– External Material – Leveraging Sustainable Systematic Literature Reviews

Abstract

This document provides supplementary information about the research process applied to the paper "Leveraging Sustainable Systematic Literature Reviews". In short, we subdivided this material in three sections: (i) Green Drivers (GD) and Sustainability Indicators; (ii) Preliminary evaluation; (iii) Leveraging Points Derivation.

I. Green Drivers (GD) and Sustainability Indicators (SI)

To define Sustainability Indicators (SI) for SLR studies, we need to define the factors that guide the understanding of what means to be sustainable in SLR context. Therefore, our first step was to adapt the 15 characteristics and 15 critical factors defined by Santos et al. [1] to transpose them into drivers that could be support the application of sustainability in practice. For this, we applied thematic analysis [2] to define which factors have major impacts on sustainability of SLR – those factors we named "Green Drivers" (GDs).

Table I illustrates the process adopted to distill the Green Drivers. For instance, the first characteristic (Table I – column 1) asserts that sustainable SLR "*should report reliable results i.e., mitigate threats to validity reducing the uncertainty level of researchers*". Once no critical factors was associated with this characteristic we defined the GD named "*study reliability*". In this context, reliability refer to the confidence that authors tackled systematically every threat to validity during the conduction process. Another example is the "*usage of iteration and pilot tests*" since "*sustainable SLR it should follow a conduction process that is iterative and concentrates on the main changes in the protocol during the pilot test*". Hence, the GD "*Iteration and pilot tests*" was defined and is reinforced by critical factors "*Usage of feasibility studies*" and "*Usage of iterative process*". In this case, we grouped both since SLR studies frequently occur while defining the SLR protocol aiding researchers to reduce the efforts fixing problems in the protocol before conduction starts.

TABLE I
Derivation of Green Drivers using Thematic Analysis (Excerpt) – for full translations See Figure 1

Sustainable SLR Characteristic	Critical Factors	Coding	Green Driver
SLR should report reliable results i.e., migate threats to validity reducing the uncertainty level of researchers	-	Reliable Results; Mitigation of threats to validity.	Study Reliability
SLR should follow a conduction process that complies with conduction standards to ensure quality	-	Follow the conduction process; complies with conduction standards	Compliance with SLR guidelines
SLR should follow a conduction process that is iterative and concentrates on the main changes in the protocol during the pilot test.	Usage of feasibility studies; Usage of iterative process	Use iterative process; Use pilot tests; Use feasibility studies	Usage of iterations and pilot tests

The complete process is illustrated by Figure 1 presents in the 1st column the characteristics, and the 3rd column presents the critical factors, we have highlighted the codes extracted in bold. In the middle (2nd column), we used juxtaposed characteristics and critical factors in 18 Green Drivers, we used the arrows to show which factors influenced the definition of the GDs. Using the 18 GDs defined in the previous step, we applied Goal-Question-Metric (GQM) [3] to derive sustainability indicators (SI) – See Table II. We applied the template of Lami et al. [4], first, defining the main Direct Effects (DE) that could impact SLR, therefore, two DE were defined as our quality focus: (i) time/effort waste; (ii) the impact of results in the target audience [5], hence, we limited our scope to this two DE. Furthermore, our viewpoint is always from a researcher point of view and in the context of scientific research.

- **Analyze**: The object under measurement (Green Driver)
- **For the purpose of**: Understanding, controlling, or improving the object
- **With respect to**: The quality focus of the object that the measurement focuses on (Direct Effect)
- **From the viewpoint of**: The people that measure the object
- **In the context of**: The environment in which measurement takes place

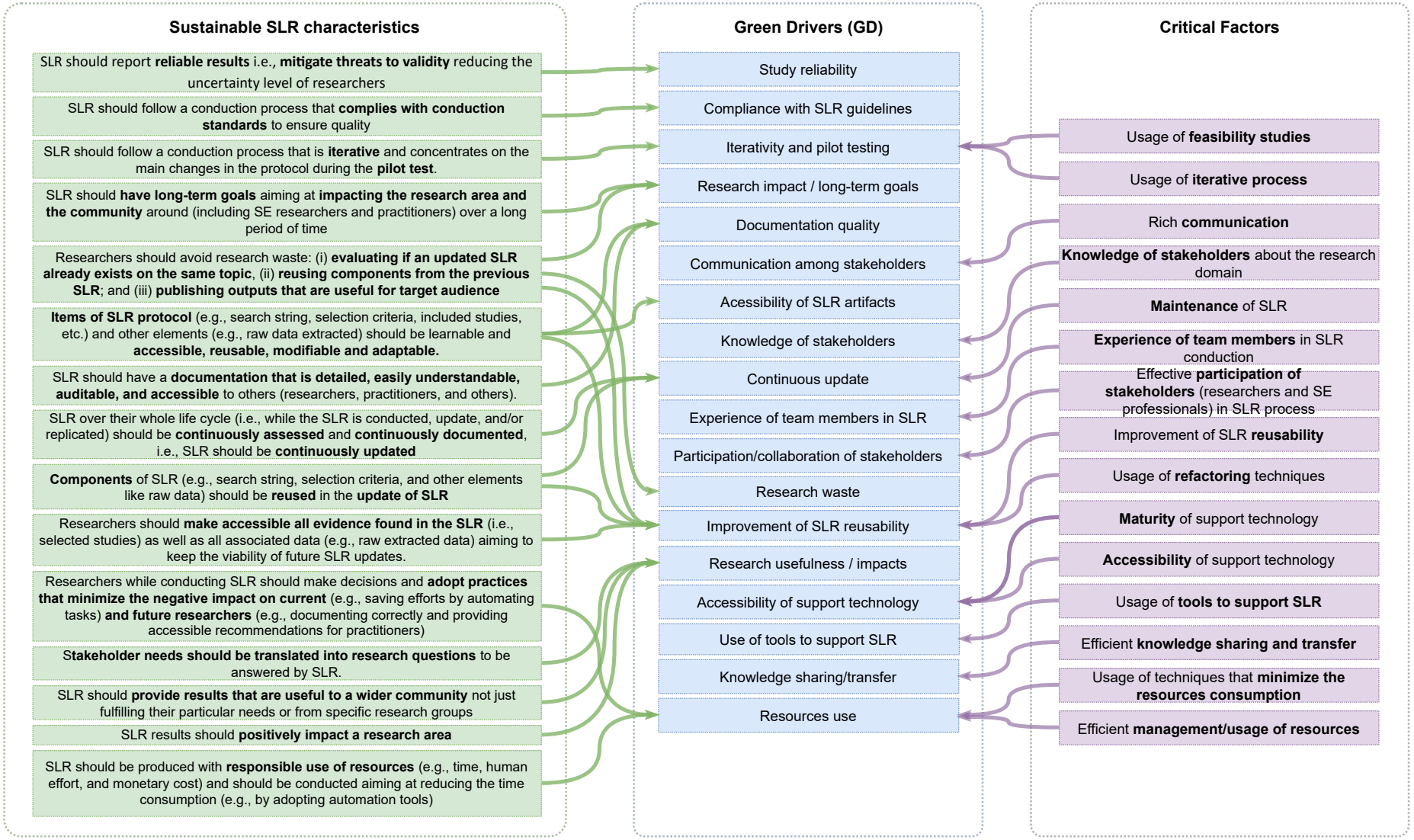

Fig. 1. Green Drivers definition using Thematic Analysis [3]

