

# Scrutinizing Systematic Literature Review Process in Software Engineering

Zlatko Stapić<sup>1</sup>, Luis de-Marcos<sup>2</sup>,  
Vjerran Strahonja<sup>1</sup>, Antonio García-Cabot<sup>2</sup>, Eva García López<sup>2</sup>

<sup>1</sup>University of Zagreb, Faculty of Organization and Informatics Varaždin Pavlinska 2, 42000 Varaždin, Croatia

<sup>2</sup>University of Alcalá, Computer Science Department Ctra. Barcelona km 33.6, 28871 Alcalá de Henares (Madrid), Spain

**Abstract:** Performing the Systematic Literature Review (SLR) in the turbulent field of Software Engineering (SE) brings different obstacles and uncertainties. The commonly used guidelines for performing the SLR in this field are adapted from health sciences and presented by Kitchenham and Charters in 2007. This paper follows the Kitchenham's three-phases-review-process and fulfils it with the findings, observations and recommendations from other influential authors in the field. The process of SLR is observed from the perspective of appliance in the field of SE and supplemented by the important precautions measures that should be undertaken by the authors performing it. Thus, this paper aims to present the state-of-the-art in performing the SLR in SE.

**Keywords:** Systematic Literature Review, SLR, Software Engineering, State-of-the-Art.

## 1. Introduction

In order to perform a comprehensive and thorough analysis of the existing knowledge in the domain of software engineering, the systematic approach should be undertaken and the existing methodologies, methods and good practices should be reviewed. Different methods and approaches can be undertaken in order to perform such analysis: *systematic literature review*, *systematic mapping studies*, *tertiary reviews* as discussed by [1], or *narrative review*, *conceptual review*, *rapid review* and several other types of review presented by [2].

Taking into consideration the undertaken initial examination of the literature, lately, a systematic literature review (SLR) has been commonly used for different analysis in the field of software engineering (SE). On the other side, the systematic mapping study should be used when a topic is either very little or very broadly covered, and tertiary reviews are the most suitable approach if several reviews in the target domain already exist and should be summarized.

A systematic literature review is trustworthy, rigorous and replicable methodology that is used to evaluate and interpret all reported research relevant to any phenomenon of interest [1]. The origins of systematic review can be traced back from the beginning of the 20<sup>th</sup> century, but during the 1980's, systematic research synthesis and meta-analysis reach an especially distinctive methodological status in the domain of health sciences [3]. During this period, and as a result of performing similar methods in other different fields, different synonyms of this method have been used in the literature. Some of those are research review, research synthesis, research integration, systematic overview et cetera [4].

In the field of software engineering, during the last years, several primary studies have been conducted and although these studies are accompanied by an increasing improvement in methodology, this field is still an area of investigation that remains to be explored and that could well bring many benefits in terms of mechanisms needed to assist practitioners to adopt appropriate technologies and methodologies [4]. Kitchenham in [5] proposed the guideline for systematic reviews in software engineering. It was created as adaptation of several existing guidelines, mainly from medicine.

As a response to several authors, like Biolchini et al. [4], Mian et al. [6] and Staples and Niazi [7] who found that Kitchenham described guidelines at a relatively high level and partially inappropriate to conduct for researchers in the field of SE, Kitchenham and Charters in 2007 published a


DOI: 10.18421/TEM51-16

<https://dx.doi.org/10.18421/TEM51-16>

**Corresponding author: Zlatko Stapić:**

**Vjerran Strahonja:** University of Zagreb, Croatia

**Luis de-Marcos, Antonio García-Cabot, Eva García López:** University of Alcalá, Spain

 © 2016 Zlatko Stapić et al, published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License. The article is published with Open Access at [www.temjournal.com](http://www.temjournal.com).

new version of technology report [1] aiming to propose more comprehensive guidelines of performing SLR for the researchers and PhD students in the field of SE. But, the basis for these guidelines remained the same: guidelines used by medical researchers, reinforced by several books and discussions with the researches from the other fields.

Taking the remarks from mentioned authors, that researchers in the field of the software engineering need more focused guidelines, this paper will cover in detail the systematic literature review methodology as it is proposed in [1] by summarizing and aggregating all phases of the methodology with additional inclusion of observations and recommendations from other influential authors in this field. Thus, the second section defines a method of SLR, while the third section, as the most important part of this paper, brings the enhanced details of performing SLR in SE. As the result of reports and opinions comparison, the fourth section discusses the advantages and disadvantages of using the SLR. Finally, the last section concludes the topic. The whole mentioned process of performing SLR is supplemented by the important precautions measures that should be undertaken by the authors performing it.

## 2. Definition of systematic literature review (SLR)

A systematic literature review (SLR) is defined by Kitchenham and Charters [1] as “a form of secondary study that uses a well-defined methodology to identify, analyze and interpret all available evidence related to a specific question in a way that is unbiased and (to a degree) repeatable”. Dybå and Dingsøy [8] define SLR as “a concise summary of the best available evidence that uses explicit and rigorous methods to identify, critically appraise, and synthesize relevant studies on a particular topic”. According to Dybå, these methods should be defined in advance and documented in a protocol so the others could critically appraise and replicate the review.

There are different reasons to perform a systematic literature review. In general, whenever a literature review is performed it could be done by a systematic (following stated procedures and steps) or unsystematic (just reading and taking notes) approach. The usual reason to use SLR is to summarize the available reports concerning a topic of interest. This is to say that for example, systematic literature review could be used to summarize the methodologies that could be used for development of mobile applications. SLR could also be used to find gaps in the focused research that could be further investigated. In addition, there are other reasons to use a systematic approach, such as the purpose of the

research, the scientific approach, the quality expectations or the existence of previous researches on the selected topic.

