Empir Software Eng (2017) 22:2852–2891
DOI 10.1007/s10664-016-9492-y

On the pragmatic design of literature studies in software engineering: an experience-based guideline

Marco Kuhrmann[1] · **Daniel Méndez Fernández**[2] · **Maya Daneva**[3]

Published online: 6 January 2017

Abstract Systematic literature studies have received much attention in empirical software engineering in recent years. They have become a powerful tool to collect and structure reported knowledge in a systematic and reproducible way. We distinguish systematic literature reviews to systematically analyze reported evidence in depth, and systematic mapping studies to structure a field of interest in a broader, usually quantified manner. Due to the rapidly increasing body of knowledge in software engineering, researchers who want to capture the published work in a domain often face an extensive amount of publications, which need to be screened, rated for relevance, classified, and eventually analyzed. Although there are several guidelines to conduct literature studies, they do not yet help researchers coping with the specific difficulties encountered in the practical application of these guidelines. In this article, we present an experience-based guideline to aid researchers in designing systematic literature studies with special emphasis on the data collection and selection procedures. Our guideline aims at providing a blueprint for a practical and pragmatic path through the plethora of currently available practices and deliverables capturing the dependencies among the single steps. The guideline emerges from various mapping studies and literature reviews

Communicated by: Jeffrey C. Carver

✉ Marco Kuhrmann
kuhrmann@acm.org

Daniel Méndez Fernández
daniel.mendez@tum.de

Maya Daneva
m.daneva@utwente.nl

[1] Mærsk Mc-Kinney Møller Institute, Section Software Engineering, University of Southern Denmark, Campusvej 55, 5230, Odense M, Denmark

[2] Institute for Informatics, Software, Systems Engineering, Technical University of Munich, Boltzmannstr. 4, 85748 Garching, Germany

[3] University of Twente, Drinerlolaan 5, 7522 AE, Enschede, The Netherlands

conducted by the authors and provides recommendations for the general study design, data collection, and study selection procedures. Finally, we share our experiences and lessons learned in applying the different practices of the proposed guideline.

Keywords Systematic literature review · Systematic mapping study · Empirical software engineering · Guideline proposal · Lessons learned

1 Introduction

Systematic literature studies have received much attention in recent years as a powerful instrument to gather and structure reported knowledge in a systematic and reproducible way. We distinguish two types of secondary studies:

A *Systematic Mapping Study* (SMS; Petersen et al. 2008) is a method to build a classification schema for topics studied in a field of interest. By counting the number of publications for categories within a schema, the coverage and maturity of the research field can be determined. Graphical maps showing the number of publications in the different categories of the schema represent the study results. Mapping studies usually cover a broader range of publications as the analysis focuses on the key terms and abstracts of publications.

A *Systematic Literature Review* (SLR; also: Systematic Review, SR; Kitchenham et al. 2015) is a means to identify, analyze and interpret reported evidence related to a set of specific research questions in a way that is unbiased and (to a degree) repeatable. In contrast to mapping studies, systematic reviews usually cover a smaller, more specific range of publications while the analysis focuses on the details of the published contributions.

A mapping study is therefore often used to provide (and visualize) a big picture of a publication space while the systematic review is additionally concerned with analyzing and integrating the knowledge contained in the reviewed publications, as well as identifying inconsistencies among results, and areas that need more investigation. Both types of secondary studies (also applicable in combination) allow to share a structured overview of the publications in a specific research area and a common understanding of the state of reported evidence in topics along a given (or emerging) classification scheme. Since the initially proposed guidelines to conduct literature studies in software engineering (Kitchenham 2004), we, as a community, could collect and systematize the procedures required, and we could see a boost of secondary studies in the various international evidence-based software engineering venues. This indicates the value of such studies to the research communities.

Problem Statement Since researchers face a variety of challenges for which available guidelines do not yet give sufficient practical advice; they either comprise generic workflows or provide methods and techniques in a compendium-like style (Kitchenham et al. 2015; Petersen et al. 2015), or elaborate selected methods only, e.g., the effectiveness of certain selection procedures (Ali and Petersen 2014; Zhang et al. 2011). Hence, conducting a literature study still depends to a large extent on the expertise of the involved researchers. Furthermore, conducting literature studies, to a large extent, still lacks tool support (Hassler et al. 2016; Carver et al. 2013; Tell et al. 2016) thus making the research process as such difficult to implement; notably for novices. While working on a number of literature studies ourselves (Section 3), we experienced the following challenges to be the most critical ones worth deeper examination:

- How do we begin a secondary study, how do we build search strings adequate for given databases, and how can we control accurate results given the dependency to the expertise, experiences, and potential subconscious bias of the researchers?
- How do we deal with a large amount of data including hundreds or even thousands of potentially relevant papers to classify and structure, and how do we efficiently filter relevant results from irrelevant ones?
- How do we efficiently work in a distributed team? Which tools can we use to organize our (potentially distributed) way of working?

We experienced those challenges to concern mainly the design of a study (Kitchenham et al. 2015) and the data collection and study selection itself (Zhang et al. 2011), notably independent of whether it is conducted as a systematic review or a mapping study. The choice of one particular study approach or a combination thereof (as for instance found in Petersen et al. 2015 oftentimes) affects subsequent data analysis where the data is structured, classified, coded, and analyzed to draw conclusions in tune with the research questions.

Despite the criticality of the initial design and data collection steps, little practical advice is given on how to effectively cope with the mentioned challenges. Existing guidelines are either too generic (Staples and Niazi 2007), or they focus on *what* a design should accomplish rather than on *how* and *why* particular practices should be executed in a cost-effective way, and how these practices are interconnected with each other (see also our discussion in Section 4). In turn, for each literature study, researchers need to carefully design and outline the process from the beginning again and again, and they need to work out or even re-invent their own set of best practices.

Contribution In this article, we report on our own experiences in conducting systematic literature studies and contribute

- A detailed blueprint for the design, data collection, and study selection procedures steered by the aforementioned challenges.
- A set of practical lessons learned and supporting material readily available for use by other researchers approaching their own systematic literature studies.

We aim at supporting researchers, who already have a basic knowledge about the general guidelines, in their literature studies by providing a practical and pragmatic, experienced-based path through the available practices and deliverables capturing the dependencies among the single steps (Section 4). Researchers can directly reuse our blueprint to design and conduct their own domain-specific literature study and build on top the data analysis to answer their individual research questions.

Outline The remainder of this article is organized as follows: In Section 2, we present our experienced-based approach to design and set up a literature study. We describe our procedures as they emerged from our previously conducted studies. We also outline the handover to the data analysis, which depends on the type of the respective study (mapping study and/or systematic review) and the research questions previously defined. Our previously conducted studies from which we distill the blueprint are discussed in Section 3 along practical lessons we learned while conducting these studies. In Section 4, we finally discuss related work and position our guideline, before concluding our article in Section 5.

In the articles's appendix, we provide exemplary integrated workflows describing reusable standard workflows, and further complementing material.

2 Study design and data collection: An experience-based approach

We provide an experienced-based guideline to support the study design, and to perform the data collection, cleaning, and study selection procedures. For each step, we provide a guideline complemented with small inline examples. The guideline is organized in the three phases *Preparation*, *Data collection and Dataset cleaning*, and *Study selection*. Figure 1 provides a big picture of the whole process including the most important inputs and outputs for the respective phases. The figure also outlines the variations in the data analysis procedures that depend (in more detail) on whether it is a mapping study or a systematic review.

Our guideline presented in this article emphasizes the early stages of a literature study and constitutes a new building block in the methodical instrumentation of evidence-based software engineering (Kitchenham et al. 2015). A detailed discussion on the relation to existing guidelines and publications is provided in the related work in Section 4.

2.1 Preparation

The study preparation phase serves the purpose of setting up the study design including, inter alia, the definition of appropriate research questions, the choice of relevant literature databases, or the development of search queries. This phase relates to the *planning* step mentioned in Kitchenham et al. (2015) where, for instance, the protocol development is described. To set the scope of the search, inclusion and exclusion criteria need to be carefully outlined, and, if necessary, preliminary studies can be carried out to, among other things, support search string development or testing and improving the study design (see also test-retest procedures as mentioned in Kitchenham et al. (2015), or the quasi-gold standard search approach from Zhang et al. 2011). In the following, we describe the individual and minimum steps to be carried out during the preparation of a literature study and give examples.

2.1.1 *Research goals and research questions*

There is no silver bullet to define the goals of a literature study, as this strongly depends on the purpose of the study. In general, the primary goal of literature studies is to systematically

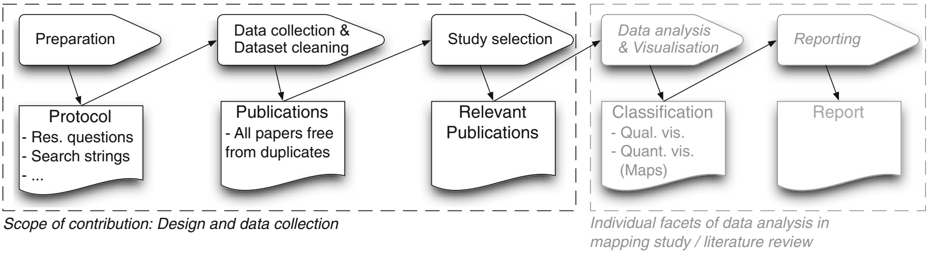


Fig. 1 Overview of the presented approach and scoping

Table 1 Exemplary standard research questions for literature studies

No.	Research question
1	Which/how many publications on [topic] are published?
2	Which/how many publications on [topic] are published over the years?
3	What is the scientific maturity of the publication set?
4	What is the contribution of the publication set?
5	What are observable mainstreams in the publication set?
6	What new approaches for [topic] are available?

collect reported knowledge in an area of interest. This can be done *in-breadth*, usually in scope of mapping studies (Petersen et al. 2008) that quantify selected aspects reported in literature, or *in-depth*, usually in scope of systematic reviews (Kitchenham et al. 2015) to analyze publications in detail. The purpose of a study eventually dictates the goals of the study, such as providing an overview of all relevant contributions dealing with a particular topic.

Independent of the respective goals, we have found some general research questions particularly worth considering in a literature study, as they help elaborating a big picture and providing relevant background information about the publication space. Table 1 summarizes such generic research questions, which could be answered in every literature study—regardless of the particular study's scope and selected topic.

The research questions in Table 1 address the general descriptive aspects present in every result set. Questions 1 and 2 aim at drawing a demographic picture to outline the current state of a field under investigation, i.e., providing information about publication quantity and frequency. This information can be instrumented to show the development over time of the studied domain and to analyze trends, for example, an emerging or a maturing domain (as exemplarily depicted in Fig. 2). The level of detail and data type (quantitative or qualitative) further depends on the respective study type.[1]

To direct the study towards its goal, i.e., a mapping study or a literature review, further standard questions can be asked that support the next steps in the study selection process. For instance, the scientific maturity addresses the classification according to the *research type facet* (Wieringa et al. 2005) to work out the level of evidence in the publications. A mature field should for example not only contain solution proposals, but also validation and evaluation research papers, and consequently experience reports (Fig. 2). The question for the result set's contribution aims at working out the different kinds of *contribution type facets* (Petersen et al. 2008) and their respective distribution in the publication population.

[1] Note that finding the "right" research question is a challenge and highly depends on the actual study type. For instance, Kitchenham et al. (2015) mention (standard) research questions for systematic reviews usually addressing the evaluation of impact and/effectiveness of certain paradigms, while mapping studies usually address more high-level questions with the purpose of providing some sort of categorization. The questions presented in Table 1 are addressing more the latter aspect, as this covers information available from all sorts of studies. Nonetheless, to plan and implement a literature study efficiently, Staples and Niazi (2007) make clear that narrowly defined research questions are key. We therefore recommend to use a combination of generic research questions (e.g., Table 1 to "get a feeling" about the result set) and specific narrow research questions—even for mapping studies.

For instance, does the result set contain models, theories, lessons learned, or frameworks? The remaining questions address further general aspects, such as observable streams in the result set. Such streams can become obvious by certain trends or accumulations of publications, e.g., outstanding number of solution proposals and, at the same time, no theories. Such a discussion can also be supported by applying further specific models, such as the *rigor-relevance model* proposed by Ivarsson and Gorschek (2011). Mainstreams can also be brought to light by studying the contents of the paper in more detail, e.g., by introducing *focus type facets* (Paternoster et al. 2014), which can also direct the in-depth investigation of a systematic review.

In summary, the standard research questions from Table 1 aim at providing a demographic overview of the study. Answering these questions shows how many publications have been published over time, about which topics they are, and which results they provide. These questions already provide a big picture of a research field, and they allow for getting a better understanding of the studies available in that field. Finally, these questions also help scoping the study and preparing the collection and selection procedures according to the overall study objectives. For example, an initial analysis of the demographic information helps checking the suitability of research questions and adjusting them if necessary.