TABLE II: Application of Goal-Question-Metric to derive Sustainability Indicators

Green Driver (GD)	Goal	Question	Indicator (SI)
GD1 - Compliance with SLR guidelines	Analyze the compliance with SLR guidelines for the purpose of understanding their alignment with guidelines and their impact target audience and the reduction of unnecessary efforts	Q1 - Do the authors use guidelines to conduct their SLR?	SI1 - A reference list indicating the guidelines used to conduct the studies
GD2 - Iterativity and pilot testing	Analyze the iterativity and pilot testing for the purpose of understanding their presence and their impacts on reducing the amount of time/effort consumption	Q2 - Does the process that was followed contain iterations?	SI2 - A boolean indicating the presence of iterations
		Q3 - Was the pilot test conducted?	SI3 - A boolean indicating the use of pilot testing
GD3 - Documentation quality	Analyze the documentation quality for the purpose of understanding the guidelines used to document and impact on documentation quality in target audience	Q4 - Which guidelines for SLR documentation were used?	SI4 - A reference indicating the use of guidelines for documentation
GD4 - Study reliability	Analyze the study reliability for the purpose of understanding the level of reliability of the study and the impacts of producing reliable studies on the target audience	Q5 - Which actions to ensure reliability were applied?	SI5 - Number of actions taken to ensure reliability.
	Analyze the study reliability for the purpose of verifying the presence of quality assessments and their impact on target audience	Q6 - Was a quality assessment conducted?	SI6 - A boolean indicating the conduction of quality assurance assessment.
GD5 - Resources use	Analyze the resources usage for the purpose of verifying their impacts on time and effort consumption	Q7 - Is the study produced making use of techniques that minimize resource consumption?	SI7 - Number of techniques used to minimize time/effort consumption.
		Q8 - How many studies were reviewed and analyzed?	SI8 - Number of studies reviewed/analyzed.
GD6 - Use of tools to support SLR	Analyze the usage of tools to support SLR for the purpose of verifying their impacts on reducing excessive time/effort consumption	Q9 - Which tools were used?	SI9 - Number of tools used during the process.
GD7 - Accessibility of support technology	Analyze the accessibility of support technology for the purpose of understanding their impacts on time/effort waste their maintenance	Q10 - Are the tools used available for use?	SI10 - A boolean indicating whether the tool used is available for use.
		Q11 - Are the tools used free?	SI11 - A qualitative analysis using the following classification: (i) closed tool; (ii) free with restrictions; (iii) free for use (open source).

Continue in the next page

Green Driver (GD)	Goal	Question	Indicator (SI)
GD8 - Communication among stakeholders	Analyze the communication among stakeholders for the purpose of verifying the effects of good internal/external communication on the impact on the target audience	Q12- Any methods/strategies used to support rich communication among researchers, share knowledge, and solve conflicts?	SI12 - Number of methods used to improve communication among stakeholders
GD8 - Participation/collaboration of stakeholders	Analyze the participation/collaboration of stakeholders for the purpose of understanding the effects of research team composition and contributions on the impact on target audience	Q13 - How is the research team composed?	SI13 - A qualitative analysis on the author profiles using the following classification: (i) Academic team; (ii) Industry; (iii) hybrid.
		Q14 - How did stakeholders contribute?	SI14 - Number of stages which stakeholders contributed.
GD10 - Knowledge of stakeholders	Analyze the knowledge of stakeholders about the research domain for the purpose of verifying the effects of the experience of stakeholders in the research domain over the impact in target audience	Q15 - How much experience does the research team have?	SI15 - Number of published papers about the topic addressed in SLR
GD11 - Experience of team members in SLR	Analyze the experience of team members in SLR conduction for the purpose of understanding the impacts on effort/time waste	Q16 - How much experience does the research team have in SLR conduction?	SI16 - Number of secondary studies conducted
GD12 - Knowledge sharing/transfer	Analyze the knowledge sharing/transfer for the purpose of understanding the effects of sharing knowledge in the impacts on target audience	Q17 - Was any method used to share knowledge between team members?	SI17 - Number of Knowledge sharing techniques used
GD13 - Accessibility of SLR artifacts	Analyze the accessibility of artifacts for the purpose of verifying the effects of accessibility of protocol in time/effort waste	Q18 - Is a protocol available including all its versions generated in SLR packaging?	SI18 - A boolean indicating whether the protocol is available.
	Analyze the accessibility of artifacts for the purpose of understanding the impacts of the availability of replication kit in time/effort waste	Q19 - Does the study provide a complete replication kit?	SI19 - A boolean indicating whether the study have a replication kit.
GD14 - Research waste	Analyze the research waste for the purpose of understanding the effects of compliance in avoiding research waste in the effort/time waste	Q20 - Do authors avoid research waste by evaluating existing SLRs on the same topic?	SI20 - A boolean indicating whether the authors perform an evaluation of similar studies.
GD15 - Continuous update	Analyze the continuous update for the purpose of understanding the impact of providing up-to-date information in target audience	Q21 - Was the study updated?	SI21 - Number of updates.
GD16 - Improvement of SLR reusability	Analyze the improvement of SLR reusability for the purpose of identifying the reuse of components in/by other SLRs and their impacts on time/effort waste	Q22 - Were study components used as a basis for other SLR replications?	SI22 - Number of studies that reuse components of the study; SI23 Number of components reused/adapted from previous SLR

Continue in the next page

Green Driver (GD)	Goal	Question	Indicator (SI)
GD17 - Research usefulness/impacts	Analyze the research usefulness and the impacts over community for the purpose of identifying practical recommendations to increase study usefulness for its target audience	Q24- Do authors present practical recommendations useful for a wider community?	SI24 - A boolean indicating the presence of recommendations for industry practitioners.
GD18 - Research impact / long-term goals	Analyze the Long-term goals and research impact over time for the purpose of understanding the relevance of the study for other researchers over time	Q25 - How many citations did the study receive in total and per year?	SI25 - Number of citations per year.

II. Preliminary Evaluation

The research method applied in our preliminary evaluation is presented in presented in Figure 2. In short, our method is composed of four main phases: the setup phase established the main goals, restrictions, and formal methods to extract information from studies. In the second phase, we performed searches in electronic databases and study selection according to criteria defined in the protocol. In the third phase, we piloted our data extraction form to confirm which information was available, next the data extraction occurred. Finally, in the fourth phase, we synthesized and reported results including an analysis using a sustainability perspective. Following, we describe in detail each of these activities.

