a post-it card: *"#4 interesting to connect with practice"* and complemented during the discussion: *"I think that item 4 there, that you are going to look for the problems and certainties that professionals have in practice. I think it's one of the most*

important types of contributions that we can make with this type of study". In this respect, researcher *FG2-02* mentioned that those questions are useful because they combine academia and industry, as we pointed out: *"Uh, also sometimes we have to combine the whatever we have in the academic with some insight from the practice"*.

The third point highlighted was that question 4 (*Table 4* of Garousi's guidelines) helps *to validate findings from academia and industry*. This reason was only mentioned in FG1. In this regard, researcher *FG1-05* using a post-it card wrote: *"#4 validate or corroborate results using evidence that may be closer to professionals"*. For *FG1-04*, question 4 is the most important, as we pointed out: *"[. . .] where we can verify if what the academy is saying is the same as the industry and vice versa"*.

Table 39 – Perceived usefulness identified in the focus groups of guidelines 2 and 3.

Reasons	# FG1	# FG2
To decide whether to include GL	5	4
To bring practical experiences	4	3
To validate findings from academia and industry	4	0
To provide additional evidence	1	0

Note: The columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

Issues

Although the identified perception of the usefulness of guidelines 2 and 3, several issues were identified related to not useful points, points of improvement, and one problem/challenge, as shown in Table 40. In the following, we discussed some of these issues.

Table 40 – Perceived issues identified in the focus groups about guidelines 2 and 3.

Issues	# FG1	# FG2
Not useful		
Unnecessary item/information	5	0
Points of improvement		
To avoid the use of vague or subjective definitions	2	5
To include prioritization criteria for each question of *Table 4*	4	1
To include a question in *Table 4* to cover recent topics	0	4
To improve Guideline 3 changing the word "systematically" to "structured"	0	1
Problem/Challenge		
Difficult to the questions of *Table 4* can be used universally	0	1

Notes:
(1) Columns with "#" show the number of researchers that mentioned a given category for each focus group.
(2) "*Table 4*" refers to a table of Garousi's study (GAROUSI; FELDERER; MÄNTYLÄ, 2019). Questions of this table are presented in Figure 24.

Source: the author.

Considering the opinions in which thought something **not useful**, we identified five researchers of FG1 considering Question 6 (*Table 4*) as an *"Unnecessary"* item. This question covers the importance of synthesizing evidence from the industrial and academic community. For *FG1-03*, this question is not necessary because *"[. . .] The results of investigations in ESE (Empirical Software Engineering) should always aim to have these results"*. Researchers *FG1-01* and *FG1-04* considered it is an obvious question. For instance, using a post-it card, researcher *FG1-04* expressed that *"[Question] #6 is self-evident"*. Another not useful point mentioned was Question 3 that is related to the context of the information. As she wrote it on a post-it card, all the contexts are important for researcher *FG1-05*. Researchers *FG1-01* and *FG1-04* agreed with this viewpoint. For *FG1-04*, it is a little weird that this item exists in a guideline. The participants of FG2 did not point to anything as not useful.

Considering the **points for improvement**, most of them were related to *"To avoid the use of vague or subjective definitions"*. In this respect, researcher *FG1-04* pointed out that *"It is confused to define what is considered as contextual information"*. Researcher *FG2-05* discussed another example of vague definition: *"Question #7, is extremely vague! So what is large is it? Are we talking about 100 hits? Five thousand knows like usually vague Google*

something and get something in the order of the billions you know. So, it is like, is it large? If it is a billion, is it large?". Researchers *FG2-01*, *FG2-02*, and *FG2-04* agreed with *FG2-05*. Researcher *FG2-04* complemented the discussion as we pointed out that: *"[. . .] I want to point is that in Question #1 also what is the definition of the complex? Exactly the same as the large and in the last question, the complex in the first question is also vague"*.

The second point of improvement most mentioned was to *"Include prioritization criteria for each question of *Table 4*"*. According to Garousi et al.'s interpretation, one single item in *Table 4*, like "Yes", suggests GL inclusion. Although, for some researchers of both focus groups, it is not a matter of the number of items with "Yes", but **the relevance** of each item, as *FG2-04* pointed out in a post-it card: *"Some questions are more important than others"*. In the discussion of FG1, researcher *FG1-02* pointed out that *"Only answering 'yes' the question #7 of *Table 4*, it is not reasonable to perform an MLR"*.

One **challenge** was also perceived related to questions presented in *Table 4* of Garousi's guidelines, as researcher *FG2-03* wrote in a post-it card: *"By being pretty straightforward, I'm not sure that these questions can really be universally applied when it comes to the details of a matter"*.

Guideline 4 - Defining research questions

Guideline 4 described setting the goal and the research questions. **No useful points were mentioned** for this guideline. For this reason, as mentioned before (see Section 6.2.2), it was not assessed by the second focus group (FG2).

In the following, we present our findings regarding the perceived issues related to Guideline 4.

Issues

For most researchers of FG1 (4 out of 5), Guideline 4 has some issues. For instance, researcher *FG1-03* considered it as a **not useful** guideline because it is *Similar considering the existence of previous studies*, as we pointed out: *"So I wonder if this should really be there. I know that Kitchenham also has this information, right, but hers was the first, right? Why didn't Garousi just reference her? What he points out in these guidelines [. . .] since the first one, I've already realized that they have extremely basic things, or that wouldn't fit there.*

Because so, imagine if I were to put in a paper everything that others have already said in full. The paper will never end [laughs]".

Other researchers considered some **points for improvement**, for instance, *FG1-02* mentioned that *several parts of this guideline could be removed*, as we pointed out: *"I agree with FG1-03. As he said, the information is already there in Kitchenham. The article ends up getting even wordy, right? For example, in my view, tables 5 and 6 should not be in the paper"*. The same opinion has the participant *FG1-01* that mentioned: *"When I looked at the sections of this guideline I saw so much that didn't need to be there"*.

Guideline 5 - Adopting various Research Question types

Guideline 5 suggested adopting various types of RQ, being aware that primary studies may not allow all the questions types to be answered.

At the end of the discussion about this guideline in FG1, we did not identify any useful point mentioned. In the beginning, researcher *FG1-05* wrote in a post-it card: *"Useful to understand the types of RQ"*. However, during the discussion with other participants, she changed her idea and agreed that this guideline was not useful. For this reason, as the same as the previous guideline, it was only assessed by the first focus group.

In the following, we present our findings regarding the perceived issues related to Guideline 5.

Issues

During the FG1, all five researchers questioned the usefulness of Guideline 5. Although they perceived the importance of understanding the diverse types of research questions, it was a consensus that the guideline is **not useful**. In this regard, researcher *FG1-04* pointed out that: *"[Guideline 5] is a basic topic of the scientific methodology"*. Researcher *FG1-01* said that *"[. . .] it does not make any sense to say that I have to adopt several types of RQs"*.

Three researchers mentioned **points of improvement** for this guideline. For instance, according to *FG1-01*, the obligation to adopt several types of RQ should be removed, as we pointed out: *"Does not make sense to say that I have to adopt several types of RQ. Can't I have just one? Justify!"*, and *FG1-02* complemented talking that *"Good RQs (research questions) does not mean they must have been of different types"*. A similar opinion had *FG1-03* when

he wrote in a post-it card: *"Why adopt multiple RQs? I think it makes no sense"*.

Guidelines 6 and 7 - Identifying and Searching for relevant Grey Literature

Guidelines 6 and 7 presented information about the search process for GL, focusing on explaining several strategies that could be used to identify relevant sources, search engines, and specialized databases and websites.

In the following, we present our findings regarding what researchers considered useful and the perceived issues related to guidelines 6 and 7.

Useful

Seven researchers perceived useful points in guidelines 6 and 7. Five researchers from the FG1 and two from the FG2. We highlighted that the most significant differences between the participants' perceptions of both focus groups were for these guidelines. The list of usefulness perceived for these guidelines is presented in Table 41. In the following, we discussed some of them.

Table 41 – Perceived usefulness identified in the focus groups about guidelines 6 and 7.

Reasons	# FG1	# FG2
To understand what are the important GL sources	5	0
To explain how to use a general web search engine for GL	4	0
To support beginners researchers	1	1
To find new/good sources contacting individuals directly or using Snowballing	1	1

Note: The columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

As shown in Table 41, guidelines 6 and 7 were considered useful *"To understand what are the important GL sources"*. This category was related to the importance of understanding GL types and better defining the scope limit. For instance, according to researcher *FG1-01*, they are essential to researchers that did not familiarize with GL, as she mentioned: *"It's very useful because it's putting things there. For example, I, who only used white literature, didn't know about it"*. Researcher *FG1-05* considered them important to identify relevant sources early on, as she wrote in a post-it card: *"Identify the origin of the information as soon as possible to*

determine or not its inclusion". Other observations were done using post-it cards. For instance, *FG1-02* pointed out that: *"It's important to define the scope that will be included as GL"*.

The second most useful point, perceived by four researchers, was that guidelines explained and gave examples of *"To explain how to use a general web search engine for GL"*. For researcher *FG1-05*, it was important to know about the possibility of using a general web search engine. Researcher *FG1-03* agreed with researcher *FG1-05*, although he also highlighted that it is important to use it with a quality assessment.

Other useful points were perceived. For instance, one researcher of each focus group considered these guidelines useful *To support beginners researchers*. In this respect, researcher *FG2-01* wrote in a post-it card that: *"It's useful for new researchers to give them, a general idea of the whole process"*. Other reasons mentioned were *To find good sources using Snowballing* and *Contacting individuals directly is useful when you have few materials*, mentioned by *FG1-04*.

Issues

Despite the perceived usefulness of guidelines 6 and 7, we also identified several issues, as shown in Table 42. In the following, we discussed some of them.

Considering the **not useful** points perceived, nine researchers defined them as *obvious guidelines*. For instance, researcher *FG1-03* wrote in a post-it card that: *"Guideline 6 seems obvious to me"*. During the discussion of this point, all FG1 researchers agreed with *FG1-03* opinion. Researcher *FG1-04* complemented that it is evident for experienced researchers. Some researchers of FG2 shared similar ideas. For instance, *FG2-05* wrote in a post-it card that *"A bit naive: it's pretty obvious that I will choose data source pretty early on"*, and *FG2-04* wrote that *"simple and obvious guideline"*. Researcher *FG1-05* pointed out that even it is an obvious guideline, it is useful for novice researchers. This opinion was shared by other researchers of FG1 (*FG1-03* and *FG1-04*) and by four researchers of FG2.

Some **points of improvement** were discussed. For instance, related to the discussion about an obvious guideline, researcher *FG2-01* proposed *To include a flow to guide different directions to novice and experienced researchers*, as we pointed out that: *"Maybe it's useful to very new researchers. So, it would be better if the guideline has a flow: if you are not familiarized with any type of secondary study, follow this one"*. All other researchers of the FG2 agreed with the proposal to use a flow.

Table 42 – Perceived issues identified in the focus groups about guidelines 6 and 7.

Issues	# FG1	# FG2
Not useful		
Obvious guideline	5	4
Points of improvement		
To include a flow to guide different directions to novice and experienced researchers	0	5
To better explain about GL definitions and types	3	1
To explain that "Contact individuals directly" is a difficulty strategy to obtain rapid results	4	0
To improve the understanding is better to join guidelines 6 and 7	3	0
To include decision-making criteria for selecting GL types and sources	3	0
To improve the readability of the guidelines	2	0
To explain that the process to identify relevant GL sources could be iterative	0	2
To propose more specific guidelines on how to identify relevant sources for a given type of problem	0	1
To improve the process of Figure 7	0	1
To improve the explanation of how to select GL sources	1	0
Problems/Challenges		
The tiers' classification depends on context	0	4
Using reference lists and backlinks are needed extra tools	2	0
Difficult to replicate the results using a general search engine	0	1

Notes:
(1) Columns with "#" show the number of researchers that mentioned a given category for each focus group.
(2) "Figure 7" refers to a figure of Garousi's study (GAROUSI; FELDERER; MÄNTYLÄ, 2019).

Source: the author.

According to four researchers, the conduction of an MLR implies *To explain GL definitions and types* better, as *FG1-05* wrote in a post-it card: *"It seems obvious, but it depends on what you understand about GL"*. Researchers *FG1-03* and *FG1-04* agreed with this point raised by *FG1-05*. Another issue was in concern of the strategy to search for GL. It was mentioned the importance *of explaining that "Contact individuals directly" is a difficult strategy to obtain rapid results*. Thus, according to *FG1-04*, this strategy should only be used if you have few

results retrieved from the other search strategies. Researchers *FG1-01*, *FG1-03*, and *FG1-05* agreed with researcher *FG1-04*'s opinion.

Some improvements regarding the readability of these guidelines were discussed. For instance, for researcher *FG1-01*, it is important *To improve the readability of the guidelines* because, according to her, only reading the boxes for these guidelines, *"I didn't quite understand what he meant there. He could have been more explicit in his proposal for that guideline summary"*. Another point to improve was to join guidelines 6 and 7, as *FG1-03* pointed out: *"Guideline 6 seems obvious to me, as I said before. But then, when you put guidelines 6 and the 7 together, there comes this idea of the definition of GL that you need to have in advance to really know what you're going to do with it"*.

Another point, mentioned only by two FG2 researchers, is related to Guideline 6, which addresses the importance of identifying GL sources early on. However, for researcher *FG2-01*, an *iterative mindset* is better, as we pointed out: *"I think it's also better always to have any iterative mindset on this kind of thing. So firstly, sub small with, let's say, Stack Overflow, and then you move on to some other site. So, you get more and more data and don't get overwhelmed"*. In this regard, researcher *FG2-04* said: *"GL types and data sources may not be clear in early stages and researchers may revisit them during the process"*.

Researchers also identified some **problems/challenges**. For instance, four FG2 researchers perceived that *the tiers' [shades of GL] classification depends on context*, as researcher *FG2-03* pointed out: *"It has like tiers [shades of GL], and it puts whitepapers in the 1st tier. But I think that's also like context-dependent because most of the time, at least for my problem, whitepapers were mostly made to sell stuff like they were selling their own platform"*. Researcher *FG2-01* agreed and complemented the discussion: *"Yeah, whitepapers should be in figure one in the third tier, I think because they are completely unreliable"*. Other problems/challenges were perceived, for instance, two researchers (*FG1-02* and *FG1-04*) discussed that it is a challenge *to use the reference lists and backlinks because it needs extra tools* that usually have a paywall, and some researchers may have no access. Another one was related to *difficulty of replicating the results using a general search engine*, pointed out only by researcher *FG2-05*.

Guideline 8 - Stopping criteria for Grey Literature searches

Guideline 8 provided information about three possible stopping criteria (theoretical saturation, effort bounded, and evidence exhaustion) to be chosen for the search process for GL

sources.

In the following, we present our findings regarding what researchers considered useful and the perceived issues related to Guideline 8.

Useful

All researchers perceived useful points in Guideline 8. The list of useful perceived for this guideline is presented in Table 44.

For most researchers (8 out of 10), the stopping criteria are important in diverse manners. For instance, in general, helping SE researchers *To decide when to stop the search*, perceived in both focus groups. For *FG1-03*, pointed out that: *"In general, the idea to bring stop criteria is very useful. It is something that I cannot see in traditional guidelines, for instance, to conduct SLR or MS"*.

In addition, the criteria of the *Effort bounded* is crucial *to avoid an exhaustive search process*, as perceived by all FG1 researchers. In this respect, *FG1-04* pointed out: *"It is particularly important [criterion 2] because I think it is the most different from the way systematic review is done in general. [...] That's why criterion 2 is important, which is you just look, I don't know, the 50 pages of Google, the 20 pages of Google, and that's good because if not, there's no end"*. In addition, *FG1-04* pointed out that criteria 2 is also important *To better deal with a general web search*, such as Google, because he was unaware that this could be used.

Table 44 – Perceived usefulness identified in the focus groups about Guideline 8.

Reasons	# FG1	# FG2
To decide when to stop the search	3	5
To avoid exhaustive search process	5	0
To better deal with a general web search	2	0

Note: The columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

Issues

As shown in Table 45, several issues were also identified for Guideline 8. The issues were related to points of improvement and challenges. In the following, we discussed some of them.

Table 45 – Perceived issues related to Guideline 8.

Issues	# FG1	# FG2
Points of improvement		
To improve the definitions of each stop criteria and its differences	5	2
To include a "Limit date" as new stop criterion	0	2
To randomly select a sample of N first results	0	1
Problems/Challenges		
Hard to define the limit of a valid effort	1	2
Hard to reach evidence exhaustion	2	1

Note:
Columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

Although all researchers considered Guideline 8, in general, useful, they also identified some **points of improvement**. For instance, seven researchers mentioned the *Vague definition of each stop criteria and its differences*. In this respect, researcher *FG2-01* using a post-it card, argued that: *"Third point [evidence exhaustion] is unclear. How do you know you extracted all the evidence?"*. Others perceived that it is not clear the difference between the "Theoretical Saturation" and "Evidence Exhaustion", as researcher *FG1-03* pointed out: *"I couldn't see the difference between criteria 1 and 3"*. During the discussion for this point mentioned by *FG1-03*, all FG1 researchers agree with this point. About the same point, researcher *FG2-05* wrote in a post-it card: *"How 'evidence exhaustion' is defined, and how it differs from 'theoretical saturation', is not clear to me"*.

We also identified other points of improvement. For instance, *To add a "Limit date" as a new stop criterion* mentioned by two researchers of FG2. At this point, researcher *FG2-03* begins saying that the "evidence exhaustion" using some GL types of sources, such as Twitter, is hard if there's new information every day. For this reason, it is important to set a limit date

to extract the evidence, as he pointed out: *"So, for instance, in a very specific case, we had to set a date right after this date. We're not going to be looking at more data, and that's not a point that's in the guidelines"*. In addition, it was mentioned the inclusion of a new stop criterion. For instance, *To randomly select a sample of N first results*, as *FG2-02* pointed out: *"Uh, but maybe we need some additional criteria. [. . .] randomly or select a sample, for example, let's say that we check the first ten pages, then after that, if there are a lot of other pages so result we select a randomly selected sample. And see whether they are a new concept or maybe need some new guideline or criteria"*.

Two **problems/challenges** were perceived. Three researchers (one from FG1 and two from FG2) believed it is *"Hard to define the limit of a valid effort"*. For researcher *FG1-02*, the "Effort bounded" stop criterion is complicated, as we pointed out: *"It's hard to say that it will only make up to page 10. How much is really an adequate number?"*. Although all FG1 researchers considered the "Effort bounded" as the most important stop criteria, they also agree with what was mentioned by researcher *FG1-02*. Researcher *FG2-04* also perceived a challenge using this criterion, as we quoted: *"I used the number 2 [effort bounded], but my problem with that criteria was it was hard to convince why I stopped at page 10 for example, so we will keep asking me, why don't you look at the rest of our results"*. *FG1-03* perceived another challenge, that is *"Hard to reach evidence exhaustion"*, as we quoted: *"And when you talk about GL, you think that extracting all the evidence is really exhausting, right? Imagine the amount of information to be analyzed from all over the internet about, for example, Agile, where both the industry community and academia are talking about it"*.

Guideline 9 - Inclusion and exclusion criteria with quality assessment

Guideline 9 informed that the combination of inclusion and exclusion criteria for GL with quality assessment criteria should be used.

In the following, we present our findings regarding what researchers considered useful and the perceived issues related to Guideline 9.

Useful

For Guideline 9, only one useful point was considered. Three researchers (two from FG1 and one from FG2) considered it useful to *avoid rework*. Researcher *FG1-01* pointed out that:

"In the first moment, I think it is useful due to effort reduced". Researcher *FG1-02* wrote in a post-it card that *"Avoid rework"* and added during the discussion the following: *"I think it's interesting to think to anticipate the decision to include or not a source, avoiding rework"*. Researcher *FG2-03* had similar positioning, pointing out that: *"[. . .] I think it could be useful, as I wrote, to avoid wasting time since you can exclude some GL sources early on"*.

Issues

Considering the issues, all the investigated researchers perceived, at least, one **point of improvement** to Guideline 9. The list of these points is presented in Table 47.

The issue most mentioned was *To improve the text, changing the term 'should' in "QA criteria selection should be used" to 'could'*. It was perceived in both focus groups. For researcher *FG1-02*, the first impression was to perceive that it was useful to reduce the effort. However, he disagrees that this guideline should be followed, as we pointed out: *"At first, I thought it would be useful in terms of effort. Nevertheless, when it comes to something as mandatory using the expression "should be used", then I disagree. I think it depends a lot on the context being applied. It may not make sense for a particular search, right?"* In the discussion about this point, all other FG1 researchers agreed with *FG1-02*. Researcher *FG1-03* added the following comment: *"I also agree. One thing to say is that it is important, that it recommends. Another thing is to say that It should be done"*.

Another point of improvement discussed in both focus groups was *To avoid the use of QA criteria in the inclusion or exclusion criteria*, which is considered important to avoid mixing different approaches in different phases. In this regard, *FG1-03* mentioned: *"If the quality criteria were used to exclude a study, especially during the selection process, then it should be called an exclusion criterion"*. The same discussion occurred in the FG2, where *FG2-05* commented the following possibility of improvement to this guideline, as we pointed out: *"[. . .] One is to move the quality assessment question to be part only of the selection process*.

The last issue mentioned by two researchers of FG2 was *Use a decisive QA question in the first position of the QA process* as a trigger, aiming to remove the source depending on the answer. In this respect, *FG2-04* discussed: *"[. . .] this guideline could have some directions according to some points. In general, if you conduct a multivocal literature mapping, you do not need to perform a quality assessment. So, you could include the quality question in the process selection. But, if you will conduct a multivocal literature review, you just include it*

in the first position as a decisive question to know if you continue to assess its quality or the source will be excluded". Researcher *FG2-05* agreed with *FG2-04*.

Table 47 – Perceived issues identified in the focus groups about Guideline 9.

Issues	# FG1	# FG2
Points of improvement		
To improve the text, changing the term 'should' in "QA criteria selection should be used" to 'could'	5	4
To avoid the use of QA criteria in the inclusion or exclusion criteria	4	3
Use a decisive QA question in the first position of the QA process	0	2

Note: Columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

Guideline 10 – Coordinated integration of the source selection processes for grey literature and formal literature

Guideline 10 informed that one should ensure a coordinated integration of the source selection processes for GL and TL in the source selection process.

During the discussion in the focus groups, most researchers considered Guideline 10 confused, despite some useful points that were also perceived. In the following, we present our findings.

Useful

Only four researchers perceived useful points in Guideline 10. Two researchers of each focus group. The list of useful perceived for this guideline is presented in Table 49 and discussed in the following.

Table 49 – Perceived usefulness identified in the focus groups about Guideline 10.

Reasons	# FG1	# FG2
It gives the same importance to select GL and TL sources	2	1
To achieve promising findings	0	2

Note: The columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

Three researchers (two of FG1, *FG1-01* and *FG1-03*, and one of FG2, *FG2-02*) considered Guideline 10 important because it *gives the same importance to select GL and TL sources*, as *FG1-01* wrote in a post-it card: *"It's a useful item for trying to show whether it's just an SLR or an MLR, both are important"*. During the discussion, she mentioned that searching for GL tends to be higher than for a TL. So, this guideline reminds you to balance this effort. Researcher *FG1-03* had a similar opinion during the discussion, as we quoted that: *"When you go to select the fonts, you have to have the notion for the source selection, Grey Literature and Traditional Literature have the same importance"*.

Another reason, *To achieve promising findings*, mentioned for two FG2 researchers, is related to the information presented about assessing the source, involving two or more researchers in the discussion to solve any disagreement. In this regard, *FG2-03* pointed out: *"The idea of voting between the researchers is interesting. I think it is a different method of reaching an agreement"*. Researcher *FG2-01* considered it a good step.

Issues

For Guideline 10, all researchers perceived some **points of improvement** to this guideline. The list of these points is presented in Table 50 and discussed in the following.

Table 50 – Perceived issues identified in the focus groups about Guideline 10.

Issues	# FG1	# FG2
Points of improvement		
To explain better the Guideline 10	5	5
To explain that to select GL requires more effort than TL	0	2
To improve the understanding is better to join Guidelines 9, 10, and 11	0	1

Note: The columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

As shown in Table 50, all researchers considered Guideline 10 *confused*. For this reason, one point of improvement was *To improve the explanation of Guideline 10*. For researcher *FG1-03*, both the text and the abstract are very summarized, making it difficult to understand them. Researcher *FG1-04* agrees with *FG1-03*, as we quoted out: *"I found it confusing and vague [. . .] What does coordinate source selection mean?"*.

During the discussion on FG2, researcher *FG2-01*, using a post-it card, wrote that: *"A little bit obscure to understand. What do they mean by coordinated integration exactly? An example would help a lot!"*. Researcher *FG2-05* had the same opinion, saying that "coordinated integration" was not straightforward for him. For researcher *FG2-02*, this guideline needs additional information or sub guidelines to explain it better.

Other issues were mentioned, although only by researchers of FG2. For instance, for researcher *FG2-05*, it appears that the guideline considered that conducting an MLR would spend the same effort to select GL and TL as well. For him, that is not what happens because selecting GL sources takes more effort. In this regard, we quoted that: *"Again, for me, it's very counterintuitive that you should put the same amount of effort. [. . .] for the grey literature, on top of that, you also have to do some sort of quality assessment [for select sources]. Uhm, so you should spend more effort in the Grey in respect of the white"*. Researcher *FG2-04* agreed with researcher *FG2-05*.

The last point, mentioned only by researcher *FG2-05*, *To improve the understanding is better to join Guidelines 9, 10, and 11*, as we pointed out: *"I can't distinguish between guidelines 9, 10, and 11. All of them are kind of talking about the same concept. Uhmmm... and I don't know why we need different guidelines"*.

Guideline 11 – Quality Assessment of sources

Guideline 11 recommended applying and adapting a set of Quality Assessment (QA) to determine the extent to which a source is valid and free of bias.

Guideline 11 brings a table, called *Table 7*. In this chapter, when we mention *Table 7*, we refer to a table of Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019), with questions covering criteria about the producer's authority, methodology, objectivity, date, novelty, impact, and outlet control to assess GL sources.

This guideline was one of the most researchers perceived as useful of Garousi's guidelines. In the following, we present our findings regarding what researchers considered useful and the perceived issues related to Guideline 11.

Useful

All the researchers perceived usefulness when Guideline 11 *provided questions to QA*. For researcher *FG1-03*, the QA checklist provided is essential, especially for beginners researchers dealing with Secondary Studies searching for GL, which do not know which criteria could be used. Researchers *FG1-01*, *FG1-04*, and *FG1-05* agreed with this point mentioned by *FG1-03*. Similar perception had researcher *FG2-03* of FG2, as he wrote in a post-it card: *"Table 7 gives some good general ideas to start making a quality assessment"*. In addition, researcher *FG1-04* considered the QA checklist provided as the most important thing, like the realization of what needs to be done. For researchers *FG2-01*, *FG2-02*, *FG2-04*, and *FG2-05*, Guideline 11 is the most useful in the whole paper.

Moreover, some researchers pointed out that these *QA questions are useful because you can adapt and use them to other contexts*. Researcher *FG2-05* using a post-it card, pointed out that: *"Can be fit to many contexts"*. Researcher *FG2-05*, during the discussion, commented that: *"Basically they say modify it if you want, we're not saying it works, might"*.

Issues

Several issues for Guideline 11 were discussed in both groups. In general, all researchers perceived some issue. For Guideline 11, the issues were related to points of improvement and challenges, as shown in Table 51.

Table 51 – Perceived issues identified in the focus groups about Guideline 11.

Issues	# FG1	# FG2
Points of improvement		
To avoid the use of vague or subjective definitions	5	3
To add more criteria to the QA checklist	0	3
To improve the QA criteria to be fit for GL and not TL	2	0
To include criteria to classify the QA score	2	0
To improve the explanation of Guideline 11	1	0
To avoid the use of boolean answers on QA	0	1
Problems/Challenges		
Leaving the researcher to decide which QA criteria to use can introduce bias	1	1
Difficult to have QA criteria that fit all the GL types used	1	0
Time-consuming	0	1

Note: Columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

Despite most of the researchers pointed out that Guideline 11 is important to provide questions to QA, several researchers perceived several **points of improvement**. Most of these points are related to the criteria and questions presented in the quality assessment checklist provided in Garousi's guidelines. For instance, most participants mentioned *avoiding vague or subjective definitions*. This issue was related to Table 7 of Garousi's guidelines. In this regard, researcher *FG2-04* wrote in a post-it card that: *"Some questions are unclear or hard to answer"*. For researcher *FG2-03*, most of the questions presented in Table 7 will depend on the context, as we pointed out: *"I also find this point a bit vague, I believe it will once again depend on the context of the problem"*. In addition, researcher *FG1-01* wrote in a post-it card: *"[. . .] It will involve a lot of subjectivity to point out the criteria to be adopted for each source"*.