2.1.2 Search strings

Once the scope of the study has been set, researchers need to reflect on proper search strings, which also depend on the domain under investigation. Depending on the precision of the search strings, the queries may produce inappropriate results, too much overhead, or just

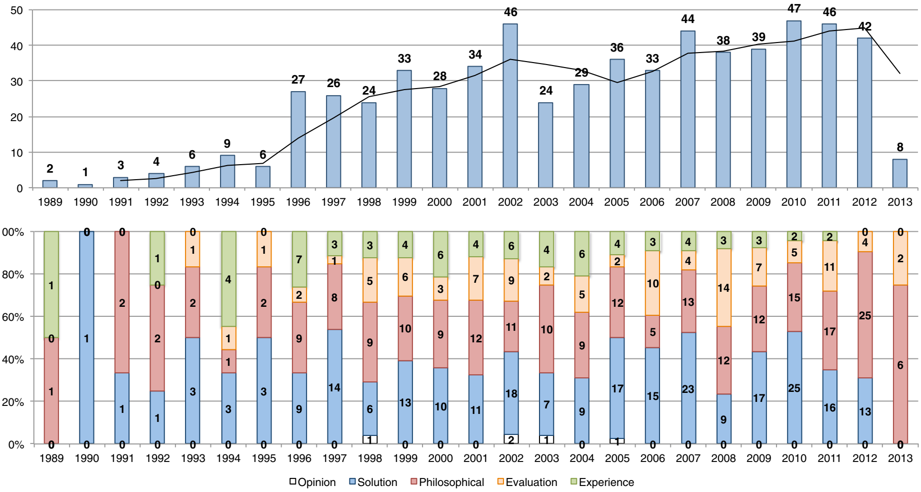


Fig. 2 Exemplary demographic distribution of publications over specific facets (addressing research questions 1, 2, and 4) as taken from Kuhrmann et al. (2015). This figure illustrates on top the number of publications over time and per year depicting publication trends. The bottom part indicates to the maturity of the result set by providing information about the research type facets

an incomplete result set. Therefore, search strings must be defined with care (Kitchenham et al. 2015), and search queries should always be tested prior to the actual search.[2] There exist some strategies to develop proper search strings, e.g.:

Snowballing One way to narrow down the search space in advance is to conduct a preliminary investigation of the field by relying on *snowballing* (Kitchenham et al. 2015). That is, the investigation starts by studying publications known in advance and by iteratively extending the known literature set by following the references provided therein. This procedure helps providing an initial overview of the publication space and key contributors, but very much depends on the expertise necessary to select an appropriate starting point (see also Section 3). However, as reported by Badampudi et al. (2015), manual search strategies compared to automatic ones are capable of producing "competitive" results regarding result set precision while, at the same time, avoiding vast overhead usually produced by automatic database searches.

Trail-and-Error Search One approach suitable to find and test search queries is the "Trail-and-Error Search". This approach relies on meta-search engines, e.g., Scopus or Google Scholar, and requires initial keywords or (partial) key phrases that are considered search query candidates for the "real" search. The purpose aims at iteratively narrowing down the list of potential candidates by checking whether:

- A search query returns a (potentially) meaningful result set.
- A keyword or a combination thereof returns hits (at all).
- A search query is of sufficient precision; for instance, if searching a particular domain, how many hits are not in the domain of interest?

Hence, a trail-and-error search serves two major purposes: First, it can be used to initially test and develop search queries, e.g., by determining which keywords might (not) generate useful results. Second, results from such test runs can be used to harvest reference publications to support manual search strategies (as for instance exercised in Theocharis et al. 2015). Although this approach can be seen as everything but a good scientific practice, it still helps taking the initial steps into the overall research design development—especially in domains in which few or no secondary studies are present to provide structure to the field of interest (as it for instance was the case in Ingibergsson et al. 2015).

2.1.3 Inclusion and exclusion criteria

Depending on the study's scope, result sets can contain a vast amount of potentially relevant publications. In the worst case, we experienced searches to yield in several thousands of

[2] Note that the construction of search strings also depends on the planned search strategy (see Section 2.2), since search stings for automated database searches have a different "layout" than those used for a curiosity-driven or trail-and-error search, e.g., using Google Scholar. Regardless of the search strategy, finding the proper key words is crucial. The most straight-forward approach to develop appropriate search strings is either to do a trail-and-error search or to call in domain experts. Alternatively, a preliminary study can be conducted to "test" the field of interest.

Table 2 Exemplary (standard) inclusion (I) and exclusion (E) criteria for literature studies

No.	Criterion	
1	I	Title, keyword list, and abstract make explicit that the paper is related to [topic].
2	I	The paper presents [topic]-related contributions, e.g., [topic list].
3	E	The paper is not in English [or any other language of interest].
4	E	The paper is not in the domain [domain name(s)].
5	E	The paper is a tutorial-, workshop-, or poster summary only.
6	E	The paper relates to [topic] in its related work only.
7	E	The paper occurs multiple times in the result set.
8	E	The paper's full text is not available for download.

hits. We doubt it should be questionable that several 10,000 hits cannot be treated seriously within an acceptable timeframe.[3] Therefore, researchers need to clean the dataset and to select the relevant studies. In order to make these procedures rigorous and reproducible, inclusion and exclusion criteria need to be defined.

Similar as with standard research questions (Table 1), we experienced some inclusion and exclusion criteria to be useful in a broad spectrum of studies. These standard criteria listed in Table 2 allow researchers to obtain an appropriate result set and to define their requirements on the objective-dependent relevance of publications retrieved. For instance, experience shows workshop- or tutorial summaries can contain a lot of relevant keywords, but might not necessarily advance the actual body of knowledge. Also, since contributions might occur multiple times or might be out of scope, those have to be eliminated as soon as possible (criterion 7). Another important criterion is the eighth, i.e., if the full text is not available, the respective publication is usually of little value (regarding possibilities to analyze them and eventually draw proper conclusions). In context of a mapping study, this issue can be compensated to a certain extent as those studies focus on an early, abstract-based analysis. However, when it comes to in-depth analyses, e.g., in a systematic review, the full text needs to be available.

Finally, Kitchenham et al. (2015) recommend aligning search strings with the research questions. We add to this the suggestion to also align the in-/exclusion criteria with the research questions. This might result in a number of "duplicated" criteria, i.e., a paper could be relevant to topic *A* or to topic *B* if the literature study aims at synthesizing knowledge thus requiring multiple topics to be addressed and analyzed together. This furthermore allows for later replication why a specific paper was in- or excluded to/from the study.

2.2 Data collection and dataset cleaning

Once the study is designed, data can be collected. In that stage, the resulting data needs to be analyzed, cleaned/harmonized, and prepared for the upcoming investigations.

[3] As it is also criticized by Staples and Niazi (2007). In Kuhrmann et al. (2015), however, we accepted this challenge. It took us about a year just to clean the data and perform the selection procedures. We do not recommend this for replication.

2.2.1 Data collection

The data collection is usually conducted as an automated search using different sources. Automated data search, however, needs careful preparation and potentially extra test runs, as every data source has a slightly different format of the query strings, or constraints regarding the queries' length and complexity (see also the discussion in Ali and Petersen 2014, Badampudi et al. 2015; Brereton et al. 2007; Kitchenham et al. 2015). In practice, we experienced the design of multiple and overlapping query strings beneficial. Although the search procedure must be executed several times and produces some overhead, simple queries are usually better accepted by the search engines (see Section 3 for a detailed discussion).

Appropriate Data Sources Depending on the particular disciplines, several standard databases or collections (so-called baskets[4]) are available. In the following, we give an exemplary discussion for software engineering. Apart from specific conference- and workshop series (so-called restricted approach Kitchenham et al. 2015), a literature search should address the most common sources. That is, instead of searching specific proceedings of a conference, search queries should be designed to work with entire digital libraries. For the more general field of software engineering, the following libraries can be considered as standard libraries (or subsets thereof if opting for the restricted approach):

- IEEE Digital Library (Xplore)
- ACM Digital Library
- SpringerLink
- ScienceDirect (Elsevier)
- Wiley Interscience
- IET (also accessible via IEEE)

However, these libraries have their "specialties", notably, regarding the search query construction. Another point to take care of when using such digital libraries is the continuous indexing, i.e., indexes will "evolve" over time, which makes it hard to reproduce searches (see Section 3.1.2).

Checking the Result Set Before conducting the data collection, we recommend to have a set of reference publications available. One criterion we found useful for checking the appropriateness of a search is if the result set contains the expected reference publications (see also, e.g., Zhang et al. 2011). If one expects a particular publication in the result set, e.g., arising from a preliminary search, but it is not contained in that set, the revision of the search might be recommendable. Options to identify reference publications can be found in Section 2.1.2.

Primary Search and Backup Search Primary searches should always be conducted using aforementioned (or comparable) standard libraries. However, for several reasons, those libraries do not always contain all relevant publications. For example, contributions relevant to the field might result from Ph.D. theses that are not published in/not indexed by the standard libraries.

[4] Such as the Senior Scholars Basket, cf. http://home.aisnet.org/displaycommon.cfm?an=1&subarticlenbr=346

Therefore, we experienced it beneficial to complement the primary search with a backup search utilizing meta-search engines to complete the result set. However, using a meta-search engine must be done carefully. Besides the standard meta-search engines[5], such as DBLP or Scopus, Google Scholar is often used to get results quickly. However, the quality of search results obtained from such engines also depends on search preferences and even trends and, thus, searches might be much less repeatable than compared to standard libraries. Also, the results might also provoke duplicates and introduce extra threats to the validity of a literature study. A Ph.D. thesis, for example, can be written in a cumulative manner where parts of it exist separately as peer-reviewed publications already present in the result set of the primary search. Hence, it is important that the results obtained via meta-search engines are not included into the main result set without crosscheck. To this end, hits produced by such engines should be included in an own category, and such searches should be discussed as part of the threats to validity to increase the transparency.

(Data) Export Practices Data obtained from a data source must be stored in a way in which it can be used for further analyses. This part can become time consuming since different databases provide different export formats, which later on need to be joined and integrated. Therefore, data should be exported in at least two formats:

– A literature management tool of choice, such as BIBTEX
– As plain or (better) comma-separated (CSV) text files

These formats have the advantage that they are easy to process and convert into spreadsheets to allow for further selection (Section 2.2.2), and later on, analysis steps.

2.2.2 Dataset Cleaning

Cleaning a result set is a demanding, time-consuming task. Usually, we find two types of papers to be removed from the result set (cf. Table 2):

1. Contributions that are out of scope, and
2. Duplicates.

Duplicates are easy to find and eliminate, yet it is hard to decide which of the duplicates should be eliminated. It often happens that one publication is listed in multiple literature databases (e.g., for cross-indexing reasons). In such cases, it needs to be decided which paper to consider for inclusion into the result set. A pragmatic approach is to include the results from the database that provides the paper for download and to remove the other occurrences; this needs, however, to be defined in the exclusion criteria for the sake of transparency. Another case for a duplicate is a conference paper, followed by a journal article, e.g., a special issue paper. In such cases, it must be decided whether the original or the extended publication should be selected for inclusion. A criterion could be to always select the higher-valued publication, i.e., journal over conference, as journal articles are expected to have a higher maturity (Paternoster et al. 2014) and level of detail.

Publications that are out of scope are, on the other hand, easy to remove, yet they are often hard to identify if part of a large result set. Since the result set might have been created

[5] Note: Apart from serving the backup search, meta-search engines can also be a useful instrument in studies that also include (continuous) updates, e.g., to monitor the development of a field over time (Kuhrmann et al. 2016).

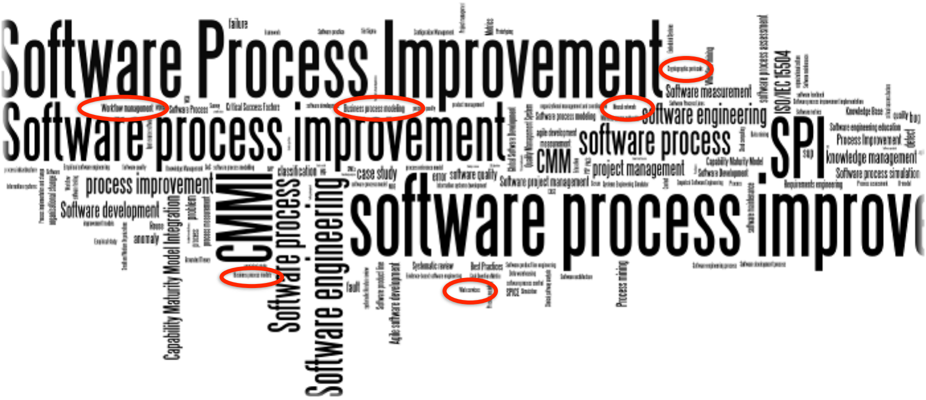


Fig. 3 Example of a word cloud from Kuhrmann et al. (2015) for visually inspecting the result set. "Outliers" to be used for excluding further papers from the result set are highlighted

from an automatic search, even out-of-scope publications that met at least one selection criterion could be present. Those publications need to be found manually and removed in the cleaning procedures.

Scoping via Word Clouds To support the identification of out-of-scope-papers, we experienced word clouds (tag clouds) to be a useful tool. Word clouds can be automatically created using keyword lists or abstracts. A word cloud is an instrument to visualize the (quantified) occurrence of a word/term in relation to other terms. They can be easily created using several publicly available tools[6], e.g., Wordl or TagCrowd.[7]

Word clouds can serve two purposes: First, word clouds can be used to analyze the appropriateness of a result set. A word cloud, which is based on the keywords, can be analyzed to work out whether the contained publications' keywords match the expectations (Fig. 3). Unexpected and/or "wrong" keywords can be easily detected and used to clean the result set. Depending on the quality, a considerable share of non-fitting papers can be removed; remaining papers (in a reduced set) are then removed during the selection phase (Section 2.3).

However, word clouds have to be used with care: even though there is research that shows word clouds providing improvement concerning the clustering and summarizing of descriptive information, such as Oosterman and Cockburn (2010), Ramage et al. (2010), Kuo et al. (2007), Schrammel et al. (2009), Rivadeneira et al. (2007), there is still the risk of eliminating relevant papers; for instance, because those papers might rely on a rarely used terminology. Therefore, eliminating papers based on word clouds only might threaten the validity of a study why we recommend that the use of word clouds must be planned with care and in detail in advance, and resulting candidates for removal require careful inspection.