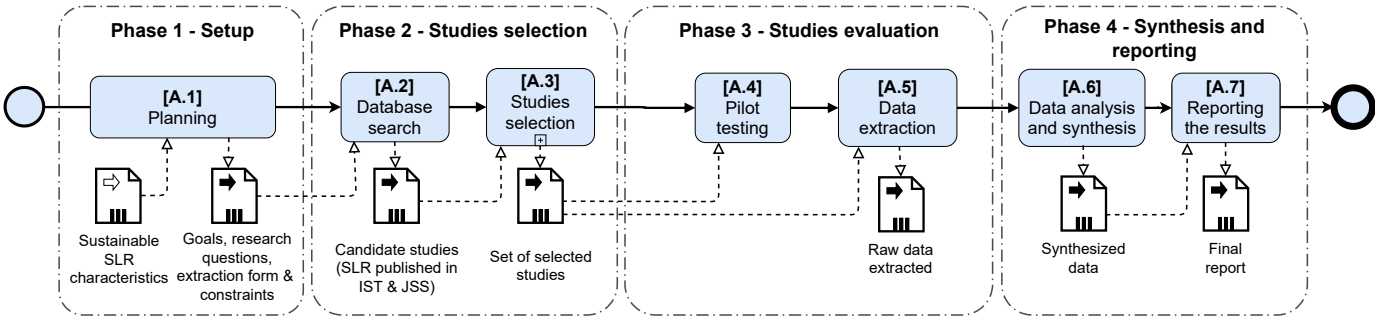

Fig. 2. Research Method followed in this paper

A.1 - *Planning*

This activity was responsible for establishing a protocol that defined the main goals, research questions, constraints, search strategies, extraction forms, and data extraction/synthesis techniques. To achieve our goals, we defined our research question:

-

1) RQ_1 – What is the state of the practice of SLR that is currently being conducted in Software Engineering?

   Rationale: despite SLR having well-established guidelines, its success does not rely only on the observance of these guidelines, instead, but also relies on multiple aspects that impact directly its whole life cycle. This question provides an overview of the studies selected regarding different aspects such as processes, documentation, resource usage, knowledge management, communication, tools, maintenance, and research usefulness/impacts likewise relevant information for analyzing sustainability.

The constraints applied for this preliminary evaluation must limit its scope and allow a deeper analysis of results. Firstly, we analyzed only studies that claim to be Systematic Literature Reviews, hence, variations likewise Systematic Mappings [6], Multivocal Reviews [7] or Rapid Reviews [8] were not considered in this analysis. Furthermore, we selected as our data source studies published in two major journals, namely Information and Software Technology (IST) and the Journal of Systems and Software (JSS). These journals appreciate and encourage the submission of SLR, consequently, these reviews were approved in a rigorous peer-review process which endorses their reliability.

A.2 - *Database Search*

This activity was responsible for executing the database searches and identifying the candidate studies for selection. For this, we used Science Direct[1] database that is the official search engine of papers published in IST and JSS considering the following metadata fields: title, abstract, and keywords. We used the following search string:*"literature review", "systematic*

[1]Science Direct: https://www.sciencedirect.com/

review", "systematic literature review" that returned a total of 412 candidate studies (see Figure 3) published between 2004 and 2022 which were submitted to the selection process.

A.3 - *Studies Selection*

In this activity, we performed the selection process systematically and extracted a sample to be further analyzed. For this, we first defined a set of two inclusion criteria (IC) and five exclusion criteria (EC) that are described below:

-

1) IC_1 – Study is a Systematic Literature Review (Systematic Mappings, MLR, Rapid Reviews, and tertiary studies are not allowed);
2) IC_2 – Study is about software engineering or any subarea of SE.

-

1) EC_1 – Study is not a Systematic Literature Review or is a different type of Secondary study;
2) EC_2 – Study discusses the Systematic Literature Review as a method and does not apply it;
3) EC_3 – Study is a shorter/older version of another study already included;
4) EC_4 – Record is only a call for papers or a collection of abstracts.

Figure 3 presents the selection process used. In A3.1, we processed metadata resulting in 412 studies, next, in A3.2 filtered these studies considering a ten-year time span resulting in 354 valid studies. In A3.3, we applied the aforementioned selection criteria considering the title, abstract, and keywords, and whenever it was necessary we considered the full text, resulting in 238 studies. Finally, in A3.4 a sample of ten studies was extracted to compose our final set. As a metric, we used the studies with the highest number of citations per year mapped by Scopus database. Our final set of studies is presented in Table III.

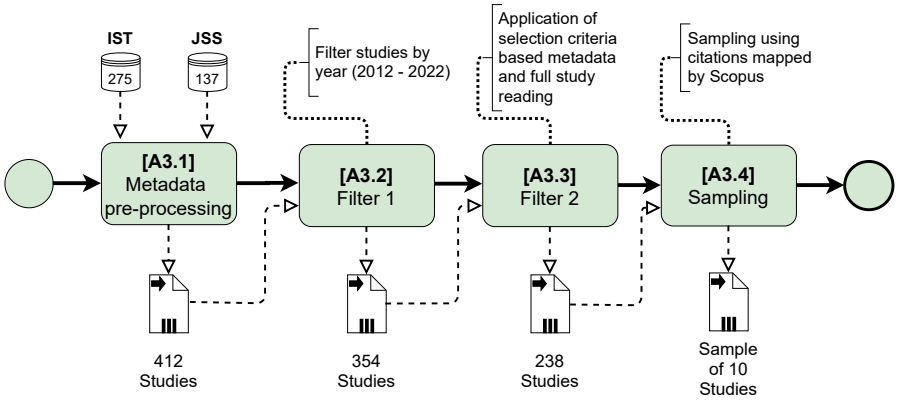

Fig. 3. [A.3] Selection process expanded

TABLE III
STUDIES SELECTED FOR ANALYSIS

ID	Title	Year	Venue	Citations (scopus)	Citations per Year	Ref.
S1	Systematic literature review of machine learning based software development effort estimation models	2012	IST	314	31.40	[9]
S2	Software fault prediction metrics: A systematic literature review	2013	IST	374	41.6	[10]
S3	A systematic literature review of software requirements prioritization research	2014	IST	276	34.5	[11]
S4	Exploring principles of user-centered agile software development: A literature review	2015	IST	195	27.8	[12]
S5	Challenges and success factors for large-scale agile transformations: A systematic literature review	2016	JSS	412	68.7	[13]
S6	Static analysis of android apps: A systematic literature review	2017	IST	198	39.6	[14]
S7	Test case prioritization approaches in regression testing: A systematic literature review	2018	IST	127	31.8	[15]
S8	Machine learning techniques for code smell detection: A systematic literature review and meta-analysis	2019	IST	100	33.3	[16]
S9	A systematic review of unsupervised learning techniques for software defect prediction	2020	IST	70	35.0	[17]
S10	A systematic literature review of blockchain and smart contract development: Techniques, tools, and open challenges	2021	JSS	36	36.0	[18]

A.4 - *Pilot testing*

This activity is responsible for running a pilot study over the data extraction procedures to understand the capability of our artifacts of capturing evidence about the current scenario of SLR. For this, the preliminary data extraction form generated in the first activity (A.1) was used to gather data from two studies (S1 and S2) from our final set. This activity was performed

iteratively i.e., we progressively adjusted the form by adding or removing questions that could not be answered only by reading study documentation. This process was repeated until we define a more stable version of the form amenable to be applied to all studies selected.

A.5 - *Data extraction*

In this activity, each study was carefully appraised to better comprehend its motivation, research method, contributions, and relevance. We answered questions defined in data extraction form and explored the studies metadata to find relevant information about authors' profiles, supplementary materials, or any other associated data that could help us to answer our questions. To support data storage, manipulation, and analysis, we generated artifacts using Excel Spreadsheets which are publicly available here [2].