According to Dybå and Dingsøy [8] the key feature that distinguishes the SLR from a traditional narrative reviews lies in their explicit attempt to minimize the chances of making wrong conclusions which could be the result of biases either in the primary studies or in the review process itself.

## 3. The review process

Although the methodology of the SLR, presented by Kitchenham and Charters [1], is considerably upgraded if compared to the first version from 2004, the main three phases remained the same. General steps to be performed in the SLR are also similar and are defined as in Table 1.

Table 1. The review process [1]

<b>Phase 1: Planning the review</b>	
	Identification of the need for a review
	Commissioning a review (optional)
	Specifying the research question(s)
	Developing a review protocol
	Evaluating the review protocol (recommended)
<b>Phase 2: Conducting the review</b>	
	Identification of research
	Selection of primary studies
	Study quality assessment
	Data extraction and monitoring
	Data synthesis
<b>Phase 3: Reporting the review</b>	
	Specifying dissemination mechanisms
	Formatting the main report
	Evaluating the report (recommended)

According to author of the review process, Kitchenham, all mentioned activities (stages) are mandatory except commissioning a review as it depends on planned commercialization of review results, evaluating the review protocol and evaluating the report which are optional as being dependent on procedures of the quality assurance chosen by author of the review. In any case, the latest activities are also recommended.

Additionally, one could conclude that above mentioned stages and phases are sequential, but even being defined as sequential, it is important to mention that some of the stages could be repeated more than once and might involve iteration or reimplementation. For example, the negative evaluation of review protocol or negative evaluation of the report might result in the need to repeat the part or the whole review process. Or, the inclusion and exclusion criteria of the relevant studies could be refined after the quality criteria are defined. The fact

that review protocol is often changed even by experienced scientists brings critics to existing reviews as not being completely objective. However, some authors, like Staples and Niazi [7] consider the review protocol necessary even if changed through the process. General conclusion is that the protocol is needed and that it increases the quality of the process.

Each stage of the SLR process will be discussed in more details in the following sections.

### *Planning the review*

Most important activities during the phase of review planning are definition of the review question(s) and creation of the review protocol, but all other activities should not be neglected and should be taken seriously. The results of this phase should include a clearly defined review protocol containing the purpose and the procedures of the review.

The summary of each stage is presented below and is based on guidelines presented in [1] and on additional discussions from other authors which are cited in the text.

#### *3.1.1. Identification of the need for a review*

The first activity of the process is to align the need for a review by performing a preliminary research. This should include the identification and review of existing SLRs on the same topic by using evaluation criteria defined in advance. Existing SLRs are commonly examined according to set of questions or even a checklist. Different authors proposed such checklists which usually focus on quality, inclusion and exclusion criteria, coverage of relevant studies and quality of included studies. An example is given by Centre for Reviews and Dissemination [9] who defines following set of questions to be used when critically examining review articles:

- ✓ Was the review question clearly defined in terms of population, interventions, comparators, outcomes and study designs (PICOS)?
- ✓ Was the search strategy adequate and appropriate? Were there any restrictions on language, publication status or on publication date?
- ✓ Were preventative steps taken to minimize bias and errors in the study selection process?
- ✓ Were appropriate criteria used in primary studies quality assessment, and were preventative steps taken to minimize bias and errors in the quality assessment process?
- ✓ Were preventative steps taken to minimize bias and errors in the data extraction process?

- ✓ Were adequate details presented for each of the primary studies?
- ✓ Were appropriate methods used for data synthesis? Were differences between studies assessed? Were the studies pooled, and if so was it appropriate and meaningful to do so?
- ✓ Do the authors' conclusions accurately reflect the evidence that was reviewed?

#### *3.1.2. Commissioning a review*

This optional task should be performed only if organization has no resource to perform a review by itself. In such case, a commissioning document containing all the important information relevant to the SLR should be created.

Scientists and PhD students will not create a commissioning document while performing a systematic literature review as a part of their own work. The only issue that should be addressed is that a dissemination strategy should be, in this case, incorporated in the review protocol.

#### *3.1.3. Specifying the research question or questions*

The most important part of the review process and the base for all other activities is the specification of the research question(s) which will define studies to include or exclude from the review and the data that should be extracted from them. The final review report will answer the stated review question(s).

The research questions in the domain of SE, according to Kitchenham [1], may concern for example the effect or impact of technology, cost and risks et cetera. The type of the question usually determines the guidelines and the procedures to be used. In contrast to defining the finite set of types of research question, our opinion is that it is better to use well defined guidelines on how to create a well formatted and structured question. Such question should be important to researches, in terms of changing the current SE practice, confirming its value or identifying discrepancies between reality and common beliefs. Finally, the right question could also serve the researches to identify and scope the future research.

Usually, authors define more than one research question or they define one high-level research question and then break it down to several more specific and concrete questions. For example, in order to characterize software architecture changes by means of a systematic review, Williams and Carver [3] created the following high-level question: Can a broad set of characteristics that encompass changes to software architectures be identified using the current software engineering body of knowledge and be used to create a comprehensive change

assessment framework? Additionally, the authors created five more specific questions along with accompanying motivation. The specific questions were:

- ✓ What are attributes of existing software change classification taxonomies?
- ✓ How are software architecture elements and relationships used when determining the effects of a software change?
- ✓ How the architecture is affected by functional and non-functional changes to the system requirements?
- ✓ How is the impact of architecture changes qualitatively assessed?
- ✓ What types of architecture changes can be made to common architectural views?

Another approach is to create a single research question, but in order to clarify its boundaries several complementary research questions could be created. An example of such approach is given by Staples and Niazi in [10].