As a second purpose to be served, a word cloud can support the later analysis of a result set during, for example, the concept classification conducted as part of a mapping study. For

[6] Note: Some of the tools have limitations regarding the amount of text they can process. Furthermore, the tools offer different features, such as thresholds, visualization and export mechanisms. Those points need to be evaluated prior to usage.

[7] Both tools are available at: http://www.wordle.net and http://tagcrowd.com/.

instance, in our study on method engineering (Kuhrmann et al. 2014), we analyzed the final word cloud to get a better understanding about which research type facets to expect from the publication set (e.g., how to interpret terms like "case study" as used in the respective community). The result of the word cloud and the result of the classification conducted in the study can then be compared to analyze the subjective authors' self-classification and the more objective one from the reviewers' classification. In another example (Kuhrmann et al. 2016), we used a word cloud to support the development of a *focus type facet* (Paternoster et al. 2014) and, furthermore, to conduct a cluster analysis.

Merging and Reducing the Integrated Dataset Depending on the particular search strategy—especially the search query construction approach—researchers have to deal with multiple (isolated) datasets. This is especially true if the work during the data collection is distributed among multiple researchers. To prepare the selection, the individual result sets need to be integrated into a holistic one. This integration constitutes a challenging task:

- If a literature database was queried multiple times (e.g., for the search string construction), the individual results need to be joined.
- Every literature database provides a slightly different export format and/or structure, e.g., CSV files obtained from Springer and from ACM have a different structure. These differences need to be reconciled.
- If duplicates were removed on a per-database basis, the integrated result set may still contain cross-database duplicates. The integrated dataset must then be cleaned again by identifying and removing duplicates.
- If the individual datasets were yet not investigated for duplicates, the respective cleaning procedures must be performed now.

The aforementioned steps can be (partially) automated (Kuhrmann et al. 2016). Nevertheless, the inclusion and exclusion criteria selected for the study should be consulted to support the compilation of the integrated dataset as well. We experienced the following procedure (Fig. 4) to be best suited for the stepwise integration:

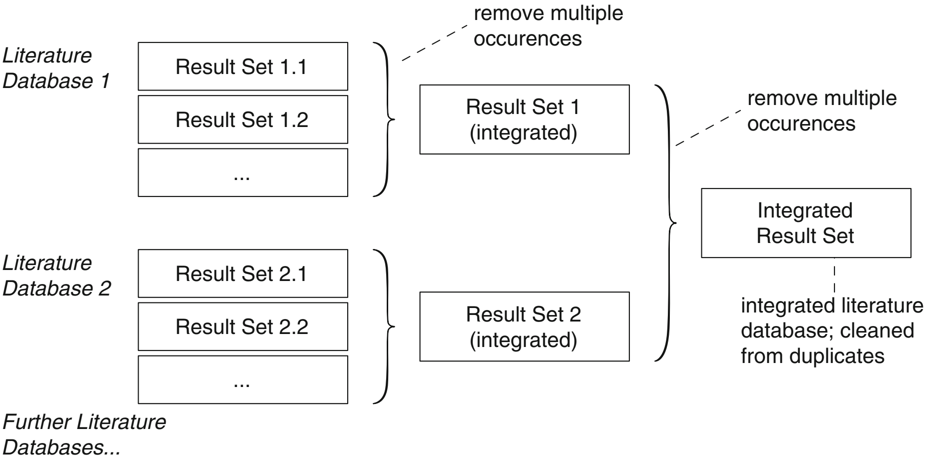

remove multiple occurences
Literature Database 1
Result Set 1.1
Result Set 1.2
...
Result Set 1 (integrated)
remove multiple occurences
Integrated Result Set
integrated literature database; cleaned from duplicates
Literature Database 2
Result Set 2.1
Result Set 2.2
...
Result Set 2 (integrated)
Further Literature Databases...

Fig. 4 Exemplary procedure of stepwise integrating and cleaning literature databases. In each integration step, reporting-relevant information needs to be recorded

1. Integrate and clean the data on a per-database level, i.e., if a database was queried multiple times, integrate the obtained sub-result-sets first.
2. Integrate all sub-result-sets into the integrated dataset and repeat the cleaning.

Eventually, we create an integrated dataset. Appendix B provides an example illustrating and explaining the minimal required data. Please note that the step of integrating and reducing the data is crucial and, therefore, needs to be documented carefully. The particular steps of the applied procedures are valuable information for other researchers to reproduce the overall study. Furthermore, the outcome of these steps forms the input for the rest of the study. Hence, researchers must ensure that no relevant publication is lost during this step.

Step-wise Dataset Completion Once an integrated data set is obtained, it should be analyzed for (sufficient) completeness. Depending on the database and the individual publications, some information might be missing, e.g., abstracts or keywords. This information needs to be collected and integrated, however, the step bears some pitfalls:

- There are abstract-free publications, e.g., magazine articles, of which the respective literature databases provide parts of the introduction section as abstract substitute. Such cases require a manual inspection and researchers need to discuss how to treat them.
- There are publications without (electronically available) keywords. These are publications that have no keywords at all, or publications that may well have defined keywords, but those were not listed in the exported data structure. For those publications, it must be defined how to treat them.
- For technical reasons, some literature databases do not provide options to export the abstracts. In such cases, manual work is required to get the abstract and integrate it into the dataset.
- Pieces of required metadata might be missing, e.g., the publication year, publication vehicle (conference, journal, etc.). This information needs to be completed.

Apart from this essential information, another aspect needs to be taken into account: the representation of the authors. Literature databases do not have a uniform representation of the author lists; for instance, authors might have varying affiliations or their first and second names are ordered differently (e.g., “J. J. Abrams” versus “Abrams, J. J.”). If researchers plan for a study, for example, to conduct some analyses on the author lists, such as by creating collaboration networks, cliques, and mainstreams, the author information must be available in a uniform way. As dataset completion can be extremely fidgety work, it should be performed iteratively and under continuous quality assurance:

1. Complete the abstracts
2. Complete the keywords
3. Complete all other required metadata
4. Ensure consistency in the author lists

Dataset Structure: A Template To support all aforementioned steps, a defined data structure needs to be in place. The particular data structure depends on the specific study. However, we recommend minimal data structure shown in Table 8 (Appendix B) as it emerges from our previously conducted studies. The table illustrates the recommended minimal data structure to organize the result set. This table serves the basic purposes and can be extended respecting the actual study’s needs, such as extra columns for classifications for mapping studies.

2.3 Study selection

In the study selection phase, the prepared dataset is analyzed for publications relevant for the actual study, i.e., researchers systematically select the relevant papers from the search results (this phase relates to the (primary) study selection in Kitchenham et al. 2015). Since result sets can comprise several hundreds or even thousands of papers, this phase requires special attention and, thus, careful planning.

2.3.1 Plan: Defining the study selection procedure

Many factors influence the actual study selection (e.g., number of researchers, degree of distribution, familiarity with the topic, etc.). In case of multiple researchers conducting the study, we consider the following aspects of the study selection necessary to be planned and agreed on in advance:

– Schedule for the study selection including workshops, regular meetings/calls for discussing intermediate selection/classification results, etc.
– Technical infrastructure (tools, data storage, file formats, etc.)
– The criteria upon which researchers decide the relevance of a publication
– The procedure to infer an agreement and the voting procedure (if applicable)

The last step, the voting, assumes that various researchers vote for in-/exclusion of a publication independently. The final decision for including the publication into the final result set then depends on the result of the voting. There are many practices that can be included into the voting procedure (e.g., veto rights) while we believe that this also much depends on the research context, e.g., researchers' experience, expertise, but also their personal preferences to conduct the study (see also Section 2.3.3).

2.3.2 Kick-off: Setting up the selection approach

Assuming a study within a group of multiple researchers, the study selection starts with a kick-off meeting in which the inclusion and exclusion criteria are recalled (Table 2), the selection/voting procedure is discussed, and a schedule for subsequent meetings is defined. In the following, every participating reviewer gets a copy of the cleaned result set, which is rated individually. That is, the study selection procedures are initiated.

2.3.3 Voting procedure

Voting is essentially a headcount procedure in which a team of researchers works out a decision whether a particular paper is considered relevant for the study or not, i.e., to eliminate those papers from the result set that are considered irrelevant. The relevance can be determined by different measures, which need to be defined in advance (e.g., title, abstract, and full text). Potential routes towards a decision are *majority votes* or *relative ratings*. The actual classification can be carried out in a group of researchers or individually, iteratively, round-based or in workshops. In the following, we focus on an individual, traditional round-based classification.

Majority Voting The voting is a headcount that aims to bring in objectivity into the study selection. Although there are in-/exclusion criteria, the final application of the criteria to the

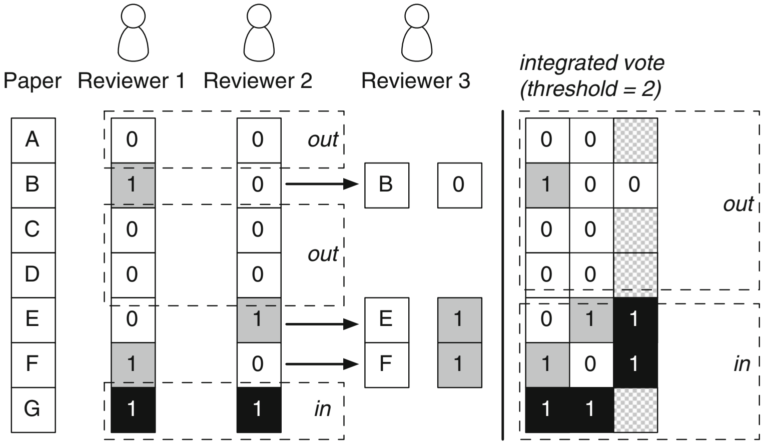


Fig. 5 Overview of the standard majority voting procedure for a 3-reviewer team

publications to be selected is in the hands of individuals thus including individual interpretations of a publication. The reason why we recommend including multiple researchers in this procedure is to overcome the inherent threat arising from this subjectivity. Hence, we also consider a *majority vote* to be the standard procedure as it is the most straightforward approach: every reviewer is provided with the integrated result set and reviews the items individually according to defined criteria, e.g., title and/or abstract. If a reviewer considers a publication relevant, 1 point is given, 0 otherwise. For n reviewers and m publications, the procedure results in an $n \times m$ voting matrix, which helps to select the relevant papers. The (final) selection is then based upon the agreements, such as a threshold or agreement statistics (e.g., Cohen's or Fleiss' κ). For example, if three reviewers participate, the voting procedure could be organized as shown in Fig. 5: two reviewers start individually. To get a paper included in the set of relevant papers, 2 points are required (threshold approach). Two reviewers can come up with the following results: 2 points = paper is relevant, 0 points = paper is irrelevant, and 1 point = paper is not yet decided. In the next step, the third reviewer[8] is called in and is presented a reduced list that only contains the papers yet not decided. The third reviewer then conducts the voting to finally decide about the papers' relevance.

Alternative Approaches Instead of calling in a third reviewer to conduct a fully independent review, a voting workshop can be organized. In such a workshop, all reviewers involved in the selection process discuss and decide the non-decided papers. We applied this approach for instance in Mendez Fernandez et al. (2014), Kuhrmann et al. (2015). Yet another approach is to provide reviewers with overlapping subsets of the whole result set, e.g., to incrementally collect three votes in just one run (Fig. 6).

Scaling So far, we performed the majority voting procedure with 2 reviewers in the workshop model, 3 and 4 reviewers, and two 2-person review teams (see Section 3). However, as we talk about simply summing up points, the approach can be scaled to an even larger

[8] Please note that a reviewer can be an internal reviewer (e.g., a co-author) as well as an external researcher or expert not involved at all in the design in case of unknown domains.

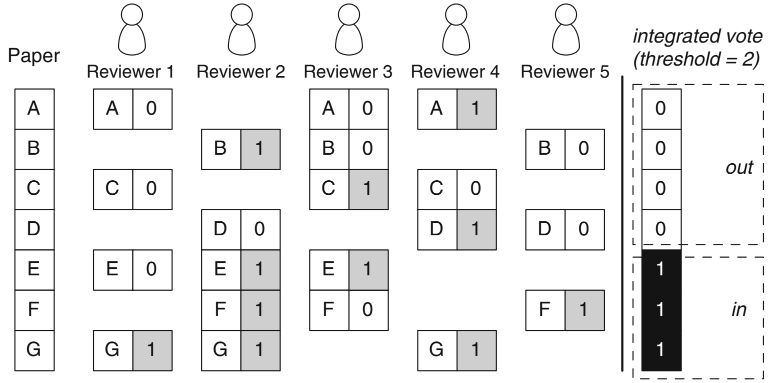


Fig. 6 Paper selection based on overlapping paper subsets (a reviewer evaluates only subset of papers, usually just one run required to find the selection)

number of reviewers. A paper's relevance is then simply defined by a function

$$relevance : \mathbb{R}^+ \times \ \mathbb{Z} \rightarrow \{0, 1, ?\} \tag{1}$$

that is used to determine the relevance of a paper p_j in relation to a threshold *th*, and to (de-)select papers or marking them for later decision:

$$relevance(rating(p_j), \text{th}) = \begin{cases} 1 & \text{if} \quad rating(p_j) > \text{th} \\ 0 & \text{if} \quad rating(p_j) < \text{th} \\ \text{toDecide} & \text{if} \quad rating(p_j) = \text{th} \end{cases} \tag{2}$$

The actual threshold *th* needs to be defined during the initialization of the selection procedure (Section 2.3.1). The rating (simple, unweighted case; Fig. 5) of a paper is then defined by the number of points that a paper received from n reviewers involved in the process:

$$rating(p_j) = \sum_{i=1}^{n} r_i(p_j) \tag{3}$$

Regardless of the number of stages and reviewers involved, rating statistics need to be carefully documented in order to be able to reproduce which paper came in in which stage and to make explicit the inter-rater agreement. Furthermore, we also suggest to document according to which criteria a paper was included or excluded after all, which can require extending the data structure of the result set to keep this information.