A.6 - *Data Synthesis*

- Reliability of SLR: to appraise the reliability of our set of studies, we reused former experiences mentioned and adapted the checklist proposed by Ampatzouglou et al. [19] establishing 49 unique actions to increase SLR reliability. Next, we appraised each study individually to collect evidence of practices adopted, not adopted, and not informed. For this, we strictly based on documentation and supplementary files to comprehend the reliability of SLR based on double-checking mechanisms implemented.
- Resource consumption: we used three items available in the final reports that were used to measure the effort level needed to conduct the study, they are (i) a number of candidate studies processed; (ii) relevant studies analyzed during SLR conduction; and, (iii) techniques used to minimize the efforts within most time-consuming activities.
- Collaboration and Communication / previous knowledge about the research domain: since this information is not usually not explicitly asserted in the final report. Hence, we used all information publicly available in the public profiles of authors and all metadata of studies to gather more evidence about stakeholders' participation. We analyzed 43 public profiles (using DBLP [3] and Google Scholar [4]) collecting 3789 studies that were used to evaluate the experience of researchers in scientific research. Using title, abstract, and keywords we filtered those which were secondary studies (i.e., Systematic Mappings, SLR, Grey Literature Reviews, etc.) to comprehend researchers' expertise in SLR conduction. Additionally, we applied the classification proposed by Budgen et al. [20] which uses three categories: inexperienced (no previous secondary studies conducted); limited experience (between 1 and 5 studies published); and, Experienced (more than five studies published).
- Update: we performed a Forward Snowballing (FS) technique [21] to identify possible updates derived from our set of studies. For this, we used the Scopus database since more generic search engines are considered a more suitable option due to their coverage [22]. Scopus mapping identified a total of 2102 references, unfortunately, 111 studies were unavailable/inaccessible, hence, our classification process started considering 1991 studies. First, we removed 39 duplication resulting in 1952 unique papers. Next, we evaluated those studies using the title, abstract, and keywords looking exclusively for SLR, in this step 514 studies remained. Next, a full-text analysis was conducted remaining only 284 studies, finally, after analyzing these papers only three papers [23], [24], [25] were considered updates.
- Usefullness to academia/industry: two metrics were used to appraise the usefulness: (i) Number of citations/year received over years; (ii) presence of practical recommendations to support practitioners in applying knowledge summarized into practice. It is worth mentioning that citation per year was already used by Garousi and Fernandes [26] to quantify the usage of the study. They collected data from 71,668 SE papers indexed by Scopus [5], aiming to analyze the top 100 most cited SE papers. Their results presented that highly cited papers have an average citation value was 6.82 per paper. This metric was also used by Mendes et al. [27] to evaluate how much interest the community has in this topic and, consequently, its impact and usefulness for the development of the research domain making it a cut-off parameter to decide the feasibility of update. Finally, another parameter used to appraise usefulness is the presence of recommendations for practitioners.

A.7 - *Reporting the Results*

This activity was responsible for creating a comprehensive report for data synthesized to disseminate results to the audience. Since our evaluation is still preliminary, we synthesized data in the main paper highlighting important details that somehow evidence the lack of compliance with Green Drivers that were found using sustainable indicators.

[2]https://www.doi.dx/1941234
[3]https://dblp.org/
[4]https://scholar.google.com/
[5]https://www.scopus.com

III. Leveraging Points Derivation

To define the leveraging points for SLR, we synthesized information collected in our data extraction. Next, we distilled this information into an insightful point that should be addressed by the authors. Following, we detail the information that led us to derive the leveraging points.

A. Research process and documentation

Regarding the guidelines adopted by researchers to conduct their SLR, Kitchenham et al [28], [29] is still considered a common basis for studies in our sample (S1, S2, S3, S4, S5, S6, S7, S8, S10), additionally, S4 and S9 also mention the usage of two additional studies that evaluate SLR conducted in SE [30] and update the original guidelines [31]. Besides, S2 and S4 also considered recommendations and lessons learned from Brereton et al. [32], Brocke et al [33], Staples and Niazi [34], Webster & Watson [35]. Finally, only S8 adopted more specific guidelines for snowballing [36].

LP1: *Improve team knowledge about best practices for SLR conduction and always prioritize the adoption of guidelines defined and validated by the community (GD1).*

Our analysis showed that iterativity was mentioned only in three studies (S2, S4, S5) and most of the refactoring occurred during the initial stages of SLR. Besides, only two studies claimed explicitly the usage of pilot testing in their review process (S1, S2). It is worth mentioning that S5 and S7 authors mentioned the conduction of "preliminary searches" which by itself does not fully comply with the rigor required of this kind of testing, however, we recognize that this assertion may indicate the informal conduction of pilot testing. Concerning parts of the protocol refined through pilot testing, the search string is the most mentioned item (S2, S3, S5, S7). Moreover, other items like selection criteria (S1, S2), database selection (S7), and data extraction strategy (S1, S2) were also refined through pilot tests. Despite results showing that researchers are using pilot tests in compliance with guidelines [29] that state the importance of evaluating mainly study selection criteria and data extraction forms before proceeding with SLR, we notice that there is still a lack of adoption of these tests within our sample since six studies (S3, S4, S6, S8, S9, S10) do not mention any kind of testing in their process.

LP2: *Design SLRs to evolve iteratively, using pilot testing (GD2).*

More specifically about documentation, all studies have used Kitchenham's guidelines [28], [29], [30], [31] to document their SLR. In addition, only S4 adopted more specific recommendations for documentation proposed by Brocke et al. [33]. Although all reviews in our studies set had good quality documentation, there is a discrepancy regarding the detail level in reported items. For instance, the identification and selection of primary studies are often quite detailed, at the same time, other processes that are equally important (e.g., data extraction and quality assessment) are described with fewer details. Good examples are found in S2, S4, and S10 which provided many details about the classification used in the data extraction, meanwhile in S3 and S7 fewer details are provided and readers only understand how data was summarized while reading the results section. Despite more updated recommendations have been published to improve SLR documentation e.g., [37], [38], in our sample they were not used.

LP3: *Provide a detailed SLR report using best documentation standards to ensure its quality (GD3).*

In Table IV we present the ten most used MA's whose five refer to study selection validity indicating that improving SLR coverage was a priority for analyzed studies. Similarly, data validity mitigation actions were also used in studies analyzed aiming to ensure the quality of studies and avoid misinterpretation of results using more than one researcher to perform analysis and apply statistical techniques. Finally, an important indicator of reliability is the fact that 9 of 10 studies analyzed conducted some kind of quality assessment, consequently, preventing biases derived from primary studies impact critically the results.

TABLE IV: Actions taken to increase reliability

ID	Actions	Studies
A1	Selection of the most-known digital libraries or specific publication venues or usage of broad search engines or indices	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10
A2	Documentation of inclusion/exclusion criteria been explicitly in the protocol	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10
A3	Usage of snowballing	S1, S2, S3, S4, S5, S6, S8, S10