Although the proposed criterion "Impact" of Guideline 11 mentioned using impact metrics, as the number of comments posted for a GL, three researchers of FG2 mentioned the necessity to *To add more criteria to the QA checklist*. In this respect, *FG2-01* wrote in a post-it card that: *"It's important to read the comments of other users on the piece of GL (or source) under*

assessment". Researchers *FG2-03* and *FG2-05* agreed with *FG2-01*. In addition, researcher *FG2-03* made an important comment about including other criteria to assess GL, avoiding considering only the reputation of the GL producer, as we pointed out: *"Not just because they have this title, or they belong to this organization, their opinion is the right one. We have to look truly at their opinion and independently of who's saying it or where it comes from, we need to look at their evidence, I think. More than their reputation"*.

Other points of improvement were also mentioned. For instance, *To improve the QA criteria to fit GL and not TL*. According to researcher *FG1-03*, it is important to revise some questions of the QA checklist because some of them seem to be assessing traditional literature instead of GL, such as some formalities of TL, like methodology. Moreover, it was mentioned *To include criteria to classify the QA score*, cited by two researchers of FG1. Researcher *FG1-02* considered the QA checklist important, although he perceived several issues. One of those is the importance of knowing how to interpret the score of QA. In this regard, researcher *FG2-01* pointed out: *"And what is the meaning of being good or being bad, do you understand?"*.

Garousi's guidelines informed that, as GL has many types, the researcher needs to choose which criteria can already be applied for the selection process. Although several researchers have considered the proposed QA checklist criteria important, as we reported previously, *Leaving the researcher to decide which QA criteria to use* was perceived as **challenging**. In FG1, it was considered a challenge that may introduce bias, as *FG1-01* pointed out: *"[. . .] who guarantees that, for example, I am not biased when doing my study and I will choose criteria that do not lower the score of GL sources since it does not have a predefined set"*. Similar perception of FG2's researcher *FG2-05*.

Guideline 14 - Reporting review

Guideline 14 recommended that the writing style of the study should match the target audience. For instance, a style with straightforward suggestions and no details about the research methodology should be chosen if targeting practitioners. In addition, the guideline strongly recommended asking the practitioners for feedback.

In the following, we present our findings regarding what researchers considered useful and the perceived issues related to Guideline 14.

Useful

For Guideline 14, all the investigated researchers considered, in general, a helpful guideline. Although, as we reported further on, some researchers also considered it an obvious guideline. In Table 53, we listed the reasons identified.

Table 53 – Perceived usefulness identified in the focus groups about Guideline 14.

Reasons	# FG1	# FG2
To think about the target audience	5	0
Ask for feedback from practitioners	1	3
To support beginners researchers	0	4
Could be applied to other contexts outside MLR	0	1

Note: The columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

The reason most perceived was that this guideline is useful to *To think about the target audience*, although it was related only to the FG1. It happened because most of the researchers of FG2 considered it an obvious guideline, as we presented in the issues part. For FG1 researchers, all of them considered this guideline useful. In this regard, researcher *FG1-04* wrote in a post-it card that: *"I think it's important to think about the target audience"*. Researcher *FG1-03* agreed with *FG1-04* and added that he strongly agreed that one should think about the target audience. Although, he mentioned that it should be used in any type of scientific research.

Another point considered useful was to *ask for feedback from practitioners* about the MLR findings. Considering the participants of each focus group, only one of FG1 mentioned that feedback is important. Researcher *FG2-01* wrote in a post-it card: *"I like that they suggest to ask for feedback from practitioners"* and complement it during the discussion, mentioning that: *"[. . .] it's very important because especially in our field we tend to overlook a lot. Uh, about this and having more input from industry, it's fundamental in our field"*. Researcher *FG2-02* wrote that ,"In general, make sense. Validation with industry is important".

According to four researchers of FG2, Guideline 14 is useful *To support beginners researchers*. Although it could be considered an obvious guideline issue for experienced researchers, we included it as perceived usefulness. In this regard, researcher *FG2-03* pointed

out that: *"Some recommendations from section 6 could be good reminders for beginners"*, and *FG2-05* wrote that: *"Could be useful for unsupervised very-early career researchers"*.

Issues

Considering the issues identified for Guideline 14, most of them are related to consider it **not useful**. For instance, five researchers (four from FG2) considered it an *Obvious guideline* for researchers. For researcher *FG2-01* is a general guideline obvious to researchers. Researchers *FG1-02*, *FG2-03*, *FG2-04*, and *FG2-05* had the same opinion. One point to consider here, as we pointed out previously, researchers mentioned that it was obvious to them as experienced researchers. Nevertheless, it could be useful for beginners researchers.

FG2-05 identified another issue. For him, this guideline is not specifically related to an MLR study type because it could be applied to any type of research.

The most common **point of improvement** was *To explain the importance of transparency of methodological aspects to the target audience*. It was mentioned only by FG2. According to *FG1-05*, using a post-it card, she wrote that independently of the target audience, it is needed that the results be transparent. During the discussion of FG1, researcher *FG1-04* pointed out that: *"For this guideline, the need for transparency was lacking"*.

Another point of improvement mentioned the importance *To think about the venue before the target audience*, as we pointed out in the discussion of *FG2-03*: *"I would you put you know, the instead of targeting practitioners or researchers, maybe venue would be the first thing to write, there, you know"*. Researcher *FG2-01* agreed with *FG2-03* and complemented that: *"Yeah because those also target specific practitioners, a specific audience"*.

Table 54 – Perceived issues identified in the focus groups about Guideline 14.

Issues	# FG1	# FG2
Not useful		
Obvious guideline	1	4
It is a general recommendation	0	2
Points of improvement		
To explain the importance of transparency of methodological aspects to the target audience	4	0
To improve the explanation that the report needs to focus first on the venue of publication before the target audience	1	2

Note: Columns with "#" show the number of researchers that mentioned a given category for each focus group.

Source: the author.

Summary of RQ4: In general, SE researchers considered Garousi's guidelines useful, showing that they are essential to support researchers, mainly the beginners, to conduct the MLR studies. The most useful considered guidelines explain identifying and searching for GL and performing quality assessments of GL sources. Despite the usefulness, we also identified several issues, most of them related to points of improvement. We also identified challenges and not useful points related to the these guidelines. For this reason, we claimed the importance of profoundly investigating these issues, although we provided some recommendations to deal with most of them.

6.4 DISCUSSION

This section discusses the main findings related to the investigation of Garousi's guidelines that caught our attention. First, we present the perceptions of the main useful points of Garousi's guidelines (Section 6.4.1). Then, we discuss the main perceived issues and, whenever applicable, their proposed solutions or suggestions to improve/solve them (Section 6.4.2). Finally, we discuss some threats to validity (Section 6.4.3).

6.4.1 Garousi's guidelines' usefulness

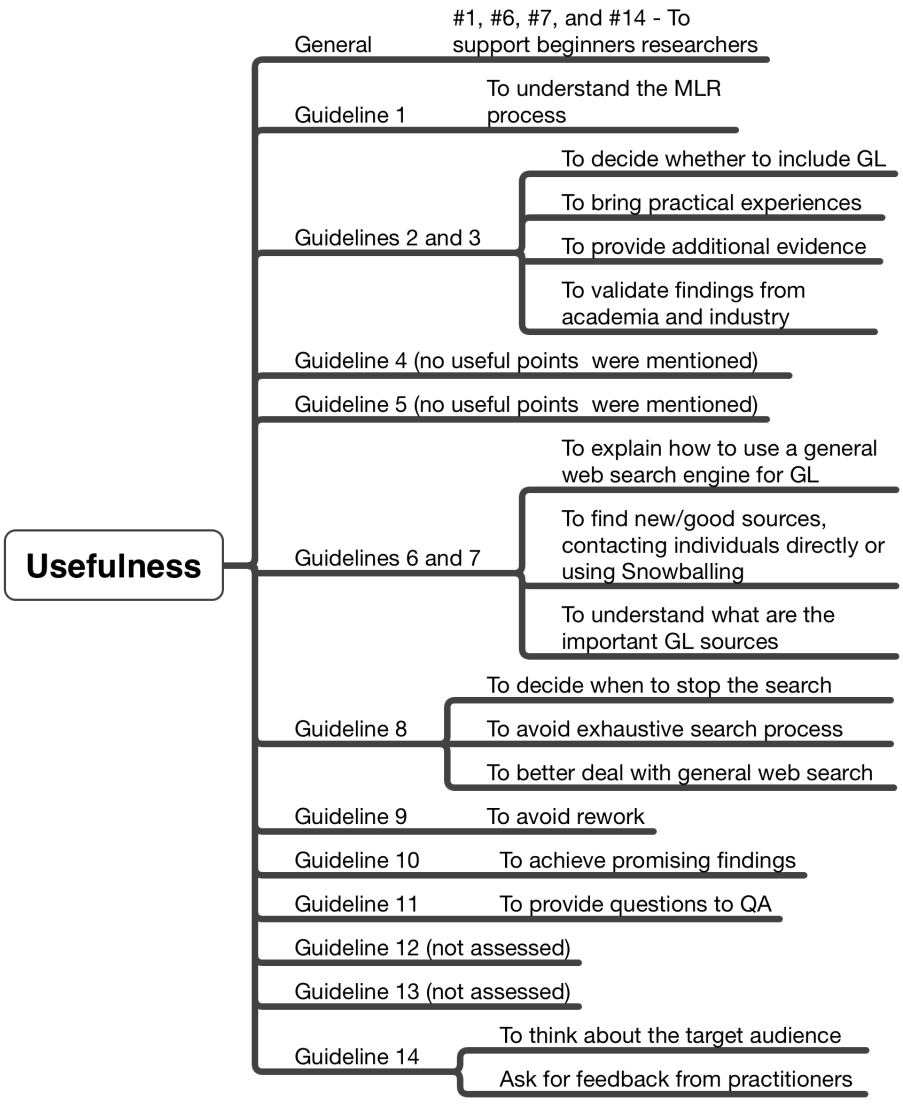
Based on the researcher's opinions, Garousi's guidelines are generally useful. Considering Table 37, Guidelines 11 and 14 were the most that individually researchers perceived anything useful. Although, during the discussions, each group pointed out what guideline they considered most useful. The FG1 participants considered Guideline 7, whereas, for FG2, Guideline 11 was chosen. This finding caught our attention because only two researchers of FG2 thought Guideline 7 was useful. Nevertheless, during the discussion about Guideline 7 in FG2, some researchers pointed out that *it is needed to understand the topic investigated and know the important sources early*. It happens in SLR or MS studies, where most of the researchers know which sources to use. In our perception, as all FG2 researchers had previously used Garousi's guidelines, this may have interfered with their opinions.

We summarized the usefulness perceptions of Garousi's guidelines in Figure 25. In this figure, we separated the points considering "why" it is useful for each guideline. We inform that all these categories emerged without any previous pre-formed category during our analysis.

Figure 25, in general, shows that the guidelines were **useful** according to the researcher's perceptions. For instance, as MLR is recent in SE research, Guideline 1 was important to *support the conduction of MLR studies, mainly to beginners researchers*. In addition, some researchers did not know how to justify whether to include GL. For this reason, researchers considered guidelines 2 and 3 very useful to help researchers *decide whether to include GL* and to support them to think about benefits on the outcome if GL was used (e.g., to bring practical experience, to validate findings from academia and industry). We agree with these claims, and they are also supported by our previous investigation (Study 1 (KAMEI et al., 2020)) that showed several researchers were skeptical and avoided using GL due to its lack of scientific value. Thus, these guidelines, especially Guideline 2, provide essential considerations to support the researcher in deciding whether GL or not.

As the process of traditional secondary studies differs from MLR, the guidelines provided information to **understand** better the process related *to searching for GL and when to stop the search*, for instance, providing examples of important GL sources and several strategies to search for GL. One of those useful strategies was to use a scope limit that avoids a time-consuming exhaustive search process, which tends to be part of the search for GL, as pointed out in our previous investigation of Study 3 (KAMEI et al., 2021b). In addition, the QA checklist provided was considered necessary to the MLR process, helping in the selection of GL sources

Figure 25 – General overview of the usefulness of Garousi's guidelines perceived in the Focus Groups.



Source: the author.

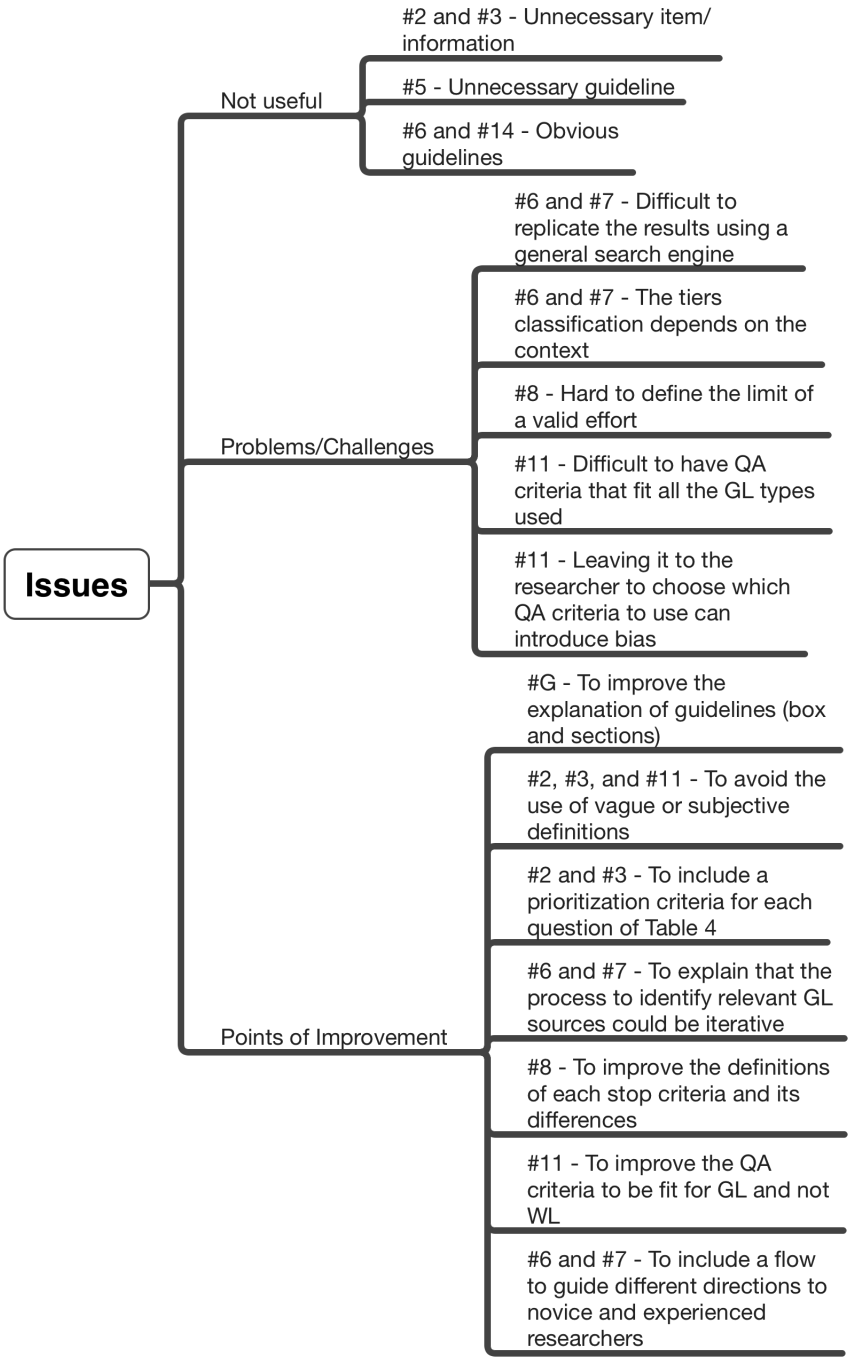
and increasing the value of the findings.

6.4.2 Garousi's guidelines' issues

Despite several useful points perceived about Garousi's guidelines, several **issues** were identified. Figure 26 summarizes the main issues perceived. In the following, we discussed some of these issues.

Figure 26 shows the three categories of the issues. The first is **not useful**, and according

Figure 26 – The main issues of Garousi's guidelines perceived in the Focus Groups.



Notes:
1) The first level indicates "Why" it is an issue. The second indicates "What", meaning the perceived issue properly.
2) "#" indicates the number of the guideline.

Source: the author.

to some researchers, *some parts could be removed* from Garousi's guidelines. For instance, questions #6 and #3 of *Table 4* in Guideline 3. Another one was to remove Guideline 5. For this last, we understand the researcher's questioning, as the type of RQ depends on the objective and the purpose of the investigation.

Guidelines 6 and 14 were considered not useful because researchers considered it *evident/obvious*. For Guideline 6, researchers pointed to the need to identify relevant GL sources early on, mainly for experienced researchers that know about this need. The same perceptions researchers had for Guideline 14. For this reason, one point of improvement was mentioned *to include a flow to guide different directions to novice and experienced researchers*. In our perspective, it is an important point that will benefit the researchers according to their experience, that may avoid reading unnecessary information as claimed by some researchers.

Several **problems/challenges** were discussed during the focus groups, as presented in Figure 26. For instance, researchers pointed out the *difficulty of replicating the results using a general search engine* as an example of search strategies for GL provided in Guideline 7. It is a valid concern mentioned in previous studies in other areas (e.g., (PIASECKI; WALIGORA; DRANSEIKA, 2018; CURKOVIC; KOšEC, 2018)). According to Piasecki et al. (PIASECKI; WALIGORA; DRANSEIKA, 2018), it occurred because, in the case of Google, it uses a search algorithm that is not known and cannot be controlled. After all, the search is adapted to each user, focusing on personalizing the result. For this reason, we believe that the SE research community will benefit from Piasecki et al.'s (PIASECKI; WALIGORA; DRANSEIKA, 2018) recommendation to consider general search engines, such as Google, only as additional search engines of information sources and should not as the only one source, to avoid the difficulty of replicability.

The *tiers of GL* could be a problem that SE researchers may face to perform QA of GL because, as pointed out, the classification into the tiers depends on the context. In our previous investigation (Survey 2 – Phase 1), we identified that in addition to the tiers (control and expertise), it is essential to consider: 1) the rigor (quality of presentation), 2) producer reputation (experience, credibility), and 3) the peer-interaction (what the others are talking about/in the GL source).

Two other problems/challenges caught our attention. Both are related to the *difficulty in performing GL quality assessment* mentioned when assessing Guideline 11. First, some researchers pointed out a difficulty in using a QA checklist considering the diversity of GL types. In our previous investigation of Study 3 (KAMEI et al., 2021b), we also identified this

challenge. Garousi et al. pointed out that it is important to select QA criteria, observing if they are adequate to the GL types. However, the option to the researcher select these criteria was perceived by the researchers as another challenge. To the best of our knowledge, no existing QA checklist fits all the GL types, until now. In this regard, the last investigation that we have known was conducted by De Angelis and Lonetti (ANGELIS; LONETTI, 2021), which proposed a methodology for managing the explicit uncertainties during the assessment of GL sources using fuzzy Likert scales. Nevertheless, Angelis and Lonetti mentioned that their proposal needs to be refined to reach the different GL types.

Considering that GL has various types, it is difficult to assess them. For this reason, Garousi and colleagues *gave the option to researcher select the appropriate QA criteria that fit the type of GL used*. However, for some researchers, it could introduce some bias. This challenge was also perceived as a threat, called "Subjective quality assessment" in Zhou et al.'s study (ZHOU et al., 2016). To mitigate these challenges, Zhou and colleagues mapped some strategies that we considered important to be adopted since the phase to decide which QA criteria will be used until performing the quality assessment with GL sources. These strategies are: *to use impartiality of the quality evaluation* (e.g., using a paired process) and *external evaluations* (e.g., inviting an external reviewer). In that case, we could invite an external researcher to validate if the chosen QA criteria cover the GL types that will be used in the studies. If the criteria could not fit all the GL types used in the research, it is essential to discuss the implication of the decision to include these GL types in the synthesis.

Considering the **points of improvement**, one that should be considered for all guidelines is *to improve the explanation of guidelines* — mainly related to the box that gives a summary of each guideline. For instance, some researchers considered that some summaries are difficult to understand without reading the entire section associated. Another point is related to *avoiding terms with no explicit meaning*, which some researchers called "vague definitions". For instance, the same examples of vague definitions were presented in Garousi's guidelines, i.e., "What is considered complex? What is considered as a reputable company?". For this reason, we agreed with the researchers that vague terms or definitions should be avoided and explained better, avoiding researcher bias of subjective interpretations that might affect the study, as discussed by Feldt and Magazinius (FELDT; MAGAZINIUS, 2010).

Another point of improvement is related to the search for GL sources, which is important to *explain that the identification of relevant GL sources could be iterative*, considering that the researcher could learn during the process. For this reason, Wen and colleagues (WEN et

al., 2020) used the "Incremental Screening Process". This method divides the search process into data source audiences and iteratively conducts the sample selection. Thus, the researcher could learn during each iteration and, for instance, discovering new sources.

6.4.3 Threats to Validity

We organized our threats according to the classification proposed by Wohlin and colleagues (WOHLIN et al., 2000): construct, internal, external, and conclusion validity.

Construct Validity. Even with our efforts to improve our focus group design, we identified two potential threats: 1) Despite Garousi's guidelines having 14 guidelines, only nine were investigated for both focus groups (FG1 and FG2). We used this strategy to reduce the duration of the meetings, aiming to avoid participants suffering from fatigue when discussions are more extended, as proposed in Nyumba and colleagues (NYUMBA et al., 2018). We inform that in the first moment, we removed only guidelines that are very similar to the proposed in previous guidelines used to conduct SLR (e.g., Kitchenham and Charters' guidelines (KITCHENHAM; CHARTERS, 2007)), and in the second moment, we removed guidelines in which the results of the discussion were consensual among FG1 researchers; and 2) We used a non-probability sample by convenience (BALTES; RALPH, 2020) to compose the FG1, that may introduce some bias.

Internal Validity. As occurred in any qualitative investigation, the personal interpretations may have introduced biases in the findings. We tried to mitigate this threat by including two researchers to revise the categories that emerged. We also paid careful attention to the validity of the interpretation performed based on the transcribed data, video record that collected what researchers wrote in the cards of the Miro tool, and notes taken by the researchers who conducted the focus groups.

Another internal threat is related to the guidelines omitted, either by our opinion that they are not relevant to be discussed or because they were removed based on the views of only one focus group.

External Validity. Before conducting this investigation, we already knew that it is difficult to generalize the results using focus group research, as pointed out by Kontion et al. (KONTIO; BRAGGE; LEHTOLA, 2008). The point here is to provide a detailed explanatory description to understand each researcher's perceptions regarding each guideline assessed. Despite it,

we attempted to mitigate external validity by inviting all the authors of the MLR and GLR identified that followed Garousi's guidelines at the beginning of this research. We invited all the 182 authors identified, with a rate of 26.9% answers. Although 15.4% were available to participate in our investigation, we could conduct the focus group involving researchers from different five countries, in addition to the other five participants using a non-probabilistic sample by convenience.

Another threat is related to the memories about how much each researcher could remember about Garousi's guidelines. For instance, we received an email rejecting our invitation, saying that two years is a long time to remember the guideline. In addition, we were aware that when the researcher was not the first or second author of the paper, they often did not recognize or have knowledge about the guideline, as answers received. For instance, some authors answered that their work was strictly to revise the paper and did not contribute to the research conduction. To mitigate this threat, in the invitations, we informed the group's purpose was to obtain the perceptions about Garousi's guidelines from a discussion among the participants. We also attached Garousi's paper in the email. Despite this threat, we have full confidence in our findings because we collected data from experienced researchers conducting empirical and secondary studies. All of them have been involved in applying SLR and MS methodology to a diverse set of topics in SE, and among them, five had experience using Garousi's guidelines.

Conclusion Validity. Even our efforts to investigate the authors of studies that followed Garousi's guidelines and other SE researchers, it is possible that our findings cannot corroborate with the perceptions of the not investigated researchers, out of the scope of this investigation. Nevertheless, we highlighted that, in general, the large majority of the interviewees reported similar opinions, even considering the different groups, which increases our confidence in the findings of this study.

6.5 SUMMARY

This chapter presented a focus group research with ten SE researchers to understand their viewpoints related to the usefulness and perceived issues about Garousi's guidelines. Our analysis consisted in analyzing the discussions between the researchers.

We identified that, in general, Garousi's guidelines appear relevant to SE researchers to

conduct MLR or GLR studies. However, some issues related to what researchers considered not useful, points of improvement, and problems/challenges were also perceived that need particular attention that could significantly improve its value as a research tool for SE researchers. We also provide a discussion and a set of recommendations to better deal with some of these issues.

By describing our findings showing the usefulness and the perceived issues of Garousi's guidelines with the potential ways to deal with them, we expect to contribute to improving the conduction of future MLR and GLR studies in SE research.

7 DISCUSSIONS

This chapter presents our discussions to answer our general research question (GRQ - "What is the Role of Grey Literature in Software Engineering Research?"). We informed that, even the processes presented in this section were elaborated from understandings obtained from a set of investigations (see Chapter 3, Chapter 4, Chapter 5, and Chapter 6), these proposals have not yet passed through community evaluation. Thus, it is a limitation that we will explore in the future works.

In Section 7.1, we provide our recommendations to decide whether to use GL in their research. These recommendations are based on the benefits, challenges, motivations to use, and reasons to avoid GL explored in our investigations. In Section 7.2, we provide a process to support researchers to conduct Secondary Studies using GL with recommendations to improve its use. Finally, in Section 7.3 we discussed some findings that caught our attention.

7.1 WHEN TO USE OR AVOID GREY LITERATURE

Based on the findings and experience conducting these investigations (Studies 1–3), we proposed a decision-making process to support SE researchers to understand when to use or avoid GL, as we presented in the following.

A Process to Decide When To Use or Avoid Grey Literature

We identified two studies providing discussions whether to include GL in a Secondary Study. The first one was Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019), that provides a questionnaire with seven items to be answered (possible answers – 'yes' and 'no') to help SE researchers to understand the need to include GL in a Secondary Study. Afterwards, Zhang and colleagues (ZHANG et al., 2021) complemented Garousi's guidelines and proposed one additional item. Table 56 shows the items of this questionnaire. Items 1–7 were proposed by Garousi's guidelines and item 8 in Zhang et al.'s study. We inform that, based on our investigations of Garousi's guidelines (Chapter 6), we improved and proposed a questionnaire to help on the decision-making process whether to use GL, as we depict in the following.

Besides these questions to help SE researchers to think about the necessity of GL in their

Table 56 – Proposed questionnaire of previous studies to help in the decision-making process on whether to include GL in SE reviews.

#	Question	Possible answers
1	Is the subject "complex" and not solvable by considering only the formal literature?	Yes/No
2	Is there a lack of volume or quality of evidence, or a lack of consensus of outcome measurement in the formal literature?	Yes/No
3	Is the contextual information important to the subject under study?	Yes/No
4	Is it the goal to validate or corroborate scientific outcomes with practical experiences?	Yes/No
5	Is it the goal to challenge assumptions or falsify results from practice using academic research, or vice versa?	Yes/No
6	Would a synthesis of insights and evidence from the industrial and academic community be useful to one or even both communities?	Yes/No
7	Is there a large volume of practitioner sources indicating high practitioner interest in a topic?	Yes/No
8	Is there a strong academia–industry interaction in the formal literature?	Yes/No

Notes:
Questions 1–7 were proposed in Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019).
Question 8 was proposed by Zhang et al.'s study (ZHANG et al., 2021).

Source: the author.