Relative Rating The *relative ratings* approach[9] as illustrated in Fig. 7 is similar to the *majority voting* where all reviewers are asked to vote a result set, but with a difference in the applied metric: Instead of a simple "Yes/No" (1/0) metric, in this approach, we use Likert scales and thresholds. The basic underlying procedure remains the same: each reviewer is provided with the integrated result set and rates a paper, but on a scale, such as a 5-point Likert scale:

– 5 points: Paper is highly relevant (must be included)

[9] So far, we did not yet apply this method to a complete study, but partially applied it during sample-based result set testing and evaluation (cf. Section 3). As this approach is quite complex compared to the majority vote, it requires sufficient tool support.

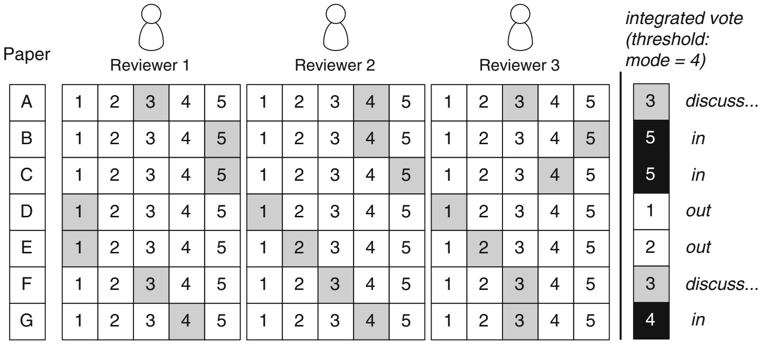


Fig. 7 Paper selection based on relative votes (final selection is, in this example, made using the *mode* value, while the "neutral" element 3 serves as marker for papers to be discussed)

- 4 points: Paper is (somewhat) relevant
- 3 points: neutral/no opinion
- 2 points: Paper is not relevant
- 1 point: Paper is absolutely irrelevant

Based on the individual ratings, relevance can be determined, e.g., using the mean value or the mode, and precision can be determined, e.g., using standard deviations or distance metrics. The inclusion criterion is then a selected value on the used scale, e.g., 4 (or better). Critical is the handling of papers that end up with the neutral value. These papers require extra handling.

Balancing Votes How reliable is this way of selecting papers? In the simple case, which is the *majority vote*, a democratic headcount is used to in-/exclude a paper. However, this procedure has some flaws. For instance, given a situation in which two reviewers ended up in a stalemate. A third reviewer is then called in to make the decision; and now scale this up to 7 reviewers: the 7th reviewer makes the final decision by outvoting 3 others. To overcome such situations, workshops can be performed to discuss critical papers (which can be unrealistic if, for instance, 250 papers need to be discussed), thresholds can be defined, or weights can be introduced, e.g., senior reviewer votes count twice. However, the basic problem remains: what is the level of agreement, i.e., the reliability of the selection? As a first step to determine the reliability, the inter-rater reliability can be calculated, e.g., using Cohen's κ (1968) for two reviewers or, more general, Fleiss' κ (1971) for more than two reviewers.[10] Furthermore, the basic agreement can be visualized (and partially automated) as shown in Fig. 10 (Appendix B).

Yet, the headcount is a fairly simple, but absolute metric. In some cases, we experienced the need for a more differentiated vote, which can be implemented, e.g., using *relative votes* with Likert scales. However, the more differentiating scale introduces a new challenge: *How to find a final and consolidated rating?* Approaching a consolidated rating via the mean or the mode might fail, because they are easy to trick or because they might be even not applicable; consider, for example, the mode of {0, 0, 1, 1}, and what a resulting mean of 0.5 even implies in relation to a $th \in \mathbb{Z}$ (cf. (1)). Again, a simple solution could be to introduce

[10] Please note that inter-rater reliability calculations also depend on the scales applied, e.g., weighted κ values when using ordinal data (cf. Kitchenham et al. 2015; Wohlin et al. 2012).

rater-specific weights. Furthermore, simple weighting methods, such as, the 3-point-method can be applied, with $V_j = \{v_{p_j}^{r_1}, ..., v_{p_j}^{r_n}\}$ being the set of n reviewer votes for a paper p_j:

$$rating(p_j) = \left| \frac{min(V_j) + 4 \cdot \bar{V_j} + max(V_j)}{6} \right| \tag{4}$$

The extended weighted rating from (4) can be used in the determination of the relevance in (2).

2.3.4 The gathering: Integrate and finalize the paper selection

Having all individual ratings conducted, the study's moderator (Kitchenham et al. 2015 speak of a team leader) collects all individual ratings and starts the integration of the results. The basic task is to, initially, integrate the individual ratings to work out the current state of selected and/or undecided papers (see also color-coding in Fig. 10 that is based on (2) and (3). Depending on the approach defined in the initialization of the selection procedure (Section 2.3.1), the moderator prepares the dataset for extra review iterations and/or organizes required workshops. In the following, the selection procedure is iterated until all papers are finally decided.

Once all papers are decided, the moderator draws a baseline and prepares the final selection of papers, i.e., a cleaned list that only contains those papers considered relevant for the study, and he finally prepares the clearing work.

2.3.5 Class dismissed: Analyze the result set and report

When the selection is done, the moderator concludes the selection process and prepares the handover to the actual analysis. This includes some standard tasks as well as some optional tasks depending on the eventually targeted study. In particular, the moderator has to prepare the study selection report and the resulting literature database. The literature database must at least contain all papers that were selected as relevant to the study. The report comprises some statistics, such as, databases, results per database from search, and elimination statistics (an example is shown in Table 3).

Depending on the intended study type, just in this step, the moderator can also provide some extra data to support the later analyses. For example, if applicable, the inter-rater agreement helps identify those publications that form the heart of the result set. Furthermore, several outputs can be generated from the result set that help finding a starting point, e.g., exports of the keyword lists and abstracts and word clouds generated thereof, and, associated with more effort, social networks (Section 3.1.4).

2.4 Concluding and handover to data analysis

The last step consists in initiating the actual data analysis, which is dictated by the research questions and eventually the type of secondary study. From the aforementioned described steps, the outcomes listed in Table 4 have to be assembled and shipped to the in-depth analyses. These deliverables can be properly integrated with the research protocols as, for instance, recommended by Kitchenham et al. (2015).

Table 3 Exemplary search and selection report (excerpt from Kuhrmann et al. 2015)

Step	IEEE	ACM	...	Total
Step 1: Search				
S_1 and (C_1 or C_2)	71	543	...	3,185
...	...	...	...	...
S_8 and C_2	114	105	...	8,374
Step 2: Removing duplicates				
Duplicates per database	1,486	566	...	16,643
Duplicates across all databases	916	551	...	5,315
Step 3: In-depth filtering				
Applying filters F_1 and F_2	578	–	...	1,562
Unfiltered	–	551	...	1,610
Result set (search process)	578	551	...	3,172
Step 4: Voting				
Final result set	283	65	...	635

3 Example studies and lessons learned

The guideline presented in this article emerges from various conducted systematic reviews and mapping studies. In this section, we provide an overview of the previously contributed studies and discuss how we applied the discussed practices and procedures so far. Table 5 provides an overview of the referred studies and relates the studies to the respective methods and techniques.

Table 4 Artifacts to be created in the early stages of literature studies to be shipped to the in-depth data analysis

Reference	Outcomes and content to be delivered
Section 2.1.2	Search terms and resulting search queries (generic terms and queries, as well as database-specific queries)
Section 2.1.3	In-/exclusion criteria used in the study
Section 2.2.1	List of selected and queried databases, and raw result sets (e.g., CSV files)
Section 2.2.2	Cleaned and integrated data sets (including all support instruments used)
Section 2.3.1	A documented study selection approach, including team setup, selection procedures, and so forth
Section 2.3.4	Decided data set (final result), statistics of the selection, further complementing report data

Table 5 Overview of the different studies utilizing the presented practices

Ref.	Title	Type (r/m)	Preliminary Study	Trail-and-Error Search	Snowballing	Search String (1/n)	Majority Voting	Relative Rating (s/f)	Workshops	Inter-rater Agreement (s/f)	Multiple Researcher Teams	Word Clouds	Social Network Analysis	Rigor-Relevance Model (Ivarsson and Gorschek 2011)
(Kuhrmann et al. 2014)	A Mapping Study on the Feasibility of Method Engineering	m	✓		✓	✓$^{(n)}$	✓$^{(3)}$		✓			✓	✓	
(Méndez Fernández et al. 2014)	Where Do We Stand in Requirements Engineering Improvement Today? First Results from a Mapping Study	m			✓	✓$^{(n)}$	✓$^{(3)}$	✓$^{(f)}$	✓					
(Kalus and Kuhrmann 2013)	Criteria for Software Process Tailoring: A Systematic Review	r			✓	✓$^{(n)}$	✓$^{(3)}$							
(Kuhrmann et al. 2013)	Systematic Software Process Development: Where Do We Stand Today?	r			✓	✓$^{(1)}$	✓$^{(3)}$							
(Kuhrmann et al. 2015)	Software Process Improvement: Where Is the Evidence?	m			✓	✓$^{(n)}$	✓$^{(2)}$	✓$^{(s)}$	✓	✓$^{(s)}$		✓		
(Kuhrmann et al. 2016)	Software process improvement: A systematic mapping study on the state of the art ☆	m	✓			✓$^{(n)}$	✓$^{(2)}$		✓	✓$^{(s)}$				
(Kuhrmann et al. 2016)	How does software process improvement address global software engineering?	m/r	✓			★	✓$^{(2)}$		✓			✓		✓
(Jacobson et al. 2016)	On the Role of Software Quality Management in Software Process Improvement	m/r	✓			★	✓$^{(2)}$		✓					✓
(Kuhrmann et al. 2013)	Towards Artifact Models as Process Interfaces in Distributed Software Projects	m/r				✓$^{(n)}$	✓$^{(2)}$		✓					
(Theocharis et al. 2015)	Is Water-Scrum-Fall Reality? On the Use of Agile and Traditional Development Practices	r		✓	✓	✓$^{(1)}$	✓$^{(2)}$		✓		✓$^{(2)}$			
(Ingibergsson et al. 2015)	On the Use of Safety Certification Practices in Autonomous Field Robot Software Development: A Systematic Mapping Study	m		✓	✓	✓$^{(n)}$	✓$^{(3)}$					✓		
(Racheva et al. 2009)	Value Creation by Agile Projects: Methodology or Mystery?	m			✓	✓$^{(n)}$	✓$^{(3)}$							
(Condori-Fernandez et al. 2009)	A Systematic Mapping Study on Empirical Evaluation of Software Requirements Specifications Techniques	m			✓	✓$^{(n)}$	✓$^{(4)}$							
(Inayat et al. 2015)	A Systematic Literature Review on Agile Requirements Engineering Practices and Challenges	r			✓	✓$^{(n)}$	✓$^{(3)}$							

Search String (1/n): The study uses 1 large or *n* smaller search strings

Relative Rating (s/f): Relative rating of the full result set or on samples thereof

✓$^{(*)}$: * number of search strings, or number of reviewers involved

☆: Study update for Kuhrmann et al. (2015);

★: Detailed study using the dataset from Kuhrmann et al. (2016)

3.1 Selected examples and lessons learned

Over the last years of working on literature studies, we collected a number of lessons learned, which we briefly summarize below. Furthermore, in order to illustrate the lessons learned with examples, in this section, we relate the lessons learned to the studies from Table 5 and provide some examples. Moreover, the practices listed in Table 5, in general, can be considered self-contained building blocks, i.e., they can be combined in different ways. However, in our experience, some combinations of practices showed especially beneficial. Those are presented in Appendix A as a blueprint. We also have to note that there might exist dependencies and/or constraints providing arguments in favor or against applying certain practices in respect of a particular context (see also Zhang et al. 2011). For example, if a preliminary study was already conducted to find the study's scope and a set of reference publications, the "Trail-and-Error" search approach will not add to the study. Another example is the combination of selection strategies, i.e., the combination of majority votes, relative ratings, and voting/rating workshops. Here, setting up workshops ("expensive" due to required human resources) should be preferably scheduled for the late selection phases when the amount of publications to be decided was reduced to a manageable number (see Section 2.3.3). The rest of this section is organized according to the stages of this guideline (cf. Fig. 1).

3.1.1 Basic planning

Regarding the general planning activities associated with a literature study, we consider the following lessons learned the most important.

Make a Cunning Plan that Cannot Fail Given the effort, duration, and the involvement of various researchers, a literature study should be built upon a concrete plan of which the research protocol (Kitchenham et al. 2015) is key. We experienced that involving all researchers at the beginning is crucial to establish a shared understanding of:

- The basic terms, concepts, and their synergies, in the field of interest, and
- The way the classification criteria should be interpreted and applied.