Continued on next page

TABLE IV – *Continued from previous page*

ID	Actions	Studies
A4	Prescription decision rules set for study inclusion/exclusion	S2, S3, S4, S6, S7, S8, S9, S10
A5	Usage of a strategy for systematic search string construction	S1, S2, S3, S4, S5, S7, S8, S10
A6	Exhaustive search for related work to (a) familiarize with the field, (b) identify comparable studies, and (c) identify relevant publication venues and influential papers	S2, S3, S4, S5, S6, S7, S8, S10
A7	Conduction of article quality assessment as inclusion/exclusion criterion	S1, S2, S3, S7, S8, S9, S10
A8	Involvement of more than one researcher	S1, S2, S3, S5, S6, S8, S10
A9	Usage of statistics to deal with quantitative data for answering research questions	S2, S3, S6, S7, S8, S9, S10
A10	Selection of existing initial classification schema	S1, S2, S4, S5, S6, S9, S10
A11	Development of a consistent strategy (e.g., keep the newer one or keep the journal version) for selecting which study should be retained in the list of primary studies	S1, S3, S5, S6, S7, S9
A12	Discussion about inclusion/exclusion of selected articles in case of conflict	S1, S2, S5, S6, S8, S10
A13	Definition of quality thresholds for inclusion/exclusion	S1, S2, S3, S6, S8, S9
A14	Execution of paper screening to cross-check data extraction	S1, S2, S3, S6, S8, S10
A15	Usage of broad search process without an initial starting date	S2, S3, S5, S6, S7, S8
A16	Comparison of findings and compliance with those of existing studies	S4, S5, S6, S7, S8
A17	Usage of broad search process in generic search engines or indices (e.g., Google Scholar) to ensure the identification of all relevant publication venues	S1, S3, S7, S9
A18	Usage of tools to facilitate the review process	S3, S4, S6, S9
A19	Evaluation of search results and documentation of the outcomes	S2, S4, S8, S9
A20	Continuous update of classification schema until it becomes stable and capable of classifying all primary studies in one or more classes	S2, S4, S5, S6
A21	Usage of a formal data synthesis method	S1, S4, S5, S8
A22	Execution of random screening of articles among authors	-
A23	Manual selection and scan venues to check if they publish articles related to your secondary study	S1, S6, S8
A24	Usage of scientific quality of primary studies to draw conclusions	S4, S8, S9
A25	Provision of public access to all gathered data	S6, S8, S9
A26	Discussion and brainstorm with authors to reach possible interpretations of the findings when there is a lack of related studies	S3, S5, S7
A27	Execution of pilot searches to train your search string	S2, S5
A28	Comparison of primary studies list to a gold standard or to other secondary studies	S2, S5
A29	Usage of summaries of candidate primary studies to guarantee the correct identification of all duplicate articles	S3, S6
A30	Discussion of inclusion/exclusion criteria revising them after pilot studies or experts' suggestions	S1, S2
A31	Selection of variables among authors to guarantee that the research questions can be answered	S1, S2
A32	Appraisal of primary studies validity and their impact applying statistics	S1, S9
A33	Execution of pilot data extraction to test agreement between researchers	S1, S2
A34	Holistic discussion and brainstorm about research questions coverage regarding the goal of the study	S3, S4
A35	Revision of independent experts over search process	S2
A36	Comparison of the number of primary studies in different languages with the population	S8
A37	Comparison of the number of studies with missing full texts with the population	S8

Continued on next page

TABLE IV – *Continued from previous page*

ID	Actions	Studies
A38	Quantification of experts' disagreement with the kappa statistic	S2
A39	Inclusion of grey literature	S2
A40	Execution of pilot data extraction to test the existence of relationships	S1
A41	Execution of pilot data analysis and interpretation	S2
A42	Conduction of reliability checks like post-SLR surveys with experts	S4
A43	Discussion of the research method used (SLR or SMS) to fit the goals/research questions of the study and justification of the purpose and scope of the methods	S7
A44	Usage of systematic voting	-
A45	Usage of sensitivity analysis	-
A46	Usage of experts or external reviewers' opinion in case of conflicts	-
A47	Development, supervision, and documentation of the protocol and their possible deviations	-
A48	Consultation of target audience for setting up study goals	-

Therefore, we derived the following leveraging point:

LP4: *Identify, prioritize, and apply a reasonable number of techniques to mitigate threats to the validity (GD4).*

B. Resources Usage

Regarding the number of studies selected by manual searches and in electronic databases, the values range from 707 to 181,829 studies. This variation drew our attention because most of the studies are composed of small teams (4 or 5 members) and predominately they did not mention the usage of tools to support nor automation techniques indicating that possibly much manual effort from researchers was consumed. Meanwhile, the number of studies considered relevant to be deeply analyzed was somehow more stable ranging from 15 to 124 studies which seems more reasonable since data collection can be very time and effort-consuming depending on the depth, and the amount of data to be extracted, tabulated, and processed. Nevertheless, as this number increases more efforts are needed to mitigate biases inserted by researchers' interpretations. We can find evidence of these additional efforts by observing that seven studies involved more than one researcher in the process (S1, S2, S3, S5, S6, S8, S10), six cross-checked data extraction among authors (S1, S2, S3, S6, S8, S10); and, nine studies conducted quality assessments over primary studies (S1, S2, S3, S4, S5, S7, S8, S9, S10).

Only two studies mentioned techniques to reduce the consumption of efforts while conducting the SLR. In S2, the authors said that they limited the use of human efforts in data extraction and quality assessment by checking only 10 randomly selected studies. In S6, the authors said that they have used Python scripts to help process data extracted from the databases, in addition, they mentioned that the workload was balanced with the authors of the work, making resource management more efficient. Broad searches in SLR are acceptable and often recommended to attend to the need for coverage. However, given the laborious nature of review studies, and the fact that SLR analyzed in this study were conducted by small teams, it is difficult to comprehend how these large amounts of data were processed without the use of tools. The lack of details about which techniques were used hindered us to comprehend exactly the amount of effort consumed in the process.

LP5: *Prioritize the preservation of resources and foster the use of techniques to reduce efforts (GD5).*

C. Tool Support

During our analysis, we noticed that only four studies mentioned the usage of some tools to support the review process. In S6 authors used Python scripts to support collecting studies from databases; S4 used MAXQDA2 [6] to process qualitative data; S9 claimed to use R scripts to support data analysis and Mendeley [7]; and, S3 claimed to use a visualization tool without mentioning which one. Considering that the studies analyzed treated a high amount of information, probably the six studies that did not mention the usage of tools to support the review (S1, S2, S5, S7, S8, S10) omitted this information in the final report.

Accessibility of techniques used in S6 and S9 is questionable since they mention the use of programming languages (Python and R) to support in SLR process, which may represent a barrier to future researchers due to the long learning curve of coding

[6]MAXQDA website: https://www.maxqda.com/

[7]Mendeley website - https://www.mendeley.com/

techniques to reproduce the results. The adoption of MAXQDA2 to summarize qualitative data is more accessible since this tool was validated by empirical studies with positive remarks about its easy and interactive design [39], additionally, it has more solid documentation [40], [41] to support researchers to use it. However, it is worth mentioning that MAXQDA2 is developed and maintained by a private company and operates only under a paid license i.e., it is not freely available for everyone which may pose as a barrier to updating/reproduction. Other tools like Mendeley are also supported by private companies (Elsevier), but it is free for usage. Meanwhile, Python and R are open-source programming languages i.e., they are free for use and are maintained by the community. Despite the existence of more specific tools to support the SLR process [42], there is evidence in the literature that these tools are not used and researchers prefer to use more generic tools like Google/Excel spreadsheets, Jabref, Mendeley, etc. [43], which aligns with our results.

LP6: *Use tools to support the SLR process to the fullest extent (GD6) and document the experience to foster the development and improvement of these tools.*

D. Communication and Collaboration

Analyzing team composition, all 43 are formally affiliated with some universities or research centers, and only one is affiliated with both university and industry. Concerning external stakeholders' collaboration in protocol elaboration or other steps of the process, the only evidence available is presented in S6 when the authors contacted primary studies authors to self-check their data extraction and point out any inaccuracies.