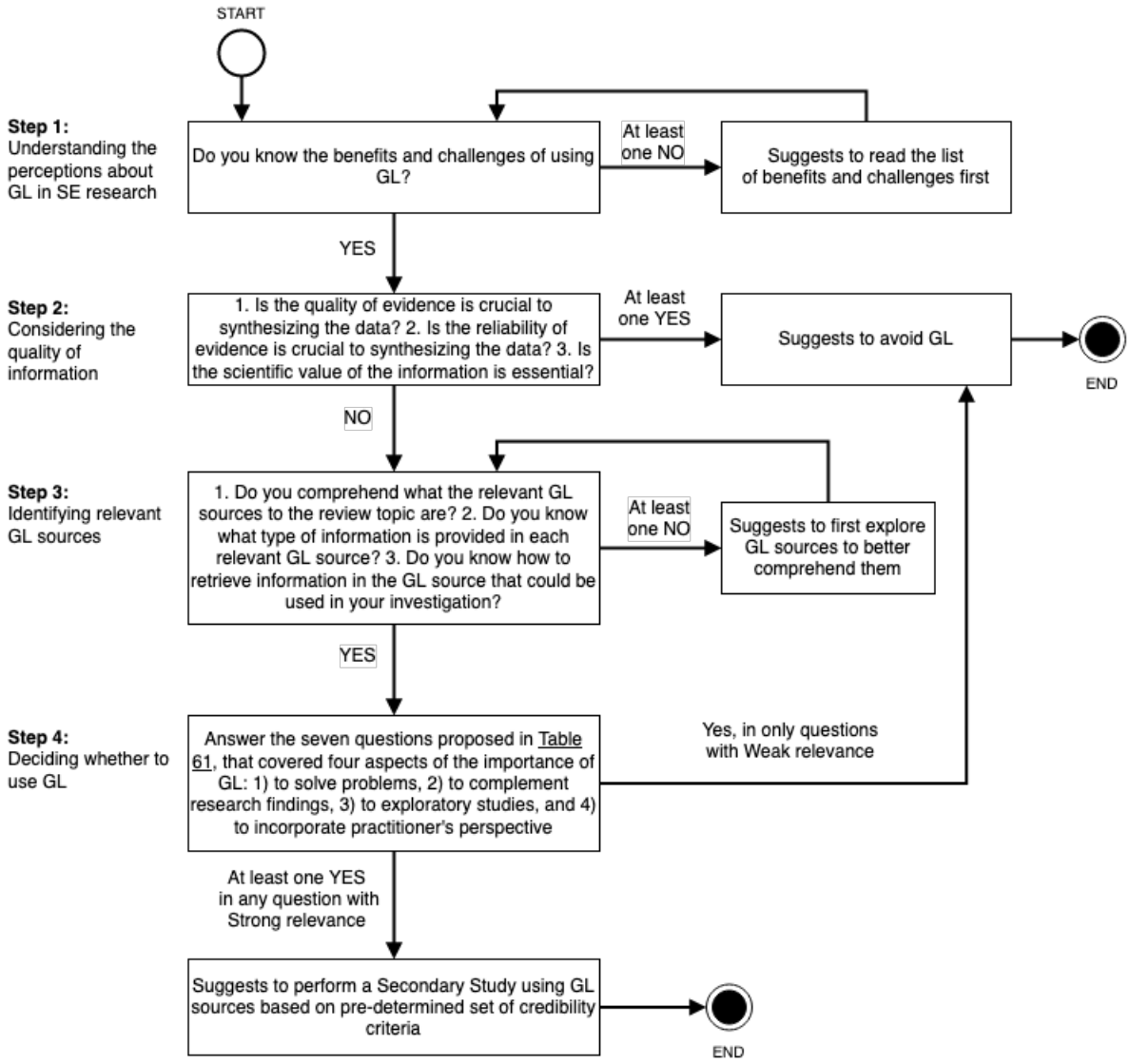
review, we considered important to think about epistemological concerns and understand about the *benefits* and *reasons* to GL use, and also considering the *challenges* and *when to avoid* GL use. For this reason, we provide a decision-making process considering these points to support SE researchers to assess whether to include GL in their Secondary Studies. Figure 27 depicted this process, that includes four steps. We inform that this process is supported by the investigations presented in this thesis. In what follows, we describe each of these steps.

Step 1: Understanding the perceptions about GL in SE research

The process starts with Step 1, which aims to make the researcher understand about the benefits and challenges of Grey Literature. In what follows, we presented some of those aiming to support the decision-making of SE researchers.

We explored the benefits and challenges in our investigations of Study 1 and Study 2 (see

Figure 27 – Process to decide whether to include Grey Literature in a Secondary Study.
Possible answers to the questions: 'YES' and 'NO'.



Source: the author.

Chapter 3 and Chapter 4). Although they are different investigations, the former was based on the survey with Brazilian researcher's opinions, and the latter was based on what researchers from diverse countries reported in Secondary Studies, we identified similar perceptions.

Table 58 summarize the most common **benefits** identified in our investigations. In this regarding, for instance, SE researchers considered GL was beneficial in *providing practical evidence* retrieved from the sources used by SE practitioners that help to understand the state of practice and answer practical and technical questions. Another perceived benefit was that GL *helps in knowledge acquisition* when the topic of the investigation is not well explored in traditional literature, and that GL *covers different results from scientific studies*, as some GL

brings results uncovered by scientific studies.

Table 58 – Main benefits of Grey Literature use identified in our Study 1, Study 2, and Study 3.

Benefits	Study 1	Study 2	Study 3
Provides practical evidence	x	x	
Helps in knowledge acquisition	x	x	
Covers different results from scientific studies	x	x	x
Up to date information	x		
Makes academic studies more interesting		x	
Easy to access and read	x	x	

Source: the author.

Despite the benefits, we also investigated the **challenges** of GL use (see Table 59), based on the findings presented in Chapter 3 and Chapter 4. The main challenge was related to the *lack of reliability* of GL sources, due to the epistemological problem of the GL information, as it is difficult to validate what is being reported in GL sources. Another challenge was the *lack of scientific value* of GL sources, making several SE researchers afraid that GL use weakens the research paper when it undergoes a peer review process. The last most common challenge was the *difficulty to search or find information* due to the non-structure information of GL sources and its diversity of types.

Table 59 – Main challenges of Grey Literature use identified in our investigations of Study 1 and Study 2.

Challenges	Study 1	Study 2
Lack of reliability	x	x
Lack of scientific value	x	x
Difficult to search/finding information	x	x

Source: the author.

Thus, before start using of GL in their investigations, researchers must comprehend these benefits and challenges, trying to better deal with each one. Then, as shown in Figure 27, if the answer to Step 1's question is 'NO', we recommended the researcher **first comprehends its benefits and challenges and then decide**. But, if the researcher answer 'YES', (s)he can proceed to the next step (Step 2).

Step 2: Considering the quality of information

Step 2 focuses on identifying if GL should be used based on the necessity of the investigation, considering the quality of the information. In our investigations of Study 1 and Study 2 (Chapter 3 and Chapter 4), we identified some **reasons to avoid GL use**, as shown in Table 60. One reason mentioned was the *lack of quality* of the GL sources, which may affect the validity of the results. Another reason was the *lack of reliability* of GL sources, related to the lack of rigor in which GL sources are written and published.

Thus, based on those reasons to avoid GL use, we proposed a set of questions to guide the SE researchers to decide whether to include GL. As shown in Figure 27, if the answer to Step 2's questions received at least one 'YES', we suggest to the researcher **avoid GL use or conduct a review selecting only GL sources that tend to have more credibility** (1st tier of the "shades of GL"), for instance, books/chapter books, thesis, magazine articles, and technical reports, as we reported in Chapter 3. Nevertheless, we do not recommend considering only the GL type because, it is difficult to analyze the credibility using only this information. For this reason, the researcher should have to assess the GL source using a set of **credibility criteria** (Chapter 3).

If the researcher answer 'NO' to all the Step 2's questions, (s)he can proceed to the next step (Step 3).

Table 60 – Main reasons to avoid Grey Literature use identified in our investigations of Study 1 and Study 2.

Reasons to avoid	Study 1	Study 2
Lack of quality	x	x
Lack of reliability	x	x
Lack of scientific value		x

Source: the author.

Step 3: Identifying relevant GL sources

Step 3 aims to make SE researchers reflect what are the relevant GL sources to be explored that are related to the topic under investigation. We identified in Study 2 (Chapter 4) that, in general, Secondary Studies did not focus on specific data sources to search for GL. Instead, the studies searched for GL using general search engines (e.g., Google). We do not recom-

mend doing this because it increases the amount of irrelevant data retrieved. For this reason, we highlight the importance to focus on specific and important sources to the topic. Thus, before start the search, due to the diversity of GL sources, it is essential that the researcher previously explore or conduct a pre-search on them, aiming to **verify the importance of each GL source to the research** and **familiarize with each environment** since they have different structure and provide different types of information. Then, as show in Figure 27, if the researcher answer 'NO' to at least one question of Step 3, we suggest to the researcher **first explore GL sources to better comprehend them**. But, if the researcher answer 'YES' to all the Step 3's questions, (s)he can proceed to the next step (Step 4).

Step 4: Deciding whether to use GL

Step 4 intends to guide SE researchers to understand the reasons to include GL in a Secondary Study, using a proposed questionnaire with seven items to be answered. This questionnaire is the result of multiple investigations: 1) On a set of motivations to use GL identified in our Study 1 and Study 2 (Chapter 3 and Chapter 4); 2) On the proposed questionnaire in Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019) and with the complement of Zhang and colleagues (ZHANG et al., 2021); and 3) On our assessment of Garousi's guidelines that derived a set of improvements to the questionnaire. Our proposed questionnaire covered four motivations for GL use: 1) To help understand or solve some problem; 2) To complement the research with findings not explored in Traditional Literature; 3) To exploratory studies focusing on answering practical and technical questions; and 4) To SE researches that intend to incorporate practitioner's perspective, as SE practitioners in the last years have produced much information using social media and communication channels.

To answer the proposed questions presented in Table 61, we highlighted the importance of performing this analysis in pairs and invoking a third opinion in case of any disagreements because some answers to the questions could be subjective. As shown in Figure 27, after the researcher answer all the proposed seven questions, if only the questions with weak relevance received 'YES', it **suggests avoiding the use of GL**. Nonetheless, if one or more 'YES' responses were obtained to questions with Strong relevance, we suggest **performing a Secondary Study using GL**.

Table 61 – Proposed questionnaire to help in the decision-making process whether to include GL in SE reviews.

#	Question	Relevance
1	Is the subject not solvable by considering only the traditional literature?	Strong
2	Is there a lack of volume or a lack of consensus on outcome measurement in the formal literature?	Weak
3	Is GL source important to understand/solve the problem of topic or to complement the research findings?	Strong
4	Is it the goal to validate or challenge assumptions or corroborate or falsify scientific outcomes with practical experiences, or vice versa?	Strong
5	Are there practitioner sources indicating high practitioner interest in the topic?	Weak
6	Is there a strong academia–industry interaction in the formal literature?	Weak
7	Is it the goal to answer an exploratory question with practical and technical questions?	Strong

Notes:
1) One or more "yes" responses in strong relevance questions suggest inclusion of GL.
2) One "yes" in only one weak relevance question did not suggest inclusion of GL.

Source: the author.

7.2 SUPPORTING THE USE OF GREY LITERATURE IN SECONDARY STUDIES

In this section, we provide a typical process showing some recommendations to better deal with the challenges identified in our investigations to improve GL use in Secondary Studies. Figure 28 presents a typical process of a Secondary Study with these recommendations, with the addition of one phase, Prior review. This process was divided into four phases: Prior review, Planning review, Conducting review, and Reporting review. In the following, we discussed each phase.

Phase 1 - Prior review

Phase 1 starts by asking the researcher to ***decide whether to use GL in a Secondary Study***. We designed a decision-making process with four steps explained in Section 7.1. We considered it essential that the process cover aspects of the benefits, challenges, and the reasons to use or avoid GL use. After going through the process, if the decision is to include GL,

it is needed **to decide which type of Secondary Study will be conducted** (Multivocal Literature Review or Grey Literature Review). To support this decision, we provided a question (*"Would a synthesis of insights and evidence from the industrial and academic community be useful to the review?"*). If the answer to this question was 'YES', it shows the necessity to aggregate information from academia and industry. Thus, we suggest performing a Multivocal Literature Review, as this type of Secondary Study intends to search for information in both Traditional Literature and Grey Literature, where this last, it is common to have information provided from the SE practice. Nonetheless, if the question received 'NO', we suggest performing a Grey Literature Review. Subsequently, we recommended that the researcher ***read Garousi's Guidelines*** (GAROUSI; FELDERER; MÄNTYLÄ, 2019) and *our assessment of these guidelines* (Chapter 6) that contains what SE researchers considered useful and what are the points of improvements that need the researcher attention, focusing on guiding the conduction of the research better.

Phase 2 - Planning the review

Phase 2 relates to the planning of the review, which usually starts by ***defining the aims*** and ***specifying the RQs*** (see Garousi's guidelines (GAROUSI; FELDERER; MÄNTYLÄ, 2019)). Then, it is needed to ***define the search string***. To compose it, we recommended using *a relaxed search string* that considers the particularity of how SE researchers and practitioners talk about the same topic, as pointed out in previous studies (WEN et al., 2020; MELEGATI; GUERRA; WANG, 2021). For instance, it is common for practitioners to refer to developers as "devs". Thus, if the search uses the term developers, consider using "devs" to search in GL sources. There are other slang words used in SE practice[1].

The next step is to ***define the GL sources to be explored***. Nevertheless, first, as we perceived several misunderstandings about GL and its types (Chapter 4), we considered it important for the researcher to have *knowledge about the GL definitions and its types for SE*. Second, as GL sources vary in their characteristics and types of information that they can contain, we recommended that the researcher *explore the GL sources before*, focusing on understanding how they work and meet the best GL sources for the topic in an investigation. Third, based on the observations of Wen and colleagues (WEN et al., 2020), they recommended

[1] https://web.archive.org/web/20220524170913/https://medium.com/transparent-data-eng/programmer-jargon-a-few-programming-slang-words-that-you-should-know-when-working-with-a-5644c256896b

observing if the research will consider both in-text content and audiovisual content. In Chapter 4, we identified studies that avoided videos, audio, or images without text. We understand the difficult to analyze these type of content. Although, in our investigations, videos showed as an important resource, for instance, to provide practical information to explain some feature in tools or solve some problem (Chapter 5). Finally, we recommend *focusing the search process on specific GL data sources* considered important to the research, avoiding using general search engines, such as Google, as we discussed in Chapter 4.

Phase 3 - Conducting the review

Phase 3 consists in conducting the review, starting by the ***search process***. In addition to search for evidence in the GL sources defined in Phase 2, based in our investigations presented in Chapter 6, the search strategies (contacting individuals directly and snowballing) proposed in Garousi's guidelines are useful. In addition, as the researchers could learn more about the important sources of GL on the topic, we considered important the recommendation proposed by Wen and colleagues (WEN et al., 2020) to *use an incremental search screening process*.

In Secondary Studies, that includes GL sources require more effort and time-spent to search for the sources, as we identified in Chapter 4. For this reason, we considered important to *define when to stop the search process* to avoid exhaustive process and better deal with general web search, as we reported in Chapter 6.

After the search process, we need to ***select sources*** based on the proposed *selection criteria*. We consider that the selection could be conducted *adopting a set of criteria to assess GL credibility*, that were explored in our Study 1 and showed in the following:

- Renowned authors
- Renowned institutions
- Cited by a renowned source
- Cites academic source
- Present empirical data
- Renowned companies
- Cites practitioner source

- Rigor in presenting information
- Describe the methods of collection

The use of those criteria could potentially increase the reliability and credibility of the findings present in the review. Even though these criteria were identified in our investigations (Chapter 3), we recommended that they should be used with attention. For instance, we did not recommend to use only one GL credibility criteria. For instance, adopting only the criterion if the GL producer is renowned, we consider it a risk that could introduce a bias. For this reason, we emphasize the importance of employing it together with analyzing the consistency and the rigor present in the information.

In addition, the review could ***assess the quality of GL sources***. To perform a quality assessment, Garousi's guidelines proposed a set of QA criteria, that in general, in our investigations of Chapter 6, SE researchers considered them important. However, several points of improvements are needed in those QA criteria, as we pointed in Chapter 6. In addition, it is important that researchers verify if QA criteria used fit for the GL type included, as we discussed in Chapter 4 and Chapter 6.

During the ***data extraction***, we recommended that the researchers consider the particularities of GL and TL sources, as the fields to be extracted could differ from each type of source, as pointed in HIQA document (HIQA, 2018). For this reason, we recommend to *separate the data extracted from TL and GL sources* (Chapter 5), considering the particularity of each type and that the data can be tabulated in a way to answer the research questions (KITCHENHAM; BUDGEN; BRERETON, 2015). For instance, the data extracted from a GL need to *provides detailed information about the source*, as we reported in Chapter 4 and Chapter 5, considering to extract, for instance, the URL of the source, the year of publication, and the names of the authors. It could improve understanding of the GL sources and their evidence. This advice is for any type of GL source because, in Chapter 4, we identified that for 25% of the GL sources, the URL was not informed and almost half of these sources were retrieved from academia.

Other recommendations are related to classify the GL sources according to its types (Chapter 4 and Chapter 5) focusing to permit a better understanding about the GL sources included, and to think about GL availability (Chapter 4), focusing on permitting the replicability or future analysis by other researchers about the findings of the review. The lack of availability of GL sources was reported in Chapter 4, showing that several GL sources used in previous Secondary Studies are unavailable. When we analyzed the total amount of a given GL source

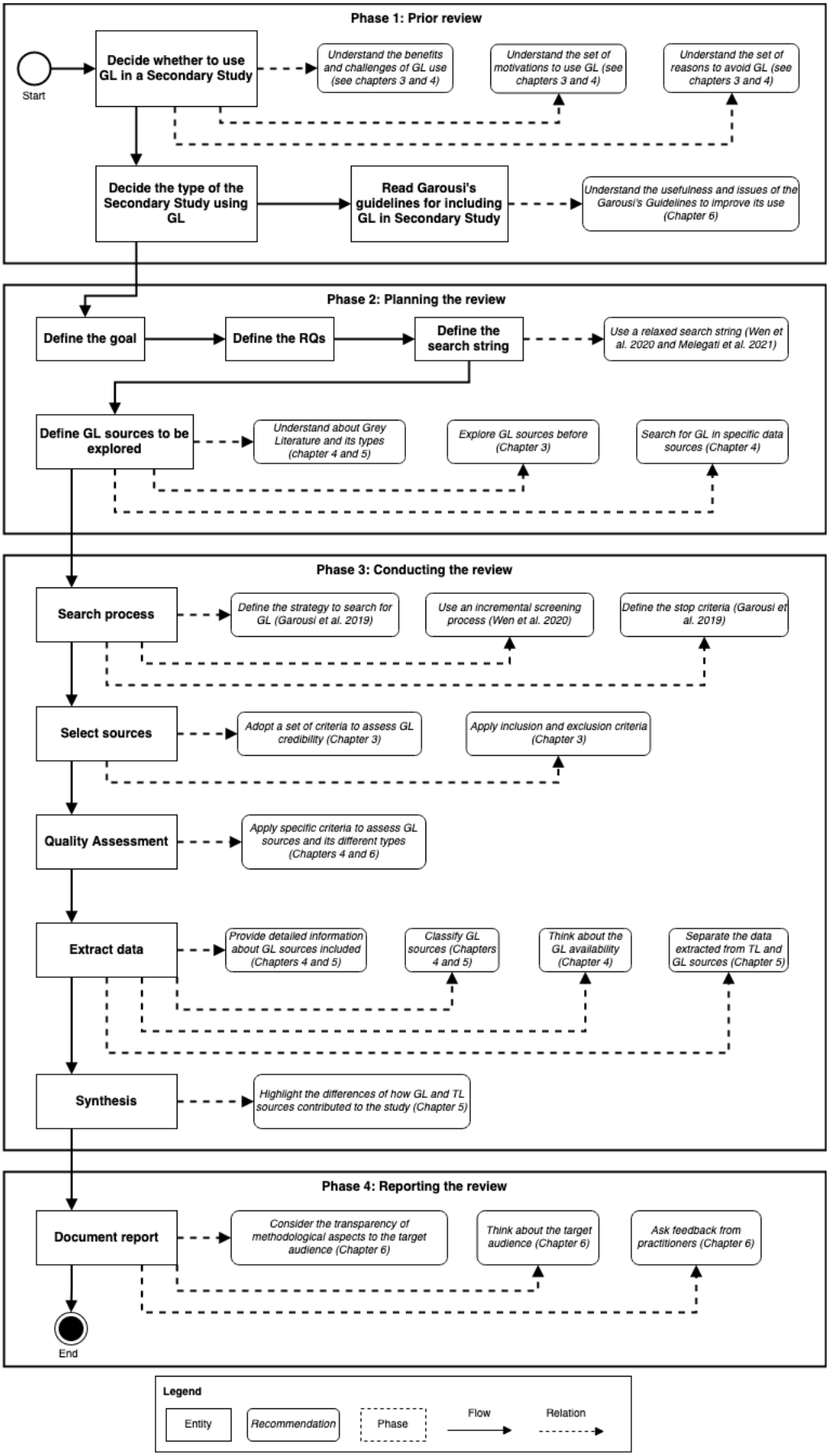
to the number of sources that are no longer available for each type of GL, we realized that web documents, tutorials, and web pages are the most unavailable. We highlighted that most of these unavailable sources were produced by companies, that in our analysis, most of the companies changed the structure of their pages, removing the old contents. To deal with this challenge, we recommended that SE researcher *store all the GL sources included in an external database or use a web archiving initiative*, as we proposed in Chapter 4.

Finally, the ***synthesis*** need to differ when dealing with MLR or GLR. Specifically talking about the MLR study, due to the epistemological concerns between GL and TL, that innately presented heterogeneous data, we recommended performing an interpretive synthesis based on a *textual narrative synthesis* (CRUZES; DYBÅ, 2011b), that can be applied to both qualitative and quantitative reviews. Textual narrative synthesis makes clearer the context and characteristics of each source, focusing on arranges them into more homogenous groups. To clear the context of each evidence or information retrieved from the GL sources, we recommended to researcher explicitly highlight the differences between the findings retrieved from GL and TL (as we proposed in Chapter 5). This procedure clears the understanding of how the results were obtained and how each source contributed to them (OGAWA; MALEN, 1991).

Phase 4 - Reporting the review

Phase 4 is the last proposed, which consists of ***reports the review***. Our recommendations to this phase are based on the assessment of Gararousi's guidelines reported in the Study 4 (Chapter 6). The first two recommendations are to think about the target audience of the review. The former is to consider the transparency of presenting the *methodological aspects*, and the last is how we *present the report* to the target audience. We could differ in presenting the findings to researchers and practitioners in both cases. Although the practitioners tend to be less rigorous with methodological aspects, we emphasize that the report could use an external link to present this information. Thus, enter the external link if the practitioner wants to read. In our opinion, it is the same as the following recommendation, that we need to think about the audience. The last recommendation is to *ask feedback from practitioners* about the review. We understand that it not always could be possible, as several reviews are conducted without any interaction with SE practitioners. However, thinking about it could increase and approximate the research from the practice.

Figure 28 – A typical process of a Secondary Study with recommendations to better deal with GL use in each phase.



Source: the author.

7.3 OTHER DISCUSSIONS

In this section, we discussed some of our findings that called our attention when comparing the findings of Study 1 (Chapter 3), based on the Brazilian researcher's perceptions, studies 2 and 3 (Chapter 5 and Chapter 4), based on what researchers of different countries presented in their investigations, and Study 4 (Chapter 6), based on the focus groups with researchers from different countries.

7.3.1 Motivations to Use and Reasons to Avoid Grey Literature

Based in our four investigations (Studies 1–4), our findings showed consistent to discuss the motivations to use and reasons to avoid GL, despite the nature of each investigation.

Considering the motivations to use GL, all of our investigations identified similar findings. For instance, showing that the main motivation to include GL in research was to *incorporate practitioners' perspectives* and *to identify more studies (in addition to traditional literature) to complement research findings*. Even in Study 4, when investigating guidelines 2 and 3 of Garousi's guidelines that explored whether to include GL in a review, most of the researchers of both focus groups considered it essential to include GL to bring practical experiences.

We identified the same consistency in investigating the reasons to avoid GL, showing that, in general, researchers tend to avoid GL due to the lack of reliability and the quality of the source, which hinders the scientific value of the information retrieved from those sources.

7.3.2 Types of Grey Literature Used

We explored the GL types used in three investigations (Studies 1–3). The Study 1 was the only one we conducted with researchers, more specifically with Brazilian SE researchers, showing that the most GL types used were those with low Credibility (e.g., community websites, blogs, survey reports, websites). This result differs from the Study 2 but is similar to the Study 3. In our opinion, it occurred due the difference between Study 1 and Study 2 existed because, in the former, prominent researchers are using GL for exploratory investigations (e.g., to understand problems, to answer practical and technical questions) but not as a source of evidence for their researches, as pointed the challenges and reasons to avoid identified. In the latter, as most of the Secondary Studies investigated were based on SLR or MS, our

findings showed that they tend to search for evidence in TL or GL sources with more credibility (e.g., books/chapter books, technical reports, theses). On the other hand, in Study 3, as were explored MLR studies that have the intention to search for GL sources, the findings related to the GL sources used are quite similar to the findings of Study 1. In addition, these differences between the Study 1 and 2 occurred because several researchers of Study 1 did not consider thesis and books/chapter books.

8 RELATED WORKS

This chapter presents the related works according to their types of investigations. Section 8.1 presents the studies that investigated the perceptions of Grey Literature (GL). Section 8.2 groups the studies that explored GL's credibility and quality assessment in Software Engineering (SE) research. Section 8.3 shows the tertiary studies that also investigated Secondary Studies that used GL. Section 8.4 discusses the studies that reported experiences conducting Grey Literature Review (GLR) or Multivocal Literature Review (MLR) studies. Finally, Section 8.5 presents a summary of our differences with the related works.

8.1 PERCEPTIONS ABOUT GREY LITERATURE

Perceptions about GL were investigated in primary studies, but were commonly investigated in Secondary Studies. The perceptions of the related works were related to the *benefits*, *challenges*, and *motivations* of GL use. **We informed that as the investigations about the *reasons to avoid* GL were explored only in our work, we did not discuss this topic in this section.**

In the following, we briefly discussed these studies, presenting the differences between our work.

Williams and Rainer's studies

Williams and Rainer conducted three studies exploring the benefits and challenges of using blog articles as a source of GL evidence for SE research. The first study (WILLIAMS; RAINER, 2017) reviewed the literature, in addition to a pilot mapping study that identified a set of criteria to assess GL credibility (discussed in Section 8.2).

In the second study (RAINER; WILLIAMS, 2018b), focusing on examining the opportunities for using blogs as a source of evidence in SE research, they informally reviewed practitioners' use of blogs and the research literature to identify the benefits and challenges of blogs use.

In the third research, Williams and Rainer (WILLIAMS; RAINER, 2019) performed a literature review (not systematic), aiming to identify benefits, challenges, and directions of using blogs as evidence for SE. The information provided in a blog was compared with the data collected

using research methods (e.g., interviews and surveys), showing the importance of a blog as an instrument in which the practitioner exposes their thoughts without any interference or an introduced bias by a researcher.

Differences with our investigations: Even though the similarity of Williams and Rainer's works with this thesis, there are differences in at least five points: (i) We did not focus on a specific type of GL source; (ii) We explored the experience of SE researchers to understand which type of GL They have used; (iii) We tried to understand what motivates and demotivates SE researchers to use GL; (iv) We found different criteria to assess GL credibility; and (v) We explored a broader population of SE researchers, not only experimental SE researchers.

Garousi et al.'s study

Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2016) expanded the investigation of GL as a source of evidence for MLR studies in SE research, conducting two investigations. The first one presented a meta-analysis from three cases in which GL was used to understand what knowledge is missed when an SLR does not consider GL. The second one investigated three MLRs to understand what the community gains when conducting multivocal literature. For the authors of this study, GL sources were important to cover technical research questions and be beneficial to practitioners, once the evidence is retrieved from the industry.

Differences with our investigations: Comparing our findings with Garousi et al.'s study, they performed an exciting investigation to show the differences between including and not GL in Secondary Studies. However, they focused on investigating the content provided in Secondary Studies, and they did not assess the perceptions of SE researchers as we performed.

Galindo Neto et al.'s study

Galindo Neto and colleagues (GALINDO NETO et al., 2019) conducted a tertiary study which focused only on the types of MLR and GLR, aiming to provide a preliminary investigation about research involving MLR and GLR studies, focused to understand their motivations to included GL (e.g., lack of academic research on the topic and evidence in GL), the GL types used (e.g.,

videos, books, blog post, and technical report), and the search engines used (e.g., Google, Google Scholar, and websites). They searched for the studies published between 2009 to April 2019, using six academic search engines. Fifty-six studies were returned, 12 of them were selected.