The research questions also depend on the type of review which according to Noblit and Hare [11] could be integrative or interpretative. According to Dybå and Dingsøy [8] the difference between integrative and interpretative reviews is that integrative reviews are concerned with combining or summarizing the data for the purpose of creating generalizations, and interpretative reviews achieve synthesis through combination of concepts identified in the primary studies into a higher-order theoretical structure. This division could be aligned with the principles of “right questions” mentioned earlier in this chapter.

According to Petticrew and Roberts [2] it is a good way to start the question writing process by breaking it down into sub-questions. If the review aims to answer a question about the effectiveness, the authors suggest using a model called PICOC, defining a population, intervention, comparison, outcomes and context. These criteria were accepted in Kitchenham’s guidelines and discussed in SE viewpoint as follows:

- ✓ Population in the terms of SE could assume wide range of roles or groups and even areas, from novice testers, experienced software architects to for example control systems. As the number of undertaken primary studies in the field of SE is relatively small (comparing to other fields), the restrictions on the population should be avoided.
- ✓ Intervention should define a software methodology/tool/technology/procedure that the authors are interested in reviewing and that

should address specific issue that is in the focus of the research. Basically, intervention is the concept that is going to be observed in the context of the planned systematic review.

- ✓ Comparison/control is any software engineering concept used to compare the intervention with. The used control must be properly described.
- ✓ Outcomes important to practitioners should be specified, without using surrogate measures that may be misleading.
- ✓ Context refers the circumstances and facts related to the research place (e.g. academia vs. industry), participants taking part (e.g. practitioners, consultants, students) and performed tasks (e.g. small scale, large scale). There are many examples of unrepresentative experiments, i.e. the experiments that are undertaken in the academia using students and small scale tasks, and these should be excluded from serious systematic reviews.

#### 3.1.4. Developing a review protocol

Detailed review protocol is the most important output of the planning phase. It should describe the methods planned to be used in the process. The creation of the protocol prior to the review itself will reduce possible bias. Also, as Staples and Niazi [7] claim, protocol often insinuates the structure of the final report.

As the protocol contains the definition of the whole systematic review process, usually it is hard to predict all obstacles in it. That is why some authors, like Staples and Niazi [7], discuss that the protocol is a subject of constant changes through the whole systematic review process, while other, like Kitchenham suggests that aspects like search terms, selection criteria and data extraction procedures should be piloted during its development.

The full list of elements of the review protocol is defined in [1] and it includes ten elements: *background, research questions, search strategy, study selection criteria, selection procedures, quality assessment elements, data extraction strategy, data synthesis, dissemination and time plan.*

Taking into consideration the discussions from the other authors, several stated elements are especially important. For example Dybå and Dingsøy [8] discuss that explicit inclusion and exclusion criteria (which should specify the types of study designs, interventions, populations and outcomes that will be included in the review) and a systematic search strategy (which should specify the keyword strings and the bibliographic sources defined in a such way to ensure good topic coverage) are of the most importance. They also state that sometimes it is even

necessary to perform a search of the key journal and conference proceedings by hand to identify relevant studies that are not fully indexed. On the other hand, some authors put focus on quality assurance elements, or on planning to be critical in order to mitigate risks of researcher bias [1] or in order to support the practical conduct of the systematic review [7].

In order to make the process of development of review protocol easier, Kitchenham gave an example protocol that can be used for a tertiary study review. On the other hand, Biolchini et al. [4] created a Systematic Review Protocol Template which, even based on the first version of the Kitchenham's guidelines, covers majority of concepts and could be used as a starting point in creating a review protocol. Except the mentioned guidelines, protocol was also based on the systematic review protocols developed in the medical area and on the example found in Protocol for Systematic Review by Mendes E. and Kitchenham B., 2004. (as cited by Biolchini). Every concept in Biolchini's template is described in detail and a pilot study was conducted in order to evaluate the developed protocol template. The results of the study showed that usage of template has significantly shortened the time spent on planning against the review execution time.

The Systematic Review Protocol Template created by Biolchini et al. [4] is composed of five main parts. The original template is presented in Table 2. without any changes.

Table 2. Systematic Review Protocol Template [4]

<b>1. Question Formularization</b>	
1.1.	Question Focus
1.2.	Question Quality and Amplitude <i>Problem, Question, Keywords and Synonyms, Intervention, Control, Effect, Outcome Measures, Population, Application, Experimental Design</i>
<b>2. Sources Selection</b>	
2.1.	Sources Selection Criteria Definition
2.2.	Studies Languages
2.3.	Sources Identification <i>Sources Search Methods, Search String, Sources List</i>
2.4.	Sources Selection after Evaluation
2.5.	References Checking
<b>3. Studies Definition</b>	
3.1.	Studies Definition <i>Studies Inclusion and Exclusion Criteria Definition, Studies Types Definition, Procedures for Studies Selection</i>
3.2.	Selection Execution <i>Initial Studies Selection, Studies Quality Evaluation, Selection Review</i>
<b>4. Information Extraction</b>	
4.1.	Information Inclusion and Exclusion Criteria Definition
4.2.	Data Extraction Forms

4.3.	Extraction Execution <i>Objective results Extraction: Study Identification, Study Methodology, Study Results, Study problems; Subjective Results Extraction: Information through authors, General Impressions and Abstractions</i>
4.4.	Resolution of divergences among reviewers
<b>5. Results Summarization</b>	
5.1.	Results Statistical Calculus
5.2.	Results Presentation in Tables
5.3.	Sensitivity Analysis
5.4.	Plotting
5.5.	Final Comments <i>Number of studies, Search, Selection and Extraction Bias, Publication Bias, Inter-Reviewers Variation, Results Application, Recommendations</i>

### 3.1.5. Evaluating the review protocol

The evaluation of the review protocol is highly recommended in order to improve its quality. There are several methods that could be in this research step:

- ✓ author's review (not recommended)
- ✓ peer review
- ✓ review by supervisor (appropriate for PhD students)
- ✓ review by external experts (the best option)
- ✓ test of protocol execution

Review by external experts is probably the best option, but it usually depends on the financial construction of the review project. In this case, the group of external experts is reviewing the protocol, and the same group could review the final outputs of the project.