If classifying the relevance or other concepts based on a pre-defined scheme, those concepts need to be clarified at the beginning.

Watch out! The Technical Infrastructure Matters One of the most important administrative tasks is to define the technical infrastructure to be used for the study. The two most important aspects are:

- Use a version control system (VCS).
- Don't mix up Microsoft Excel and OpenOffice.org/LibreOffice.

The VCS is crucial to create baselines of the study, e.g., raw data or tentative result sets. Furthermore, a VCS allows for distributed collaborative and concurrent work, and it ensures that results are not accidentally overwritten. The second aspect is caused by practical experience: In several studies, some researchers just took the pre-configured Microsoft Excel file (see Appendix B) and worked on it with OpenOffice.org/LibreOffice, so that many scripts and auto-formatting configurations did not further work, or that other researchers could simply not open it anymore with the respectively other tool (e.g., as happened in Kuhrmann et al. 2015). Fixing those situations is time-intensive and avoidable.

3.1.2 Search Strings and Search Engines

Regarding the construction of proper search strings, we consider the following lessons learned the most important ones.

One Search String or Multiple Ones? Applying the introduced search strategies may result in more than one search string, which then can be customized for the different search engines. A practical problem remains: the length and complexity of the search strings, and the ability or limitations of literature databases to process search queries of and above a certain complexity (as observed when trying to replicate (Schramm et al. 2014)). That is, the major question is which alternative is better: One integrated long search string or multiple shorter ones, as exemplarily shown in Table 6.

An integrated search string has the advantage of (relative) high precision. Furthermore, it allows for capturing the entire domain in only one query. However, many literature databases, such as IEEE Xplore, have some limitations regarding length and complexity. Furthermore, the syntax of the search queries differs from database to database, thus, requiring database-specific instances of the query anyway. In contrast, multiple shorter search queries bypass database limitations by providing simpler structures (also recommended by Kitchenham et al. 2015) and, furthermore, those strings are easier to adapt to specific database requirements. On the other hand, in order to ensure search precision, multiple search strings require more effort in their design. For instance, to get a maximum of publications, multiple search strings require some overlap to avoid “losses at the borders”. This, however, may cause some overhead in the result set and multiply occurring publications that have to be identified and removed later on. Furthermore, due to the simpler structure, such search strings are prone to attract unwanted publications (Zhang et al. 2011) thus requiring extra context selectors and filter constructs (Kuhrmann et al. 2015).

Don’t trust Old Result Sets When it comes to updating or replicating a literature study, one problem is the literature database as such. For example, in a student study activity, we aimed at replicating and updating a previously conducted SLR (Schramm et al. 2014) of which we had the full research protocol available. The replication package also included text files containing the database-specific search strings. In an initial test run, we encountered the following to happen: IEEE Xplore rejected the search query stating it was too complex having more than 50 terms. Transferring the (general master) search string to Scopus (to test if it will trigger any papers at all) and configuring the search properly (limiting the venues and publishers etc.), we found 215 instead of 125 papers matching the search criteria. So far, we couldn’t sufficiently elaborate what happened exactly, but argue this being one of the effects coming along with continuously updating indexes (see also Brereton et al. 2007, who mention indexing of current digital libraries inappropriate).

Table 6 Exemplary search strings for an automated database search (excerpt from Kuhrmann et al. 2015)

	Search string	Addresses...
S_1	(life-cycle or lifecycle or life cycle) and (management or administration or development or description or authoring or deployment)	process management: general life cycle
	...	...
S_8	(feasibility or experience) and (study or report)	reported knowledge and empirical research

In short, over the time, search queries age and literature databases evolve. There is no guarantee that a result set obtained at one point in time will be re-constructible some time later. There is no mitigation strategy for this problem, except to increase the transparency of the data collection by reporting a timestamp for the searches to support the reproducibility and thereby the validity. Therefore, search queries as well as raw result sets (Section 2.2) should be stored—at least to reproduce the findings from the raw data.

3.1.3 Data collection and cleaning

Regarding the data collection and cleaning, we consider the following lessons learned the most important ones.

Find the Right Scope In some studies, we saw an explicit and intentional limitation of the search; for instance, instead of searching a whole library, authors of a study limited themselves to particular conferences or journals (Schramm et al. 2014). Such an approach promises the advantage of having a more focused result set by avoiding overhead (Kitchenham et al. 2015). However, this may come possibly at the price of information loss, because many relevant publications might not be found. Such procedure is of course possible, but not recommended; yet, if conducted that way, it should be explicitly mentioned in the threats to validity to increase the transparency and reproducibility. Finally, if the ultimate goal is a systematic mapping study, however, this approach cannot be applied, as the limitation of the search scope hampers the overall result set quality and also the quality and reliability of the conclusions.

What Publication Type to Include? Besides the used search engines, researchers need to clarify what types of publications can/cannot be included into the result set. We consider, for example, including textbooks and edited chapters as a viable option in case the study is about the analysis of definitions, e.g., to understand the meaning of a particular concept as used by authors in a field. The choice of certain books can and should be justified based on their popularity in a community; for example by including well-established textbooks as used for teaching, or books that have a high number of citations in empirical papers in the area. Master theses in turn should be avoided given their missing peer-review process. Involving Ph.D. theses, however, depends on various contextual characteristics; for instance, whether they passed a peer-review process or whether they are cumulative ones (which might, of course, lead to duplicates in the result set given that the content is previously published material, see also Section 2.2.1).

How Valid is the Paper Selection Process? In the previous sections, we described different voting procedures that can be applied. With every voting procedure comes different ways of increasing the validity of the methods applied and the results obtained. The least common denominator of all procedures, however, is the inter-rater agreement (Kitchenham et al. 2015). We postulate the use of inter-rater agreements especially if used in a multi-staged voting procedure as they serve as a constructive quality assurance measure between the stages; for example, to clarify misconceptions, misinterpretations of research questions, misinterpretations of classification schemes, and different understanding of the relevance of publications. Besides the value of inter-rater agreements for constructive quality assurance, it also increases the transparency to the reader and, therefore, the conclusion validity.

However, such an agreement makes only sense if the voting procedure is not conducted iteratively over incomplete result sets whereby it is impossible to use the agreement as a means to improve the classification between stages (if not used in a training/test phase). Hence, there is a trade-off regarding the purpose and the effort of using the inter-rater agreement, which needs to be clarified in advance.

How Much is Enough? As a matter of fact, there is no meaningful metric that could be used to indicate whether the result set is sufficiently large or not, let alone because the size of a dataset provides no indicator to the quality of its content (Wohlin et al. 2013; Zhang et al. 2011; Badampudi et al. 2015). For example, in Kalus and Kuhrmann (2013) and Ingibergsson et al. (2015), we performed the data search, but then capped the result sets to include only the first 50 hits per query result. Is this enough? What is the risk of loosing relevant papers? As there is no common ground, such a decision must be taken on a per-study basis. Yet, it needs to be ensured that the result set of papers obtained is of high quality, i.e., representative for the field of investigation and the research questions formulated. This means to ensure an accurate result set and a detailed and validated review protocol including a search string potentially adapted to the particularities of the search engines, and detailed inclusion and exclusion criteria.

3.1.4 Preparing the handover

Although a study selection might be completed, more activities can be carried out before entering the in-depth analysis. The final dataset provides already data that can be used early in the overall literature study process to help researches finding appropriate points to start with the analysis. From our so far conducted studies, we consider the following lessons learned helpful.

Exporting Keyword Lists, Abstracts, and Word Clouds From the result set, keyword lists and abstracts can be easily harvested and prepared to support the beginning of the analysis. We can create, for example, word clouds from these lists to get a quick visual inspection where a striking keyword could indicate to a set of publications to start with. However, what seems easy to generate and use can eventually turn out to be difficult or even misleading: several tools for word cloud generation have limitations regarding the amount of text they can process. A solution is to perform a keyword coding, which serves three purposes (as used in Kuhrmann et al. 2015): first, the list of keywords is shortened; second, the used terminology is harmonized (e.g., “small-to-medium-sized companies” and “small and medium enterprises” are coded to “SME”); third, the keyword coding can be considered a first step towards full coding, which is normally performed in the context of a mapping study to work out the classification schemas. If a keyword (and/or abstract) coding would be performed, the outcomes of the activities would comprise the respective keyword lists, abstract lists, the mapping files containing the codes and all synonyms, and optionally generated visuals.

Utilizing Social Network Analysis as a Means for Pre-Selection A social network is a graph that provides an overview of subjects and their relationships (see for instance Hanneman and Riddle 2005; Scott 2000; Wasserman and Faust 1994). Right in the early stages, even before the actual study begins, a social network graph can be generated from the result set. Such a graph can serve multiple purposes. For instance, a social network graph highlights cooperation cliques, i.e., authors that collaborate and contribute a considerable share of the result set, thus, forming the "community leaders". When it eventually comes to begin with the result set analysis, researchers can face the problem to find a proper starting point. Potentially identified clusters can provide some guidance through the result set. Another option is to look for domain-shaping key contributions, which are potentially highlighted by a citation network.[11] Beyond the analysis preparation, a social network is also a supportive means within a study. For example, in Kuhrmann et al. (2014), we used a collaboration network to study if a found trend in the publication space is just because of the result set's background noise. Therefore, we generated the social network to identify the key contributors and created a sub-map, which was based on the respective publications only, and compared whether the general trends differed.

4 Related work and discussion

This article complements a number of existing guidelines and initiatives for conducting literature studies. In this section, we provide an overview of related work including approaches, methods, experiences, and tools to support literature studies and position our contribution in context of the current publication landscape. Table 7 summarizes the body of knowledge in existing guidelines we found particularly relevant and adds how our contribution at hand deviates (i.e., adapts/extends) from existing contributions.

Approaches We deliberately use the term "approaches" to subsume all the different processes and methods utilized in literature studies. One prominent approach in context of literature studies is the *systematic review* process as initiated for software engineering by Kitchenham (2004) and continuously improved, e.g., Kitchenham and Charters (2007), eventually leading to a consolidated guideline (Kitchenham et al. 2015), as well as the *systematic mapping study* made popular for software engineering by Petersen et al. (2008) (updated in Petersen et al. 2015).

These general guidelines, which—despite of their value to provide a common structure and consistent terminology—have been experienced as too generic for direct practical application (Petersen et al. 2015; Staples and Niazi 2007). They still serve as an umbrella and a multitude on fine-grained methods and models, and advice and best practices can be embodied by the guidelines. For example, a challenge in literature studies is the development of proper classification schemas. In literature, we find, for instance, the *research type facet* classification schema developed by Wieringa et al. (2005) and the *contribution type facet* schema as illustrated by Petersen et al. (2008) (adopted from Shaw's work Shaw 2003) serving as generic classification patterns for studies (Petersen et al. 2015). Another perspective is provided by Paternoster et al. (2014), who utilize a *focus type facet* and a *pertinence facet*. Furthermore, Paternoster et al. (2014) include a model for determining *rigor and relevance* of the involved studies (based on a model proposed by Ivarsson and Gorschek 2011) to support the determination of the result set's reliability. However, Petersen et al. (2015) mention those classification schemas critical. The reason is that such schemas, as the one by Wieringa et al. (2005), leave room for interpretation. As a matter of fact, we can find "tailored" variants of this schema in a number of studies (see also Wohlin et at. 2013). It

[11] This approach needs to be considered with care, as for instance newer publications may have a high-quality contribution, but don't have a high citation count (e.g., compared to a 10-year old publication). Therefore, citation networks only deliver initial indication and trends shouldn't be taken for granted.

Table 7 Relation of the present guideline with further established guidelines

Ref.	Key Contributions	Adaptation/Extension
Kitchenham et al. (2015)	Kitchenham et al. provide a well elaborated overview of the systematic literature study processes. To this end, they introduce a conceptual description of what to do in a systematic review or in a systematic mapping study, and an explanation of why these steps should be carried out. The aim is to provide a generalized view on *what to do* while concrete advice of *how to operationalize* the respective steps in a specific context is out of scope.	Our guideline emphasizes the operationalization of the particular steps in the data collection and study selection phase, and the guide provides examples and critical discussion of lessons we learned. Furthermore, our guideline describes activities as building blocks and offers exemplary workflow templates for literature studies of different complexity and size.
Petersen et al. (2015)	Peterson et al. propose a guideline, which extends their original one (Petersen et al. 2008) grounded in evidence obtained from analyzing 52 mapping studies and comparing the guidelines used therein. The guideline provides a checklist of activities and refers to articles that used those to select data for the study. It further proposes a more detailed classification schema (compared to (Petersen et al. 2008; Wieringa et al. 2005)) and comprises small examples for illustration.	Our guideline has a different scope compared to Petersen et al. (2015) as we focus on the relatively unexplored early stages only. That is, our guideline focuses on the data collection and study selection process, whereas we pay little attention to the data extraction and analysis which we believe to be already well elaborated. Yet, our guideline provides a more detailed perspective, e.g., on the different practices and how to combine them, how to utilize techniques such as word clouds or social networks to aid the selection process (both not mentioned in Petresen et al. 2015). Therefore, our guideline is a pragmatic complementation of the *study identification* phase from Petersen et al. (2015).
Zhang et al. (2011)	Zhang et al. describe a “quasi-gold standard” to find an effective study selection strategy. Among other things, Zhang et al. define a search process to achieve high sensitivity and precision of the searches.	Similar to Zhang et al., our guideline recommends utilizing different search engines. Yet, our guideline provides more details regarding actual practices to analyze and clean the result (sub-)sets obtained from different search runs, and we also provide recommendations to develop an integrated result set to be evaluated in the actual study selection process. Therefore, our guideline complements (Zhang et al. 2011) and provides recommendations to fill gaps, such as missing information concerning the steps required to get from step 4 (conduct automated search) to step 5 (evaluate search performance).