Differences with our investigations: Galindo Neto et al.'s study focused only on the MLR studies to explore the perceptions of GL use. Differently, in our investigations, we explored any type of Secondary Study (while we investigated 126 studies, Galindo Neto et al.'s explored 12). In addition, we explored other context of perceptions, for instance, the reasons of SE researchers tend to avoid GL.

Zhang et al.'s study

Zhang and colleagues (ZHANG et al., 2020), using a mixed-methods approach, investigated GL, by conducting a tertiary study and a survey with SE researchers that used GL in the Secondary Studies. They aimed to obtain an overview of the research community's understanding of: (i) The possible definitions of GL in SE; (ii) Reasons for including GL from the perspectives of both literature users and community experts; (iii) Proposing a conceptual model for how SE researchers work with GL in the research life-cycle; and (iv) Identifying the significant challenges of GL use in SE.

Differences with our investigations: Considering Zhang et al.'s study (ZHANG et al., 2020), our works differs at least in three dimensions: (i) We conducted a broader search to retrieve the highest number of studies published on premier SE conferences and journals; (ii) We explored waters not chartered by previous studies (e.g., we investigated the inclusion/exclusion criteria, the perspective of the GL use according to the types of Secondary Studies, and the availability of the GL data, investigating the extent to which the research questions are answered using GL); and (iii) We found some different findings that need further investigation. For instance, Zhang et al. found that only 25% of researchers used GL to evaluate their conclusions and our results found that more than 50% of the Secondary Studies that used GL used it to support their findings. Moreover, the definitions of GL found in our work were slightly different.

8.1.1 Summarizing the differences between the perceptions of our investigations with related works

This section summarizes the similarities and differences between our investigations and related works concerning the perceptions about GL: benefits, challenges, and motivations to use.

Table 63 summarized the ***benefits*** perceived. As shown, in total, ten benefits were identified in our investigations and related works. The main of them was that GL *covers different results from scientific studies*. In addition, four out of ten benefits identified were exclusively found in our investigations (*Advance the state of the art/practice, Easy to access and read, Knowledge acquisition*, and *Makes academic studies more interesting*), and the others were found both in previous literature as in our studies.

Our works and previous literature identified eleven ***challenges*** of GL use, as showed in Table 68. The most identified challenges among the studies were the *Difficulty in quality assessment* and the *Diversity of content* of GL sources. Three challenges were exclusively found in our investigations (*Lack of scientific value, Time-consuming*, and *Uncertain availability of GL*).

Eleven ***motivations to use*** GL were identified. Four of them were exclusively identified in our works (*To answer practical and technical questions, To explore trends, To conduct government studies,* and *To prepare classes*). In our opinion, we considered the two last are very specific to the community explored. We highlighted that the main motivation to use GL was *To incorporate practitioners' point of view*, mentioned in all studies. The other main motivations to use were to *Lack of academic research on the topic* and *To complement research findings*.

Table 63 – Related works: Comparing the findings related to the Benefits of GL use.

Benefit perceived	(1)	(2)	(3)	(3)	(5)	(6)	This thesis
Advance the state of the art/practice							x
Covers different results from scientific studies	x	x	x	x	x	x	x
Easy to access and read							x
Knowledge acquisition							x
Makes academic studies more interesting							x
Multi–method triangulation		x	x			x	x
Provide practical evidence		x	x	x	x	x	x
To reduce publication bias		x	x			x	x
Trend analysis	x	x					x
Up to date information	x	x		x	x		x

(1) (WILLIAMS; RAINER, 2017), (2) (RAINER; WILLIAMS, 2018b),
(3) (WILLIAMS; RAINER, 2019), (4) (GAROUSI; FELDERER; MÄNTYLÄ, 2016),
(5) (GALINDO NETO et al., 2019), (6) (ZHANG et al., 2020).
Source: the author.

Table 64 – Related works: Comparing the findings related to the Challenges of GL use.

Challenge perceived	(1)	(2)	(3)	(4)	(5)	(6)	This thesis
Different understanding on the definition of GL						x	x
Difficult to search/find information		x	x				x
Difficulty in quality assessment	x	x	x		x	x	x
Diversity of content	x	x	x		x	x	x
Lack of formality of language they use			x				x
Lack of Reliability / Credibility		x	x	x		x	x
Lack of scientific value							x
Lack of standard for annotations/structure	x	x	x				x
Lack of transparency and reproducibility of searches			x				x
Time-consuming							x
Uncertain availability of GL							x

(1) (WILLIAMS; RAINER, 2017), (2) (RAINER; WILLIAMS, 2018b),
(3) (WILLIAMS; RAINER, 2019), (4) (GAROUSI; FELDERER; MÄNTYLÄ, 2016),
(5) (GALINDO NETO et al., 2019), (6) (ZHANG et al., 2020).
Source: the author.

Table 65 – Related works: Comparing the findings related to the Motivations to Use GL.

Motivations perceived	(1)	(2)	(3)	This thesis
Emerging research on this topic		x		x
Lack of academic research on the topic	x	x	x	
Provide different perspectives between researchers and practitioners	x		x	
To answer practical and technical questions				x
To complement research findings	x		x	x
To conduct government studies				x
To explore trends	x			x
To incorporate practitioners' point of view	x	x	x	x
To prepare classes				x
To reduce publication bias	x		x	x
To understand the problems				x

(1) (WILLIAMS; RAINER, 2019), (2) (GALINDO NETO et al., 2019),
(3) (ZHANG et al., 2020).

Source: the author.

8.2 GREY LITERATURE CREDIBILITY

We identified two studies that also investigated GL credibility criteria. They were conducted by Williams and Rainer (WILLIAMS; RAINER, 2017; WILLIAMS; RAINER, 2019) that conducted two studies to investigate how to improve the quality and credibility assessment of blog articles in SE research.

In what follows, we discussed these studies.

Williams and Rainer's studies

The first study (WILLIAMS; RAINER, 2017) examined some criteria to evaluate blog articles as a source of SE research evidence through two pilot studies (a systematic mapping study and preliminary analyses of blog posts). The findings showed some criteria for selecting a blog article's content (e.g., authentic, informative).

The second study (WILLIAMS; RAINER, 2019) focused on finding credibility criteria to assess blog posts by selecting 88 candidate credibility criteria from a previous Mapping Study (WILLIAMS; RAINER, 2017). Then, to gather opinions on a blog post to evaluate those credibility crite-

ria, they surveyed 43 SE researchers. Some criteria were found, such as reasoning, reporting empirical data, and writing data collection methods.

Differences with our investigations: Even though the similarity of these works with our works, there are differences in at least four points: (i) We found some different credibility criteria: the source needed to be provided by renowned institutions, renowned companies, cited by others, and derived from academia; (ii) We did not focus on a specific type of GL source; (iii) We explored the experience of SE researchers to understand the perspectives on the credibility of different GL types and how SE researchers assess them; and (iv) We investigated a set of prioritization criteria used to assess GL credibility.

8.2.1 Summarizing the differences between the GL credibility criteria identified in our investigations with related works

This section summarizes the similarities and differences of the findings between our investigations and related works concerning the Credibility criteria.

Table 66 summarized the Credibility criteria identified in Williams and Rainer (WILLIAMS; RAINER, 2017; WILLIAMS; RAINER, 2019) and our investigations. As shown in this table, 13 credibility criteria were identified. Among them, three were exclusively found in our investigations (*Renowned institutions, Renowned companies,* and *Peer interaction*). Interesting to note that, in our follow-up survey (S1), *Renowned institutions* and *Renowned companies* are among the criteria that researchers considered most important.

Table 66 – Related works: Comparing the findings related to the GL Credibility criteria.

Credibility criteria	(1)	(2)	This thesis
Citations to practitioner sources		x	
Citations to research sources		x	
Cited by a renowned source		x	x
Peer interaction			x
Presence of reasoning		x	
Presenting empirical evidence	x	x	
Relevance	x		
Renowned authors		x	x
Renowned companies			x
Renowned institutions			x
Reporting methods of data collection		x	
Research experience		x	x
Rigor	x	x	x

(1) (WILLIAMS; RAINER, 2017), (2) (WILLIAMS; RAINER, 2019).
Source: the author.

8.3 GREY LITERATURE IN SECONDARY STUDIES

We identified a set of tertiary studies that provided an overview and explored the GL evidence in Secondary Studies, which we present in the following.

Kitchenham et al.'s study

Kitchenham and colleagues (KITCHENHAM et al., 2009) conducted one of the first studies using the multivocal approach in SE, comparing manual and automated searches and assessing the importance and breadth of GL. Their findings showed the importance of GL, especially, to investigate research questions that need practical and technical answers. For instance, when comparing two technologies. Although they recognized that, in general, the quality of GL studies is lower than TL.

Differences with our investigations: We consider this study a starting point for SE researchers to think about other sources of evidence, different from traditional literature, that could contribute to Secondary Studies. However, our investigations expand this notion of the contribution of GL in several manners by exploring the opinions of researchers and deeply assessing how GL contributed to MLR studies.

Garousi et al.'s study

As mentioned before, Garousi et al's study focused on investigating a set of Secondary Studies to identify what knowledge was missed when the study did not include GL, what was gained with its inclusion. Their findings highlighted the importance of the GL inclusion to the Secondary Studies incorporate more evidence retrieved from the practice of SE.

Differences with our investigations: Even the similarities of our works, we deeply explored the evidence in which GL contributed with the findings of the Secondary Studies, by investigating to which GL type and producer each type of contribution are related, and proposed a process to be used to identify GL contributions and its impact on the study.

Galindo Neto et al.'s study

As mentioned early, Galindo and colleagues were the first that focused on explore MLR studies in SE. It is a short investigation that did not make an in-depth investigation about the GL use of Secondary Studies.

Differences with our investigations: The main difference between our research and theirs is that we investigated any Secondary Study to understand GL in these studies. Beyond that, we also investigated the studies' motivations that did not use GL in research and conducted an in-depth investigation of GL use in the Secondary Studies (GL definitions used, where they searched for GL, method used to analyze the data).

Yasin et al.'s study

Yasin and colleagues (YASIN et al., 2020) investigated the evidence of GL use in Secondary Studies published until 2012. Their investigations found GL in 76% of the Secondary Studies, and that 9% of the GL sources included were used as evidence on the synthesis discussion of the works. This work employed some GL definitions, for instance, the Luxembourg definition. Although, GL was also referred to as "fugitive literature" and semi-published work. Regarding the sources, they considered theses, conference proceedings, technical reports, official documents, company whitepapers, discussion boards, and blogs as GL types.

Differences with our investigations: Comparing this study with our investigations, we expand this investigation to cover the studies published until 2018. We expanded our investigation's scope (we explored the methods of studies used to collect and assess GL's quality, motivations for use and reasons to avoid it, and perceived benefits and challenges of its use). In addition, we covered more sources to search for the studies and employed different interpretations about GL types. This work is particularly interesting to provide the first overview about GL in Secondary Studies in SE.

Zhang et al.'s study

Zhang et al.'s study (ZHANG et al., 2020) is the closest to our work to the best of our knowledge. They investigated GL using a mixed-methods approach by conducting a tertiary study and a survey with SE researchers that used GL in Secondary Studies. They aimed to obtain an overview of the research community's understanding of (i) the possible definitions of GL in SE (they did not find a standard definition), (ii) the reasons for including GL from the perspectives of both literature users and community experts, (iii) proposing a conceptual model for how SE researchers work with GL in the research life-cycle, and (iv) identifying the significant challenges of GL use in SE.

Differences with our investigations: our work differs in at least three dimensions: (i) we conducted a broader search to retrieve the highest number of studies published on premier SE conferences and journals; (ii) we explored waters not chartered by previous studies (e.g., we investigated the inclusion/exclusion criteria, the perspective on GL use according to the types of Secondary Studies, and the availability of the GL data, investigating the extent to which the research questions are answered using GL); and (iii) we found some different findings that need further investigation. For instance, Zhang and colleagues (ZHANG et al., 2020) found that only 25% of researchers used GL to evaluate their conclusions, and we found that more than 50% of the secondary studies used GL to support their findings. Moreover, the definitions of GL found in our work were slightly different.

8.3.1 Summarizing the differences between our investigations and related works that explored GL in Secondary Studies

This section summarizes the similarities and differences between our investigations and related works concerning the investigations of GL use in Secondary Studies.

Table 67 summarized and compared the related works that investigated GL in Secondary Studies. In the first row, we showed the time-spam investigated by each study. Then, the second row shows how each study interpreted the GL types. It caught our attention because different interpretations occurred; although even more recent was the period investigated, the GL interpretation employed was most similar. For instance, the Kitchenham et al.'s (KITCHENHAM et al., 2009) and Yasin et al.'s studies (YASIN et al., 2020) interpreted workshop papers as a GL type. Yasin et al. still considered conferences papers as GL type. We did not have the same interpretation in this thesis. We considered these types as Traditional Literature, and we considered the GL types according to the "Shades of GL", considering GL into different tiers, as we explained in Section 2.3.

In the third row, we showed that these studies have similarities and several distinctions in the investigated topics. For instance, Kitchenham et al.'s study (KITCHENHAM et al., 2009) focused on the GL contributions and the quality of evidence of GL sources. The only study investigated it, showing that including GL, the quality of evidence of the Secondary Studies tends to decrease. It differs from our analysis because we did not focus on the quality but on what information could be missed, showing that some information did not exist for various

studies if GL was not considered. In addition, our investigations covered more topics, as shown in Table 67, column seven.

In the fourth row, we present the information about the usage of GL in the investigations. The studies of Kitchenham and colleagues (KITCHENHAM et al., 2009) and Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2016) were the one that related the value in terms of primary sources included into three MLR studies. Considering the others, the study of Zhang and colleagues (ZHANG et al., 2020) presented similar value with our investigation.

Table 67 – Comparing the related works that explored Grey Literature in Secondary Studies.

Subject	(1)	(2)	(3)	(4)	(5)	This thesis (S2)
Time-spam	01/2004 to 06/2007	Not determined	01/2009 to 04/2019	01/2004 to 06/2012	Until 06/2019	01/2004 to 12/2019*
GL types considered	Workshop papers, book chapters, technical reports	Books, videos, whitepapers, webinars, Q&A websites, technical reports, blog posts, slide presentation	Web articles, videos, whitepapers, blog posts, books, magazines, technical reports	Conference and workshop papers, technical reports, thesis, preprints, lecture notes, guidelines	"Shades of GL"	"Shades of GL"

Topics investigated	GL contributions, Quality of source	GL usage, GL contributions.	GL types, motivations to use, procedures used to search for GL.	GL usage, GL types, Using Google to search for GL.	GL definitions, GL usage, GL types, benefits, challenges, motivations to use.	GL definitions, GL usage, GL contributions, GL types, benefits, challenges, motivations to use, reasons to avoid, procedures used to search, select, and QA of GL, GL availability.
Usage of GL	26.5% (sources)	66%, 28%, 100% (sources) (**)	Not mentioned	76% (Secondary Studies)	26% (Secondary Studies)	28.2% (Secondary Studies)

Notes:

(1) (KITCHENHAM et al., 2009), (2) (GAROUSI; FELDERER; MÄNTYLÄ, 2016),

(3) (GALINDO NETO et al., 2019), (4) (YASIN et al., 2020),

(5) (ZHANG et al., 2020), (*) We searched for MLR and GLR studies up to 12/2019.

The other types (SLR, MS), until 12/2018, (**) Were reported three MLR studies.

Source: the author.

8.4 CHALLENGES, LESSONS LEARNED, AND RECOMMENDATIONS TO DEAL WITH GREY LITERATURE IN SECONDARY STUDIES

Similarly, as we have done in our investigations, some studies exposed challenges, lessons learned, and provide some recommendations with their experience investigating or dealing with GL in Secondary Studies. In the following, we briefly discussed these studies, comparing their findings with what we found in our investigations.

Soldani et al.'s study

As mentioned before, Soldani et al. (SOLDANI; TAMBURRI; VAN DEN HEUVEL, 2018) conducted a GLR. This study observed that due to the diversity of GL forms, it is difficult to use a unique measure to assess its quality. Another difficulty to conduct a GLR was related, specifically to a more time-consuming to assess video content when compared with blog posts and whitepapers, although the authors considered videos as important sources that provide more specific and rich information.

Differences with our investigations: Comparing this study with our investigations, we perceived that the first issue presented was also the same as we presented in Study 3 (tertiary study) and Study 4 (focus group). We inform that our investigations showed additional issues, for instance, the challenges of the lack of GL information and several points of improvements perceived to Garousi's guidelines.

Wen et al.'s study

Wen and colleagues (WEN et al., 2020) conducted a GLR to understand the Free/Libre/Open Source Software (FLOSS) phenomenon, assessing community publications. In this research, they highlighted the challenges, specificities, adaptations, and lessons learned from their experiences in conducting a Grey Literature Review (GLR). For instance, they reported four essential aspects to ensure the quality of the selected documents: (1) outline the problem and define the research question, (2) define the inclusion and exclusion criteria, (3) develop a relaxed search string, and (4) define the resource-types to consider. In addition, they developed and applied additional methods to refine the screening step and complete the GL search: (5)

data source evaluation and the selection and (6) incremental screening process.

Differences with our investigations: When comparing this study with our investigations, we perceived that interesting findings were reported that we did not identify. For instance, using a relaxed search string appears as an important step that a researcher could consider. Although, we also identified different challenges (e.g., lack of GL definition, grey literature availability), and points of improvement (issues and problems of Garousi's guidelines) in Study 3 (tertiary study) and 4 (focus group), which were not reported in Wen et al.'s study.

Zhang et al.'s study

The study of Zhang and colleagues (ZHANG et al., 2021) is an experience report based on one single case (an GLR on DevSecOps). The authors distill ten challenges into nine activities of a GLR process and further suggest ways to tackle each challenge. The study also discusses the decision process for selecting a suitable review methodology amongst SLR, MLR, and GLR and elaborates the impacts of GL on review results. Some challenges reported were: (1) Difficult to confirm the need for GL in emerging fields, e.g., DevSecOps; (2) Necessity of re-scoping the research questions (RQ) under the uncertainty of GL; (3) The lack of well-structure of GL sources make it need to conduct pre-search research to scope the proposed RQ from GL properly; and (4) The necessity to develop a quality assessment framework suitable for most GL types.

Differences with our investigations: Comparing our findings, we confirmed that several challenges showed in Zhang et al. were identified in our investigations (e.g., the lack of well-structure of GL and lack of QA criteria that cover most of GL types). We inform that we find other challenges (and proposed recommendations to deal with them), such as the lack of GL definition and search efforts for Grey Literature. In addition, we provided various challenges and points of improvement based on the researcher's opinions identified in our Study 3 (focus group) about Garousi's guidelines.

Melegati et al.'s study

Melegati et al. (MELEGATI; GUERRA; WANG, 2021) performed a GLR to understand what activities were used in software startups, are entailed to handle hypotheses, facilitating the comparison, creation, and evaluation of relevant techniques. In this study, the authors reported that the high number of on GL sources retrieved and how to select them were identified as challenging activities, since in GL there is no metadata information, such as title, venue, year, and abstract. For this challenge, they proposed using separated search strings to compare the results with different synonyms.

Differences with our investigations: Comparing our findings, we perceived that the challenge reported in Melegati et al.'s stuy is quite similar to what we identified in our Study 1 (survey), in which six SE researchers pointed out the lack of writing pattern identified in GL sources. We inform that we find other challenges, lessons learned, and recommendations were investigated in our research.

8.4.1 Summarizing the differences between our investigations and the related works that explored challenges and recommendations to deal with GL in Secondary Studies

This section summarizes the similarities and differences between our investigations and related works concerning the Challenges and Recommendations of GL use in Secondary Studies.

Table 68 present 15 challenges identified. From them, 12 were also found in our investigations, as we checked in the column six ("This thesis"). As we can see, three challenges were only identified in our investigations. Among them, the "lack of classification of GL sources included" and "information about GL" identified in our investigations of Study 2 and Study 3. The last implies to make it difficult a deeply analyze of GL sources. For instance, who are the producers of the included sources. Another challenge was related to the "misunderstanding of a GL type" (Study 2), as we perceived that some researchers considered some types a GL type, and others did not.

Table 68 – Comparing the related works that inform Challenges dealing with GL in Secondary Studies.

Challenges	(1)	(2)	(3)	(4)	(5)	This thesis	Part of process

Differences on how industry and academia express about the same topic.	x	x					Search process
Different understanding of GL definition.			x			x	General
Difficult to set appropriate stopping criteria to the search process.				x		x	Search process
GL has long been troublesome to search and locate.				x		x	Search process
Huge amount of data to be analyzed				x	x	x	General
Lack of classification of GL sources included in Secondary Studies.						x	Data Extraction
Lack of information about GL included in Secondary Studies.						x	Data Extraction
Lack of structured information.					x	x	Data Extraction and Analysis
Missing reliable GL sources in SE			x			x	General
Misunderstanding of what is a GL type						x	General
Performing a thematic analysis is more challenging with GL than TL.					x		Data Extraction and Analysis

The diversity of GL sources makes it difficult to perform a unique measure of QA.	x		x	x		x	QA
The necessity to re-scope the research questions after raising the initial ones using a pre-search.				x			Pre-search
Time-consuming to assess video content.	x					x	Data Extraction and Analysis
Uncertain availability of a GL source along the years.			x			x	Replicability

Notes:

Type of Issue: C = Challenge, R = Recommendation

Studies: (1) (SOLDANI; TAMBURRI; VAN DEN HEUVEL, 2018), (2) (WEN et al., 2020), (3) (ZHANG et al., 2020), (4) (ZHANG et al., 2021), (5) (MELEGATI; GUERRA; WANG, 2021).

Source: the author.

Table 69 – Comparing the related works that provided recommendations to better deal with GL in Secondary Studies.

Recommendations	(1)	(2)	(3)	(4)	(5)	This thesis	Part of process
A pre-search may bring help while understanding the important GL sources on the topic.						x	Search process

A quality assessment framework suitable for most types of gray literature and the assorted improved methods based on specific criteria need to be proposed.	x			x		x	QA
Adopting the consensus models or principles in specific fields is an effective method to reduce the effort of qualitative coding and synthesis.				x			Data Extraction and Analysis
Apply snowballing to search for GL.		x				x	Search process
Choose a suitable order of search hit.		x					Search process
Classify the GL data sources when conducting the search.						x	Search process
Consider searching by tags/categories.		x					Search process
Consult experts to search for GL.		x					Search process
Define stopping criteria influenced by weak search engines or large volumes of data.		x					Search process

Define the resource types to consider (in-text content and audiovisual content).		x					Search process
Detailed records of events/controversies in the research process are beneficial to the final report writing.				x		x	Data Analysis
Develop a relaxed search string.		x			x		Search process
Avoid bias excluding part of product marketing articles.				x			Selection process
A continuous search within a search period is recommended for an emerging topic.		x				x	Search process
To include practitioners as a project contributor.				x			Report
Performed a data extraction and synthesis separated.						x	Data Extraction and Analysis
Save the search results as PDF to deal with the problem of reproducibility.				x		x	Data Extraction and Analysis

Store the records of original text and causal connections to achieving traceability.				x		x	Data Extraction and Analysis
The researcher should map data sources and types, search terms, selection criteria, and boundaries		x					General
To include all the relevant synonyms in the search string and use related terms support in the search engines.	x	x					Search process
To use a set of credibility criteria to help in selecting GL sources.						x	Data Selection process
Use an incremental screening process.		x					Search process
Use targeted source to search for GL.						x	Search process
Validate the initial protocol with a pilot study.				x			General

Notes:

Type of Issue: C = Challenge, R = Recommendation

Studies: (1) (SOLDANI; TAMBURRI; VAN DEN HEUVEL, 2018), (2) (WEN et al., 2020)

(3) (ZHANG et al., 2020), (4) (ZHANG et al., 2021), (5) (MELEGATI; GUERRA; WANG, 2021).

Source: the author.

8.5 SUMMARY

This chapter presented the related works according to the similarities of investigations with our work. In the following, we summarized (in-depth differences were previously reported) our main differences:

1. We did not limit our investigation to a specific type of GL;
2. We explored the experience of SE researchers to understand which type of GL they have used;
3. We explored what motivates and demotivates SE researchers to use GL;
4. We found different criteria to assess GL credibility;
5. We considered the different GL types to assess their credibility;
6. We explored a broader population of SE researchers that was not investigated before and not focused only on experimental SE researchers;
7. We explored waters not chartered by previous studies (e.g., we investigated the inclusion/exclusion criteria, the perspective of the GL use according to the types of Secondary Studies, the availability of the GL data, analyzing the extent to which the research questions are answered using GL, and the consistency of the evidence in GL sources in SE);
8. We expanded and deepened the investigations of what the MLR studies gain when GL is considered by investigating all GL evidence included to understand and classify their contributions in MLR studies;
9. We designed a process to support SE researchers that intend to investigate the contributions of GL in Secondary Studies;
10. We identified a set of challenges that SE researchers may face dealing with GL in Secondary Studies. In addition, we provided a set of potential recommendations to deal with each one. Some challenges are similar to the related works, although some others are new;
11. We designed a process to support the decision whether to include GL in Secondary Studies;

12. We designed a process including recommendations to better use GL in Secondary Studies;

13. Until the moment, to the best of our knowledge, no one study has yet investigated the use of the most known guidelines proposed by Garousi et al. (GAROUSI; FELDERER; MÄNTYLÄ, 2019), used to conduct secondary studies using GL.

9 CONCLUDING REMARKS

This chapter concludes this thesis. Section 9.1 presents the problems explored and summarizes the results obtained from each phase conducted (P1–P4). Then, Section 9.2 exposes the proposed future works.

9.1 CONCLUSIONS

Grey Literature (GL) is not recent in diverse areas of knowledge, and recently, its use has increased in Software Engineering (SE) research. Even with this increase, as we previously described in Chapter 1, this thesis was guided to tackle the following problems:

1. The little evidence about the use and the importance of GL in SE research;
2. The lack of investigations of GL credibility criteria to assess a diversity of GL types.
3. The little evidence of how GL use could contribute to Secondary Studies;
4. The lack of assessing Garousi's guidelines used to conduct Secondary Studies using GL.

In this work, we focused on four core investigations to explore these problems, as we discussed in the following.

First, in Phase 1 (P1), our goal was to understand the *perceptions of SE researchers about GL*. To achieve this goal, we performed a survey investigation. Firstly, investigating 76 Brazilian SE researchers to understand better their perceptions about the benefits, motivations to use, challenges, and reasons to avoid GL. From the collected answers of 53 SE researchers, we identified that most researchers are using GL to understand new topics, find information about practical and technical questions, and complement research findings. Its use is most common, for instance, related to community websites, blogs, and technical reports. Nevertheless, reasons to avoid its use were also mentioned, such as its lack of reliability and scientific value. The major of these reasons are related to the perceived challenges of its use. Then, we conducted a follow-up survey with those researchers that answered using GL. We received answers from 34 Brazilian SE researchers to understand how they assess GL credibility. The production rigor, producer's reputation, the permission of peer interaction, and the researcher experience are the most important criteria employed to assess GL credibility.

Second, in Phase 2 (P2), we aimed to *understand the use of GL in Secondary Studies*. Thus, we performed a tertiary study by exploring 446 Secondary Studies. We found that only 126 used GL, showing that GL is not extensively used, although we identified a growth over the years. Despite this low use, we identified that most of the Secondary Studies (n=126) used GL as source of information to answer at least one research question. Among the studies that searched or included GL, we identified a lack of specific procedures to search for GL and assess its quality. The use of GL in those studies were most related to books or book chapter, technical reports, theses, and web articles. Most of these sources were produced by consultants and academia. We also identified that almost 50% of the GL sources assessed are now unavailable.

Third, in Phase 3 (P3), we explored *how GL contributed to MLR studies*. We also conducted another tertiary study by exploring nine MLR studies. Our findings showed that miscellaneous information provided in MLR studies were exclusively retrieved from GL sources, showing that, for some studies, some information would be missed if the study did not consider GL. Among the contributions, GL was mainly used to explain some topics, classify findings, and provide recommendations. These contributions were most related to the use of blog posts, slide presentations, and project or software descriptions, which were produced mainly by SE practitioners, consultants or companies, and tool vendors. Although our investigations identified that the inclusion of GL increased the additional evidence and information provided, we can not claim the quality of the data retrieved, only that some information was not found in TL sources and that some RQs would not be answered without the use of GL sources.

Fourth, in Phase 4 (P4), our last investigation, we *assessed Garousi's guidelines* from the perspective of SE researchers that read and have used them. In general, the investigated researchers considered useful the proposed guidelines to help in the conduction of Secondary Studies using GL. However, many improvements are necessary to solve the issues identified. By discovering these issues, we provided a set of potential points of improvements, recommendations, and strategies to deal with these issues based on the researcher's opinions, our investigations, and previous studies.