To support collaboration among researchers the most popular technique was consensus meetings (S1, S2, S3, S5, S6, S8, S10), nevertheless, this statement is quite generic since it does not provide clues about which and how decisions were taken. In some cases, studies combined consensus meetings with more pragmatic techniques, such as Kappa Coefficient [44] (S2) or Krippendorff Alpha Kr_a [45] (S8) or even data extraction cards (S1). In three studies nothing is mentioned about how authors communicated internally, being possible that it still occurs informally or is very limited to some participants, which makes this aspect a threat to validity that is omitted by authors.

In S2 and S4 reported team participation through the suggestion of important studies or in activities related to SLR planning, such as, database selection, piloting selection criteria, or protocol review. In conduction stage, researchers reported more intensive team collaboration in studies selection (S1, S2, S5, S6, S8), data extraction (S1, S2, S3, S6, S8) and quality assessment (S1, S3, S8). In reporting stage only S10 stated that all authors collaborated to write the final report.

LP7: *Foster participation and collaboration of stakeholders (GD8) by defining and documenting the roles played by each one, always making use of strategies to improve internal and external communication (GD9).*

E. Knowledge Management

Our results revealed that the research teams were predominantly heterogeneous regarding their experience in research work. In eight studies analyzed (S1, S2, S3, S4, S5, S8, S9, S10) the first author had the lowest number of studies published (avg. of six studies published) indicating a relative lack of experience. At the same time, the remainder of the authors were more experienced (avg. 194 studies published) indicating that most of the SLR incorporated into the research team experienced researchers to avoid bias in the process. Our results are in consonance with those presented in Budgen et al. [20] observed the number of inexperienced authors conducting SLR increased over time due to the value of beginning postgraduate study by conducting a formal literature review. Furthermore, experienced researchers are often involved as additional authors in studies led by inexperienced authors, probably being a combination of student and supervisor in postgraduation.

Concerning the experience of authors in SLR conduction, our results revealed that 10 (out of 43) authors could be considered experienced in SLR conduction and the remainder of the 33 authors were considered with limited experience in SLR conduction. Additionally, we noticed that eight studies (S2, S3, S4, S5, S6, S7, S8, S9) have in their team at least one author which was considered experienced in SLR conduction, once again, reaffirming the heterogeneity of research teams.

LP8: *Prefer hybrid teams combining experienced stakeholders in the SLR process (GD9) and stakeholders with knowledge in the topic addressed in the SLR (GD10).*

Few details were provided about internal knowledge sharing/transfer among researchers. Consensus meetings were used by 7 out of 10 studies (S1, S2, S3, S5, S6, S8, S10) as the main technique to exchange information, though no additional methods to exchange information internally or externally to the research group were mentioned. Another indicator of sustainability analyzed was the information exchange with future generations of researchers by analyzing the availability of SLR artifacts which are essential for updates, replications, and audition of results. In this sense, the availability of the SLR protocol including

all versions and clear statements of changes that occurred during the review is essential to ensure the exchange of know-how about the process executed. Nonetheless, in all studies, the protocol is presented only in its final version and no additional details are provided in external repositories, supplementary materials, or associated technical reports. This can be considered a problem when compared to SLR conducted in other areas (e.g., medicine), whose researchers are encouraged to register their research protocol in open science tools like PROSPERO[8], or Open Science Framework[9]. Hence, these approaches may foster researchers' collaboration, avoid wasting efforts on research that is already being conducted, and reuse elements from other reviews. Additionally, we noticed that raw data provided by authors are often incomplete and include only few artifacts generated by SLR conduction, for example, all studies included the list of studies selected for data analysis, but none provided a complete list of excluded studies which makes it more difficult to reproduce, update or audition of the study.

LP9: *Ensure that knowledge acquired in the SLR process is shared with interested parties (GD12) and all artifacts are fully accessible for readers, fostering open science (GD13).*

Our analysis showed that few studies reported the reuse of previous SLR elements. Most of the reuse is related to punctual artifacts of SLR, for instance, in S4 authors reused a set of keywords and S7 reused a database set. Only in S8 authors use previous results to compare their findings and draw more reliable conclusions. In spite of the results indicating a lack of reuse of elements, we noticed that researchers are aware of the importance of not duplicating SLR studies. Our analysis concluded that nine studies (S1, S2, S3, S4, S5, S6, S7, S8, S10) assert that there are no previous SLR already existing addressing the same topic. However, a common flaw in studies is a lack of more concrete evidence i.e., a systematic search to ensure that authors have not missed any similar SLR that could be reused.

LP10: *Make an effort to avoid unreasonably duplicating SLRs so reducing research waste (GD14). Reuse as many elements as possible from prior SLRs (GD16).*

F. Update/Maintenance

Forward snowballing process identified over two thousand candidate studies that after a classification process two studies [23], [25] remained as updates.

Ali et al. [23] performed an SLR based on the empirical studies published in the time period of January 1991 to December 2017. Authors mention that their coverage of studies allows them to update results already provided by [9]. Due to the substantial amount of work done in the last 7 to 8 years based on software effort estimation using ML methods, the main objective of Ali et al. [23] is to review studies that used and discussed the software effort estimation models built using ML techniques outlining conclusions that changed in further years. Five research questions proposed by [9] were maintained and some additional ones are proposed to provide a new analysis of trends of applying ML techniques for effort estimation.

Castro-Cabrera et al. [25] perform a systematic literature review (SLR) on the test case prioritization (TCP) that updates the results provided in S7. Authors provide the latest developments in TCP specifically using the taxonomy proposed in S7 including the most important publications from 2017 to 2019. Results showed an increasing interest in the topic analyzing 320 papers published in this short period of time dealing with many different TCP approaches.

Despite the last years providing more guidelines to update SLR (e.g., [46], [27], [22], [47]) both studies [23], [25] analyzed that claim to update S1 and S7 respectively do not use a systematic approach to execute the update. According to [48] SLR updates should justify any changes in the original SLR protocol, in addition, they should be made only when really necessary. In both studies analyzed there is a lack of information about the update process, Ali et al. [23] only state to cover the same period and answer the same research questions and it is not clear which modifications were done in the protocol. Meanwhile in [25] authors states that years covered in their review are complementary to S7, but it does not describe how much of former study was reused.

Our investigation also revealed that studies selected inspired somehow at least 284 other SLR (not including Systematic Mappings, Ad-hoc literature reviews, or Grey Literature Reviews). These inspirations are presented in different manners across studies, for instance, Tieppo et al. [24] performed minor adaptations in protocol proposed by S1, and Pan et al. [49] claim the whole research process was inspired by S5. Meanwhile, other studies just mention studies in related work or reused smaller components of previous SLR, for instance, S1 inspired the search strategy [50], data extraction [50], [51], quality criterion [50], [51], and results analysis [50]. Nevertheless, we did not considered reuse of components as updates as concluded Nepomuceno et al. [48].