Table 7 (continued)

Ref.	Key Contributions	Adaptation/Extension
Wieringa et al. (2005)	The work by Wieringa et al. has become representative for developing classification schemas based on a well elaborated reference (see also Petersen et al. 2008, Peternoster et al. 2014 or Petersen et al. 2015).	In the present guideline, we explicitly do not aim to support schema development. However, when providing a data structure template, we leave room for classification schemas. Furthermore, grounded in our experience, we also propose considering free metadata to be collected, since we found strict classification schemas not well-applicable in all setups.
Ivarsson and Gorschek (2011)	The rigor-relevance model by Ivarsson et al. provides a scale-based approach to determine the relevance to industry and the rigorousness of the research conducted. Hence, this model can support the paper selection process.	In our guideline, we utilize the rigor-relevance model exactly as proposed as an explicit extra dimension to support the classification, because we experienced it to be of particularly high value. We therefore recommend to use a combination of “standard schemas” (e.g., Wieringa et al. 2005, Ivarsson and Gorschek 2011, Petersen et al. 2008, Peternoster et al. 2014) complemented with study-specific schemas, e.g., those developed from free metadata.

also remains a challenge to construct a schema in a proper and efficient manner, and a number of strategies are available for this purpose (Petersen and Ali 2011). For instance, in our study (Kuhrmann et al. 2015), we used the *focus type facet* concept finding the described construction procedure from Paternoster et al. (2014) inappropriate for the following reasons: if one has to deal with a very large number of papers, a manual coding-based schema construction is too costly. Moreover, it is challenging to clearly define the elements of such a schema, as indicated by Portillo-Rodríguez et al. (2012). This is because not all papers have sufficient information in title, keywords, and abstracts to conduct a proper and fine-grained classification (Brereton et al. 2007), and if the purpose of the study is to capture an entire domain, developing a precise classification is close to impossible, as many publications address multiple topics, which makes a unique classification hard or even impossible. Therefore, in previous work (Kuhrmann et al. 2015), we started collecting “free” metadata instead of providing a big picture of the domain, but leaving the full classification to the fine-grained analyses of selected topics. As outcome, in Kuhrmann et al. (2016), Kuhrmann et al. (2016) we used the metadata to generate *heat maps* (as also done in Penzenstadler et al. 2014) to work out trends worth further investigation.

Constructing a classification schema requires data to apply the schema. In this respect, Petersen et al. (2015) found 15 ways to collect and identify relevant studies. Data search is mainly done using manual and database searches, and snowballing. Yet, it is currently subject to discussion which of the practices (or combinations thereof) result in datasets of sufficient quality *and* what is considered a sufficient dataset after all (Wohlin et al. 2013).

Ali and Petersen (2014) review strategies to select studies in systematic reviews and formulate a selection process. They conclude that a good-enough sample could be obtained by following a less inclusive but more efficient strategy. Zhang et al. (2011) present a "quasi-gold standard" to identify relevant studies and Badampudi et al. (2015) show that snowballing also leads to an appropriate result set. That is, all the different search strategies used so far produce sufficient datasets. Up to now, however, little has been reported on the complementary use of the different search strategies, costs and benefits associated with such a combination. In the present article, similar to Dybå et al. (2007), we stress this aspect by presenting the combined use, and we also demonstrate how a search can be complemented by further techniques, such as *social network analysis* (Scott 2000; Wasserman and Faust 1994) or *word clouds* (Kuhrmann et al. 2015; Kuhrmann et al. 2016), to support pre-selection, analysis scoping, and dataset/result visualization.

The search and selection procedures also include the definition and use of inclusion and exclusion criteria. However, Petersen et al. (2015) found only five out of 10 guidelines explicitly addressing this topic, but there was so far no attempt to craft a set of standard in-/exclusion criteria. Similarly to standard research questions, standard data collection workflows, and standard study selection procedures, we have proposed a set of standard inclusion and exclusion criteria to support a quick start of the study and to lay the foundation for the development of further study-specific criteria.

Experiences Regarding the (generic) guidelines used by empirical software engineering researchers, Petersen et al. (2015) found and compared in total 10 guidelines used, whereas the (more general) ones by Kitchenham and Charters (2007) and Peterson et al. (2008) were identified as the most frequently used. Furthermore, their findings include identified gaps in the individual guidelines, such as missing practical advice on how to do self-evaluation, justification and motivation of the research question chosen regarding the demographic overview of a planned study, or missing shared practices from personal accounts of designing systematic reviews and mapping studies by following specific guidelines. Petersen et al. (2015) add to a series of meta-studies that aim to monitor the guidelines' application and to collect lessons learned and best practices is a required step to consolidate experience. For instance, Kitchenham and Brereton (2013) analyzed 68 studies and found that the time required to conduct a systematic review and difficulties regarding quality assessment are problematic. This finding provides extra arguments for sophisticated tool support. In their study, authors also found current digital libraries not appropriate for broad literature searches. This is also supported by Brereton et al. (2007), who specifically found the indexing of those digital libraries inadequate and also mention that the quality of paper abstracts is too poor, e.g., to judge upon the relevance of a paper based on its abstract only. This provides a rationale for different search and selection strategies (Ali and Petersen 2014; Badampudi et al. 2015; Zhang et al. 2011). A more general discussion is raised by Staples and Niazi (2007), who generally recommend using guidelines, but also mention a need to optimize the process as such (e.g., narrowly defined research questions, improved selection procedures, and improved data extraction) to reduce the effort needed to conduct such a study. However, exemplary research questions to start a literature study are only provided by Petersen et al. (2015) as part of the analysis of other studies, thus, being focused to the respective study subjects—the presented list of quoted research questions does not serve the generalization. Dybå et al. (2007) consider "normal" meta-analytic approaches to be of limited use for software engineering only and, hence, report their experience from applying diverse study types in a systematic review; a mixed-method approach similar to the practices reported in the present article. Riaz et al. (2010) provide a different perspective in their report and mention experts and novices having a different perception of the systematic review process and its challenges. The present article also addresses this point by providing examples, reusable assets like research questions or in-/exclusion criteria, and a detailed elaboration on selected practices and a demonstration of their use. Such challenges are also addressed by Fabbri et al. (2013), who provide an experienced-based guideline that comes as integrated process with the purpose of externalizing tacit knowledge about the process and its implementation. In contrast to Fabbri et al. (2013), the present article is not supposed to be a self-contained comprehensive guideline covering the process of conducting a literature study as a whole. Instead, we focus on the early stages and provide a limited, but interlinked set of practices illustrated by examples and reusable building blocks, which we also compile into reference workflows to follow.

Tools The body of knowledge in software engineering is growing and, thus, literature studies are likely to grow in size and complexity as well. Tool support has therefore become crucial to collect, manage, and evaluate data. However, the question of what can be considered as proper tool support has puzzled researchers for years (Hassler et al. 2016; Tell et al. 2016). A group around Marshall conducted research on tool support for literature studies (Marshall and Brereton 2013, 2015; Marshall et al. 2014, 2015). Among others, they provided a feature analysis to define basic requirements (Marshall et al. 2014), and in Carver et al. (2013), authors found a strong need to provide support for planning and teamwork when conducting a literature study. In Marshall et al. (2015), the same author group concluded a recommended list of requirements, which was generated based on 13 semi-structured interviews. Yet, the requirements list only provides a high-level overview of features that opens a fairly large design space that should be carefully considered when designing tools. The challenges coming along with this large design space were explicitly addressed in Tell et al. (2016) in which we, based on a shared set of requirements, independently developed two tools—both realizations with different features emphasized and implementing different work and collaboration patterns. Over the years, few tools dedicated to support researchers performing systematic reviews have been proposed; notable examples are SLuRp (Bowes et al. 2012), SESRA (Molléri and Benitti 2015), and StArt Fabbri et al. (2016). These tools were analyzed in Marshall et al. (2014), yet those are not ranked with flying colors. Still, the classic spreadsheet application (quite often) in combination with so-called reference managers (e.g., EndNote, Mendeley, Papers, and Zotero) seem to build the standard tooling for literature studies.

Summary of Related Works The present article contributes to the body of knowledge by stressing the need for more concrete advice to complement the generic guidelines, and by offering an experience-based guideline especially to perform the steps in the early stages. Although for instance Petersen et al. (2015) provide a comprehensive selection of practices used for these stages, a streamlined approach to presenting, explaining, and linking these steps to each other is not in scope of their contribution. In a nutshell, most of the available guidelines are focused on *what* a design should accomplish rather than on *how* and *why* a particular step should be executed in a cost-effective way. For example, we found no guideline explaining what pieces of information are worthwhile including and what justified particular configurations of descriptive data pieces to be taken care of by the researchers. Our recommended minimal data structure (Table 8) can be directly used by researchers facing this question. Furthermore, no guideline so far discussed in detail the ways to run a voting procedure. We provided an operationalized description on how to do this in a systematic way, along with a discussion on a research team model, a scaling/vote calculation schema and a demonstration of a potential technical realization based on the suggested minimal data structure (Fig. 10). Based on our reported experience, we also provide a description of the work deliverables that are produced during a literature study process and the dependencies among the deliverables (Section 2.4), and we shared our lessons learned regarding the issues coming along with handling search engines, which are barely discussed in available guidelines.

5 Conclusion

Systematic literature studies have become a powerful means to elaborate and structure the state of reported knowledge. Especially in the software engineering community, they have received much attention in recent years. Despite their relevance to the community and first valuable proposals of guidelines, they are difficult to conduct, require a lot of effort and depend on experiences and expertise of the researchers involved. Especially the latter decides often over the success of a study, depending on aspects such as

- Appropriateness of the research questions and value to the community,
- Accuracy of the design, or
- Reproducibility of the data collection.

When conducting literature studies, there are various challenges all concerning the initial stages of the data collection rather than the particularities of the later analytical phase, and there are challenges that concern the organization of such a study.

In this article, we reported on our experiences made in the course of various literature studies and contributed an experience-based guideline that puts strong emphasis on tackling some practical challenges. Our aim was to specifically support young scholars facing their first literature study and to provide them with a pragmatic and easy-to-enact guideline. To this end, we collected and structured our experiences, and we also shared our experiences in utilizing different tools to support the data collection, the dataset cleaning, or the study selection procedures. Furthermore, we provided some generalized blueprint-style workflows to follow in a particular study, also increase the efficiency in the way study designs are reported in papers within the space limitations given for conference submissions so that the used approaches don’t have to be justified from scratch all the time drowning the presentation of the results out.

While compiling this guideline, we also realized again the need for fine-grained guidelines and, moreover, the need for a sophisticated tool support. As a matter of fact, all our studies were conducted utilizing fairly simple tools, such as spreadsheets or plain text files to feed further external tools, e.g., word cloud generators. However, having conducted the data collection, research teams have rich data available, which could be used for extensive tool-support. Yet, comprehensive tools are not yet publicly available or, if at all, in their early stages of their development as for instance (Tell et al. 2016). This indicates to a strong need to (1) increasing the effort spent on developing applicable procedures and fine-grained reference workflows from the available knowledge and experience, and (2) to put effort into the development of tools to support literature studies. These tools need to support the collection of data, their storage and organization, the management of in-/exclusion criteria, support to implement workflows for paper selection and classification, which also includes the management of classification schemas, and, eventually, supporting the connection to further tools, e.g., word cloud generators, statistics software, and social network analysis tools.

Acknowledgments We want to thank Roel Wieringa for fruitful discussions on previous versions of this article and all our students, especially those who contributed to our previously conducted literature studies over the past years. Finally, we are grateful for the constructive feedback provided by the anonymous reviewers of this article who helped improving it substantially.

Appendix: A Study Workflow Templates

In this appendix, we provide selected workflow templates, which we inferred from experiences (Table 5) for simple reuse in research method descriptions of scientific papers. The provided templates can be used to inspire or shorten the description of research methods, which especially in conference papers consumes much precious space. For each model in subsequent sections, we provide a brief context description, an exemplary workflow, and textual description.

A.1 Template 1: 2 Researcher Workshop Model with Snowballing

Context This model addresses smaller literature studies in which just two researchers collaborate, thus, having no option to implement more comprehensive study selection procedures, such as majority votes. Our experience shows this model to be well-applicable in settings with up to approximately 50 papers, two senior or one senior and one junior researcher, and in distributed settings. Apart from an initial research objective and/or a set of research questions and a (small) set of reference publications, no extra entry conditions need to be fulfilled.

Workflow Figure 8 illustrates the basic workflow for this model including some notes emphasizing the most relevant points to be considered.

Workflow Description The *2 Researcher Workshop Model with Snowballing* is implemented as follows: Right in the beginning of the study, a snowballing-based preliminary study is conducted. For this pre-study, a set of reference papers is selected to lay the foundation for an (incremental) snowballing search. When the snowballing is done, the obtained papers are analyzed for keywords, which are used to construct the search queries for an automated database search. As the last preparation steps, the data sources of interest are selected and the inclusion and exclusion criteria are defined.