Finally, based on the findings of the four phases conducted, we proposed two processes. One *to help SE researchers decide whether to use GL in Secondary Studies*, and the other one *to provide recommendations to conduct Secondary Studies using GL*. We inform that a limitation is that these proposals have not yet passed through community evaluation. For this reason, we intend to perform this assessment in future work.

To conclude, GL shows as an important source of information that could contribute to SE's Secondary Studies in different ways. However, despite the perceived motivations and benefits of using GL in SE research, we identified a set of challenges that SE researchers may face and reasons to avoid its use that are important to the researcher considering deciding whether to use GL in their Secondary Study. For this reason, we highlighted the importance of GL being used with additional care, for instance, by considering first, when to use and what GL sources to use. In addition, we recommended using GL with some credibility criteria and considering the recommendations proposed in our investigations and previous literature to better deal with the use of GL in Secondary Studies.

9.2 FUTURE WORKS

The investigations present in this work are the initial step toward improving the use of GL in SE research. In the following, we present some open challenges to guide the directions of future works:

- **GL types vs Research Question types**. Some challenges identified in our investigations are related to the diversity of types and the high amount of data produced in GL. For this reason, we consider it essential to understand the best GL types to be used with specific types of research questions. It will help future researchers keep focused on investigating GL sources that can contribute more to the research;
- **Incorporate a practitioner's perspective to assess the credibility of GL**. In this work and previous studies, the criteria used to assess GL credibility were retrieved from the investigations through SE researchers. However, Williams (WILLIAMS, 2019) pointed out the importance of evaluating the GL credibility from SE practitioners' perspective. This investigation becomes even more important because Marculescu et al. (MARCULESCU; JABBARI; MOLLÉRI, 2016) identified that SE practitioner's and researcher's perceptions of what each one considered as evidence are different. Thus, they could assess GL differently. For this reason, we considered that a merge of both views could improve the GL credibility criteria to be used in SE research;
- **Quality Assessment to cover most of the GL types**. Our investigations showed a lack of guidelines covering most GL types, mainly related to the Quality Assessment

instrument. For this reason, we emphasize the importance of future investigations to explore this gap;

- **GL availability**. In our investigations, we identified that most GL sources are unavailable. This problem hinders the Secondary Studies' replicability. Thus, for future work, investigations must explore how to maintain the content GL sources used in the Secondary Studies to be future assessed by other SE researchers;

- **Rapid Reviews using GL**. As we identified, the inclusion of GL in Secondary Studies increases the time spent conducting the research. In addition, to better focus on the SE practitioners' demand, one possible way to deal with it is to conduct Rapid Reviews using GL sources;

- **Update the version of the existing guidelines to include GL in Secondary Studies**. Based on our investigation assessing the researcher's perceptions about Garousi's guidelines, our investigations to deal with Secondary Studies that used GL, and previous studies that provided recommendations and lessons learned, we consider it critical that the SE research community discuss and propose an updated version of the Garousi's guidelines;

- **Evaluate the proposed process to help decide whether to use GL**. In this work, we provided two processes to help SE researchers decide whether to use GL to perform a Secondary Study and support the researchers with recommendations to conduct Secondary Studies better using GL. However, these processes' proposals were not evaluated in the research community. For this reason, we intend to perform these evaluations.

REFERENCES

ADAMS, J.; HILLIER-BROWN, F. C.; MOORE, H. J.; LAKE, A. A.; ARAUJO-SOARES, V.; WHITE, M.; SUMMERBELL, C. Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. *Systematic Reviews*, v. 5, n. 1, p. 164, 2016. Available at: <https://doi.org/10.1186/s13643-016-0337-y>.

ADAMS, R. J.; SMART, P.; HUFF, A. S. Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews*, v. 19, n. 4, p. 432–454, 2017. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ijmr.12102>.

AMPATZOGLOU, A.; BIBI, S.; AVGERIOU, P.; CHATZIGEORGIOU, A. Guidelines for managing threats to validity of secondary studies in software engineering. In: _____. *Contemporary Empirical Methods in Software Engineering*. Cham: Springer International Publishing, 2020. p. 415–441. ISBN 978-3-030-32489-6. Available at: <https://doi.org/10.1007/978-3-030-32489-6_15>.

ANGELIS, G. D.; LONETTI, F. About the assessment of grey literature in software engineering. In: *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2021. (EASE '21), p. 373–378. ISBN 9781450390538. Available at: <https://doi.org/10.1145/3463274.3463362>.

ANICHE, M.; TREUDE, C.; STEINMACHER, I.; WIESE, I.; PINTO, G.; STOREY, M.-A.; GEROSA, M. A. How modern news aggregators help development communities shape and share knowledge. In: *Proceedings of the 40th International Conference on Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2018. (ICSE '18), p. 499–510. ISBN 9781450356381. Available at: <https://doi.org/10.1145/3180155.3180180>.

AURUM, A.; WOHLIN, C. Applying decision-making models in requirements engineering. In: *Proceedings of the 8th International Working Conference on Requirements Engineering: Foundation for Software Quality*. [S.l.: s.n.], 2002. (REFSQ '02).

BABAR, M. A.; ZHANG, H. Systematic literature reviews in software engineering: Preliminary results from interviews with researchers. In: *Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement*. [s.n.], 2009. (ESEM '09), p. 346–355. Available at: <https://doi.org/10.1109/ESEM.2009.5314235>.

BALTES, S.; RALPH, P. *Sampling in Software Engineering Research: A Critical Review and Guidelines*. 2020. Available at: <https://arxiv.org/abs/2002.07764>.

BASILI, V. R. Editorial. *Empirical Software Engineering*, v. 1, n. 2, p. 105–108, jan 1996. ISSN 1382-3256 (print), 1573-7616 (electronic). Available at: <http://link.springer.com/accesspage/article/10.1007/BF00368700>.

BERG, V.; BIRKELAND, J.; NGUYEN-DUC, A.; PAPPAS, I. O.; JACCHERI, L. Software startup engineering: A systematic mapping study. *Journal of Systems and Software*, v. 144, p. 255–274, 2018. ISSN 0164-1212. Available at: <https://www.sciencedirect.com/science/article/pii/S0164121218301286>.

BONATO, S. *A Handbook for Searching Reports, Working Papers, and Other Unpublished Research*. [S.l.]: Rowman & Littlefield, 2018.

BRAUN, V.; CLARKE, V. Using thematic analysis in psychology. *Qualitative Research in Psychology*, Routledge, v. 3, n. 2, p. 77–101, 2006.

CARTAXO, B.; PINTO, G.; VIEIRA, E.; SOARES, S. Evidence briefings: Towards a medium to transfer knowledge from systematic reviews to practitioners. In: *Proceedings of the ACM/IEEE 10th International Symposium on Empirical Software Engineering and Measurement*. New York, NY, USA: Association for Computing Machinery, 2016. (ESEM '16). ISBN 9781450344272. Available at: <https://doi.org/10.1145/2961111.2962603>.

COSTA, M.; GOMES, D.; SILVA, M. The evolution of web archiving. *International Journal on Digital Libraries*, v. 18, p. 191–205, 09 2017. Available at: <https://doi.org/10.1007/s00799-016-0171-9>.

CRUZES, D. S.; DYBÅ, T. Recommended steps for thematic synthesis in software engineering. In: *Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement*. USA: IEEE Computer Society, 2011. (ESEM '11), p. 275–284. ISBN 9780769546049. Available at: <https://doi.org/10.1109/ESEM.2011.36>.

CRUZES, D. S.; DYBÅ, T. Research synthesis in software engineering: A tertiary study. *Information and Software Technology*, v. 53, n. 5, p. 440–455, 2011. ISSN 0950-5849. Available at: <http://www.sciencedirect.com/science/article/pii/S095058491100005X>.

CURKOVIC, M.; KOšEC, A. Bubble effect: Including internet search engines in systematic reviews introduces selection bias and impedes scientific reproducibility. *BMC Medical Research Methodology*, v. 18, 11 2018.

DANCEY, C. P.; REIDY, J. *Statistics Without Maths for Psychology: Using Spss for Windows*. USA: Prentice-Hall, Inc., 2004. ISBN 013124941X.

EASTERBROOK, S.; SINGER, J.; STOREY, M.-A.; DAMIAN, D. Selecting empirical methods for software engineering research. In: _____. *Guide to Advanced Empirical Software Engineering*. London: Springer London, 2008. p. 285–311. ISBN 978-1-84800-044-5. Available at: <https://doi.org/10.1007/978-1-84800-044-5_11>.

ECK, M.; PALOMBA, F.; CASTELLUCCIO, M.; BACCHELLI, A. Understanding flaky tests: The developer's perspective. In: *Proceedings of the 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. New York, NY, USA: ACM, 2019. (ESEC/FSE '19), p. 830–840. ISBN 9781450355728. Available at: <https://doi.org/10.1145/3338906.3338945>.

FARACE, D.; SCHÖPFEL, J. *Grey Literature in Library and Information Studies*. Berlin, Boston: De Gruyter Saur, 2010. ISBN 978-3-598-44149-3. Available at: <https://www.degruyter.com/view/title/34553>.

FELDERER, M.; TRAVASSOS, G. H. The evolution of empirical methods in software engineering. In: _____. *Contemporary Empirical Methods in Software Engineering*. Cham: Springer International Publishing, 2020. p. 1–24. ISBN 978-3-030-32489-6. Available at: <https://doi.org/10.1007/978-3-030-32489-6_1>.

FELDT, R.; MAGAZINIUS, A. Validity threats in empirical software engineering research - an initial survey. In: *Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering*. [S.l.]: Knowledge Systems Institute Graduate School, 2010. (SEKE '10), p. 374–379.

FIDEL, R.; GREEN, M. The many faces of accessibility: Engineers' perception of information sources. *Inf. Process. Manage.*, Pergamon Press, Inc., USA, v. 40, n. 3, p. 563–581, jan 2004. ISSN 0306-4573. Available at: <https://doi.org/10.1016/S0306-4573(03)00003-7>.

FISCHER, F.; BÖTTINGER, K.; XIAO, H.; STRANSKY, C.; ACAR, Y.; BACKES, M.; FAHL, S. Stack overflow considered harmful? the impact of copy paste on android application security. In: *Proceedings of the IEEE Symposium on Security and Privacy*. [S.l.: s.n.], 2017. (SP '17), p. 121–136. ISSN 2375-1207.

FLYNN, V.; WILLIAMS, D. W. A framework for understanding the factors that influence the decision-making behavior of it managers as an aid to improved competence in decision-making. In: *Proceedings of the 4th International Competence-Based Management Conference*. [S.l.: s.n.], 1998.

FRANÇA, B. B. N. de; RIBEIRO, T. V.; SANTOS, P. S. M. dos; TRAVASSOS, G. H. Using focus group in software engineering: lessons learned on characterizing software technologies in academia and industry. In: *Proceedings of the 18th Ibero-American Conference on Software Engineering*. [S.l.]: Curran Associates, Inc., 2015. (CIbSE '15), p. 351.

GALINDO NETO, G. T.; SANTOS, W. B.; ENDO, P. T.; FAGUNDES, R. A. A. Multivocal literature reviews in software engineering: Preliminary findings from a tertiary study. In: *Proceedings of the ACM/IEEE 13th International Symposium on Empirical Software Engineering and Measurement*. [s.n.], 2019. (ESEM '19), p. 1–6. ISSN 1949-3770. Available at: <https://doi.org/10.1109/ESEM.2019.8870142>.

GAROUSI, V.; FELDERER, M.; MÄNTYLÄ, M. V. The need for multivocal literature reviews in software engineering: Complementing systematic literature reviews with grey literature. In: *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: ACM, 2016. (EASE '16), p. 26:1–26:6. ISBN 978-1-4503-3691-8. Available at: <http://doi.acm.org/10.1145/2915970.2916008>.

GAROUSI, V.; FELDERER, M.; MÄNTYLÄ, M. V. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology*, v. 106, p. 101 – 121, 2019. ISSN 0950-5849. Available at: <http://www.sciencedirect.com/science/article/pii/S0950584918301939>.

GAROUSI, V.; FELDERER, M.; MÄNTYLÄ, M. V.; RAINER, A. Benefitting from the grey literature in software engineering research. In: _____. *Contemporary Empirical Methods in Software Engineering*. Cham: Springer International Publishing, 2020. p. 385–413. ISBN 978-3-030-32489-6. Available at: <https://doi.org/10.1007/978-3-030-32489-6_14>.

GAROUSI, V.; FELDERER, M.; MäNTYLä, M. V. *Guidelines for including grey literature and conducting multivocal literature reviews in software engineering*. 2017.

GAROUSI, V.; KüçüK, B. Smells in software test code: A survey of knowledge in industry and academia. *Journal of Systems and Software*, v. 138, p. 52 – 81, 2018. ISSN 0164-1212. Available at: <http://www.sciencedirect.com/science/article/pii/S0164121217303060>.

GODIN, K.; KIRKPATRICK, S.; HANNING, R.; STAPLETON, J.; LEATHERDALE, S. T. Examining guidelines for school-based breakfast programs in canada: A systematic review of the grey literature. *Canadian Journal of Dietetic Practice and Research*, v. 78, p. 1–9, feb 2017.

GUL, S.; SHAH, T.; AHMAD, S.; SHABIR, T. Is grey literature really grey or a hidden glory to showcase the sleeping beauty. *Collection and Curation*, ahead-of-print, 02 2020.

GUYATT, G.; CAIRNS, J.; CHURCHILL, D. Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA*, v. 268, n. 17, p. 2420–2425, 1992. Available at: <http://dx.doi.org/10.1001/jama.1992.03490170092032>.

HIQA. *Evidence Synthesis Process; Methods in the development of National Standards, Guidance and Recommendations for the Irish health and social care sector.* [S.l.], 2018.

HOLLOWAY, I. *Basic concepts for qualitative research*. 7. ed. [S.l.]: Blackwell Science, 1997.

HOSSEINZADEH, S.; RAUTI, S.; LAURéN, S.; MäKELä, J.-M.; HOLVITIE, J.; HYRYNSALMI, S.; LEPPäNEN, V. Diversification and obfuscation techniques for software security: A systematic literature review. *Information and Software Technology*, v. 104, p. 72–93, 2018. ISSN 0950-5849. Available at: <https://www.sciencedirect.com/science/article/pii/S0950584918301484>.

KAMEI, F.; PINTO, G.; WIESE, I.; RIBEIRO, M.; SOARES, S. What evidence we would miss if we do not use grey literature? In: *Proceedings of the ACM/IEEE 15th International Symposium on Empirical Software Engineering and Measurement*. New York, NY, USA: Association for Computing Machinery, 2021. (ESEM '21). ISBN 9781450386654. Available at: <https://doi.org/10.1145/3475716.3475777>.

KAMEI, F.; WIESE, I.; LIMA, C.; POLATO, I.; NEPOMUCENO, V.; FERREIRA, W.; RIBEIRO, M.; PENA, C.; CARTAXO, B.; PINTO, G.; SOARES, S. Grey literature in software engineering: A critical review. *Information and Software Technology*, p. 106609, 2021. ISSN 0950-5849. Available at: <https://www.sciencedirect.com/science/article/pii/S0950584921000860>.

KAMEI, F.; WIESE, I.; PINTO, G.; RIBEIRO, M.; SOARES, S. On the use of grey literature: A survey with the brazilian software engineering research community. In: *Proceedings of the 34th Brazilian Symposium on Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2020. (SBES '20), p. 183–192. ISBN 9781450387538. Available at: <https://doi.org/10.1145/3422392.3422442>.

KAMEI, F. K. The use of grey literature review as evidence for practitioners. *SIGSOFT Softw. Eng. Notes*, Association for Computing Machinery, New York, NY, USA, v. 44, n. 3, p. 23, Nov. 2019. ISSN 0163-5948. Available at: <https://doi.org/10.1145/3356773.3356797>.

KAMEI, F. K.; SOARES, S.; PINTO, G. The use of grey literature review as evidence for software engineering. In: *Anais Estendidos da X Conferência Brasileira de Software: Teoria e Prática*. Porto Alegre, RS, Brasil: SBC, 2019. (CBSoft '19), p. 56–63. ISSN 2177-9384. Available at: <https://sol.sbc.org.br/index.php/cbsoft_estendido/article/view/7656>.

KAMEI, F. K.; WIESE, I.; PINTO, G.; FERREIRA, W.; RIBEIRO, M.; SOUZA, R.; SOARES, S. Assessing the credibility of grey literature: A study with brazilian software engineering researchers. *Journal of Software Engineering Research and Development*, v. 10, p. 9:1 –

9:20, Jun. 2022. Available at: <https://sol.sbc.org.br/journals/index.php/jserd/article/view/1897>.

KITCHENHAM, B.; BRERETON, O. P.; BUDGEN, D.; TURNER, M.; BAILEY, J.; LINKMAN, S. Systematic literature reviews in software engineering – a systematic literature review. *Information and Software Technology*, v. 51, n. 1, p. 7–15, 2009. ISSN 0950-5849. Special Section - Most Cited Articles in 2002 and Regular Research Papers. Available at: <http://www.sciencedirect.com/science/article/pii/S0950584908001390>.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. [S.l.], 2007. Available at: <http://www.dur.ac.uk/ebse/resources/Systematic-reviews-5-8.pdf>.

KITCHENHAM, B.; PRETORIUS, R.; BUDGEN, D.; BRERETON, O. P.; TURNER, M.; NIAZI, M.; LINKMAN, S. Systematic literature reviews in software engineering - a tertiary study. *Information and Software Technology*, Butterworth-Heinemann, Newton, MA, USA, v. 52, n. 8, p. 792–805, aug 2010. ISSN 0950-5849. Available at: <http://dx.doi.org/10.1016/j.infsof.2010.03.006>.

KITCHENHAM, B. A.; AL-KILIDAR, H.; BABAR, M. A.; BERRY, M.; COX, K.; KEUNG, J.; KURNIAWATI, F.; STAPLES, M.; ZHANG, H.; ZHU, L. Evaluating guidelines for reporting empirical software engineering studies. *Empir. Softw. Eng.*, v. 13, n. 1, p. 97–121, 2008. Available at: <https://doi.org/10.1007/s10664-007-9053-5>.

KITCHENHAM, B. A.; BRERETON, P.; TURNER, M.; NIAZI, M.; LINKMAN, S.; PRETORIUS, R.; BUDGEN, D. The impact of limited search procedures for systematic literature reviews - a participant-observer case study. In: *Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement*. IEEE, 2009. (ESEM '09). Available at: <https://doi.org/10.1109/ESEM.2009.5314238>.

KITCHENHAM, B. A.; BUDGEN, D.; BRERETON, P. *Evidence-Based Software Engineering and Systematic Reviews*. [S.l.]: Chapman & Hall/CRC, 2015. ISBN 1482228653, 9781482228656.

KITCHENHAM, B. A.; DYBÅ, T.; JØRGENSEN, M. Evidence-based software engineering. In: *Proceedings of the 26th International Conference on Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2004. (ICSE '04), p. 273–281. ISBN 0-7695-2163-0. Available at: <http://dl.acm.org/citation.cfm?id=998675.999432>.

KITCHENHAM, B. A.; PFLEEGER, S. L. Personal opinion surveys. In: _____. *Guide to Advanced Empirical Software Engineering*. London: Springer London, 2008. p. 63–92. ISBN 978-1-84800-044-5. Available at: <https://doi.org/10.1007/978-1-84800-044-5_3>.

KONTIO, J.; BRAGGE, J.; LEHTOLA, L. The focus group method as an empirical tool in software engineering. In: _____. *Guide to Advanced Empirical Software Engineering*. London: Springer London, 2008. p. 93–116. ISBN 978-1-84800-044-5. Available at: <https://doi.org/10.1007/978-1-84800-044-5_4>.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, [Wiley, International Biometric Society], v. 33, n. 1, p. 159–174, 1977. ISSN 0006341X, 15410420. Available at: <http://www.jstor.org/stable/2529310>.

LINÅKER, J.; SULAMAN, S.; Maiani de Mello, R.; MARTIN, H. *Guidelines for Conducting Surveys in Software Engineering*. [S.l.], 2015.

MACLEOD, L.; BERGEN, A.; STOREY, M.-A. Documenting and sharing software knowledge using screencasts. *Empirical Software Engineering*, v. 22, n. 3, p. 1478–1507, Jun 2017. ISSN 573-7616. Available at: <https://doi.org/10.1007/s10664-017-9501-9>.

MACLEOD, L.; STOREY, M.-A.; BERGEN, A. Code, camera, action: How software developers document and share program knowledge using youtube. In: *Proceedings of the IEEE 23rd International Conference on Program Comprehension*. [s.n.], 2015. (ICPC '15), p. 104–114. ISSN 1092-8138. Available at: <https://doi.org/10.1109/ICPC.2015.19>.

MARCULESCU, B.; JABBARI, R.; MOLLÉRI, J. S. Perception of scientific evidence: Do industry and academia share an understanding? In: *Lärarlärdom, Kristianstad*. [S.l.]: Kristianstad University Press, 2016. p. 121–133.

MARO, S.; STEGHÄFER, J.-P.; STARON, M. Software traceability in the automotive domain: Challenges and solutions. *Journal of Systems and Software*, v. 141, p. 85 – 110, 2018. ISSN 0164-1212. Available at: <http://www.sciencedirect.com/science/article/pii/S0164121218300608>.

MELEGATI, J.; GUERRA, E.; WANG, X. Understanding hypotheses engineering in software startups through a gray literature review. *Information and Software Technology*, v. 133, p. 106465, 2021. ISSN 0950-5849. Available at: <https://www.sciencedirect.com/science/article/pii/S0950584920302111>.

MENDEZ, D.; GRAZIOTIN, D.; WAGNER, S.; SEIBOLD, H. Open science in software engineering. In: _____. *Contemporary Empirical Methods in Software Engineering*. Cham: Springer International Publishing, 2020. p. 477–501. ISBN 978-3-030-32489-6. Available at: <https://doi.org/10.1007/978-3-030-32489-6_17>.

MORGAN, D. L. Focus groups. *Annual Review of Sociology*, v. 22, n. 1, p. 129–152, 1996. Available at: <https://doi.org/10.1146/annurev.soc.22.1.129>.

NYUMBA, T. O.; WILSON, K.; DERRICK, C. J.; MUKHERJEE, N. The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and Evolution*, v. 9, n. 1, p. 20–32, 2018. Available at: <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12860>.

OGAWA, R. T.; MALEN, B. Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. *Review of Educational Research*, v. 61, n. 3, p. 265–286, 1991. Available at: <https://doi.org/10.3102/00346543061003265>.

PAEZ, A. Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, v. 10, n. 3, p. 233–240, 2017.

PETTICREW, M.; ROBERTS, H. *Systematic Reviews in the Social Sciences: A Practical Guide*. [S.l.: s.n.], 2006.

PIASECKI, J.; WALIGORA, M.; DRANSEIKA, V. Google search as an additional source in systematic reviews. *Science and Engineering Ethics*, v. 24, n. 2, p. 809–810, Apr 2018. ISSN 1471-5546. Available at: <https://doi.org/10.1007/s11948-017-0010-4>.

PINTO, G.; FERREIRA, C.; SOUZA, C.; STEINMACHER, I.; MEIRELLES, P. Training software engineers using open-source software: The students' perspective. In: *Proceedings of IEEE/ACM 41st International Conference on Software Engineering: Software Engineering Education and Training*. IEEE, 2019. (ICSE-SEET '19), p. 147–157. Available at: <https://doi.org/10.1109/ICSE-SEET.2019.00024>.

RAINER, A. Using argumentation theory to analyse software practitioners' feasible evidence, inference and belief. *Information and Software Technology*, v. 87, p. 62–80, 2017. ISSN 0950-5849. Available at: <http://www.sciencedirect.com/science/article/pii/S0950584917300769>.

RAINER, A.; WILLIAMS, A. *Technical Report: Do software engineering practitioners cite research on software testing in their online articles? A structured search of grey data*. [S.l.], 2018.

RAINER, A.; WILLIAMS, A. Using blog articles in software engineering research: Benefits, challenges and case–survey method. In: *Proceedings of the 25th Australasian Software Engineering Conference*. [s.n.], 2018. (ASWEC '18), p. 201–209. Available at: <https://doi.org/10.1109/ASWEC.2018.00034>.

RAINER, A.; WILLIAMS, A. Heuristics for improving the rigour and relevance of grey literature searches for software engineering research. *Information and Software Technology*, v. 106, p. 231–233, 2019. ISSN 0950-5849. Available at: <http://www.sciencedirect.com/science/article/pii/S0950584918302192>.

RAINER, A.; WILLIAMS, A. Using blog-like documents to investigate software practice: Benefits, challenges, and research directions. *Journal of Software: Evolution and Process*, v. 31, n. 11, p. e2197, 2019. E2197 smr.2197. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/smr.2197>.

RAULAMO-JURVANEN; PÄIVI; MÄNTYLÄ, M.; GAROUSI, V. Choosing the right test automation tool: A grey literature review of practitioner sources. In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2017. (EASE '17), p. 21–30. ISBN 978-1-4503-4804-1. Available at: <http://doi.acm.org/10.1145/3084226.3084252>.

RODRíGUEZ-PéREZ, G.; ROBLES, G.; GONZáLEZ-BARAHONA, J. M. Reproducibility and credibility in empirical software engineering: A case study based on a systematic literature review of the use of the szz algorithm. *Information and Software Technology*, v. 99, p. 164–176, 2018. ISSN 0950-5849. Available at: <https://www.sciencedirect.com/science/article/pii/S0950584917304275>.

SALAH, D.; PAIGE, R. F.; CAIRNS, P. A systematic literature review for agile development processes and user centred design integration. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2014. (EASE '14). ISBN 9781450324762. Available at: <https://doi.org/10.1145/2601248.2601276>.

SALTAN, A. Do we know how to price saas: A multi-vocal literature review. In: *Proceedings of the 2nd ACM SIGSOFT International Workshop on Software-Intensive Business: Start-Ups, Platforms, and Ecosystems*. ACM, 2019. (IWSiB '19), p. 7–12. Available at: <https://doi.org/10.1145/3340481.3342731>.

SCHöPFEL, J.; PROST, H. How scientific papers mention grey literature: a scientometric study based on scopus data. *Collection and Curation*, ahead-of-print, 03 2020.

SCHUM, D. A. *The Evidential Foundations of Probabilistic Reasoning*. first. [S.l.]: Northwestern University Press, 2001.

SHARMA, T.; SPINELLIS, D. A survey on software smells. *Journal of Systems and Software*, v. 138, p. 158–173, 2018. ISSN 0164-1212. Available at: <http://www.sciencedirect.com/science/article/pii/S0164121217303114>.

SILVA, F. Q. B. da; SANTOS, A. L.; SOARES, S.; FRANćA, A. C. C.; MONTEIRO, C. V. F.; MACIEL, F. F. Six years of systematic literature reviews in software engineering: An updated tertiary study. *Information and Software Technology*, v. 53, n. 9, p. 899–913, 2011. ISSN 0950-5849. Studying work practices in Global Software Engineering. Available at: <http://www.sciencedirect.com/science/article/pii/S0950584911001017>.

SILVA, F. S.; SOARES, F. S.; PERES, A. L.; AZEVEDO, I. M. de; VASCONCELOS, A. P. L.; KAMEI, F. K.; MEIRA, S. R. Using cmmi together with agile software development: A systematic review. *Information and Software Technology*, v. 58, p. 20–43, 2015. ISSN 0950-5849. Available at: <http://www.sciencedirect.com/science/article/pii/S0950584914002110>.

SINGER, L.; FILHO, F. F.; STOREY, M.-A. Software engineering at the speed of light: How developers stay current using twitter. In: *Proceedings of the 36th International Conference on Software Engineering*. New York, NY, USA: ACM, 2014. (ICSE '14), p. 211–221. ISBN 978-1-4503-2756-5. Available at: <http://doi.acm.org/10.1145/2568225.2568305>.

SJØBERG, D. I. K.; DYBÅ, T.; JØRGENSEN, M. The future of empirical methods in software engineering research. In: *Future of Software Engineering*. USA: IEEE Computer Society, 2007. (FOSE '07), p. 358–378. ISBN 0769528295. Available at: <https://doi.org/10.1109/FOSE.2007.30>.

SOLDANI, J.; TAMBURRI, D. A.; VAN DEN HEUVEL, W.-J. The pains and gains of microservices: A systematic grey literature review. *Journal of Systems and Software*, v. 146, p. 215–232, 2018. ISSN 0164-1212. Available at: <http://www.sciencedirect.com/science/article/pii/S0164121218302139>.