Test of protocol execution is a good and widely used alternative method. In this case, the review of protocol is executed by performing a full cycle of systematic review (following the protocol) but on a reduced set of selected sources. If gained results are not suitable, or if any phase of the review reveals unexpected problems, the new version of the protocol must be created.

### Conducting the review

According to Kitchenham's guidelines, conducting the review phase consists of five obligatory stages. This phase takes most of the researcher's time, and although all five stages are important, identification of research and selection of primary studies will determine the rest of the reviewing process. In this phase the predefined protocol should be followed and

the phase should result in data extracted, summarized and ready for dissemination.

The summary of each stage is presented below and is based on guidelines presented in [1] and on additional discussions from other authors which are cited in the text.

### 3.1.6. Identification of research

The result of identification of available research will be a list of all reported publications relevant to defined research questions and obtained by performing defined search strategy.

Search strategy, defined in review protocol, should be open for external review and be stated to allow research replication. It should break down a research question identify initial search strings by taking into account the population, intervention, control, outcomes, context and the design of the study. Also, strategy should include synonyms, abbreviations and different spelling options. Finally results should be gained from digital libraries, but also from journals, grey literature, research registers and the Internet should be obtained.

The process of definition of search strategy is usually iterative and should benefit from preliminary searches, trial searches and consultations with experts in the field.

In order to address publication bias (the problem of researchers tend not to publish negative results) and not to allow it to become a systematic bias, Kitchenham suggests that it is important to take appropriate steps. For example scanning the grey literature, conference proceedings and contacting domain experts could result in addition of studies with “negative” results.

As the number of identified primary studies could be large (some authors, for example Unterkalmsteiner et al. [12] have identified more than 10.800 publications), the appropriate reference management software should be used to keep a record on all of them along with the links to the potentially useful full papers.

In order to be transparent and replicable, the whole SLR process should be documented. All search results, starting from initial unfiltered set should be backed up for possible reanalysis. As presented in Table 3., Kitchenham [1] proposed the documenting procedures in accordance to data source.

Table 3. Search process documentation procedures

Data source	Documentation
Digital Library	Database name Search strategy Date Covered years
Journal hand Searches	Journal name Covered years Issues excluded from search
Conference proceedings	Proceedings title Conference name (if different) Title translation (if necessary) Journal name (if published in a journal)
Efforts to identify unpublished studies	Researches contacted (contact details) Research web sites searched (date and URL)
Other sources	Date of search URL Any other specific information important to the search

Source: [1]

When attempting to perform a comprehensive search Brereton et al. [13] identified seven electronic sources as the most relevant sources to Software Engineers, and they also discuss about considering the use of additional sources (\*) from publishers or bibliographical databases:

- ✓ IEEEExplore
- ✓ ACM Digital library
- ✓ Google scholar
- ✓ Citeseer library
- ✓ INSPEC
- ✓ ScienceDirect
- ✓ EI Compendex
- ✓ \*SpringerLink
- ✓ \*Web of Science
- ✓ \*Scopus

Unfortunately, the search of many relevant journals should be performed manually, but is also important as a part of a search process. The usual way to identify relevant journals is to read papers reference lists or by searching the Internet. Several authors also tried to identify a list of relevant journals and conferences in the field of software engineering. For example, combining the recommendations from [14], [1], the list of relevant journals and conferences (ordered alphabetically) could be:

- ✓ ACM Transactions on Software Engineering Methodology (TOSEM)
- ✓ ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)
- ✓ Empirical Software Engineering (EMSE)
- ✓ Evaluation and Assessment in Software Engineering (EASE)
- ✓ IEEE Computer
- ✓ IEEE Software
- ✓ IEEE Transaction on Software Engineering (TSE)
- ✓ Information and Software Technology (IST)
- ✓ International Conference on Software Engineering (ICSE)
- ✓ Journal of Software: Evolution and Process (JSEP)
- ✓ Journal of Software: Practice and Experience (SP&E)
- ✓ Journal of Systems and Software (JSS)

### 3.1.7. Selection of primary studies

As the initial identification of available research usually ends up with a large number of irrelevant articles, the application of the inclusion criteria will result in relevant ones, and the application of exclusion criteria on this smaller set will identify those that do not meet these extra conditions. This process is called selection of primary studies. Although the inclusion and exclusion criteria are commonly based on research question they can be defined based on study types, and change of selection criteria defined in review protocol should be avoided if possible.

Study selection is a multistage and iterative process. If the number of initially obtained studies is large, the authors usually start with simple criteria and, for example, in the first iteration include/exclude studies only by reading a title. In the second iteration the abstract is read and finally, full papers are read. Two study selection processes are shown in Figure 1.[12] and Figure 2.[8].

However, some authors advocate more strict approach. For example, Brereton et al. [13] advice the researchers to exclude studies by means of reading the title and the abstract only if there are no doubts that study can be excluded. Otherwise, they point out that they have learned out of the experience that “the standard of IT and software engineering abstract is too poor to rely on when selecting primary studies”, and they advise on reviewing the conclusions as well. Of course, final set of selected papers should be reviewed in detail.