The data collection is performed (according to the search strategy, Section 2.2.1). After the search, the dataset is cleaned (Section 2.2.2), e.g., by a stepwise integration of individual datasets. The kick-off meeting is—on the one hand—closing the data collection and cleaning phase and—on the other hand—starts the study selection phase. In the kick-off meeting, both researchers reflect on all the criteria, inspect and prepare the dataset for the rating, and agree on a schedule. According to the procedure illustrated in Fig. 5, each researcher gets a copy of the dataset and carries out the individual rating. When the rating is done, both datasets are integrated and checked for consensus. In a rating workshop (or multiple workshops), both researchers iterate through the dataset discussing all items that are not yet decided to find an agreement. When the concluding integration is done, the study selection phase is closed and the result set is transferred to the main study (Section 2.4). For handing over the result set, a copy of the fully rated result set is created for archiving, and the actual

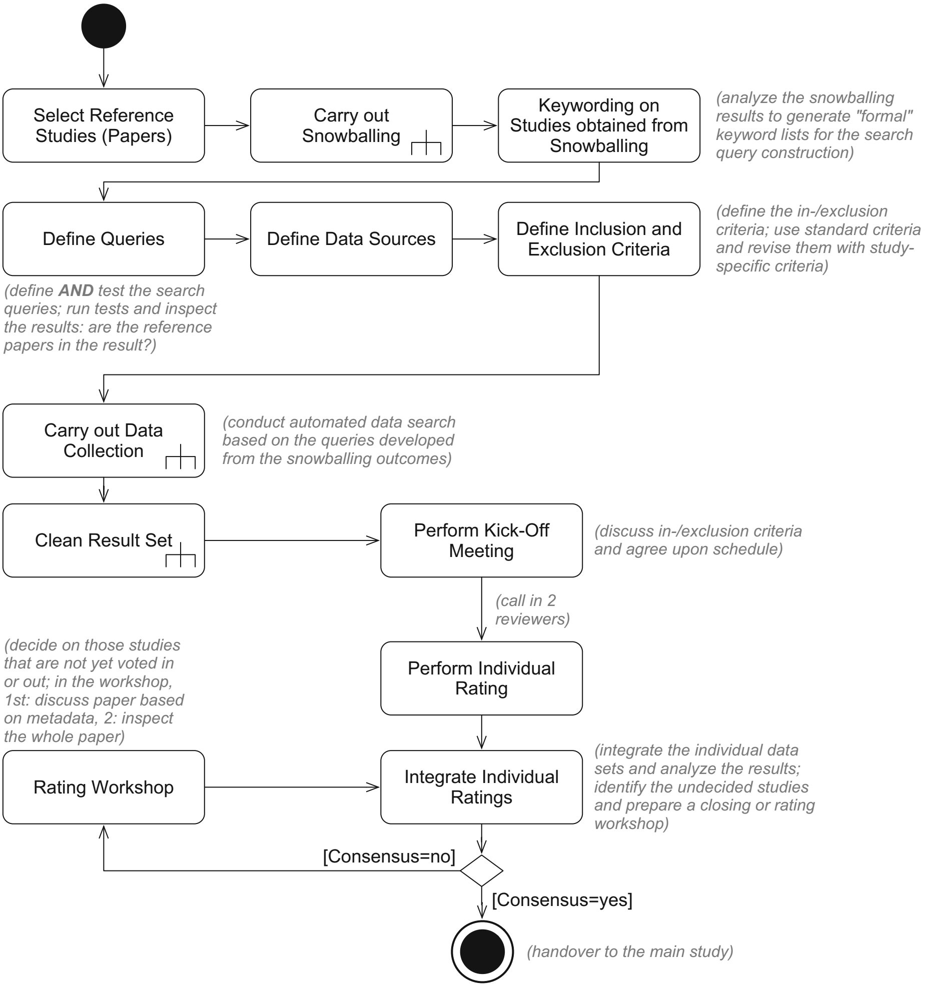

Select Reference Studies (Papers)
Carry out Snowballing
Keywording on Studies obtained from Snowballing
(analyze the snowballing results to generate "formal" keyword lists for the search query construction)
Define Queries
Define Data Sources
Define Inclusion and Exclusion Criteria
(define the in-/exclusion criteria; use standard criteria and revise them with study-specific criteria)
(define AND test the search queries; run tests and inspect the results: are the reference papers in the result?)
Carry out Data Collection
(conduct automated data search based on the queries developed from the snowballing outcomes)
Clean Result Set
Perform Kick-Off Meeting
(discuss in-/exclusion criteria and agree upon schedule)
(call in 2 reviewers)
Perform Individual Rating
(decide on those studies that are not yet voted in or out; in the workshop, 1st: discuss paper based on metadata, 2: inspect the whole paper)
Rating Workshop
Integrate Individual Ratings
(integrate the individual data sets and analyze the results; identify the undecided studies and prepare a closing or rating workshop)
[Consensus=no]
[Consensus=yes]
(handover to the main study)

Fig. 8 Exemplary workflow for the 2 researcher workshop model with a snowballing-based preliminary study

result set is reduced, i.e., those dataset items that were rated as irrelevant for the main study are removed from the dataset so that only relevant data finds its way into the analysis.

A.2 Template 2: 3 Researcher Voting-only Model

Context This model addresses literature studies in which three researchers collaborate and implement a voting-based study selection procedure. Our experience shows this model to be well-applicable in the majority of all literature study settings. This model supports mixed and distributed teams, whereas at least one senior researcher has to be involved to guide the study project. Our standard implementation of the *3 Researcher Voting-only Model* follows the 2+1 approach (Fig. 5, p. 15), i.e., the voting procedure to select relevant papers is organized by two researchers carrying out the full voting independently and calling in a third researcher to make the final decisions. In order to set up a study following this model, research objectives and questions, keyword lists and accordingly derived search queries have to be in place; optionally, a (small) set of reference publications is available.

Workflow Figure 9 illustrates the basic workflow for this model including some notes emphasizing the most relevant points to be considered.

Workflow Description The *3 Researcher Voting-only Model* is implemented as follows: After defining the search queries, data sources of interest, and the required inclusion and exclusion criteria, actual data collection is performed (Section 2.2.1). After the data collection, the data sets are cleaned (Section 2.2.2), e.g., via a stepwise integration of individual datasets.

In the kick-off meeting, the team of researchers nominates two researchers who will conduct the initial rating. According to the procedure illustrated in Fig. 5, each of the two selected researchers gets a copy of the integrated dataset for carrying out the individual rating. When both researchers have rated the dataset, one of them integrates both

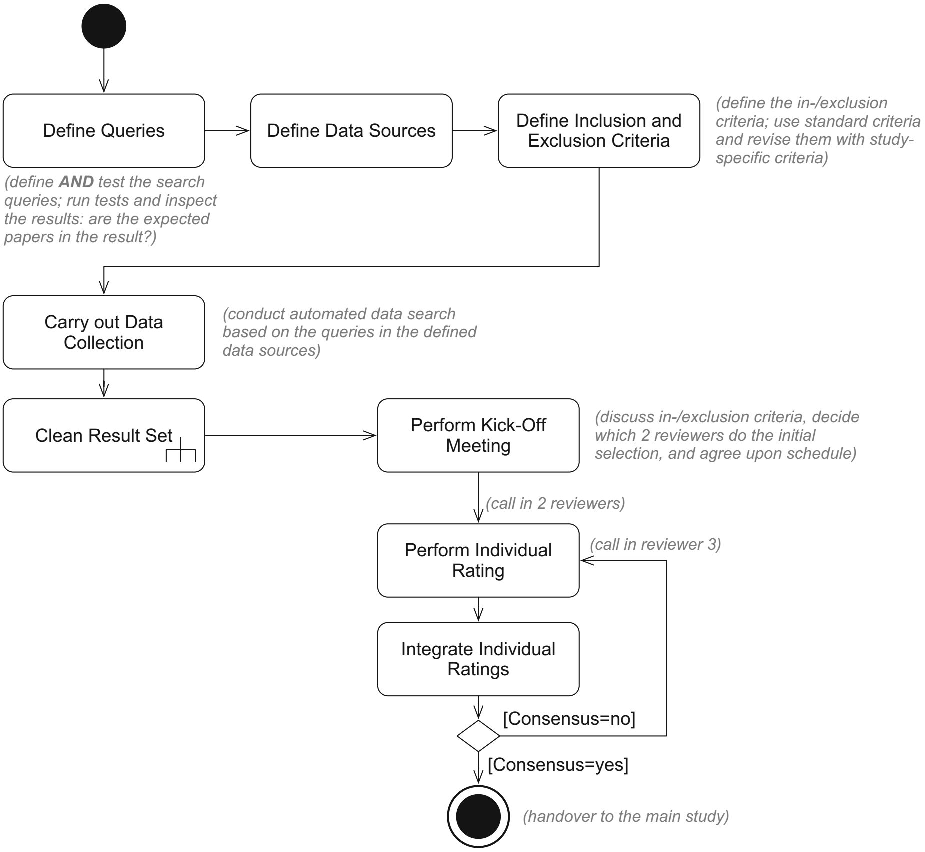

Define Queries
Define Data Sources
Define Inclusion and Exclusion Criteria
(define the in-/exclusion criteria; use standard criteria and revise them with study-specific criteria)
(define **AND** test the search queries; run tests and inspect the results: are the expected papers in the result?)
Carry out Data Collection
(conduct automated data search based on the queries in the defined data sources)
Clean Result Set
Perform Kick-Off Meeting
(discuss in-/exclusion criteria, decide which 2 reviewers do the initial selection, and agree upon schedule)
(call in 2 reviewers)
Perform Individual Rating
(call in reviewer 3)
Integrate Individual Ratings
[Consensus=no]
[Consensus=yes]
(handover to the main study)

Fig. 9 Exemplary workflow for a data collection and study selection approach for 3 reviewers using a voting-only approach

and analyzes the integrated result set for the agreement. Those dataset items that are not yet decided are selected and exported in a reduced dataset, which is given to the third reviewer. The third reviewer then performs a rating on the reduced dataset and, eventually, integrates the outcome with the full dataset. After performing this third rating, the dataset is now fully decided and can be prepared to be transferred to the main analysis (Section 2.4). If using a tool-supported approach as, for instance, shown in Fig. 10, the different stages can be supported by simple calculation, scripts, and conditional formatting (color coding).

Appendix: B Recommended Data Structure

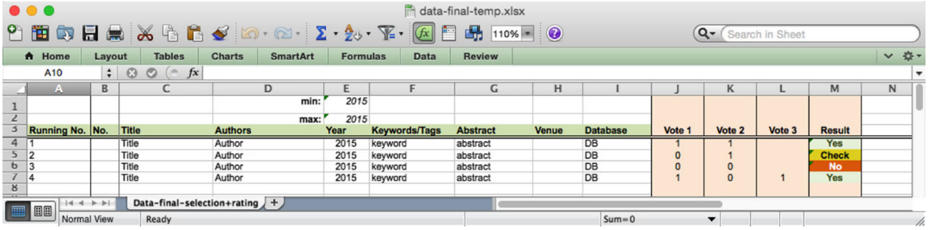
In this section, we present a recommendation of a data structure to store data obtained by a manual/automatic literature search. Table 8 presents this recommended data structure, which emerges from several literature studies (Table 5), and the table explains the meaning of the different fields. Note: We consider the presented data structure to be *minimal*, i.e., specific studies will require further data fields. However, due to the absence of comprehensive and mature tools to support mapping studies, the normal would be to set up a simple spreadsheet. Examples of such spreadsheets (Fig. 10) can be obtained from http://goo.gl/PBylsn.

The data structure as presented in Table 8 only contains a minimal set of data, which needs to be extended according to the study's scope. For systematic mapping studies, the following extra data should be contained:

- Generic/reused classification schemas, such as research/contribution type facet (Wieringa et al. 2005, Petersen et al. 2008)
- Study-specific classification schemas, such as focus type facets (Paternoster et al. 2014) or rigor/relevance models (Ivarsson and Gorschek 2011)
- In-/exclusion criteria to document, why a paper was in-/excluded (cf. Table 2)

Furthermore, grounded in our experience from Kuhrmann et al. (2015), we also recommend adding "dynamic metadata" to the data structure (as already mentioned in Table 8). Such metadata can be added on-the-fly and can support the enhancement of the dataset. From our experience (Kuhrmann et al. 2016), we recommend to collect metadata at least from the dimensions *Study* and *Context*.

The dimension *Study* covers the overall research approach followed in a particular paper, e.g., is a particular paper a primary study, a replication, or even a secondary study, and it



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				min:	2015									
2				max:	2015									
3	Running No.	No.	Title	Authors	Year	Keywords/Tags	Abstract	Venue	Database	Vote 1	Vote 2	Vote 3	Result	
4	1		Title	Author	2015	keyword	abstract		DB	1	1		Yes	
5	2		Title	Author	2015	keyword	abstract		DB	0	1		Check	
6	3		Title	Author	2015	keyword	abstract		DB	0	0		No	
7	4		Title	Author	2015	keyword	abstract		DB	1	0	1	Yes	
8														

Fig. 10 Example of a color-coded voting spreadsheet. The sheet shows different combinations of a 3-person majority vote (2 reviewers + 1 extra reviewer for final decisions)

Table 8 Recommended minimal data structure

Field	Cardinality	Description
No.	1	The overall publication number in the integrated dataset.
DB-No.	1	The database-specific number if a paper from the individual literature database to allow for linking an entry to the originating dataset.
Title	1	Title of the publication.
Authors	1, 1..n	Authors of the publication; either integrated in one cell and separated by special characters (e.g., ";"), or converted into a one-author-per-cell pattern, i.e. there are n columns to represent the author list.
Keywords	1	List of keywords separated by special characters (e.g., "," or ";").
Abstract	1	Abstract of the paper.
Year	1	Year of publication (note: e.g., for journals, there might be multiple dates, such as accepted, online available, preprint, published, etc.—it is required to define which of these is the one that makes it into the dataset).
Publisher/Database	1	Which database created this item? In case of cross-indexing, publisher and originating database can differ, e.g., IEEE Xplore also lists IET papers.
Source/Venue	1	Which source or venue published this paper? In case of a conference, this field should contain the conference name and/acronym, in case of a journal, the name/acronym of the journal should be contained, and so forth.
Publication vehicle	1..n	For every publication vehicle, an individual column should be present, e.g., journal, magazine, conference, workshop, book, chapter, misc, and so forth. Experience shows individual columns beneficial for later analyses.
General comments	1	Provide some space for general comments.
Metadata classes (optional)	0..n	It was shown beneficial to provide some space for metadata, for example, this is a survey, a literature review, this deals with Agile, and so forth. The number of metadata is not limited and can be extended during analysis. Furthermore, metadata should allow for categorization, that is, one column per metadata class should be provided.

can even contain the research methods used, such as interview research or grounded theory analyses. Metadata from this category supports a more detailed classification and analysis of papers regarding the research and contribution type facets. The dimension *Context* aims at collecting as much context information from the selected papers as possible, such as the software engineering lifecycle phase addressed by a paper (e.g., design, coding, test), the organizational context in which the research was conducted (e.g., SMEs, global players etc.), and the application domain of a paper (e.g., automotive software or software for the healthcare domain).

