Software Process Improvement: Where Is the Evidence?

Initial Findings from a Systematic Mapping Study

Marco Kuhrmann
University of Southern Denmark
Odense, Denmark
kuhrmann@mmmi.sdu.dk

Claudia Konopka
4Soft GmbH
Munich, Germany
claudia.konopka@4soft.de

Peter Nellemann
University of Southern Denmark
Odense, Denmark
peter@dineitkonsulenter.dk

Philipp Diebold
Fraunhofer IESE
Kaiserslautern, Germany
philipp.diebold@iese.fhg.de

Jürgen Münch
University of Helsinki
Helsinki, Finland
Juergen.Muench@cs.helsinki.fi

ABSTRACT

Software process improvement (SPI) is around for decades: frameworks are proposed, success factors are studied, and experiences have been reported. However, the sheer mass of concepts, approaches, and standards published over the years overwhelms practitioners as well as researchers. What is out there? Are there new emerging approaches? What are open issues? Still, we struggle to answer the question for what is the current state of SPI and related research? In this paper, we present initial results from a systematic mapping study to shed light on the field of SPI and to draw conclusions for future research directions. An analysis of 635 publications draws a big picture of SPI-related research of the past 25 years. Our study shows a high number of solution proposals, experience reports, and secondary studies, but only few theories. In particular, standard SPI models like CMMI and ISO/IEC 15504 are analyzed, enhanced, and evaluated for applicability, whereas these standards are critically discussed from the perspective of SPI in small-to-medium-sized companies, which leads to new specialized frameworks. Furthermore, we find a growing interest in success factors to aid companies in conducting SPI.

Categories and Subject Descriptors

D.2.9 [**Software Engineering Management**]: Software process models

General Terms

Experimentation, Measurement

Keywords

software process, software process improvement, systematic mapping study

1. INTRODUCTION

Software process improvement (SPI; [12]) aims to improve software processes and comprises a variety of tasks, such as scoping, assessment, design and realization, and continuous improvement (e.g., [15, 17]). A number of SPI models competes for the companies' favor, success factors to support SPI implementation at the large and the small scale are studied, and numerous publications report on experiences in academia and practice. SPI is considered an important topic (according to Horvat et al. [10] regardless of company size), as many companies put emphasis on the software process and its adaptation to the company context [6, 19, 40] to improve the quality of software products or to accelerate software development.

However, SPI is a diverse field: On the one hand, a number of standards is available, e.g., ISO/IEC 15504 [13] or CMMI [4] but, on the other hand, these standards are criticized oftentimes, e.g., [3, 5, 34]. In response, tailored SPI models are proposed, inter alia, to better address needs of small and very small companies, e.g., [14, 27, 29, 30].

Moreover, since SPI is mainly a human endeavor, much research was spent to study human factors, e.g., [1, 35, 39]. Beyond, we find numerous experience reports, guidelines, and tools—all together providing a huge body of knowledge on SPI.

However, despite this comprehensive body of knowledge, we still struggle to answer questions like: What is out there? What are open issues? Are there new emerging approaches? What is the current state of SPI and related research?

Problem Statement. The field of SPI evolved for decades and provides a vast amount of publications addressing a huge variety of topics. Still, we see new method proposals, research on success factors, and experience reports. However, missing is a big picture that illustrates where SPI gained a certain level of saturation and where are still hot topics and unresolved issues calling for more investigation.

Objectives. To better understand the state of the art in SPI, we aim to analyze the whole publication flora to draw a big picture on SPI. Our overall goal is *not* to judge particular SPI research directions, but to provide the focus points of the past and to illustrate emerging/unresolved areas.

Contribution. In this paper, we present initial findings from a comprehensive systematic mapping study. We con-

ducted a broadband search in six literature databases to harvest SPI-related publications, and we analyzed the resulting 635 publications for publication frequency, research type facet, contribution type facet, and focus type facet. We draw a big picture that shows that the majority of publications on SPI either proposes new approaches (i.e., models or frameworks) or is of philosophical nature (i.e., collecting, structuring, and analyzing knowledge). Our results show continuous publication of new approaches while evaluation research regarding these proposals is scarcely available. Furthermore, our data shows rare evidence and, notably, missing long-term and independently replicated studies. However, our data also reveals some (still) emerging topics, e.g., SPI for small and very small companies, and SPI in the context of lean and agile methods.

Outline. The remainder of the paper is organized as follows. Section 2 summarizes the related work. In Section 3, we describe the research design, and discuss the results in Section 4. We conclude the paper in Section 5.

2. RELATED WORK

Literature on software process improvement is rich and addresses a variety of topics. Yet, available secondary studies mainly focus on investigating success factors, e.g., Monteiro and Oliveira [21], Bayona-Oré [2], and Dybå [7]. Some studies provide insights into selected topics. For example, Helgesson et al. [9] review maturity models, and Hull et al. [11] and El-Emam and Goldenson [8] review different assessment models. Pino et al. [26] contribute a review on SPI in the context of small and very small companies, and Staples and Niazi [32] study motivating factors to adopt CMMI for improvement programs, while Müller et al. [22] study SPI in general from the perspective of organizational change. All these representatively selected studies address specific topics, yet, they do not contribute to a more general perspective on SPI.