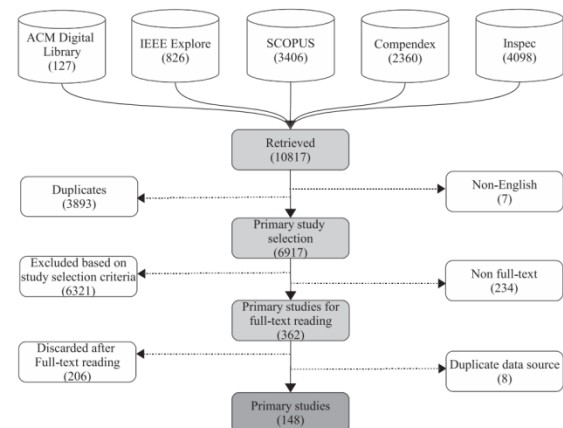


Figure 1. Example of study selection process [12]

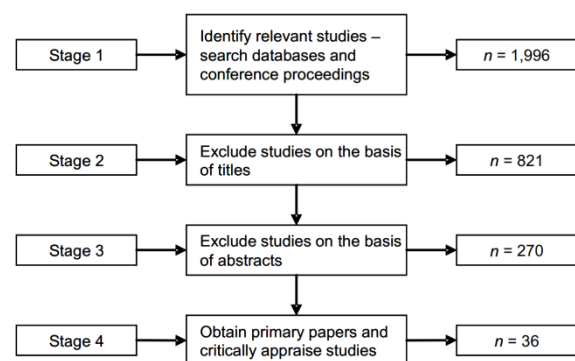


Figure 2. Example of study selection process [8]

Kitchenham is familiar with general instructions on keeping the list of the excluded papers, but she suggests that totally irrelevant papers should be excluded first (for example, papers that have nothing to do with Software Engineering) and then, while analyzing other papers, the list of exclusions should be kept updated along with the reasons of exclusion.

In order to increase the reliability of inclusion decisions it is possible to perform the same process by two or more researches. The Cohen Kappa coefficient [15] could be used to measure the agreement between the researches. If there is a disagreement then it should be discussed and resolved, but the initial value of Kappa statistics should be preserved in the final report and used for discussion and conclusions. Alternatively, using test-retest approach second and other researches could evaluate a random set of the primary studies.

On the other side, a PhD student can use one of the following methods to increase the reliability of inclusion decisions:

- ✓ consultation with advisor
- ✓ consultation with expert panel or other researcher
- ✓ re-evaluation of a random set of the primary studies by the test-retest approach
- ✓ re-evaluation of a random sample by other researcher while publishing a paper on the subject

Advisor usually helps student to choose an appropriate method and if decided so, the advisor can review the inclusion decisions or help student in order to find external experts or perform other stated methods.

### 3.1.8. Study quality assessment

The quality assessment of the studies is also important stage in this phase. To make sure that the study findings are relevant and unbiased, the quality of each primary study is analyzed and assessed in order to be finally included in data extraction and reporting process. However, this is not a trivial task as there is no agreement on the definition of study quality which according to Centre for Reviews and Dissemination [9] mainly depends on the type of the study. However, the same guidelines also state that the following elements should be assessed regardless of the study type:

- ✓ appropriateness of study design to the research objective
- ✓ risk of bias
- ✓ choice of outcome measure
- ✓ statistical issues
- ✓ quality of reporting and intervention
- ✓ generalizability

Mentioned elements don't have the same importance in every case, but the authors usually agree that the risk of bias (also known as internal validity) is pernicious as it can easily obscure intervention effects. Generalizability (also known as applicability or external validity) considers the extent to which a study is generalizable and how closely a study reflects a practice [9]. Additionally, Kitchenham states that quality assessment should be used to:

- ✓ provide more detailed inclusion/exclusion criteria
- ✓ provide explanation for differences in study results
- ✓ weight the importance of individual studies for overall synthesis
- ✓ guide the interpretation and further research.

In this process, Kitchenham also finds that three concepts are important and most closely related to the study quality. She defines them as follows:

Table 4. Quality concept definitions[1]

Term / Synonyms	Definition
Bias / Systematic error	Biased results systematically differ from 'true' results. Valid results are unbiased results.
Internal validity / Validity	The degree of prevention of systematic error by means of study design and conduct. Internal validity precedes external validity.
External validity / Generalizability, Applicability	The degree of applicability of the observed effects outside of the study.

Checklist is usually used quality assess instrument as it ensures that all studies are analyzed in standardized way. When creating a checklist, in the focus of current research it should cover the bias and validity. There are several types of biases that should be addressed in a checklist. The identified types of biases along with definition and protection mechanisms are defined in [1] and include *Selection/allocation bias*, *performance bias*, *measurement/ detection bias*, *attrition/exclusion bias*. Some of the protection mechanisms can include randomization by computer program instead of experimenter choice, the use of different experiments in order to replicate the studies, outcome assessors to the treatment could be blinded, reporting the withdraw reasons, running sensitivity analysis et cetera [1].

In addition to these, Higgins and Green [16] emphasize reporting bias and also recognize other biases. By reporting bias they discuss systematic differences between reported and unreported findings, and by other biases they presume other sources of bias that are relevant in certain circumstances (for example language etc.).

According to Kitchenham, checklist should also include consideration of biases and problems of validity that can occur at any research stage. Reviewing available papers on the subject of checklists creation for quantitative studies, and noticing that authors focus on different set of questions, Kitchenham and Charters [1] created an accumulated list of 59 questions and organized it with respect to a study stage and a study type. These questions cover four mentioned stages and could be used for quantitative empirical studies, correlation (observational) studies, surveys and experiments. The same process was conducted on qualitative studies, and resulted in 18 questions that could be used. These example checklists, which we highly recommend, should not be used literally, but rather as a pool of questions. The appropriate questions could be taken from the pool for each specific study.



Quality instruments as well as the use of quality data should be defined in review protocol. In general, there are two rather different but not mutually exclusive ways: (1) to assist primary study selection and (2) to assist data analysis and synthesis.