LP11: *Design SLRs to enable their continuous update (GD15).*

[8]https://www.crd.york.ac.uk/PROSPERO/
[9]https://osf.io/

G. Research Impact and Usefulness

TABLE V
STUDIES CITATIONS OVER YEARS

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Mean	Std. Dev.
S1	1	9	9	17	29	27	32	47	**52**	46	32	27,4	16,9
S2	-	6	12	33	30	32	46	**57**	50	45	35	34,6	16,1
S3	-	-	10	14	20	35	42	**49**	30	34	34	29,8	12,8
S4	-	-	-	5	18	29	**34**	23	33	26	21	23,6	9,4
S5	-	-	-	-	3	26	50	64	82	**90**	81	56,6	32,3
S6	-	-	-	-	-	6	16	34	34	**45**	28	27,2	14,0
S7	-	-	-	-	-	-	6	24	24	31	**33**	23,6	10,6
S8	-	-	-	-	-	-	-	7	24	27	**35**	23,3	11,8
S9	-	-	-	-	-	-	-	-	7	28	**34**	23,0	14,2
S10	-	-	-	-	-	-	-	-	-	14	**19**	16,5	3,5

Table V presents that the average of citations per year ranged between 16.5 and 56.6 citations, which means that studies analyzed are far beyond from cut-off metric previously established by Garousi and Fernandes [26] making them useful for their respective areas. Another important aspect is the timespan that information remains useful for SE community. Overall, studies published remain widely used by the academic community over the years indicating that SLR has a long-term impact on SE the community, and its information remains as a knowledge base for future researchers. In addition, some studies (S1, S2, S3, S5) achieved their peak of citations between 5 and 8 years after being published, after that, there is a tendency for decay of citation numbers indicating that an update may be necessary.

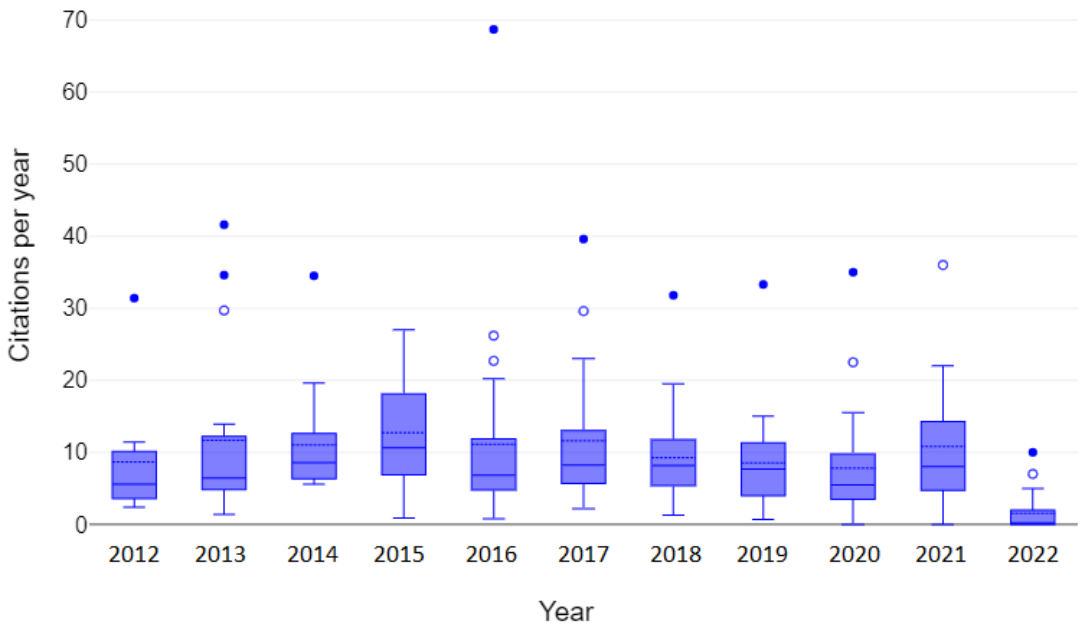

Fig. 4. Citations/Year

Finding clear practical recommendations for industry practitioners was quite difficult for most of the studies. Only in S1 a section was especially dedicated to providing insights for practitioners making it clear that researchers have the industry as a target audience. Overall, studies dedicated a few text passages to communicate with practitioners e.g., S2 suggested useful studies and included in their classification scheme means for flagging research created from industry data that may be relevant for these practitioners. Similarly, S3 mentioned that results provide insight for both academia/industry; S4 claimed that their results might contribute to support practitioners applying techniques described in practice delivery of useful and usable software; and, S5 defined its research questions believing that they represent the viewpoints that are likely to provide insights to practitioners as well as researchers. In the remainder of the studies, there are no clear indications about which piece of evidence is particularly valuable for practitioners, however, it doesn't necessarily mean that the results provided are useless, but the final report does not communicate with this audience.

LP12: *Improve the usefulness of SLR results by designing SLRs that cover multiple perspectives that can benefit the software engineering community (GD17) and have long-term impacts (GD18).*

REFERENCES

[1] V. dos Santos, A. Y. Iwazaki, K. R. Felizardo, E. F. de Souza, and E. Y. Nakagawa, "Sustainable systematic literature reviews," *Information and Software Technology*, vol. 176, p. 107551, 2024.

[2] D. S. Cruzes and T. Dybå, "Research synthesis in software engineering: A tertiary study," *Information and Software Technology*, vol. 53, pp. 440–455, May 2011.

[3] R. van Solingen, V. Basili, G. Caldiera, and H. D. Rombach, *Goal Question Metric (GQM) Approach*. Wiley, Jan. 2002.

[4] G. Lami, F. Fabbrini, and M. Fusani, "A methodology to derive sustainability indicators for software development projects," in *Proceedings of the 2013 International Conference on Software and System Process*, vol. 3 of *ICSSP '13*, p. 70–77, ACM, May 2013.

[5] V. dos Santos, A. Y. Iwazaki, K. R. Felizardo, É. F. de Souza, and E. Y. Nakagawa, "Towards sustainability of systematic literature reviews," in *15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pp. 1–6, 2021.

[6] K. Petersen and N. Ali, "Identifying strategies for study selection in systematic reviews and maps," in 5^{th} *International Symposium on Empirical Software Engineering and Measurement*, (Banff, Alberta, Canada), pp. 351–354, ACM/IEEE, 2011.

[7] V. Garousi, M. Felderer, and M. V. Mäntylä, "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering," *Information and Software Technology*, vol. 106, pp. 101–121, Feb. 2019.

[8] B. Cartaxo, G. Pinto, and S. Soares, "The role of rapid reviews in supporting decision-making in software engineering practice," in *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*, pp. 24–34, ACM, June 2018.

[9] J. Wen, S. Li, Z. Lin, Y. Hu, and C. Huang, "Systematic literature review of machine learning based software development effort estimation models," *Information and Software Technology*, vol. 54, pp. 41–59, Jan. 2012.

[10] D. Radjenović, M. Heričko, R. Torkar, and A. Živkovič, "Software fault prediction metrics: A systematic literature review," *Information and Software Technology*, vol. 55, pp. 1397–1418, Aug. 2013.

[11] P. Achimugu, A. Selamat, R. Ibrahim, and M. N. Mahrin, "A systematic literature review of software requirements prioritization research," *Information and Software Technology*, vol. 56, pp. 568–585, June 2014.

[12] M. Brhel, H. Meth, A. Maedche, and K. Werder, "Exploring principles of user-centered agile software development: A literature review," *Information and Software Technology*, vol. 61, pp. 163–181, May 2015.

[13] K. Dikert, M. Paasivaara, and C. Lassenius, "Challenges and success factors for large-scale agile transformations: A systematic literature review," *Journal of Systems and Software*, vol. 119, pp. 87–108, Sept. 2016.