References

Ali NB, Petersen K (2014) Evaluating strategies for study selection in systematic literature studies. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM. ACM, New York, pp 45:1–45:4. doi:10.1145/2652524.2652557

Badampudi D, Wohlin C, Petersen K (2015) Experiences from using snowballing and database searches in systematic literature studies. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 17:1–17:10. doi:10.1145/2745802.2745818

Bowes D, Hall T, Beecham S (2012) SLurp: a tool to help large complex systematic literature reviews deliver valid and rigorous results. In: Proceedings of the International Workshop on Evidential Assessment of Software Technologies. ACM, NY, USA, pp 33–36

Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M (2007) Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 80(4):571–583. doi:10.1016/j.jss.2006.07.009

Carver JC, Hassler E, Hernandes E, Kraft NA (2013) Identifying barriers to the systematic literature review process. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM. IEEE, Washington, DC, pp 203–212. doi:10.1109/ESEM.2013.28

Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213–220

Condori-Fernandez N, Daneva M, Sikkel K, Wieringa R, Dieste O, Pastor O (2009) A systematic mapping study on empirical evaluation of software requirements specifications techniques. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM. IEEE, Washington, DC, pp 502–505. doi:10.1109/ESEM.2009.5314232

Dybå T, Dingsøyr T, Hanssen GK (2007) Applying systematic reviews to diverse study types: An experience report. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM. IEEE, Washington, pp 225–234. doi:10.1109/ESEM.2007.21

Fabbri S, Silva C, Hernandes E, Octaviano F, Di Thommazo A, Belgamo A (2016) Improvements in the start tool to better support the systematic review process. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 21:1–21:5. doi:10.1145/2915970.2916013

Fabbri SCPF, Felizardo KR, Ferrari FC, Hernandes ECM, Octaviano FR, Nakagawa EY, Maldonado JC (2013) Externalising tacit knowledge of the systematic review process. IET Softw 7(6):298–307. doi:10.1049/iet-sen.2013.0029

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382

Hanneman A, Riddle M (2005) Introduction to social network methods Online http://faculty.ucr.edu/∼hanneman/

Hassler E, Carver JC, Hale D, Al-Zubidy A (2016) Identification of slr tool needs – results of a community workshop. Inf Softw Technol 70:122–129. doi:10.1016/j.infsof.2015.10.011

Inayat I, Salim SS, Marczak S, Daneva M, Shamshirband S (2015) A systematic literature review on agile requirements engineering practices and challenges. Comput Hum Behav 51, Part B:915–929. doi:10.1016/j.chb.2014.10.046

Ingibergsson J, Schultz U, Kuhrmann M (2015) On the use of safety certification practices in autonomous field robot software development: a systematic mapping study. In: Proceedings of the International Conference on Product Focused Software Development and Process Improvement, Lecture Notes in Computer Science, vol 9459. Springer, Berlin Heidelberg, pp 335–352

Ivarsson M, Gorschek T (2011) A method for evaluating rigor and industrial relevance of technology evaluations. Empir Softw Eng 16(3):365–395. doi:10.1007/s10664-010-9146-4

Jacobson JW, Kuhrmann M, Münch J, Diebold P, Felderer M (2016) On the role of software quality management in software process improvement. In: Proceedings of the International Conference on Product-Focused Software Process Improvement, Lecture Notes in Computer Science, vol 10027. Springer, Berlin, Heidelberg, pp 327-343

Kalus G, Kuhrmann M (2013) Criteria for software process tailoring: a systematic review. In: Proceedings of the International Conference on Software and System Process, ICSSP. ACM Press, New York, pp 171–180

Kitchenham B (2004) Procedures for performing systematic reviews. Technical Report. TR/SE-0401 Keele University

Kitchenham B, Brereton P (2013) A systematic review of systematic review process research in software engineering. Inf Softw Technol 55(12):2049–2075. doi:10.1016/j.infsof.2013.07.010

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Technical Report. EBSE-2007-01 Keele University

Kitchenham BA, Budgen D, Brereton P (2015) Evidence-Based Software engineering and systematic reviews. CRC Press

Kuhrmann M, Diebold P, Münch J (2016) Software process improvement: A systematic mapping study on the state of the art. Peer J Comput Sc 2:e62

Kuhrmann M, Diebold P, Münch J, Tell P (2016) How does software process improvement address global software engineering? In: International Conference on Global Software Engineering, ICGSE. IEEE, Washington, DC, pp 89–98

Kuhrmann M, Fernández DM, Gröber M (2013) Towards artifact models as process interfaces in distributed software projects. In: Proceedings of the International Conference on Global Software Engineering, ICGSE. IEEE, Washington, DC, pp 11–20

Kuhrmann M, Fernández DM, Steenweg R (2013) Systematic software process development: Where do we stand today? In: Proceedings of the International Conference on Software and System Process, ICSSP. ACM Press, New York, pp 166–170

Kuhrmann M, Fernández DM, Tiessler M (2014) A mapping study on the feasibility of method engineering. J Softw: Evol Process 26(12):1053–1073

Kuhrmann M, Konopka C, Nellemann P, Diebold P, Münch J (2015) Software process improvement: Where is the evidence? In: Proceedings of the International Conference on Software and Systems Process, ICSSP. ACM, New York, pp 107–116

Kuo BYL, Hentrich T, Good BM, Wilkinson MD (2007) Tag clouds for summarizing web search results. In: Proceedings of the International Conference on World Wide Web, WWW. ACM, New York, pp 1203–1204. doi:10.1145/1242572.1242766

Marshall C, Brereton P (2013) Tools to support systematic literature reviews in software engineering: A mapping study. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM. IEEE, Washington, DC, pp 296–299. doi:10.1109/ESEM.2013.32

Marshall C, Brereton P (2015) Systematic review toolbox: a catalogue of tools to support systematic reviews. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 23:1–23:6

Marshall C, Brereton P, Kitchenham B (2014) Tools to support systematic reviews in software engineering: a feature analysis. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 13:1–13:10

Marshall C, Brereton P, Kitchenham B (2015) Tools to support systematic reviews in software engineering: a cross-domain survey using semi-structured interviews. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 26:1–26:6

Méndez Fernández D, Ognawala S, Wagner S, Daneva M (2014) Where do we stand in requirements engineerign improvement today? first results from a mapping study. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM. ACM, New York, pp 58:1–58:4

Molléri J. S, Benitti FBV (2015) SESRA: a web-based automated tool to support the systematic literature review process. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 24:1–24:6

Oosterman J, Cockburn A (2010) An empirical comparison of tag clouds and tables. In: Proceedings of the Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction, OZCHI. ACM, New York, pp 288–295. doi:10.1145/1952222.1952284

Paternoster N, Giardino C, Unterkalmsteiner M, Gorschek T, Abrahamsson P (2014) Software development in startup companies: A systematic mapping study. Inf Softw Technol 56(10):1200–1218. doi:10.1016/j.infsof.2014.04.014

Penzenstadler B, Raturi A, Richardson D, Calero C, Femmer H, Franch X (2014) Systematic mapping study on software engineering for sustainability (SE4S). In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 14:1–14:14. doi:10.1145/2601248.2601256

Petersen K, Ali NB (2011) Identifying strategies for study selection in systematic reviews and maps. In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement, ESEM. IEEE, Washington DC, pp 351–354. doi:10.1109/ESEM.2011.46

Petersen K, Feldt R, Mujtaba S, Mattson M (2008) Systematic mapping studies in software engineering. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, New York, pp 68–77

Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 64:1–18

Portillo-Rodríguez J, Vizcaíno A, Piattini M, Beecham S (2012) Tools used in global software engineering: A systematic mapping review. Inf Softw Technol 54(7):663–685. doi:10.1016/j.infsof.2012.02.006

Racheva Z, Daneva M, Sikkel K (2009) Value creation by agile projects: Methodology or mystery? In: Product-Focused Software Process Improvement, Lecture Notes in Business Information Processing, vol 32. Springer, Berlin Heidelberg, pp 141–155. doi:10.1007/978-3-642-02152-7_12

Ramage D, Dumais S, Liebling D (2010) Characterizing microblogs with topic models. In: Proceedings of the International AAAI Conference on Weblogs and Social Media. Association for the advancement of artificial intelligence, pp 130–137

Riaz M, Sulayman M, Salleh N, Mendes E (2010) Experiences conducting systematic reviews from novices' perspective. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. British Computer Society, Swinton, UK, pp 44–53

Rivadeneira AW, Gruen DM, Muller MJ, Millen DR (2007) Getting our head in the clouds: Toward evaluation studies of tagclouds. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI. ACM, New York, pp 995–998. doi:10.1145/1240624.1240775

Schramm J, Dohrmann P, Rausch A, Ternité T (2014) Process model engineering lifecycle: Holistic concept proposal and systematic literature review. In: Proceedings of the Euromicro Conference on Software Engineering and Advanced Applications, SEAA. IEEE, Washington, DC, pp 127–130

Schrammel J, Leitner M, Tscheligi M (2009) Semantically structured tag clouds: An empirical evaluation of clustered presentation approaches. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI. ACM, New York, pp 2037–2040. doi:10.1145/1518701.1519010

Scott J (2000) Social network analysis: A handbook, 2nd edn. ISBN-13: 978-0761963394. SAGE Publications

Shaw M (2003) Writing good software engineering research papers: Minitutorial. In: International Conference on Software Engineering, ICSE. IEEE, DC, USA, pp 726–736

Staples M, Niazi M (2007) Experiences using systematic review guidelines. J Syst Softw 80(9):1425–1437. doi:10.1016/j.jss.2006.09.046

Tell P, Cholewa J, Nellemann P, Kuhrmann M (2016) Beyond the spreadsheet: Reflections on tool support for literature studies. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, EASE. ACM, NY, USA, pp 22:1–22:5

Theocharis G, Kuhrmann M, Münch J, Diebold P (2015) Is Water-Scrum-Fall reality? on the use of agile and traditional development practices, vol 9459. Springer, Berlin, Heidelberg

Wasserman S, Faust K (1994) Social network analysis: Methods and applications, 2nd edn. University Press, Cambridge

Wieringa R, Maiden N, Mead N, Rolland C (2005) Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. Requir Eng 11(1):102–107. doi:10.1007/s00766-005-0021-6

Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer

Wohlin C, Runeson P, Da Mota Silveira Neto PA, Engströmb E, Do Carmo Machado I, De Almeida ES (2013) On the reliability of mapping studies in software engineering. J Syst Softw 86(10):2594 – 2610. doi:10.1016/j.jss.2013.04.076

Zhang H, Babar MA, Tell P (2011) Identifying relevant studies in software engineering. Inf Softw Technol 53(6):625–637. doi:10.1016/j.infsof.2010.12.010

Marco Kuhrmann is an associate professor of software engineering at the University of Southern Denmark, Odense. His research is focused on software quality and process management and improvement, in particular on hybrid software systems development. He was part of the core development team of the V-Modell XT – the German standard IT system development process, and supported several software process improvement initiatives in industry and public sector. He is a member of the ACM, IEEE Computer Society, Gesellschaft für Informatik (GI) e.V., and the German association of university professors and lecturers.

Daniel Méndez Fernández is a senior research fellow in software & systems engineering at the Technical University of Munich, Germany. He is further director of the junior research groups at the Centre Digitisation.Bavaria. His research is on (empirical) software and systems engineering with a particular focus on interdisciplinary, qualitative research in Requirements Engineering and its quality improvement. He is a member of the ACM, the IEEE Computer Society, and the German association of university professors and lecturers.

Maya Daneva, PhD, is Senior Member of Scientific Staff in the Services, Cybersecurity and Safety group at the University of Twente, the Netherlands. Her key research interests are empirical research methods, evidence-based software engineering, requirements engineering for large systems, requirements-based project estimation, and user feedback analytics. Maya has a strong international exposure having spent two years of her career in Germany at the University of Saarbrücken and in the IDS Scheer, and 9 years as a business process analyst at TELUS Corporation, Canadas second largest telecommunication company. At Twente, she is a leading researcher in seven industry-university research projects. Maya serves as the Academic Liaison to the Dutch Industrys Software Measurement Association and as the University of Twentes representative to ISERN, the International Empirical Software Engineering Research Network.