There are several limitations that the authors should be aware of when attempting to perform a quality analysis of different studies. First, primary studies could be poorly reported, but the lack of report doesn't necessarily mean a leak in a procedure. According to Petticrew and Roberts [2] the quality checklists should address methodological quality and not reporting quality. If reporting quality is poor, the researchers should contact the authors of the study and try to obtain more information. Additionally, Kitchenham discusses that limitation could be a poor evidence of connection between factors affecting validity and actually study outcomes, and that sometimes the statistical analysis is not possible to correct as there is usually no access to the original data.

Finally, authors usually point out all undertaken quality assessment procedures and measures, but only to the level of details that is suitable for the targeting publication. For the further reading, we recommend some simple examples of quality assessment of SE studies presented in [8], [17], [18] or [19] and especially [12].

### 3.1.9. Data extraction and monitoring

This phase aims to record the appropriate information from selected papers by means of use of data extraction forms. These forms should be designed to collect information accurately and without bias. Extracted information is needed in addressing the review questions as well as the study quality criteria. As the quality criteria could be used to identify inclusion/exclusion criteria or/and as a part of the data analysis, in the first case, the data extraction forms should be separated, and in the second case, a single form can be used [1]. In any case, the same authors recommend that the forms should be piloted during the protocol definition phase, and all researchers who will use the forms should take part in the pilot study in order to assess completeness of the forms along with possible technical issues.

There are several elements that are considered to be common to all forms in order to provide standard information. According to Kitchenham [1] these elements could be:

- ✓ name of the reviewer
- ✓ date of data extraction
- ✓ title, authors, journal, publications details
- ✓ space for additional notes

Combining the examples presented in [1] and [20] we can conclude that in general, data extraction form could include parts (sections) as presented in Table 5. It is important to notice that the column Additional notes was used to present additional info on template elements, but it should also be used in extraction forms to present additional info on the extracted data.

Similarly as in the process of applying inclusion and exclusion criteria, there are different methods that could be performed to extract the data and to fill the extraction forms. In guidelines Kitchenham recommends that data extraction should be performed by two or more researchers, but as stated in [17], she finds that it is in practice useful that one researcher extracts the data and the other one checks the extraction. If several researchers are performing a data extraction, the results should be compared, aligned and if necessary discussed. On the other side, if researchers are performing extraction on different sets of primary studies, it is important to ensure that it is done in a consistent manner by employing some cross-checking activities. Additionally, Staples and Niazi [7] recommend that the whole process should be done in an iterative manner. PhD students will usually need some help from advisor or other experts to randomly check their extracted data or they will perform a re-test on a part of primary studies.

Table 5. Data collection form template

Data item	Value	Additional notes
<b>Extraction information</b>		
Data extractor		
Data checker		
Date of ext.		
<b>General study information</b>		
Study id		
Title		
Publication details		Including authors, journal etc.
<b>Questions to answer review questions</b>		
Question 1		These questions could aim to obtain numerical or descriptive data. Each review question could be covered by more questions in data extraction form.
Question 2		
Question <i>n</i>		
<b>Questions to assess study quality</b>		
Question 1		These questions should be related ONLY to data analysis. Questions related to inclusion/exclusion criteria should be stated on separate form.
Question 2		
Question <i>m</i>		
<b>Data summary</b>		
Question 1		These questions could aim to collect summary information from the observed study.
Question 2		
Question <i>p</i>		

It is very important not to include multiple studies with the same data as well as to contact the authors for more information when analyzing poorly reported studies.

Finally, the authors should consider using electronic forms as they can be useful in subsequent data analysis, especially if the extracted data will be set of numerical values and if statistical or meta-analysis will be performed.

An interesting example of data extraction process could be found in [12], the example of filled extraction forms could be found in [20] and [21] and an example of data extraction forms with a short review on the process can be found in almost all papers mentioned in this chapter.

### 3.1.10. Data synthesis

During data synthesis step, the extracted data are summarized. The synthesis, according to [9], can be descriptive (narrative) and quantitative, and these two are not mutually exclusive. According to Brereton et al. [13] the synthesis in SE reviews is likely to be qualitative. The synthesis should take into consideration the strength of evidence and consistency in order to draw reliable conclusions. The approach used in synthesis is also defined in research protocol, and it depends on the type of research question as well as on type of the available studies and the quality of data.

Regardless of its type, the synthesis usually includes a summary of included studies. The table form is commonly used in presenting the studies included in the review as it structures all important details. In same or in another table the elements of study quality and risk of bias should be presented as well. This process, even descriptive, should be explicit and rigorous, and should determine if studies are reliable for synthesization [9]. The extracted data should be presented in a manner that is consistent with review questions and focusing the outcomes [1].

Synthesizing results of qualitative studies means an integration of materials written in natural language, with significant possibility of having to understand different meanings of the same concepts as they were used by different researchers [1]. In [11] the authors propose three approaches to the synthesis of qualitative studies:

- ✓ Reciprocal transaction – translation of cases of studies with similar objective into each of other cases in order to create an additive summary.
- ✓ Refutational synthesis – translation of studies along with corresponding refutational studies in order to analyze the refutations in detail.
- ✓ Line of argument synthesis – first, the individual studies which focus the part of some problem

are analyzed and then the set is analyzed as a whole in order to get broader conclusion on the addressed problem.