[14] L. Li, T. F. Bissyandé, M. Papadakis, S. Rasthofer, A. Bartel, D. Octeau, J. Klein, and L. Traon, "Static analysis of android apps: A systematic literature review," *Information and Software Technology*, vol. 88, pp. 67–95, Aug. 2017.

[15] M. Khatibsyarbini, M. A. Isa, D. N. Jawawi, and R. Tumeng, "Test case prioritization approaches in regression testing: A systematic literature review," *Information and Software Technology*, vol. 93, pp. 74–93, Jan. 2018.

[16] M. I. Azeem, F. Palomba, L. Shi, and Q. Wang, "Machine learning techniques for code smell detection: A systematic literature review and meta-analysis," *Information and Software Technology*, vol. 108, pp. 115–138, Apr. 2019.

[17] N. Li, M. Shepperd, and Y. Guo, "A systematic review of unsupervised learning techniques for software defect prediction," *Information and Software Technology*, vol. 122, p. 106287, June 2020.

[18] A. Vacca, A. D. Sorbo, C. A. Visaggio, and G. Canfora, "A systematic literature review of blockchain and smart contract development: Techniques, tools, and open challenges," *Journal of Systems and Software*, vol. 174, p. 110891, Apr. 2021.

[19] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies," *Information and Software Technology*, vol. 106, pp. 201–230, 2019.

[20] D. Budgen and P. Brereton, "Short communication: Evolution of secondary studies in software engineering," *Information and Software Technology*, vol. 145, p. 106840, May 2022.

[21] E. Mourão, J. F. Pimentel, L. Murta, M. Kalinowski, E. Mendes, and C. Wohlin, "On the performance of hybrid search strategies for systematic literature reviews in software engineering," *Information and Software Technology*, vol. 123, p. 106294, July 2020.

[22] K. R. Felizardo, A. Y. I. da Silva, É. F. de Souza, N. L. Vijaykumar, and E. Y. Nakagawa, "Evaluating strategies for forward snowballing application to support secondary studies updates," in *Proceedings of the XXXII Brazilian Symposium on Software Engineering*, ACM, Sept. 2018.

[23] A. Ali and C. Gravino, "A systematic literature review of software effort prediction using machine learning methods," *Journal of Software: Evolution and Process*, vol. 31, July 2019.

[24] E. Tieppo, R. R. dos Santos, J. P. Barddal, and J. C. Nievola, "Hierarchical classification of data streams: a systematic literature review," *Artificial Intelligence Review*, vol. 55, pp. 3243–3282, Oct. 2021.

[25] M. d. C. de Castro-Cabrera, A. García-Dominguez, and I. Medina-Bulo, "Trends in prioritization of test cases: 2017-2019," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, (New York, NY, USA), p. 2005–2011, Association for Computing Machinery, 2020.

[26] V. Garousi and J. M. Fernandes, "Highly-cited papers in software engineering: The top-100," *Information and Software Technology*, vol. 71, pp. 108–128, Mar. 2016.

[27] E. Mendes, C. Wohlin, K. Felizardo, and M. Kalinowski, "When to update systematic literature reviews in software engineering," *Journal of Systems and Software*, vol. 167, p. 110607, Sept. 2020.

[28] B. Kitchenham, "Procedures for performing systematic reviews," Tech. Rep. TR/SE–0401, Keele University, 2004.

[29] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," tech. rep., Keele University, 2007.

[30] B. Kitchenham, O. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and software technology*, vol. 51, no. 1, pp. 7–15, 2009.

[31] B. A. Kitchenham, D. Budgen, and P. Brereton, *Evidence-based software engineering and systematic reviews*, vol. 4. New York, USA: CRC press, 2015.

[32] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of Systems and Software*, vol. 80, pp. 571–583, Apr. 2007.

[33] J. v. Brocke, A. Simons, B. Niehaves, B. Niehaves, K. Reimer, R. Plattfaut, and A. Cleven, "Reconstructing the giant: On the importance of rigour in documenting the literature search process," in *17th European Conference on Information Systems (ECIS)*, pp. 2206–2217, 2009.

[34] M. Staples and M. Niazi, "Experiences using systematic review guidelines," *Journal of Systems and Software*, vol. 80, pp. 1425–1437, Sept. 2007.

[35] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly*, pp. xiii–xxiii, 2002.

[36] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *18th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pp. 1–10, 2014.

[37] D. Budgen, P. Brereton, S. Drummond, and N. Williams, "Reporting systematic reviews: Some lessons from a tertiary study," *Information and Software Technology*, vol. 95, pp. 62–74, Mar. 2018.

[38] B. A. Kitchenham, L. Madeyski, and D. Budgen, "SEGRESS: Software engineering guidelines for REporting secondary studies," *IEEE Transactions on Software Engineering*, pp. 1–1, 2022.

[39] E. Kuş Saillard, "Systematic versus interpretive analysis with two caqdas packages: Nvivo and maxqda," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. Vol 12, p. No 1 (2011): The KWALON Experiment: Discussions on Qualitative Data Analysis Software by Developers and Users, 2011.

[40] U. Kuckartz and S. Rädiker, *Analyzing Qualitative Data with MAXQDA*. Springer International Publishing, 2019.

[41] M. C. Gizzi and S. Rädiker, *The Practice of Qualitative Data Analysis: Research Examples Using MAXQDA*. BoD–Books on Demand, 2021.

[42] C. Marshall and P. Brereton, "Systematic review toolbox," in *International Conference on Evaluation and Assessment in Software Engineering*, pp. 1–6, ACM Press, 2015.

[43] A. Al-Zubidy and J. C. Carver, "Identification and prioritization of SLR search tool requirements: an SLR and a survey," *Empirical Software Engineering*, vol. 24, pp. 139–169, May 2018.

[44] J. Pérez, J. Díaz, J. Garcia-Martin, and B. Tabuenca, "Systematic literature reviews in software engineering—enhancement of the study selection process using cohen's kappa statistic," *Journal of Systems and Software*, vol. 168, p. 110657, Oct. 2020.

[45] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[46] C. Wohlin, E. Mendes, K. R. Felizardo, and M. Kalinowski, "Guidelines for the search strategy to update systematic literature reviews in software engineering," *Information and Software Technology*, vol. 127, p. 106366, Nov. 2020.

[47] L. Garcés, K. Felizardo, L. Oliveira, and E. Nakagawa, "An experience report on update of systematic literature reviews," in *International Conferences on Software Engineering and Knowledge Engineering*, KSI Research Inc. and Knowledge Systems Institute Graduate School, 2017.

[48] V. Nepomuceno and S. Soares, "On the need to update systematic literature reviews," *Information and Software Technology*, vol. 109, pp. 40–42, May 2019.

[49] R. Pan, M. Bagherzadeh, T. A. Ghaleb, and L. Briand, "Test case selection and prioritization using machine learning: a systematic literature review," *Empirical Software Engineering*, vol. 27, Dec. 2021.

[50] A. Idri, M. Hosni, and A. Abran, "Systematic literature review of ensemble effort estimation," *Journal of Systems and Software*, vol. 118, pp. 151–175, Aug. 2016.

[51] E. Papavasileiou, J. Cornelis, and B. Jansen, "A systematic literature review of the successors of "NeuroEvolution of augmenting topologies"," *Evolutionary Computation*, vol. 29, pp. 1–73, Mar. 2021.