SPENCER, D. *Card sorting: Designing usable categories*. [S.l.]: Rosenfeld Media, 2009.

STOREY, M.-A.; SINGER, L.; CLEARY, B.; FILHO, F. F.; ZAGALSKY, A. The (r) evolution of social media in software engineering. In: *Proceedings of the Future of Software Engineering*. New York, NY, USA: ACM, 2014. (FOSE '14), p. 100–116. ISBN 978-1-4503-2865-4. Available at: <http://doi.acm.org/10.1145/2593882.2593887>.

STOREY, M.-A. D.; ZAGALSKY, A.; FILHO, F. M. F.; SINGER, L.; GERMáN, D. M. How social and communication channels shape and challenge a participatory culture in software development. *IEEE Trans. Software Eng.*, v. 43, n. 2, p. 185–204, 2017. Available at: <http://dblp.uni-trier.de/db/journals/tse/tse43.html#StoreyZFSG17>.

VIERA, A. J.; GARRETT, J. M. Understanding interobserver agreement: the kappa statistic. *Family Medicine*, v. 37, n. 5, p. 360–363, 2005.

WEN, M.; LEITE, L.; KON, F.; MEIRELLES, P. Understanding floss through community publications: Strategies for grey literature review. In: *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results*. New York, NY, USA: Association for Computing Machinery, 2020. (ICSE-NIER '20), p. 89–92. ISBN 9781450371261. Available at: <https://doi.org/10.1145/3377816.3381729>.

WILLIAMS, A. Using reasoning markers to select the more rigorous software practitioners' online content when searching for grey literature. In: *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering*. ACM Press, 2018. (EASE '18). Available at: <https://doi.org/10.1145/3210459.3210464>.

WILLIAMS, A. *Finding high-quality grey literature for use as evidence in software engineering research*. Phd Thesis (PhD Thesis) — University of Canterbury, Christchurch, New Zealand, 2019.

WILLIAMS, A.; RAINER, A. Toward the use of blog articles as a source of evidence for software engineering research. In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: ACM, 2017. (EASE '17), p. 280–285. ISBN 978-1-4503-4804-1. Available at: <http://doi.acm.org/10.1145/3084226.3084268>.

WILLIAMS, A.; RAINER, A. How do empirical software engineering researchers assess the credibility of practitioner-generated blog posts? In: *Proceedings of the 23rd International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2019. (EASE '19), p. 211–220. ISBN 978-1-4503-7145-2. Available at: <http://doi.acm.org/10.1145/3319008.3319013>.

WOHLIN, C. An evidence profile for software engineering research and practice. In: _____. *Perspectives on the Future of Software Engineering: Essays in Honor of Dieter Rombach*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 145–157. ISBN 978-3-642-37395-4. Available at: <https://doi.org/10.1007/978-3-642-37395-4_10>.

WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2014. (EASE '14). ISBN 9781450324762. Available at: <https://doi.org/10.1145/2601248.2601268>.

WOHLIN, C.; RUNESON, P.; HöST, M.; OHLSSON, M. C.; REGNELL, B.; WESSLÉN, A. *Experimentation in Software Engineering: An Introduction*. USA: Kluwer Academic Publishers, 2000. ISBN 0792386825.

YASIN, A.; FATIMA, R.; WEN, L.; AFZAL, W.; AZHAR, M.; TORKAR, R. On using grey literature and google scholar in systematic literature reviews in software engineering. *IEEE Access*, v. 8, p. 36226–36243, 2020. Available at: <https://doi.org/10.1109/ACCESS.2020.2971712>.

YITZHAKI, M.; HAMMERSHLAG, G. Accessibility and use of information sources among computer scientists and software engineers in israel: Academy versus industry. *Journal of the American Society for Information Science and Technology*, v. 55, n. 9, p. 832–842, 2004. Available at: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.20026>.

ZAHEDI, M.; RAJAPAKSE, R. N.; BABAR, M. A. Mining questions asked about continuous software engineering: A case study of stack overflow. In: *Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2020. (EASE '20), p. 41–50. Available at: <https://doi.org/10.1145/3383219.3383224>.

ZHANG, H.; BABAR, M. A. Systematic reviews in software engineering: An empirical investigation. *Information and Software Technology*, v. 55, n. 7, p. 1341–1354, 2013. ISSN 0950-5849. Available at: <http://www.sciencedirect.com/science/article/pii/S0950584912002029>.

ZHANG, H.; MAO, R.; HUANG, H.; DAI, Q.; ZHOU, X.; SHEN, H.; RONG, G. Processes, challenges and recommendations of gray literature review: An experience report. *Information and Software Technology*, v. 137, p. 106607, 2021. ISSN 0950-5849. Available at: <https://www.sciencedirect.com/science/article/pii/S0950584921000847>.

ZHANG, H.; ZHOU, X.; HUANG, X.; HUANG, H.; BABAR, M. A. An evidence-based inquiry into the use of grey literature in software engineering. In: *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2020. (ICSE '20), p. 1422–1434. ISBN 9781450371216. Available at: <https://doi.org/10.1145/3377811.3380336>.

ZHOU, X.; JIN, Y.; ZHANG, H.; LI, S.; HUANG, X. A map of threats to validity of systematic literature reviews in software engineering. In: *Proceedings of the 23rd Asia-Pacific Software Engineering Conference*. [s.n.], 2016. (APSEC '16), p. 153–160. Available at: <https://doi.org/10.1109/APSEC.2016.031>.

APPENDIX A – QUESTIONNAIRE TO IDENTIFY THE PERCEPTIONS OF THE USE OF GREY LITERATURE IN SURVEY 1

The purpose of this survey is to understand the perceptions about the use of Grey Literature (GL) by Brazilian Software Engineering Researchers.

This is a research project being conduct by Fernando Kenji (UFPE), Gustavo Pinto (UFPA), Márcio Ribeiro (UFAL), Igor Wiese (UTFPR), and Sérgio Soares (UFPE). You are invited to participate in this research because you participated in the Brazilian Conference on Software: Practice and Theory (CBSoft) in 2019.

Your participation in this research study is voluntary. You may choose not to participate. If you decide to participate in this research survey, you may withdraw at any time. If you decide not to participate in this study or if you withdraw from participating at any time, you will not be penalized.

The procedure involves filling an online survey that will take approximately 8 minutes. Your responses will be confidential, and we do not collect identifying information such as your name, email address or IP address. The survey questions will be about your experience and knowledge about Grey Literature.

All the collected information will be used anonymously and all the results will be used for scholarly purposes and may be shared in proceedings of conferences or journals. We will do our best to keep your information confidential.

If you have any questions about the research study, please contact us. This research has been reviewed according to Research Ethics Committee of UFPE procedures for research involving human subjects.

In this appendix, we present the questionnaire used as instrument to conduct a survey with Brazilian Software Engineering Researchers (P1). This questionnaire was conducted in Portuguese and translated to English to compose this thesis. The questions are grouped in two sections. Required questions are marked with an asterisk "*".

PERSONAL INFORMATION

Q1: What is your e-mail?

Q2: What is your gender?*

◯ Woman ◯ Man ◯ Other

Q3: Please list the highest academic degree you have received*.

◯ High school ◯ Technical education
◯ University graduate ◯ Specialization
◯ Master's degree ◯ Doctorate

USE OF GREY LITERATURE

Q4: Have you used Grey Literature? If you never used, go to question *ten**.

◯ Yes ◯ No

Q5: What sources of Grey Literature did you use?

Q6: In which conditions do you use Grey Literature?

◯ Academic ◯ Industry ◯ Other

Q7: In which conditions do you do not use Grey Literature?

Q8: Could you list any benefits in using Grey Literature?

Q9: Could you list any challenges in using Grey Literature?

Q10: If you answered no in the question *four*, please state why did you never use or avoid use Grey Literature?

Q11: What would be a reliable source of Grey Literature for you?

APPENDIX B – QUESTIONNAIRE TO IDENTIFY THE PERCEPTIONS OF THE GREY LITERATURE CREDIBILITY CRITERIA AND ITS TYPES IN SURVEY 2)

The purpose of this survey is to investigate a set of Credibility criteria to assess the different types of GL by Brazilian Software Engineering Researchers.

This is a research project being conduct by Fernando Kenji (UFPE), Gustavo Pinto (UFPA), Márcio Ribeiro (UFAL), Igor Wiese (UTFPR), and Sérgio Soares (UFPE). You are invited to participate in this research because you answered that had used Grey Literature in your SE researches in our last survey conducted with the participants of the Brazilian Conference on Software: Practice and Theory (CBSoft) in 2019.

Your participation in this research study is voluntary. You may choose not to participate. If you decide to participate in this research survey, you may withdraw at any time. If you decide not to participate in this study or if you withdraw from participating at any time, you will not be penalized.

The procedure involves filling an online survey that will take approximately 10 minutes. Your responses will be confidential, and we do not collect identifying information such as your name, email address or IP address. The survey questions will be about your experience and knowledge about credibility of Grey Literature and its types.

All the collected information will be used anonymously and all the results will be used for scholarly purposes and may be shared in proceedings of conferences or journals. We will do our best to keep your information confidential.

If you have any questions about the research study, please contact us. This research has been reviewed according to Research Ethics Committee of UFPE procedures for research involving human subjects.

In this appendix, we present the questionnaire used as an instrument to conduct the Survey 2 with Brazilian Software Engineering Researchers answered that had used Grey Literature in your SE researches. This questionnaire was conducted in Portuguese and translated to English to compose this thesis. The questions are grouped in two sections. Required questions are marked with an asterisk "*".

PERSONAL INFORMATION

Q1: Please list the highest academic degree you have received*.

◯ High school ◯ University graduate ◯ Master's degree
◯ Technical education ◯ Specialization ◯ Doctorate

Q2: How many years of experience do you have conducting Software Engineering research*?

◯ Until 1 year ◯ Between 4 and 6 years ◯ More than 10 years
◯ Between 1 and 3 years ◯ Between 7 and 9 years

CREDIBILITY AND CONTROL OF GREY LITERATURE SOURCES

Instructions: The definition of Gray Literature (or Gray Literature) is any material that is not formally published or peer-reviewed (Garousi, 2019). Grey Literature can be classified according to what we call "Shades of Grey". Please watch the video (2 min) below[1] to learn more about grayscale. The concepts presented will be important to answer questions 3 to 5.

Q3: How many scientific articles have you conducted using any source of Grey Literature as evidence?*

◯ I do not know ◯ Only one ◯ Between 6 and 10
◯ No one ◯ Between 2 and 5 ◯ More than 10

[1] Video explaining the "shades of GL" (in Portuguese): https://youtu.be/hGMkVXIApR0

Q4: We are aware that the *level of Control* varies from source to source. For this reason, we ask you to consider your experience more frequent in relation to each source type in relation to the level of *Control* dimension of the production*.

	Options of answers				
Type of source	High control	Moderate control	Low control	No opinion	✖ GL
Thesis	◯	◯	◯	◯	◯
Patents	◯	◯	◯	◯	◯
Books/Book chapters	◯	◯	◯	◯	◯
Magazine articles	◯	◯	◯	◯	◯
Case/Serv. desc	◯	◯	◯	◯	◯
Manuals	◯	◯	◯	◯	◯
Materials training	◯	◯	◯	◯	◯
Software repositories	◯	◯	◯	◯	◯
Blog posts	◯	◯	◯	◯	◯
Forums / Lists	◯	◯	◯	◯	◯
News articles	◯	◯	◯	◯	◯
Slide presentations	◯	◯	◯	◯	◯
Keynote speeches	◯	◯	◯	◯	◯
Videos	◯	◯	◯	◯	◯
Technical reports	◯	◯	◯	◯	◯
Q&A websites	◯	◯	◯	◯	◯
Guidelines	◯	◯	◯	◯	◯
Tutorials	◯	◯	◯	◯	◯
White papers	◯	◯	◯	◯	◯

Q5: Please explain what you considered when classifying each data type against the control criteria presented in Question 4.

Q6: We are aware that the *level of Credibility* varies from source to source. For this reason, we ask you to consider your experience more frequent in relation to each source type in relation to the *Credibility* dimension of the production*.

	Options of answers				
Type of source	High credibility	Moderate credibility	Low credibility	No opinion	✖ GL
Thesis	◯	◯	◯	◯	◯
Patents	◯	◯	◯	◯	◯
Books/Book chapters	◯	◯	◯	◯	◯
Magazine articles	◯	◯	◯	◯	◯
Case/Serv. desc	◯	◯	◯	◯	◯
Manuals	◯	◯	◯	◯	◯
Materials training	◯	◯	◯	◯	◯
Software repositories	◯	◯	◯	◯	◯
Blog posts	◯	◯	◯	◯	◯
Forums / Lists	◯	◯	◯	◯	◯
News articles	◯	◯	◯	◯	◯
Slide presentations	◯	◯	◯	◯	◯
Keynote speeches	◯	◯	◯	◯	◯
Videos	◯	◯	◯	◯	◯
Technical reports	◯	◯	◯	◯	◯
Q&A websites	◯	◯	◯	◯	◯
Guidelines	◯	◯	◯	◯	◯
Tutorials	◯	◯	◯	◯	◯
White papers	◯	◯	◯	◯	◯

Q7: Please explain what you considered when classifying each data type against the credibility criteria presented in Question 6.

Q8: Considering a GL source with important information to your research, would you include a GL source if it is produced by/with*.

	Options of answers		
Credibility criteria	No opinion	No	Yes
Be produced by a renowned author	◯	◯	◯
Be produced by a renowned institution	◯	◯	◯
Be produced by a renowned company	◯	◯	◯
Be cited by others renowned sources	◯	◯	◯
Describe the methods of collection	◯	◯	◯
Cites an academic reference	◯	◯	◯
Cites a practitioner source	◯	◯	◯
Presents information with rigor	◯	◯	◯
Presents empirical data	◯	◯	◯

Q9: Could you cite any additional potential aspect to assess the credibility of a GL source that was not mentioned before?

Q10: We are planning to conduct a future research about Quality Assessment in Grey Literature. Please, could you inform your mail to future contact?

APPENDIX C – SECONDARY STUDIES INCLUDED IN TERTIARY STUDY 1 (REFERENCES)

In this appendix, we present the bibliographic information about the included studies (n=126) related to the Study 3.

Study ID	Reference
SS1	S. Shahrivar, S. Elahi, A. Hassanzadeh, G. Montazer, A business model for commercial open source software: A systematic literature review, Information and Software Technology 103 (December 2017) (2018) 202–214. doi:10.1016/j.infsof.2018.06.018.
SS2	B. K. Olorisade, E. De Quincey, P. Andras, P. Brereton, A critical analysis of studies that address the use of text mining for citation screening in systematic reviews, ACM International Conference Proceeding Series 01-03-June-2016 (2016). doi:10.1145/2915970.2915982.
SS3	B. Ulziit, Z. A. Warraich, C. Gencel, K. Petersen, A conceptual framework of challenges and solutions for managing global software maintenance, Journal of Software: Evolution and Process 27 (10) (2015) 763–792. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/smr.1720, doi:10.1002/smr.1720.
SS4	D. Haselberger, A literature-based framework of performance-related leadership interactions in ICT project teams, Information and Software Technology 70 (2016) 1–17. doi:10.1016/j.infsof.2015.09.003.
SS5	S. Tiwari, A. Gupta, A systematic literature review of use case specifications research, Information and Software Technology 67 (2015) 128–158. doi:10.1016/j.infsof.2015.06.004.
SS6	K. Patel, R. M. Hierons, A mapping study on testing non-testable systems, Software Quality Journal 26 (4) (2018) 1373–1413. doi:10.1007/s11219-017-9392-4.
SS7	R. E. Lopez-Herrejon, L. Linsbauer, A. Egyed, A systematic mapping study of search-based software engineering for software product lines, Information and Software Technology 61 (2015) 33–51. doi:10.1016/j.infsof.2015.01.008.
SS8	B. Rizvi, E. Bagheri, D. Gasevic, A systematic review of distributed agile software engineering, J. Softw. Evol. Process 27 (10) (2015) 723–762. doi:10.1002/smr.1718.
SS9	S. Nadal, V. Herrero, O. Romero, A. Abelló, X. Franch, S. Vansummeren, D. Valerio, A software reference architecture for semantic-aware Big Data systems, Information and Software Technology 90 (2017) 75–92. doi:10.1016/j.infsof.2017.06.001.
SS10	W. Martin, F. Sarro, Y. Jia, Y. Zhang, M. Harman, A survey of app store analysis for software engineering, IEEE Transactions on Software Engineering 43 (9) (2017) 817–847. doi:10.1109/TSE.2016.2630689.
SS11	K. Mao, L. Capra, M. Harman, Y. Jia, A survey of the use of crowdsourcing in software engineering, Journal of Systems and Software 126 (2017) 57–84. doi:10.1016/j.jss.2016.09.015.
SS12	S. Segura, G. Fraser, A. B. Sanchez, A. Ruiz-Cortes, A Survey on Metamorphic Testing, IEEE Transactions on Software Engineering 42 (9) (2016) 805–824. doi:10.1109/TSE.2016.2532875.
SS13	T. Sharma, D. Spinellis, A survey on software smells, Journal of Systems and Software 138 (2018) 158–173. doi:10.1016/j.jss.2017.12.034.

Study ID	Reference
SS14	E. Gonçalves, J. Castro, J. Araújo, T. Heineck, A Systematic Literature Review of iStar extensions, Journal of Systems and Software 137 (2018) 1–33. doi:10.1016/j.jss.2017.11.023.
SS15	M. Zarour, A. Abran, J. M. Desharnais, A. Alarifi, An investigation into the best practices for the successful design and implementation of lightweight software process assessment methods: A systematic literature review, Journal of Systems and Software 101 (2015) 180–192. doi:10.1016/j.jss.2014.11.041.
SS16	E. Tüzün, B. Tekinerdogan, Analyzing impact of experience curve on ROI in the software product line adoption process, Information and Software Technology 59 (2015) 136–148. doi:10.1016/j.infsof.2014.09.008.
SS17	G. Lewis, P. Lago, Architectural tactics for cyber-foraging: Results of a systematic literature review, Journal of Systems and Software 107 (2015) 158–186.
SS18	D. Heaton, J. C. Carver, Claims about the use of software engineering practices in science: A systematic literature review, Information and Software Technology 67 (2015) 207–219. doi:10.1016/j.infsof.2015.07.011.
SS19	S. Mahdavi-Hezavehi, V. H. Durelli, D. Weyns, P. Avgeriou, A systematic literature review on methods that handle multiple quality attributes in architecture-based self-adaptive systems, Information and Software Technology 90 (2017) 1–26. doi:10.1016/j.infsof.2017.03.013.
SS20	P. Heck, A. Zaidman, A systematic literature review on quality criteria for agile requirements specifications, Software Quality Journal 26 (1) (2018) 127–160. doi:10.1007/s11219-016-9336-4.
SS21	T. Tahir, G. Rasool, C. Gencel, A systematic literature review on software measurement programs, Information and Software Technology 73 (2016) 101–121. doi:10.1016/j.infsof.2016.01.014.
SS22	J. Zhi, V. Garousi-Yusifoglu, B. Sun, G. Garousi, S. Shahnewaz, G. Ruhe, Cost, benefits and quality of software development documentation: A systematic mapping, Journal of Systems and Software 99 (2015) 175–198. doi:10.1016/j.jss.2014.09.042.
SS23	J. M. Sierra, A. Vizcaíno, M. Genero, M. Piattini, A systematic mapping study about sociotechnical congruence, Information and Software Technology 94 (September) (2018) 111–129. doi:10.1016/j.infsof.2017.10.004.
SS24	S. Zein, N. Salleh, J. Grundy, A systematic mapping study of mobile application testing techniques, Journal of Systems and Software 117 (2016) 334–356.
SS25	J. Kabbedijk, C. P. Bezemer, S. Jansen, A. Zaidman, Defining multi-tenancy: A systematic mapping study on the academic and the industrial perspective, Journal of Systems and Software 100 (2015) 139–148. doi:10.1016/j.jss.2014.10.034.
SS26	E. Tüzün, B. Tekinerdogan, M. E. Kalender, S. Bilgen, Empirical evaluation of a decision support model for adopting software product line engineering, Information and Software Technology 60 (2015) 77–101. doi:10.1016/j.infsof.2014.12.007.
SS27	O. Pedreira, F. García, N. Brisaboa, M. Piattini, Gamification in software engineering - A systematic mapping, Information and Software Technology 57 (1) (2015) 157–168. doi:10.1016/j.infsof.2014.08.007.
SS28	V. Garousi, Y. Amannejad, A. B. Can, Software test-code engineering: A systematic mapping, Information and Software Technology 58 (2015) 123–147. doi:10.1016/j.infsof.2014.06.009.
SS29	R. A. Silva, S. d. R. Senger de Souza, P. S. Lopes de Souza, A systematic review on search based mutation testing, Information and Software Technology 81 (2017) 19–35. doi:10.1016/j.infsof.2016.01.017.

Study ID	Reference
SS30	D. Tosi, S. Morasca, Supporting the semi-automatic semantic annotation of web services: A systematic literature review, Information and Software Technology 61 (2015) 16–32. doi:10.1016/j.infsof.2015.01.007.
SS31	D. Ståhl, K. Hallén, J. Bosch, Achieving traceability in large scale continuous integration and delivery deployment, usage and validation of the eiffel framework, Empirical Software Engineering 22 (3) (2017) 967–995. doi:10.1007/s10664-016-9457-1.
SS32	A. Nguyen-duc, D. S. Cruzes, R. Conradi, The impact of global dispersion on coordination, team performance and software quality – A systematic literature review 57 (2015) 277–294. doi:10.1016/j.infsof.2014.06.002.
SS33	U. Abelein, B. Paech, Understanding the Influence of User Participation and Involvement on System Success – a Systematic Mapping Study, Empirical Software Engineering 20 (1) (2013) 28–81. doi:10.1007/s10664-013-9278-4.
SS34	N. Tripathi, E. Klotins, R. Prikladnicki, M. Oivo, L. B. Pompermaier, A. S. Kudakacheril, M. Unterkalmsteiner, K. Liukkunen, T. Gorschek, An anatomy of requirements engineering in software startups using multi-vocal literature and case survey, Journal of Systems and Software 146 (2018) 130–151. doi:10.1016/j.jss.2018.08.059.
SS35	L. Ochoa, T. Degueule, J. Vinju, An empirical evaluation of OSGi dependencies best practices in the eclipse IDE, Proceedings - International Conference on Software Engineering (2018) 170–180doi:10.1145/3196398.3196416.
SS36	F. Selleri Silva, F. S. F. Soares, A. L. Peres, I. M. D. Azevedo, A. P. L. Vasconcelos, F. K. Kamei, S. R. D. L. Meira, Using CMMI together with agile software development: A systematic review, Information and Software Technology 58 (2015) 20–43. doi:10.1016/j.infsof.2014.09.012.
SS37	D. Salah, R. F. Paige, P. Cairns, A systematic literature review for Agile development processes and user centred design integration, ACM International Conference Proceeding Series (2014). doi:10.1145/2601248.2601276.
SS38	J. E. González, N. Juristo, S. Vegas, A systematic mapping study on testing technique experiments: Has the situation changed since 2000?, International Symposium on Empirical Software Engineering and Measurement (2014) 3–6. doi:10.1145/2652524.2652569.
SS39	M. Leitner, S. Rinderle-Ma, A systematic review on security in Process-Aware Information Systems - Constitution, challenges, and future directions, Information and Software Technology 56 (3) (2014) 273–293. doi:10.1016/j.infsof.2013.12.004.
SS40	F. J. Vasconcellos, G. B. Landre, J. A. O. Cunha, J. L. Oliveira, R. A. Ferreira, A. M. Vincenzi, Approaches to strategic alignment of software process improvement: A systematic literature review, Journal of Systems and Software 123 (2017) 45–63. doi:10.1016/j.jss.2016.09.030.
SS41	I. M. del Águila, J. del Sagrado, Bayesian networks for enhancement of requirements engineering: a literature review, Requirements Engineering 21 (4) (2016) 461–480. doi:10.1007/s00766-015-0225-3.
SS42	M. Irshad, R. Torkar, K. Petersen, W. Afzal, Capturing cost avoidance through reuse: Systematic literature review and industrial evaluation, ACM International Conference Proceeding Series 01-03-June (2016). doi:10.1145/2915970.2915989.
SS43	K. Dikert, M. Paasivaara, C. Lassenius, Challenges and success factors for large-scale agile transformations: A systematic literature review, Journal of Systems and Software 119 (2016) 87–108. doi:10.1016/j.jss.2016.06.013.

Study ID	Reference
SS44	M. Salam, S. U. Khan, Challenges in the development of green and sustainable software for software multisourcing vendors: Findings from a systematic literature review and industrial survey, Journal of Software: Evolution and Process 30 (8) (2018) 19–22. doi:10.1002/smr.1939.
SS45	C. A. González, J. Cabot, Formal verification of static software models in MDE: A systematic review, Information and Software Technology 56 (8) (2014) 821–838. doi:10.1016/j.infsof.2014.03.003.
SS46	S. Matalonga, F. Rodrigues, G. H. Travassos, Characterizing testing methods for contextaware software systems: Results from a quasi-systematic literature review, Journal of Systems and Software 131 (2017) 1–21. doi:10.1016/j.jss.2017.05.048.
SS47	P. Raulamo-Jurvanen, M. Mäntylä, V. Garousi, Choosing the right test automation tool: A grey literature review of practitioner sources, ACM International Conference Proceeding Series Part F128635 (2017) 21–30. doi:10.1145/3084226.3084252.
SS48	D. Ståhl, J. Bosch, Modeling continuous integration practice differences in industry software development, Journal of Systems and Software 87 (1) (2014) 48–59. doi:10.1016/j.jss.2013.08.032.
SS49	F. D. Giraldo, S. España, Ó. Pastor, W. J. Giraldo, Considerations about quality in modeldriven engineering: Current state and challenges, Software Quality Journal 26 (2) (2018) 685–750. doi:10.1007/s11219-016-9350-6.
SS50	I. D. C. Machado, J. D. McGregor, Y. C. Cavalcanti, E. S. De Almeida, On strategies for testing software product lines: A systematic literature review, Information and Software Technology 56 (10) (2014) 1183–1199. doi:10.1016/j.infsof.2014.04.002.
SS51	L. Madeyski, W. Orzeszyna, R. Torkar, M. Jozala, Overcoming the equivalent mutant problem: A systematic literature review and a comparative experiment of second order mutation, IEEE Transactions on Software Engineering 40 (1) (2014) 23–42. doi:10.1109/TSE.2013.44.
SS52	D. Carrizo, O. Dieste, N. Juristo, Contextual attributes impacting the effectiveness of requirements elicitation Techniques: Mapping theoretical and empirical research, Information and Software Technology 92 (2017) 194–221. doi:10.1016/j.infsof.2017.08.003.
SS53	M. Oriol, J. Marco, X. Franch, Quality models for web services: A systematic mapping, Information and Software Technology 56 (10) (2014) 1167–1182. doi:10.1016/j.infsof.2014.03.012.
SS54	P. A. Souza Neto, G. Vargas-Solar, U. S. Da Costa, M. A. Musicante, Designing service-based applications in the presence of non-functional properties: A mapping study, Information and Software Technology 69 (2016) 84–105. doi:10.1016/j.infsof.2015.09.004.
SS55	V. Anu, W. Hu, J. C. Carver, G. S. Walia, G. Bradshaw, Development of a human error taxonomy for software requirements: A systematic literature review, Information and Software Technology 103 (May) (2018) 112–124. doi:10.1016/j.infsof.2018.06.011.
SS56	S. Hosseinzadeh, S. Rauti, S. Laurén, J. M. Mäkelä, J. Holvitie, S. Hyrynsalmi, V. Leppänen, Diversification and obfuscation techniques for software security: A systematic literature review, Information and Software Technology 104 (July) (2018) 72–93. doi:10.1016/j.infsof.2018.07.007.
SS57	U. Kanewala, J. M. Bieman, Testing scientific software: A systematic literature review, Information and Software Technology 56 (10) (2014) 1219–1232. arXiv:1804.01954, doi:10.1016/j.infsof.2014.05.006.
SS58	O. Al-Baik, J. Miller, The kanban approach, between agility and leanness: a systematic review, Empirical Software Engineering 20 (6) (2015) 1861–1897. doi:10.1007/s10664-014-9340-x.