According to Petticrew and Roberts [2] the narrative synthesis can be performed in several ways, but most common one is to separate it into three distinct steps by: (1) organizing the description into logical categories, (2) analyzing the findings within each of the categories and (3) synthesizing the findings across all included studies. The mentioned authors argue that there is no firm guidance on how to organize the categories and that this could be done according to: intervention, population, design, outcomes etc. The second step involves a narrative description of the findings for each study. This description may vary in length and in the level of details. Finally, authors discuss the cross-study synthesis and state that it usually starts with a simple description of the uncovered information, then the summary information on the effect of mediating variables (if any) can be presented, and at the end it describes the results of the individual studies. The main goal of cross-study synthesis is to produce an overall summary of study findings taking into considerations the quality and other variations.

Additionally, some authors describe several other synthesis methods which could be used:

- ✓ Best evidence synthesis – “combines the meta-analytic approach of extracting quantitative information in a common standard format from each study with a systematic approach to the assessment of study quality and study relevance”.
- ✓ Vote counting – the easiest approach which simply compares the number of positive and negative results on specific issue. This approach is usually inappropriate to use as it has many disadvantages.
- ✓ Cross-design synthesis – in theory combines the complementary strengths of experimental and non-experimental research – for example by adjusting the results of random controlled trials (RCTs) by standardizing RCT results to the distributions obtained from database analyses.

An example of applying a narrative synthesis is presented in [9] and can be seen in Fig. 3.

Quantitative data (as well as qualitative) should be presented in tabular form. The data should be comparable, and according to Kitchenham, it should include:

- ✓ sample size for each intervention,
- ✓ estimated effect size for intervention with standard error for each effect,

- ✓ difference and confidence interval in it regarding mean values for each intervention,
- ✓ units used for measuring the effect.

Different effect measures addressing different types of outcome are proposed in literature. Kitchenham refers to medical literature and she presents binary outcomes (which can be measured by effect measures like odds, risk, odds ratio (OR), relative risk (RR), absolute risk reduction (ARR)) and continuous data (which can be measured by mean difference, weighted mean difference (WMD) or standardized mean difference (SMD)).

Apart from narrative description of results, qualitative results are usually presented and summarized in a table. It is important to explain how the data answers the research questions [13]. On the other hand, quantitative results are usually presented by forest plot [1] and, of course, additionally narratively discussed and related to the research questions. Different approaches and methods of systematized data presentation could be found in [9].

### ***Reporting the review***

The systematic review report should be written in a form suitable for selected dissemination channels and target audience. The summary of possible activities in the reporting phase is presented below and is based on guidelines presented in [1] and on additional discussions from other authors which are cited in the text.

#### ***3.1.11. Specifying dissemination strategy and mechanisms***

Specifying dissemination strategy and mechanisms is usually performed during the project commissioning activities, or if there was no commissioning phase, then dissemination strategy and mechanisms should be defined in the review protocol. Kitchenham discuss that apart from disseminating the results in academic journals and conferences, scientists should consider performing other dissemination activities that might include direct communication with affected bodies, publishing the results on web pages, posters or practitioner-oriented magazines etc.

If the results are to be published in a conference or journal, or any other publication with restricted number of pages, then the reference to a document (technical report, PhD thesis or similar) that contains all information should be provided.

#### ***3.1.12. Formatting the main report***

Kitchenham adopted the suggested structure of systematic review report given in CRD's guidelines from 2001. Although the original guidelines (from 2001) are updated in [9], the version presented by Kitchenham is sufficient in the field of software engineering. She also distinguishes the report which is to be published in technical reports and journals from the report which is to be published in PhD theses. The report structure proposed by Kitchenham is presented in detail in [1]. It distinguishes the element used both in publications and in PhD theses from those used only in publications. Table 6. brings an overview of Kitchenham's report structure.

*Table 6. Report structure [1]*

<b>Section</b>	<b>Subsection / Scope</b>
Title*	
Authorship*	
Executive summary / Structured abstract*	The context of the research questions.
	Objectives are addressed in questions.
	Methods explanation for review phases.
	Main findings including analyses.
	Conclusions and future research.
Background	Reasoning the need for the review. Report on previous reviews.
Review questions	Review questions should be specified and explained.
Review methods	Data sources and search strategy
	Study selection
	Study quality assessment
	Data extraction
	Data synthesis
In/excluded studies	Criteria for inclusion and exclusion. Excluded studies with exclusion reasons.
Results	Main findings include primary studies description, quantitative summaries and meta-analysis results.
	Sensitivity analysis.
Discussion	Principal findings
	Strengths and limitation of the review. Reasoning on differences in relation to other reviews.
	Meaning, magnitude and applicability of findings.
Conclusions	Practical implications and recommendations for SE.
	Implications for future research.
Acknowledgements*	Credits to all other contributors.
Conflicts of interests	
References and Appendices	

### 3.1.13. Evaluating the report

The final step in the SLR process concerns the report evaluation activities. The type of the publication mostly influences the type of evaluations that are to be performed. Independent peer reviews, reviews by the supervisor and doctoral committee are the common evaluation methods for scientific papers and doctoral publications respectively. Similarly, if the publication is technical review it should be submitted to an independent evaluation by field experts. In this case the same expert panel used to evaluate review protocol can do the review of the final report as well. If the review results are negative, the corresponding activities or the whole review process should be repeated in order to get unbiased results.

## 4. Advantages and disadvantages of SLR

The main advantages of systematic literature review method are its definition, universality and its generality. It means that the method is well defined, it can be applied to any research problem and includes all possible study sources and it provides techniques of generalization such as meta-analysis that will extract information normally unavailable in singly study [1]. Throughout the paper, we emphasized other advantages of SLR over simple review, making this method result in more reliable and unbiased findings.

At the same time, the method comprehensiveness becomes its major disadvantage especially in terms of time. The method requires much more effort and time if compared to other simpler review methods. Also, there is a large number of review points which normally include third persons and this significantly contributes the overall process duration [7]. Moreover, the usage of meta-analysis tends to detect small or not important biases [1]; the authors should perform and understand rather complex activities, which according to Biolchini, make the SLR process specially complicated in domain of SE [4]. Same authors, in order to help other researchers, created the process description and protocol template as they find the overall process too difficult to conduct.

Further, solid literature coverage of the observed phenomenon is prerequisite for this method, thus making it unusable when exploring new or revolutionary phenomena. Finally, the fact that even experienced authors often change review protocol during the method conduction are the target of critics and also makes it hard to document the whole process.

## 5. Conclusions on SLR

This paper aims to present the state-of-the-art in performing a complex and time consuming method of Systematic Literature Review in the field of Software Engineering. The process of Systematic Literature Review is not easy to perform, but the general opinion of the authors whose findings, recommendations and conclusions are presented in this paper, is that this method is useful and could be used to decrease the biases and to increase the review quality. Authors also note that the usage of this method has significant obstacles in the field of software engineering in comparison to other fields. The main differences are the mainly qualitative studies to be reviewed in the SE, the lack of centralized index of existing systematic reviews and the overall literature searching problem raised by many different sources, with a different and questionable quality. In order to overcome the mentioned obstacles, the authors who performed the SLR in the field of the SE suggest, and completely agree, that the scope of the review should be limited by choosing a clear and a narrow research questions and that the whole process should be in advance well defined by putting a considerable effort into creating a feasible review protocol.

As the method of SLR still emerges in the field of software engineering, the idea of publishing the replications of existing systematic reviews is welcomed, along with the idea of creation of centralized index of existing literature reviews.

The paper covers the whole SLR process as defined by Kitchenham and Charters [1], including all three phases: review plan, review conduction and review report. The findings, observations and recommendations from other influential authors in this field are also summarized and aggregated in this paper. The mentioned process is observed from the perspective of appliance in the field of SE and supplemented by the important precautions measures that should be undertaken by the authors performing it.

Finally, this paper could also be a guide to those who are performing a SLR for the first time, as it gives the definitions, directions and examples and references that are sufficient for the different types of reviews typically performed by the software engineers.

## References

- [1] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3,” Keele University and University of Durham, Technical report EBSE-2007-01, 2007.
- [2] M. Petticrew and H. Roberts, *Systematic reviews in the social sciences: a practical guide*. Malden, MA: Blackwell Pub., 2005.
- [3] B. J. Williams and J. C. Carver, “Characterizing software architecture changes: A systematic review,” *Inf. Softw. Technol.*, vol. 52, no. 1, pp. 31–51, 2010.
- [4] J. Biolchini, P. Gomes Mian, A. Candida Cruz Natali, and G. HortaTravassos, “Systematic Review in Software Engineering,” PESC, Rio de Janeiro, Technical report RT - ES 679 / 05, 2005.
- [5] B. Kitchenham, “Procedures for Performing Systematic Reviews,” Software Engineering Group; National ICT Australia Ltd., Keele; Eversleigh, Technical report Keele University Technical Report TR/SE-0401; NICTA Technical Report 0400011T.1, 2004.
- [6] P. Mian, T. Conte, A. Natali, J. Biolchini, and G. Travassos, “A Systematic Review Process for Software Engineering,” in *ESELaw '05: 2nd Experimental Software Engineering Latin American Workshop*, 2005.
- [7] M. Staples and M. Niazi, “Experiences using systematic review guidelines,” *J. Syst. Softw.*, vol. 80, no. 9, pp. 1425–1437, 2007.
- [8] T. Dybå and T. Dingsøy, “Strength of evidence in systematic reviews in software engineering,” 2008, pp. 178–187.
- [9] Centre for Reviews and Dissemination, University of York, *Systematic reviews: CRD’s guidance for undertaking reviews in health care*. York: Centre for Reviews and Dissemination, 2009.
- [10] M. Staples and M. Niazi, “Systematic review of organizational motivations for adopting CMM-based SPI,” *Inf. Softw. Technol.*, vol. 50, no. 7–8, pp. 605–620, 2008.
- [11] G. W. Noblit and R. D. Hare, *Meta-ethnography: synthesizing qualitative studies*. London: SAGE, 1988.
- [12] M. Unterkalmsteiner, T. Gorschek, A. K. M. M. Islam, C. K. Cheng, R. B. Permadi, and R. Feldt, “Evaluation and Measurement of Software Process Improvement - A Systematic Literature Review,” *IEEE Trans. Softw. Eng.*, 2012.
- [13] 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,” *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, 2007.
- [14] J. Hannay, D. Sjöberg, and T. Dyba, “A Systematic Review of Theory Use in Software Engineering Experiments,” *IEEE Trans. Softw. Eng.*, vol. 33, no. 2, pp. 87–107, 2007.
- [15] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,” *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968.
- [16] J. P. Higgins and S. Green, Eds., *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from <http://www.cochrane-handbook.org/>, 2011.
- [17] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering – A systematic literature review,” *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.
- [18] B. Kitchenham, P. Brereton, M. Turner, M. Niazi, S. Linkman, R. Pretorius, and D. Budgen, “The impact of limited search procedures for systematic literature reviews - A participant-observer case study,” 2009, pp. 336–345.
- [19] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, and S. Linkman, “Systematic literature reviews in software engineering - A tertiary study,” *Inf. Softw. Technol.*, vol. 52, no. 8, pp. 792–805, 2010.
- [20] M. Jørgensen, “Estimation of Software Development Work Effort: Evidence on Expert Judgment and Formal Models,” *Int. J. Forecast.*, vol. 23, no. 3, pp. 449–462, 2007.
- [21] T. Dybå and T. Dingsøy, “Empirical studies of agile software development: A systematic review,” *Inf. Softw. Technol.*, vol. 50, no. 9–10, pp. 833–859, 2008.