Study ID	Reference
SS59	R. Özakinci, A. Tarhan, Early software defect prediction: A systematic map and review, Journal of Systems and Software 144 (October 2017) (2018) 216–239. doi:10.1016/j.jss.2018.06.025.
SS60	O. S. Gómez, N. Juristo, S. Vegas, Understanding replication of experiments in software engineering: A classification, Information and Software Technology 56 (8) (2014) 1033–1048. doi:10.1016/j.infsof.2014.04.004.
SS61	D. M. Fernández, S. Ognawala, S. Wagner, M. Daneva, Where do we stand in requirements engineering improvement today?: First results from a mapping study, International Symposium on Empirical Software Engineering and Measurement (c) (2014). arXiv:1701.05497, doi:10.1145/2652524.2652555.
SS62	B. Costa, P. F. Pires, F. C. Delicato, P. Merson, Evaluating REST architectures - Approach, tooling and guidelines, Journal of Systems and Software 112 (2016) 156–180. doi:10.1016/j.jss.2015.09.039.
SS63	M. Kuhrmann, D. M. Fernández, M. Tiessler, A mapping study on method engineering - First results, ACM International Conference Proceeding Series (2013) 165–170 doi:10.1145/2460999.2461023.
SS64	V. Cosentino, J. Luis, C. Izquierdo, J. Cabot, Findings from GitHub: Methods, datasets and limitations, Proceedings - 13th Working Conference on Mining Software Repositories, MSR 2016 (2016) 137–141 doi:10.1145/2901739.2901776.
SS65	B. Morschheuser, L. Hassan, K. Werder, J. Hamari, How to design gamification? A method for engineering gamified software, Information and Software Technology 95 (April 2017) (2018) 219–237. doi:10.1016/j.infsof.2017.10.015.
SS66	C. Fernández-Sánchez, J. Garbajosa, A. Yagüe, J. Perez, Identification and analysis of the elements required to manage technical debt by means of a systematic mapping study, Journal of Systems and Software 124 (2017) 22–38. doi:10.1016/j.jss.2016.10.018.
SS67	E. Tom, A. Aurum, R. Vidgen, An exploration of technical debt, Journal of Systems and Software 86 (6) (2013) 1498–1516. doi:10.1016/j.jss.2012.12.052.
SS68	T. N. Ferreira, S. R. Vergilio, J. T. de Souza, Incorporating user preferences in search-based software engineering: A systematic mapping study, Information and Software Technology 90 (2017) 55–69.
SS69	Z. Li, P. Liang, P. Avgeriou, Application of knowledge-based approaches in software architecture: A systematic mapping study, Information and Software Technology 55 (5) (2013) 777–794. doi:10.1016/j.infsof.2012.11.005.
SS70	B. Mohabbati, M. Asadi, D. Gaševic, M. Hatala, H. A. Müller, Combining service-orientation and software product line engineering: A systematic mapping study, Information and Software Technology 55 (11) (2013) 1845–1859. doi:10.1016/j.infsof.2013.05.006.
SS71	K. Rievičs, R. Torkar, Equality in cumulative voting: A systematic review with an improvement proposal, Information and Software Technology 55 (2) (2013) 267–287. doi:10.1016/j.infsof.2012.08.004.
SS72	T. Besker, A. Martini, J. Bosch, Managing architectural technical debt: A unified model and systematic literature review, Journal of Systems and Software 135 (2018) 1–16. doi:10.1016/j.jss.2017.09.025.

Study ID	Reference
SS73	B. Uzun, B. Tekinerdogan, Model-driven architecture based testing: A systematic literature review, Information and Software Technology 102 (July 2017) (2018) 30–48. doi:10.1016/j.infsof.2018.05.004.
SS74	Z. Li, H. Zhang, L. O'Brien, R. Cai, S. Flint, On evaluating commercial Cloud services: A systematic review, Journal of Systems and Software 86 (9) (2013) 2371–2393. arXiv:1708.01412, doi:10.1016/j.jss.2013.04.021.
SS75	D. Rattan, R. Bhatia, M. Singh, Software clone detection: A systematic review, Vol. 55, Elsevier B.V., 2013. doi:10.1016/j.infsof.2013.01.008.
SS76	O. Franco-Bedoya, D. Ameller, D. Costal, X. Franch, Open source software ecosystems: A Systematic mapping, Information and Software Technology 91 (2003) (2017) 160–185. doi:10.1016/j.infsof.2017.07.007.
SS77	D. Radjenovic, M. Hericko, R. Torkar, A. Živkovic, Software fault prediction metrics: A systematic literature review, Information and Software Technology 55 (8) (2013) 1397–1418. doi:10.1016/j.infsof.2013.02.009.
SS78	J. Axelsson, M. Skoglund, Quality assurance in software ecosystems: A systematic literature mapping and research agenda, Journal of Systems and Software 114 (2016) 69–81. doi:10.1016/j.jss.2015.12.020.
SS79	L. Garcés, A. Ampatzoglou, P. Avgeriou, E. Y. Nakagawa, Quality attributes and quality models for ambient assisted living software systems: A systematic mapping, Information and Software Technology 82 (2017) 121–138. doi:10.1016/j.infsof.2016.10.005.
SS80	W. K. Assunção, R. E. Lopez-Herrejon, L. Linsbauer, S. R. Vergilio, A. Egyed, Reengineering legacy applications into software product lines: a systematic mapping, Empirical Software Engineering 22 (6) (2017) 2972–3016. doi:10.1007/s10664-017-9499-z.
SS81	S. T. Acuña, J. W. Castro, O. Dieste, N. Juristo, A systematic mapping study on the open source software development process, IET Seminar Digest 2012 (1) (2012) 42–46. doi:10.1049/ic.2012.0005.
SS82	N. S. M. Yusop, J. Grundy, R. Vasa, Reporting Usability Defects: A Systematic Literature Review, IEEE Transactions on Software Engineering 43 (9) (2017) 848–867. doi:10.1109/TSE.2016.2638427.
SS83	G. Rodríguez-Pérez, G. Robles, J. M. González-Barahona, Reproducibility and credibility in empirical software engineering: A case study based on a systematic literature review of the use of the SZZ algorithm, Information and Software Technology 99 (March) (2018) 164–176. doi:10.1016/j.infsof.2018.03.009.
SS84	G. Holl, P. Grünbacher, R. Rabiser, A systematic review and an expert survey on capabilities supporting multi product lines, Information and Software Technology 54 (8) (2012) 828–852. doi:10.1016/j.infsof.2012.02.002.
SS85	E. Domínguez, B. Pérez, Á. L. Rubio, M. A. Zapata, A systematic review of code generation proposals from state machine specifications, Information and Software Technology 54 (10) (2012) 1045–1066. doi:10.1016/j.infsof.2012.04.008.
SS86	H. P. Breivold, I. Crnkovic, M. Larsson, A systematic review of software architecture evolution research, Information and Software Technology 54 (1) (2012) 16–40. doi:10.1016/j.infsof.2011.06.002.

Study ID	Reference
SS87	B. Wang, R. Peng, Y. Li, H. Lai, Z. Wang, Requirements traceability technologies and technology transfer decision support: A systematic review, Journal of Systems and Software 146 (2018) 59–79. doi:10.1016/j.jss.2018.09.001.
SS88	M. Kalenda, P. Hyna, B. Rossi, Scaling agile in large organizations: Practices, challenges, and success factors, Journal of Software: Evolution and Process 30 (10) (2018) 1–24. doi:10.1002/smr.1954.
SS89	M. El-Attar, J. Miller, Constructing high quality use case models: A systematic review of current practices, Requirements Engineering 17 (3) (2012) 187–201. doi:10.1007/s00766-011-0135-y.
SS90	V. Garousi, B. Küćuk, Smells in software test code: A survey of knowledge in industry and academia, Journal of Systems and Software 138 (2018) 52–81. doi:10.1016/j.jss.2017.12.013.
SS91	A. Ahmad, M. A. Babar, Software architectures for robotic systems: A systematic mapping study, Journal of Systems and Software 122 (2016) 16–39. doi:10.1016/j.jss.2016.08.039.
SS92	L. Kaur, A. Mishra, Software component and the semantic web: An in-depth content analysis and integration history, Journal of Systems and Software 125 (2017) 152 – 169. doi:https://doi.org/10.1016/j.jss.2016.11.028.
SS93	A. L. Mesquida, A. Mas, E. Amengual, J. A. Calvo-Manzano, IT service management process improvement based on ISO/IEC 15504: A systematic review, Information and Software Technology 54 (3) (2012) 239–247. doi:10.1016/j.infsof.2011.11.002.
SS94	M. Riaz, Maintainability prediction of relational database-driven applications: A systematic review, IET Seminar Digest 2012 (1) (2012) 263–272. doi:10.1049/ic.2012.0034.
SS95	I. Santiago, Á. Jiménez, J. M. Vara, V. De Castro, V. A. Bollati, E. Marcos, Model-Driven Engineering as a new landscape for traceability management: A systematic literature review, Information and Software Technology 54 (12) (2012) 1340–1356. doi:10.1016/j.infsof.2012.07.008.
SS96	J. Li, H. Zhang, L. Zhu, R. Jeffery, Q. Wang, M. Li, Preliminary results of a systematic review on requirements evolution, IET Seminar Digest 2012 (1) (2012) 12–21. doi:10.1049/ic.2012.0002.
SS97	V. Berg, J. Birkeland, A. Nguyen-Duc, I. O. Pappas, L. Jaccheri, Software startup engineering: A systematic mapping study, Journal of Systems and Software 144 (February) (2018) 255–274. doi:10.1016/j.jss.2018.06.043.
SS98	V. Garousi, M. Felderer, T. Hacaloglu, Software test maturity assessment and test process improvement: A multivocal literature review, Information and Software Technology 85 (2017) 16–42. doi:10.1016/j.infsof.2017.01.001.
SS99	S. Maro, J. P. Steghöfer, M. Staron, Software traceability in the automotive domain: Challenges and solutions, Journal of Systems and Software 141 (2018) 85–110. doi:10.1016/j.jss.2018.03.060.
SS100	P. R. C. Rolando, O. Dieste, R. F. C. Efraín, N. Juristo, Statistical errors in software engineering experiments: A preliminary literature review, Proceedings - International Conference on Software Engineering (2018) 1195–1206doi:10.1145/3180155.3180161.
SS101	A. A. Khan, J. Keung, M. Niazi, S. Hussain, A. Ahmad, Systematic literature review and empirical investigation of barriers to process improvement in global software development: Client–vendor perspective, Information and Software Technology 87 (2017) 180–205. doi:10.1016/j.infsof.2017.03.006.
SS102	A. Idri, M. Hosni, A. Abran, Systematic literature review of ensemble effort estimation, Journal of Systems and Software 118 (2016) 151–175.

Study ID	Reference
SS103	M. Szvetits, U. Zdun, Systematic literature review of the objectives, techniques, kinds, and architectures of models at runtime, Software and Systems Modeling 15 (1) (2016) 31–69. doi:10.1007/s10270-013-0394-9.
SS104	M. A. P. Araújo, V. F. Monteiro, G. H. Travassos, Towards a model to support in silico studies of software evolution, International Symposium on Empirical Software Engineering and Measurement (2012) 281–289doi:10.1145/2372251.2372303.
SS105	M. Soualhia, F. Khomh, S. Tahar, Task Scheduling in Big Data Platforms: A Systematic Literature Review, Journal of Systems and Software 134 (2017) 170–189. doi:10.1016/j.jss.2017.09.001.
SS106	A. B. Soomro, N. Salleh, E. Mendes, J. Grundy, G. Burch, A. Nordin, The effect of software engineers' personality traits on team climate and performance: A Systematic Literature Review, Information and Software Technology 73 (2016) 52–65. doi:10.1016/j.infsof.2016.01.006.
SS107	P. A. Da Mota Silveira Neto, I. D. Carmo MacHado, J. D. McGregor, E. S. De Almeida, S. R. De Lemos Meira, A systematic mapping study of software product lines testing, Information and Software Technology 53 (5) (2011) 407–423. doi:10.1016/j.infsof.2010.12.003.
SS108	E. Barreiros, A. Almeida, J. Saraiva, S. Soares, A systematic mapping study on software engineering testbeds, International Symposium on Empirical Software Engineering and Measurement (2011) 107–116doi:10.1109/esem.2011.19.
SS109	L. Chen, M. Ali Babar, A systematic review of evaluation of variability management approaches in software product lines, Information and Software Technology 53 (4) (2011) 344–362. doi:10.1016/j.infsof.2010.12.006.
SS110	T. Yue, L. C. Briand, Y. Labiche, A systematic review of transformation approaches between user requirements and analysis models, Requirements Engineering 16 (2) (2011) 75–99. doi:10.1007/s00766-010-0111-y.
SS111	J. F. Bastos, P. A. da Mota Silveira Neto, E. S. de Almeida, S. R. de Lemos Meira, Adopting software product lines: A systematic mapping study, in: 15th Annual Conference on Evaluation Assessment in Software Engineering (EASE 2011), 2011, pp. 11–20.
SS112	V. Garousi, M. Felderer, M. V. Mäntylä, The need for multivocal literature reviews in software engineering: Complementing systematic literature reviews with grey literature, ACM International Conference Proceeding Series 01-03-June (2016). doi:10.1145/2915970.2916008.
SS113	J. Soldani, D. A. Tamburri, W. J. Van Den Heuvel, The pains and gains of microservices: A Systematic grey literature review, Journal of Systems and Software 146 (2018) 215–232. doi:10.1016/j.jss.2018.09.082.
SS114	M. Sulayman, E. Mendes, An extended systematic review of software process improvement in small and medium web companies, IET Seminar Digest 2011 (1) (2011) 134–143. doi:10.1049/ic.2011.0017.
SS115	M. J. Escalona, J. J. Gutierrez, M. Mejías, G. Aragón, I. Ramos, J. Torres, F. J. Domínguez, An overview on test generation from functional requirements, Journal of Systems and Software 84 (8) (2011) 1379–1393. doi:10.1016/j.jss.2011.03.051.
SS116	N. Salleh, E. Mendes, J. C. Grundy, Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review, IEEE Transactions on Software Engineering 37 (4) (2011) 509–525. doi:10.1109/TSE.2010.59.

Study ID	Reference
SS117	S. U. Khan, M. Niazi, R. Ahmad, Factors influencing clients in the selection of offshore software outsourcing vendors: An exploratory study using a systematic literature review, Journal of Systems and Software 84 (4) (2011) 686–699. doi:10.1016/j.jss.2010.12.010.
SS118	N. Kratzke, P. C. Quint, Understanding cloud-native applications after 10 years of cloud computing - A systematic mapping study, Journal of Systems and Software 126 (2017) 1–16. doi:10.1016/j.jss.2017.01.001.
SS119	L. Major, T. Kyriacou, O. P. Brereton, Systematic literature review: Teaching novices programming using robots, IET Seminar Digest 2011 (1) (2011) 21–30. doi:10.1049/ic.2011.0003.
SS120	O. Dieste, N. Juristo, Systematic review and aggregation of empirical studies on elicitation techniques, IEEE Transactions on Software Engineering 37 (2) (2011) 283–304. doi:10.1109/TSE.2010.33.
SS121	M. Palacios, J. García-Fanjul, J. Tuya, Testing in Service Oriented Architectures with dynamic binding: A mapping study, Information and Software Technology 53 (3) (2011) 171–189. doi:10.1016/j.infsof.2010.11.014.
SS122	A. Williams, Using reasoning markers to select the more rigorous software practitioners' online content when searching for grey literature, ACM International Conference Proceeding Series Part F1377 (2018). doi:10.1145/3210459.3210464.
SS123	L. B. Lisboa, V. C. Garcia, D. Lucrédio, E. S. de Almeida, S. R. de Lemos Meira, R. P. de Mattos Fortes, A systematic review of domain analysis tools, Information and Software Technology 52 (1) (2010) 1–13.
SS124	M. Turner, B. Kitchenham, P. Brereton, S. Charters, D. Budgen, Does the technology acceptance model predict actual use? A systematic literature review, Information and Software Technology 52 (5) (2010) 463–479. doi:10.1016/j.infsof.2009.11.005.
SS125	R. Rabiser, P. Grünbacher, D. Dhungana, Requirements for product derivation support: Results from a systematic literature review and an expert survey, Information and Software Technology 52 (3) (2010) 324–346. doi:10.1016/j.infsof.2009.11.001.
SS126	V. Garousi, M. V. Mäntylä, When and what to automate in software testing? A multi-vocal literature review, Information and Software Technology 76 (2016) 92–117. doi:10.1016/j.infsof.2016.04.015.

APPENDIX D – SECONDARY STUDIES INCLUDED IN TERTIARY STUDY 1 (CHARACTERISTICS)

In this appendix, we present the characteristics of the included studies (n=126) related to the Study 3.

Study ID	Year	Source Type	Review Type	Search Method	% of GL
SS1	2018	J	SLR	A	12.9
SS2	2016	C	SLR	A	4.5
SS3	2015	J	SLR	A	2.3
SS4	2016	J	SLR	A + M	6.5
SS5	2015	J	SLR	A + S	12.6
SS6	2018	J	MS	A	4.4
SS7	2015	J	MS	A + M + S	7.8
SS8	2015	J	SLR	A	19.0
SS9	2017	J	SLR	A	23.8
SS10	2017	J	Others	A	5.3
SS11	2017	J	Others	A + M + S	12.4
SS12	2016	J	SLR	A	2.5
SS13	2018	J	MS	A + M	1.3
SS14	2018	J	SLR	A + M	3.6
SS15	2015	J	SLR	A + M	13.8
SS16	2015	J	SLR	A + M	0.0
SS17	2015	J	SLR	A + S	25.9
SS18	2015	J	SLR	A	2.3
SS19	2017	J	SLR	A + M	20.4
SS20	2018	J	SLR	A + M + S	6.3
SS21	2016	J	SLR	A + S	1.6
SS22	2015	J	MS	A + M + S	10.1
SS23	2018	J	MS	A + M + S	7.5
SS24	2016	J	MS	A	2.5
SS25	2015	J	MS	A	68.4
SS26	2015	J	SLR	A + M + S	22.6
SS27	2015	J	MS	A	6.9
SS28	2015	J	MS	A + M	6.7
SS29	2017	J	SLR	A + S	3.1

Source Type: J – Journal; C – Conference.

Review Type: SLR – Systematic Literature Review; MLR – Multivocal Literature Review; MS – Mapping Study; GLR – Grey Literature Review

Search Method: A – Automatic; M – Manual; S - Snowballing.

Study ID	Year	Source Type	Review Type	Search Method	% of GL
SS30	2015	J	SLR	A + M	2.3
SS31	2017	J	SLR	A	20.0
SS32	2015	J	SLR	A + M	0.0
SS33	2015	J	MS	A	4.0
SS34	2018	J	MLR	A	41.7
SS35	2018	C	SLR	A + S	100
SS36	2015	J	SLR	A + M	8.6
SS37	2014	C	SLR	A + M	5.6
SS38	2014	J	MS	A + M	7.1
SS39	2014	J	SLR	A + M + S	1.5
SS40	2017	J	SLR	A + S	3.3
SS41	2016	J	SLR	A	10.0
SS42	2016	C	SLR	A + S	100
SS43	2016	J	SLR	A	5.8
SS44	2018	J	SLR	A + S	1.9
SS45	2014	J	MS	A	4.2
SS46	2017	J	Others	A + M	100
SS47	2017	C	GLR	A	100
SS48	2014	J	SLR	A	2.2
SS49	2018	J	SLR	A + S	30.1
SS50	2014	J	SLR	A + M + S	8.2
SS51	2014	J	SLR	A + M + S	4.5
SS52	2017	J	MS	A + S	5.6
SS53	2014	J	MS	A + M + S	7.7
SS54	2016	J	MS	A + S	0.0
SS55	2018	J	SLR	A	2.6
SS56	2018	J	SLR	A + S	63.2
SS57	2014	J	SLR	A + S	3.2
SS58	2014	J	SLR	A + M	45.9
SS59	2018	J	Others	A + S	7.7
SS60	2014	J	MS	A	4.8
SS61	2014	J	MS	A + S	4.4
SS62	2016	J	MS	A	36.0
SS63	2013	C	MS	A + M + S	1.6

Source Type: J – Journal; C – Conference.

Review Type: SLR – Systematic Literature Review; MLR – Multivocal Literature Review; MS – Mapping Study; GLR – Grey Literature Review

Search Method: A – Automatic; M – Manual; S - Snowballing.

Study ID	Year	Source Type	Review Type	Search Method	% of GL
SS64	2016	C	MS	A + M + S	7.5
SS65	2018	J	SLR	A + S	0.0
SS66	2017	J	MS	A + S	7.9
SS67	2013	J	MLR	A	39.1
SS68	2017	J	MS	A + S	7.5
SS69	2013	J	MS	A + M	1.8
SS70	2013	J	MS	A + M	12.3
SS71	2013	J	SLR	A + S	5.0
SS72	2018	J	SLR	A + S	4.8
SS73	2018	J	SLR	A + M	9.7
SS74	2013	J	SLR	A + S	8.5
SS75	2013	J	SLR	A + M	0.0
SS76	2017	J	MS	A + M	11.5
SS77	2013	J	SLR	A + S	15.1
SS78	2016	J	MS	A + M	16.7
SS79	2017	J	MS	A + S	3.7
SS80	2017	J	MS	A + S	7.6
SS81	2012	C	MS	A	27.6
SS82	2017	J	SLR	A + S	1.8
SS83	2018	J	SLR	A	13.9
SS84	2012	J	SLR	A + S	2.7
SS85	2012	J	SLR	A + M	18.9
SS86	2012	J	SLR	A	6.1
SS87	2018	J	SLR	A + S	0.9
SS88	2018	J	MS	A	0.0
SS89	2012	J	SLR	A	17.9
SS90	2018	J	MLR	A + S	80.1
SS91	2016	J	MS	A	3.6
SS92	2017	J	SLR	A + S	2.1
SS93	2012	J	SLR	A	3.6
SS94	2012	C	SLR	A + S	0.0
SS95	2012	J	SLR	A	13.8
SS96	2012	C	SLR	A + M	0.0
SS97	2018	J	MS	A + M + S	7.4

Source Type: J – Journal; C – Conference.

Review Type: SLR – Systematic Literature Review; MLR – Multivocal Literature Review; MS – Mapping Study; GLR – Grey Literature Review

Search Method: A – Automatic; M – Manual; S - Snowballing.

Study ID	Year	Source Type	Review Type	Search Method	% of GL
SS98	2017	J	MLR	A	28.2
SS99	2018	J	MLR	A + S	78.8
SS100	2018	C	Others	M	12.5
SS101	2017	J	SLR	A	7.1
SS102	2016	J	SLR	A	4.2
SS103	2016	J	SLR	A + S	0.8
SS104	2012	J	Others	A	9.1
SS105	2017	J	MS	A + M + S	0.3
SS106	2016	J	SLR	A	5.7
SS107	2011	J	MS	A + S	4.4
SS108	2011	J	MS	A + M	15.4
SS109	2011	J	SLR	A + M	1.0
SS110	2011	J	SLR	A + M	5.0
SS111	2011	C	MS	A + M + S	8.8
SS112	2016	C	MLR	A	28.2
SS113	2018	J	GLR	A	100
SS114	2011	C	SLR	A	12.5
SS115	2011	J	Others	A + S	18.8
SS116	2011	J	SLR	A	6.8
SS117	2011	J	SLR	A	13.9
SS118	2017	J	MS	A	40.8
SS119	2011	C	SLR	A + M	2.9
SS120	2011	J	SLR	A + M + S	3.8
SS121	2011	J	MS	A + M	6.1
SS122	2018	C	GLR	A	100
SS123	2010	J	MS	A + M + S	31.6
SS124	2010	J	SLR	A	1.4
SS125	2010	J	MS	A + M + S	2.5
SS126	2016	J	MLR	A	66.7

Source Type: J – Journal; C – Conference.

Review Type: SLR – Systematic Literature Review; MLR – Multivocal Literature Review; MS – Mapping Study; GLR – Grey Literature Review

Search Method: A – Automatic; M – Manual; S - Snowballing.

APPENDIX E – SECONDARY STUDIES NOT INCLUDED IN TERTIARY STUDY 2 (REFERENCES)

In this appendix, we present the bibliographic information about the not included studies (n=122) related to the Study 3. Although these studies did not include Grey Literature in their datasets, they presented a Grey Literature definition or justified at some level the reasons not to include it in the research, which served as a basis to partially answer the research questions RQ2.1 and RQ2.5.

Study ID	Reference
RQSS01	V. Garousi, K. Petersen, B. Ozkan, Challenges and best practices in industry-academia collaborations in software engineering: A systematic literature review, Information and Software Technology 79 (September 2014) (2016) 106–127. doi:10.1016/j.infsof.2016.07.006.
RQS002	E. N. Alkhanak, S. P. Lee, R. Rezaei, R. M. Parizi, Cost optimization approaches for scientific workflow scheduling in cloud and grid computing: A review, classifications, and open issues, Journal of Systems and Software 113 (2016) 1–26. doi:10.1016/j.jss.2015.11.023.
RQSS03	S. Sobernig, B. Hoisl, M. Strembeck, Extracting reusable design decisions for UML-based domain-specific languages: A multi-method study, Vol. 113, Elsevier Inc., 2016. doi:10.1016/j.jss.2015.11.037.
RQSS04	L. Montalvillo, O. Díaz, Requirement-driven evolution in software product lines: A systematic mapping study, Journal of Systems and Software 122 (2016) 110–143. doi:10.1016/j.jss.2016.08.053.
RQSS05	L. E. G. Martins, T. Gorschek, Requirements engineering for safety-critical systems: A systematic literature review, Information and Software Technology 75 (2016) 71–89. doi:10.1016/j.infsof.2016.04.002.
RQSS06	K. Tuma, G. Calikli, R. Scandariato, Threat analysis of software systems: A systematic literature review, Journal of Systems and Software 144 (February) (2018) 275–294. doi:10.1016/j.jss.2018.06.073.
RQSS07	D. Tofan, M. Galster, P. Avgeriou, W. Schuitema, Past and future of software architectural decisions - A systematic mapping study, Information and Software Technology 56 (8) (2014) 850–872. doi:10.1016/j.infsof.2014.03.009.
RQSS08	S. Stevanetic, U. Zdun, Software metrics for measuring the understandability of architectural structures - A systematic mapping study, ACM International Conference Proceeding Series 27-29-Apri(2015). doi:10.1145/2745802.2745822.
RQSS09	D. Torre, Y. Labiche, M. Genero, M. Elaasar, A systematic identification of consistency rules for UML diagrams, Journal of Systems and Software 144 (October 2017) (2018) 121–142. doi:10.1016/j.jss.2018.06.029.
RQSS10	S. Mahdavi-Hezavehi, M. Galster, P. Avgeriou, Variability in quality attributes of service-based software systems: A systematic literature review, Information and Software Technology 55 (2)(2013) 320–343. doi:10.1016/j.infsof.2012.08.010.
RQSS11	E. M. Arvanitou, A. Ampatzoglou, A. Chatzigeorgiou, M. Galster, P. Avgeriou, A mapping study on design-time quality attributes and metrics, Journal of Systems and Software 127 (2017) 52–77. doi:10.1016/j.jss.2017.01.026.
RQSS12	D. Torre, Y. Labiche, M. Genero, UML consistency rules: A systematic mapping study, ACM International Conference Proceeding Series (2014). doi:10.1145/2601248.2601292.

Study ID	Reference
RQSS13	M. Galster, D. Weyns, D. Tofan, B. Michalik, P. Avgeriou, Variability in software systems-A systematic literature review, IEEE Transactions on Software Engineering 40 (3) (2014) 282–306. doi:10.1109/TSE.2013.56.
RQSS14	S. Lane, I. Richardson, Process models for service-based applications: A systematic literature review, Information and Software Technology 53 (5) (2011) 424–439. doi:10.1016/j.infsof.2010.12.005.
RQSS15	K. Petersen, Measuring and predicting software productivity: A systematic map and review, Information and Software Technology53 (4) (2011) 317–343. doi:10.1016/j.infsof.2010.12.001.
RQSS16	C. Jia, Y. Cai, Y. T. Yu, T. H. Tse, 5W+1H pattern: A perspective of systematic mapping studies and a case study on cloud software testing, Journal of Systems and Software 116 (2016) 206–219. doi:10.1016/j.jss.2015.01.058.
RQSS17	P. Lenberg, R. Feldt, L. G. Wallgren, Behavioral software engineering: A definition and systematic literature review, Journal of Systems and Software 107 (2015) 15–37. doi:10.1016/j.jss.2015.04.084.
RQSS18	R. E. Lopez-Herrejon, S. Illescas, A. Egyed, A systematic mapping study of information visualization for software product line engineering, Journal of Software: Evolution and Process 30 (2) (2018) 1–18. doi:10.1002/smr.1912.
RQSS19	W. N. Behutiye, P. Rodríguez, M. Oivo, A. Tosun, Analyzing the concept of technical debt in the context of agile software development: A systematic literature review, Information and Software Technology 82 (2017) 139–158. doi:10.1016/j.infsof.2016.10.004.
RQSS20	P. Diebold, M. Dahlem, Agile practices in practice - A mapping study, ACM International Conference Proceeding Series (2014). doi:10.1145/2601248.2601254.
RQSS21	R. Verdecchia, I. Malavolta, P. Lago, Architectural technical debt identification: The research landscape, Proceedings - International Conference on Software Engineering (2018) 11–20. doi:10.1145/3194164.3194176.
RQSS22	J. Vilela, J. Castro, L. E. G. Martins, T. Gorschek, Integration between requirements engineering and safety analysis: A systematic literature review, Journal of Systems and Software 125 (2017) 68–92. doi:10.1016/j.jss.2016.11.031.
RQSS23	A. Mehmood, D. N. A. Jawawi, Aspect-oriented model-driven code generation: A systematic mapping study, Information and Software Technology 55 (2) (2013) 395–411. doi:10.1016/j.infsof.2012.09.003.
RQSS24	J. Wen, S. Li, Z. Lin, Y. Hu, C. Huang, Systematic literature review of machine learning based software development effort estimation models, Information and Software Technology 54 (1) (2012)41–59. doi:10.1016/j.infsof.2011.09.002.
RQSS25	R. Wendler, The maturity of maturity model research: A systematic mapping study, Information and Software Technology54 (12) (2012) 1317–1339. doi:10.1016/j.infsof.2012.07.007.
RQSS26	E. Engström, P. Runeson, Software product line testing - A systematic mapping study, Information and Software Technology 53 (1)(2011) 2–13. doi:10.1016/j.infsof.2010.05.011.
RQSS27	H. reza Bazi, A. Hassanzadeh, A. Moeini, A comprehensive framework for cloud computing migration using Meta-synthesis approach, Journal of Systems and Software 128 (2017) 87–105. doi:10.1016/j.jss.2017.02.049.

Study ID	Reference
RQSS28	C. Carroll, D. Falessi, V. Forney, A. Frances, C. Izurieta,C. Seaman, A Mapping Study of Software Causal Factors for Improving Maintenance, International Symposium on Empirical Software Engineering and Measurement 2015-Novem (805) (2015) 235–238. doi:10.1109/ESEM.2015.7321183.
RQSS29	Z. Sharafi, Z. Soh, Y. G. Guéhéneuc, A systematic literature review on the usage of eye-tracking in software engineering, Information and Software Technology 67 (2015) 79–107. doi:10.1016/j.infsof.2015.06.008.
RQSS30	Z. Li, P. Avgeriou, P. Liang, A systematic mapping study on technical debt and its management, Journal of Systems and Software101 (2015) 193–220. doi:10.1016/j.jss.2014.12.027.
RQSS31	M. Bano, D. Zowghi, A systematic review on the relationship between user involvement and system success, Information and Software Technology 58 (2015) 148–169. doi:10.1016/j.infsof.2014.06.011.
RQSS32	A. S. Campanelli, F. S. Parreiras, Agile methods tailoring - A systematic literature review, Journal of Systems and Software 110(2015) 85–100. doi:10.1016/j.jss.2015.08.035.
RQSS33	M. Irshad, K. Petersen, S. Poulding, A systematic literature review of software requirements reuse approaches, Information and Software Technology 93 (September 2017) (2018) 223–245. doi:10.1016/j.infsof.2017.09.009.
RQSS34	I. Hydara, A. B. M. Sultan, H. Zulzalil, N. Admodisastro, Current state of research on cross-site scripting (XSS) - A systematic literature review, Information and Software Technology 58 (2015) 170–186. doi:10.1016/j.infsof.2014.07.010.
RQSS35	B. J. Da Silva Estácio, R. Prikladnicki, Distributed pair programming: A systematic literature review, Information and Software Technology 63 (2015) 1–10. doi:10.1016/j.infsof.2015.02.011.
RQSS36	C. Yang, P. Liang, P. Avgeriou, A systematic mapping study on the combination of software architecture and agile development,Journal of Systems and Software 111 (2016) 157–184. doi:10.1016/j.jss.2015.09.028.
RQSS37	M. Riaz, T. Breaux, L. Williams, How have we evaluated software pattern application? A systematic mapping study of research design practices, Information and Software Technology 65 (2015)14–38. doi:10.1016/j.infsof.2015.04.002.
RQSS38	S. Mahmood, S. Anwer, M. Niazi, M. Alshayeb, I. Richardson, Identifying the factors that influence task allocation in global software development: Preliminary results, ACM International Conference Proceeding Series 27-29-Apri (2015). doi:10.1145/2745802.2745831.
RQSS39	É. F. De Souza, R. D. A. Falbo, N. L. Vijaykumar, Knowledge management initiatives in software testing: A mapping study, Information and Software Technology 57 (1) (2015) 378–391. doi:10.1016/j.infsof.2014.05.016.
RQSS40	S. Jayatilleke, R. Lai, A systematic review of requirements change management, Information and Software Technology 93(2018) 163–185. doi:10.1016/j.infsof.2017.09.004.
RQSS41	A. Abdelmaboud, D. N. Jawawi, I. Ghani, A. Elsafi, B. Kitchenham, Quality of service approaches in cloud computing: A systematic mapping study, Journal of Systems and Software 101 (2015) 159–179. doi:10.1016/j.jss.2014.12.015.
RQSS42	T. Mariani, S. R. Vergilio, A systematic review on search-based refactoring, Information and Software Technology 83 (2017) 14–34. doi:10.1016/j.infsof.2016.11.009.

Study ID	Reference
RQSS43	A. R. Santos, R. P. De Oliveira, E. S. De Almeida, Strategies for consistency checking on software product lines: A mapping study, ACM International Conference Proceeding Series 27-29-Apri (Section 3) (2015). doi:10.1145/2745802.2745806.
RQSS44	J. A. M. Santos, J. B. Rocha-Junior, L. C. L. Prates, R. S.do Nascimento, M. F. Freitas, M. G. de Mendonça, A systematic review on the code smell effect, Journal of Systems and Software144 (March) (2018) 450–477. doi:10.1016/j.jss.2018.07.035.
RQSS45	A. Silva, T. Araújo, J. Nunes, M. Perkusich, E. Dilorenzo, H. Almeida, A. Perkusich, A systematic review on the use of Definition of Done on agile software development projects, ACM International Conference Proceeding Series Part F1286 (2017) 364–373. doi:10.1145/3084226.3084262.
RQSS46	A. Sadeghi, H. Bagheri, J. Garcia, S. Malek, A Taxonomy and Qualitative Comparison of Program Analysis Techniques for Security Assessment of Android Software, IEEE Transactions on Software Engineering 43 (6) (2017) 492–530. doi:10.1109/TSE.2016.2615307.
RQSS47	A. Ampatzoglou, A. Ampatzoglou, A. Chatzigeorgiou, P. Avgeriou, The financial aspect of managing technical debt: A systematic literature review, in: Information and Software Technology, Vol. 64,Elsevier B.V., 2015, pp. 52–73. doi:10.1016/j.infsof.2015.04.001.
RQSS48	M. Younas, D. N. Jawawi, I. Ghani, T. Fries, R. Kazmi, Agile development in the cloud computing environment: A systematic review, Information and Software Technology 103 (December 2017)(2018) 142–158. doi:10.1016/j.infsof.2018.06.014.
RQSS49	C. Ayora, V. Torres, B. Weber, M. Reichert, V. Pelechano, VIVACE: A framework for the systematic evaluation of variability support in process-aware information systems, Information and Software Technology 57 (1) (2015) 248–276. doi:10.1016/j.infsof.2014.05.009.
RQSS50	P. Achimugu, A. Selamat, R. Ibrahim, M. N. R. Mahrin, A systematic literature review of software requirements prioritization research, Information and Software Technology 56 (6) (2014) 568–585. doi:10.1016/j.infsof.2014.02.001.
RQSS51	M. A. Javed, U. Zdun, A systematic literature review of traceability approaches between software architecture and source code, ACM International Conference Proceeding Series (2014). doi:10.1145/2601248.2601278.
RQSS52	N. B. Ali, K. Petersen, C. Wohlin, A systematic literature review on the industrial use of software process simulation, Journal of Systems and Software 97 (2014) 65–85. doi:10.1016/j.jss.2014.06.059.
RQSS53	A. Santos, O. S. Gomez, N. Juristo, Analyzing Families of Experiments in SE: a Systematic Mapping Study, IEEE Transactions on Software Engineering (July) (2018) 1–18. arXiv:1805.09009,doi:10.1109/TSE.2018.2864633.
RQSS54	M. Shahin, P. Liang, M. A. Babar, A systematic review of software architecture visualization techniques, Journal of Systems and Software 94 (2014) 161–185. doi:10.1016/j.jss.2014.03.071.
RQSS55	S. Nair, J. L. De La Vara, M. Sabetzadeh, L. Briand, An extended systematic literature review on provision of evidence for safety certification, Information and Software Technology 56 (7)(2014) 689–717. doi:10.1016/j.infsof.2014.03.001.
RQSS56	S. W. Chuang, T. Luor, H. P. Lu, Assessment of institutions,scholars, and contributions on agile software development (2001-2012), Journal of Systems and Software 93 (2014) 84–101. doi:10.1016/j.jss.2014.03.006.

Study ID	Reference
RQSS57	C. Yang, P. Liang, P. Avgeriou, Assumptions and their management in software development: A systematic mapping study, Information and Software Technology 94 (February) (2018) 82–110. doi:10.1016/j.infsof.2017.10.003.
RQSS58	A. Haghighatkhah, A. Banijamali, O. P. Pakanen, M. Oivo,P. Kuvaja, Automotive software engineering: A systematic mapping study, Journal of Systems and Software 128 (2017) 25–55. doi:10.1016/j.jss.2017.03.005.
RQSS59	H. Munir, M. Moayyed, K. Petersen, Considering rigor and relevance when evaluating test driven development: A systematic review, Information and Software Technology 56 (4) (2014) 375–394. doi:10.1016/j.infsof.2014.01.002.
RQSS60	A. Tarhan, O. Turetken, H. A. Reijers, Business process maturity models: A systematic literature review, Information and Software Technology 75 (2016) 122–134. doi:10.1016/j.infsof.2016.01.010.
RQSS61	I. ul Hassan, N. Ahmad, B. Zuhaira, Calculating completeness of software project scope definition, Information and Software Technology 94 (October 2017) (2018) 208–233. doi:10.1016/j.infsof.2017.10.010.
RQSS62	D. Landman, A. Serebrenik, J. J. Vinju, Challenges for static analysis of Java reflection-literature review and empirical study, Proceedings - 2017 IEEE/ACM 39th International Conference on Software Engineering, ICSE 2017 (2017) 507–518. doi:10.1109/ICSE.2017.53.
RQSS63	N. Taušan, J. Markkula, P. Kuvaja, M. Oivo, Choreography in the embedded systems domain: A systematic literature review, Information and Software Technology 91 (2017) 82–101. doi:10.1016/j.infsof.2017.06.008.
RQSS64	M. Franzago, D. D. Ruscio, I. Malavolta, H. MucCini, Collaborative model-driven software engineering: A classification framework and a research map, IEEE Transactions on Software Engineering 44 (12) (2018) 1146–1175. doi:10.1109/TSE.2017.2755039.
RQSS65	P. Rodríguez, A. Haghighatkhah, L. E. Lwakatare, S. Teppola,T. Suomalainen, J. Eskeli, T. Karvonen, P. Kuvaja, J. M. Verner, M. Oivo, Continuous deployment of software intensive products and services: A systematic mapping study, Journal of Systems and Software 123 (2017) 263–291. doi:10.1016/j.jss.2015.12.015.
RQSS66	R. Ros, P. Runeson, Continuous experimentation and A/B testing: A mapping study, Proceedings - International Conference on Software Engineering (2018) 35–41. doi:10.1145/3194760.3194766.
RQSS67	S. Meldrum, S. A. Licorish, B. T. R. Savarimuthu, Crowdsourced knowledge on stack overflow: A systematic mapping study, ACM International Conference Proceeding Series Part F1286 (June)(2017) 180–185. doi:10.1145/3084226.3084267.
RQSS68	N. Paternoster, C. Giardino, M. Unterkalmsteiner, T. Gorschek, P. Abrahamsson, Software development in startup companies: A systematic mapping study, Information and Software Technology 56 (10) (2014) 1200–1218. doi:10.1016/j.infsof.2014.04.014.
RQSS69	L. García-Borgoñón, M. A. Barcelona, J. A. García-García, M. Alba, M. J. Escalona, Software process modeling languages: A systematic literature review, Information and Software Technology56 (2) (2014) 103–116. doi:10.1016/j.infsof.2013.10.001.
RQSS70	T. Kosar, S. Bohra, M. Mernik, Domain-Specific Languages: A systematic Mapping Study, Information and Software Technology71 (2016) 77–91. doi:10.1016/j.infsof.2015.11.001.

Study ID	Reference
RQSS71	T. Ambreen, N. Ikram, M. Usman, M. Niazi, Empirical research in requirements engineering: trends and opportunities,Requirements Engineering 23 (1) (2018) 63–95. doi:10.1007/s00766-016-0258-2.
RQSS72	S. Kirbas, T. Hall, A. Sen, Evolutionary coupling measurement:Making sense of the current chaos, Science of Computer Programming 135 (2017) 4–19. doi:10.1016/j.scico.2016.10.003.
RQSS73	A. Dikici, O. Turetken, O. Demirors, Factors influencing the understandability of process models: A systematic literature review,Information and Software Technology 93 (2018) 112–129. doi:10.1016/j.infsof.2017.09.001.
RQSS74	V. H. S. Durelli, R. F. Araujo, M. A. G. Silva, R. A. P. D. Oliveira, J. C. Maldonado, M. E. Delamaro, A scoping study on the 25 years of research into software testing in Brazil and an outlook on the future of the area, Journal of Systems and Software 86 (4) (2013)934–950. doi:10.1016/j.jss.2012.10.012.
RQSS75	M. A. Laguna, Y. Crespo, A systematic mapping study on software product line evolution: From legacy system reengineering to product line refactoring, Science of Computer Programming 78 (8)(2013) 1010–1034. doi:10.1016/j.scico.2012.05.003.
RQSS76	A. S. Nascimento, C. M. Rubira, R. Burrows, F. Castor, A systematic review of design diversity-based solutions for fault-tolerant SOAs, ACM International Conference Proceeding Series (2013) 107–118. doi:10.1145/2460999.2461015.
RQSS77	A. Kasoju, K. Petersen, M. V. Mäntylä, Analyzing an automotive testing process with evidence-based software engineering,Information and Software Technology 55 (7) (2013) 1237–1259. doi:10.1016/j.infsof.2013.01.005.
RQSS78	H. Edison, X. Wang, R. Jabangwe, P. Abrahamsson, Innovation Initiatives in Large Software Companies: A Systematic Mapping Study, Information and Software Technology 95 (February 2018)(2018) 1–14.arXiv:1802.05951,doi:10.1016/j.infsof.2017.12.007.
RQSS79	A. M. Fernández-Sáez, M. Genero, M. R. Chaudron, Empirical studies concerning the maintenance of UML diagrams and their use in the maintenance of code: A systematic mapping study, Information and Software Technology 55 (7) (2013) 1119–1142. doi:10.1016/j.infsof.2012.12.006.
RQSS80	H. Wu, L. Shi, C. Chen, Q. Wang, B. Boehm, Maintenance effort estimation for open source software: A systematic literature review, Proceedings - 2016 IEEE International Conference on Software Maintenance and Evolution, ICSME 2016 (2017) 32–43. doi:10.1109/ICSME.2016.87.
RQSS81	M. Staron, W. Meding, MeSRAM - A method for assessing robustness of measurement programs in large software development organizations and its industrial evaluation, Journal of Systems and Software 113 (2016) 76–100. doi:10.1016/j.jss.2015.10.051.
RQSS82	P. H. Nguyen, S. Ali, T. Yue, Model-based security engineering for cyber-physical systems: A systematic mapping study, Information and Software Technology 83 (2017) 116–135. doi:10.1016/j.infsof.2016.11.004.
RQSS83	H. G. Gurbuz, B. Tekinerdogan, Model-based testing for software safety: a systematic mapping study, Software Quality Journal26 (4) (2018) 1327–1372. doi:10.1007/s11219-017-9386-2.
RQSS84	M. Niazi, S. Mahmood, M. Alshayeb, A. A. B. Baqais, A. Q.Gill, Motivators for adopting social computing in global software development: An empirical study, Journal of Software: Evolution and Process 29 (8) (2017). doi:10.1002/smr.1872.

Study ID	Reference
RQSS85	A. Aleti, B. Buhnova, L. Grunske, A. Koziolek, I. Meedeniya,Software architecture optimization methods: A systematic literature review, Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI) P227 (5) (2014) 77–78.
RQSS86	A. Tarhan, G. Giray, On the use of ontologies in software process assessment: A systematic literature review, ACM International Conference Proceeding Series Part F1286 (2017) 2–11. doi:10.1145/3084226.3084261.
RQSS87	C. Roda, E. Navarro, U. Zdun, V. López-Jaquero, G. Simhandl,Past and future of software architectures for context-aware systems:A systematic mapping study, Journal of Systems and Software 146(2018) 310–355. doi:10.1016/j.jss.2018.09.074.
RQSS88	R. D. Santos Rocha, M. Fantinato, The use of software product lines for business process management: A systematic literature review, Information and Software Technology 55 (8) (2013) 1355–1373. doi:10.1016/j.infsof.2013.02.007.
RQSS89	E. Laukkanen, J. Itkonen, C. Lassenius, Problems, causes and solutions when adopting continuous delivery—A systematic literature review, Information and Software Technology 82 (2017) 55–79. doi:10.1016/j.infsof.2016.10.001.
RQSS90	H. Washizaki, Y. G. Guéhéneuc, F. Khomh, ProMeTA: a taxonomy for program metamodels in program reverse engineering,Empirical Software Engineering 23 (4) (2018) 2323–2358. doi:10.1007/s10664-017-9592-3.
RQSS91	G. White, V. Nallur, S. Clarke, Quality of service approaches in IoT: A systematic mapping, Journal of Systems and Software 132(2017) 186–203. doi:10.1016/j.jss.2017.05.125.
RQSS92	M. Anjum, D. Budgen, A mapping study of the definitions for service oriented architecture, IET Seminar Digest 2012 (1) (2012)57–61. doi:10.1049/ic.2012.0008.
RQSS93	C. Pacheco, I. Garcia, A systematic literature review of stakeholder identification methods in requirements elicitation, Journal of Systems and Software 85 (9) (2012) 2171–2181. doi:10.1016/j.jss.2012.04.075.
RQSS94	Z. S. H. Abad, G. Ruhe, M. Noaeen, Requirements Engineering Visualization: A Systematic Literature Review, Proceedings - 2016 IEEE 24th International Requirements Engineering Conference, RE2016 (September) (2016) 6–15. doi:10.1109/RE.2016.61.
RQSS95	S. Sepúlveda, A. Cravero, C. Cachero, Requirements modeling languages for software product lines: A systematic literature review, Information and Software Technology 69 (2016) 16–36. doi:10.1016/j.infsof.2015.08.007.
RQSS96	M. Vierhauser, R. Rabiser, P. Grünbacher, Requirements monitoring frameworks: A systematic review, Information and Software Technology 80 (2016) 89–109. doi:10.1016/j.infsof.2016.08.005.
RQSS97	A. Fernandez, S. Abrahão, E. Insfran, A systematic review on the effectiveness of web usability evaluation methods, IET SeminarDigest 2012 (1) (2012) 52–56. doi:10.1049/ic.2012.0007.
RQSS98	R. Weinreich, I. Groher, Software architecture knowledge management approaches and their support for knowledge management activities: A systematic literature review, Information and Software Technology 80 (2016) 265–286. doi:10.1016/j.infsof.2016.09.007.
RQSS99	M. Unterkalmsteiner, T. Gorschek, A. K. Islam, C. K. Cheng,R. B. Permadi, R. Feldt, Evaluation and measurement of software process improvement - A systematic literature review, IEEE Transactions on Software Engineering 38 (2) (2012) 398–424. doi:10.1109/TSE.2011.26.

Study ID	Reference
RQSS100	K. Alkharabsheh, Y. Crespo, E. Manso, J. A. Taboada, Software Design Smell Detection: a systematic mapping study, Software Quality Journal 27 (3) (2019) 1069–1148. doi:10.1007/s11219-018-9424-8.
RQSS101	R. Jabangwe, H. Edison, A. N. Duc, Software engineering process models for mobile app development: A systematic literature review, Journal of Systems and Software 145 (July) (2018) 98–111. doi:10.1016/j.jss.2018.08.028.
RQSS102	B. Kitchenham, P. Brereton, D. Budgen, Mapping study completeness and reliability - A case study, IET Seminar Digest 2012 (1)(2012) 126–135. doi:10.1049/ic.2012.0016.
RQSS103	T. Vale, E. S. de Almeida, V. Alves, U. Kulesza, N. Niu,R. de Lima, Software product lines traceability: A systematic mapping study, Information and Software Technology 84 (2017) 1–18. doi:10.1016/j.infsof.2016.12.004.
RQSS104	F. Febrero, C. Calero, M. Á. Moraga, Software reliability modeling based on ISO/IEC SQuaRE, Information and Software Technology 70 (2016) 18–29. doi:10.1016/j.infsof.2015.09.006.
RQSS105	W. Afzal, S. Alone, K. Glocksien, R. Torkar, Software test process improvement approaches: A systematic literature review and an industrial case study, Journal of Systems and Software 111 (2016)1–33. doi:10.1016/j.jss.2015.08.048.
RQSS106	Z. Li, H. Zhang, L. O'Brien, S. Jiang, Y. Zhou, M. Kihl,R. Ranjan, Spot pricing in the Cloud ecosystem: A comparative investigation, Journal of Systems and Software 114 (2016) 1–19.arXiv:1708.01401, doi:10.1016/j.jss.2015.10.042.
RQSS107	F. Hujainah, R. B. Abu Bakar, B. Al-haimi, M. A. Abdul-gabber, Stakeholder quantification and prioritisation research: A systematic literature review, Information and Software Technology 102 (May) (2018) 85–99. doi:10.1016/j.infsof.2018.05.008.
RQSS108	L. Li, T. F. Bissyandé, M. Papadakis, S. Rasthofer, A. Bartel,D. Octeau, J. Klein, L. Traon, Static analysis of android apps: A systematic literature review, Information and Software Technology88 (2017) 67–95. doi:10.1016/j.infsof.2017.04.001.
RQSS109	D. Falessi, W. Smith, A. Serebrenik, STRESS: A Semi-Automated, Fully Replicable Approach for Project Selection, International Symposium on Empirical Software Engineering and Measurement 2017-Novem (2017) 151–156. doi:10.1109/ESEM.2017.22.
RQSS110	R. Vallon, B. J. da Silva Estácio, R. Prikladnicki,T. Grechenig, Systematic literature review on agile practices in global software development, Information and Software Technology 96 (December 2017) (2018) 161–180. doi:10.1016/j.infsof.2017.12.004.
RQSS111	R. O. Spínola, G. H. Travassos, Towards a framework to characterize ubiquitous software projects, Information and Software Technology 54 (7) (2012) 759–785. doi:10.1016/j.infsof.2012.01.009.
RQSS112	M. Usman, R. Britto, J. Börstler, E. Mendes, Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method, Information and Software Technology 85 (2017) 43–59. doi:10.1016/j.infsof.2017.01.006.
RQSS113	M. Khatibsyarbini, M. A. Isa, D. N. Jawawi, R. Tumeng, Test case prioritization approaches in regression testing: A systematic literature review, Information and Software Technology 93 (2018)74–93. doi:10.1016/j.infsof.2017.08.014.
RQSS114	V. Garousi, M. Felderer, Ç. M. Karapıçak, U. Yılmaz, Testing embedded software: A survey of the literature, Information and Software Technology 104 (May) (2018) 14–45. doi:10.1016/j.infsof.2018.06.016.

Study ID	Reference
RQSS115	T. B. Callo Arias, P. Van Der Spek, P. Avgeriou, A practice-driven systematic review of dependency analysis solutions, Empirical Software Engineering 16 (5) (2011) 544–586. doi:10.1007/s10664-011-9158-8.
RQSS116	B. Bafandeh Mayvan, A. Rasoolzadegan, Z. Ghavidel Yazdi,The state of the art on design patterns: A systematic mapping of the literature, Journal of Systems and Software 125 (2017) 1339–1351. doi:10.1016/j.jss.2016.11.030.
RQSS117	A. H. Ghapanchi, A. Aurum, Antecedents to IT personnel's intentions to leave: A systematic literature review, Journal of Systems and Software 84 (2) (2011) 238–249. doi:10.1016/j.jss.2010.09.022.
RQSS118	R. Jabbari, N. bin Ali, K. Petersen, B. Tanveer, Towards a benefits dependency network for DevOps based on a systematic literature review, Journal of Software: Evolution and Process 30 (11)(2018) 1–26. doi:10.1002/smr.1957.
RQSS119	C. Becker, R. Chitchyan, S. Betz, C. McCord, Trade-off decisions across time in technical debt management: A systematic literature review, Proceedings - International Conference on Software Engineering (2018) 85–94 doi:10.1145/3194164.3194171.
RQSS120	T. Vale, I. Crnkovic, E. S. De Almeida, P. A. D. M. Silveira Neto, Y. C. Cavalcanti, S. R. D. L. Meira, Twenty-eight years of component-based software engineering, Journal of Systems and Software 111 (2016) 128–148. doi:10.1016/j.jss.2015.09.019.
RQSS121	B. Kitchenham, P. Brereton, Z. Li, D. Budgen, A. Burn, Repeatability of systematic literature reviews, IET Seminar Digest2011 (1) (2011) 46–55. doi:10.1049/ic.2011.0006.
RQSS122	A. Fernandez, E. Insfran, S. Abrahão, Usability evaluation methods for the web: A systematic mapping study, Information and Software Technology 53 (8) (2011) 789–817. doi:10.1016/j.infsof.2011.02.007.

APPENDIX F – MULTIVOCAL LITERATURE REVIEWS INCLUDED IN TERTIARY STUDY 2 (REFERENCES)

In this appendix, we present the bibliographic information about the multivocal studies included (n=09) related to the Study 4 (S4).

Study ID	Reference
MLR1	V. Garousi, M. Borg, M. Oivo, Cut to the chase: Revisiting the relevance of software engineering research, CoRR abs/1812.01395 (2018). http://arxiv.org/abs/1812.01395
MLR2	O. Plant, Devops under control: development of a framework for achieving internal control and effectively managing risks in a devops environment, Master's thesis, Enschede, NLD (2019).
MLR3	R. Verdecchia, I. Malavolta, P. Lago, Guidelines for architecting android apps: A mixed-method empirical study, in: Proceedings of the IEEE International Conference on Software Architecture, ICSA '19.
MLR4	R. Bhandari, R. Colomo-Palacios, Holacracy in software development teams: A multivocal literature review, ICCSA '19, 2019, pp. 140–145. doi:10.1109/ICCSA.2019.00013.
MLR5	V. Garousi, B. Küçük, Smells in software test code: A survey of knowledgein industry and academia, Journal of Systems and Software 138 (2018) 52–81. doi: 10.1016/j.jss.2017.12.013.
MLR6	S. Maro, J.-P. Steghöfer, M. Staron, Software traceability in the automotive domain: Challenges and solutions, Journal of Systems and Software 141 (2018) 85–110. doi:10.1016/j.jss.2018.03.060.
MLR7	D. L. Freire, R. Z. Frantz, F. Roos-Frantz, S. Sawicki, Survey on the runtime systems of enterprise application integration platforms focusing on performance, Software: Practice and Experience 49 (3) (2019) 341–360. doi:10.1002/spe.2670.
MLR8	A. Saltan, K. Smolander, Towards a SaaS pricing cookbook: A multi-vocal literature review, Software Business (2019) 114–129.
MLR9	A. Ram, A. A. Sawant, M. Castelluccio, A. Bacchelli, What makes a codechange easier to review: An empirical investigation on code change reviewability (2018) 201–212. doi:10.1145/3236024.3236080.