Such general studies are scarcely to find. For instance, Rainer and Hall [28] analyze some 'core' studies on SPI for the purpose to work out addressed topics and gaps in the domain. However, they select only few studies of which they assume to be good representatives thus providing a limited picture only.

In terms of analyzing the entire domain and providing new (generalizable) knowledge, Unterkalmsteiner et al. [38] contribute a systematic review on the state of the art of evaluation and measurement in SPI. They conduct a systematic literature review for the purpose of synthesizing a list of evaluation and measurement approaches, which they also analyze for the practical application.

The study at hand does not aim at generating generalizable knowledge for one or more SPI-related topics. The purpose of the present study is to draw a big picture of the current state of the art of SPI in general. That is, as there is no comparable study available, this paper closes a gap in literature by providing a comprehensive picture of the development of the field of SPI over time and by summarizing the current state of the art. Other than, e.g., [28] or [38], we use the mapping study instrument according to Petersen et al. [25] as research method and to present our results. Therefore, our study does not address selected details, but aims to draw a general picture from a "bird's-eye perspective."

3. RESEARCH DESIGN

In this section, we present the study design. After describing the research questions, we describe the overall research design, and the case study instrument including case selection, data collection, and analysis procedures.

3.1 Research Method

In this study, we followed the general approach as applied in [18] in which we applied different methods from systematic literature reviews (SLR; Kitchenham et al. [16]) and systematic mapping studies (SMS; Petersen et al. [25]). Figure 1 shows the overall research approach.

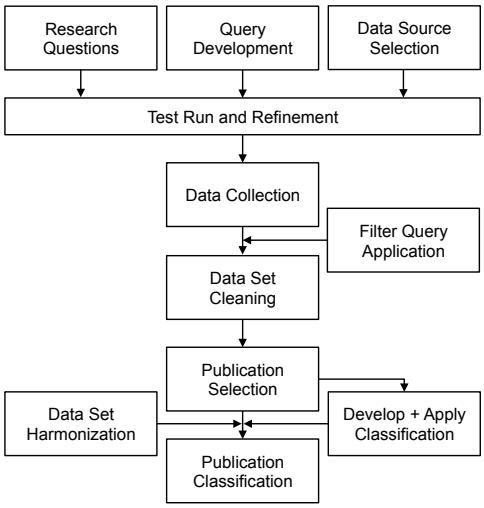

Figure 1: Overall research approach.

The overall study was designed as a breadth-first search to capture the SPI domain as complete as possible. In February 2013, we performed the study preparation (definition of research questions and search queries, and the selection of data sources), conducted a series of test runs, and refined the search queries iteratively. End of April 2013, we conducted the main search, which results in about 85,000 hits. As we expected this large number of results and in order to support the data set cleaning, we defined filter questions, which we applied to the initial result set. When the initial result set was cleaned, we performed a voting procedure to select the relevant publications from the result set. Based on this selection, we developed the classification schemas (by manual sampling as well as tool-supported) and harmonized the data set (e.g., completion of keyword lists).

In subsequent sections, we detail the single steps of the previously described procedures. Section 3.2 presents the research questions, before we describe the detailed data collection (Sect. 3.3) and analysis procedure (Sect. 3.4).

3.2 Research Questions

Our objective is to capture the domain of Software Process Improvement (SPI), to provide a snapshot of the available publication flora, and to investigate research trends. Therefore, we define the following research questions:

RQ 1: *What is the general publication population on SPI?* This research question aims to get an overview of the general publication flora in SPI. We are interested into getting information regarding publication count, frequency and, even-

Table 1: Final search strings used for the automatic database search.

	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_2	(life-cycle **or** lifecycle **or** life cycle) **and** (design **or** modeling **or** modelling **or** analysis **or** training)	phases of the software process's life cycle
S_3	modeling **or** modelling **or** model-based **or** approach **or** variant	process modeling
S_4	optimization **or** optimisation **or** customization **or** customisation **or** tailoring	process customization and tailoring
S_5	(measurement **or** evaluation **or** approach **or** variant **or** improvement)	general measurement and improvement
S_6	reference model **or** quality management **or** evaluation **or** assessment **or** audit **or** CMMI **or** Capability Maturity Model Integration	reference models and quality management
S_7	SCAMPI **or** Standard CMMI Appraisal Method for Process Improvement **or** SPICE **or** ISO/IEC 15504 **or** PSP **or** Personal Software Process **or** TSP **or** Team Software Process	reference models and assessment approaches
S_8	(feasibility **or** experience) **and** (study **or** report)	reported knowledge and empirical research
C_1	software process **and** (software development model **or** process model)	*context definition:* software processes
C_2	SPI **or** software process improvement	*context definition:* SPI
F_1	(SPI **or** software process improvement) **and** (approach **or** practice **or** management)	SPI approaches, practices, and SPI management
F_2	(SPI **or** software process improvement) **and** report **and** (feasibility **or** experience)	evaluation research on SPI, e.g., studies, reports, etc.

tually, an overview of the different research type facets addressed by the found publications.

RQ 2: *What is the contribution population?* Based on the found publications, we are interested into the addressed topics and major contributions (e.g., SPI models, theories, secondary studies, and lessons learned) to work out the fields in SPI to which research contributed so far.

RQ 3: *What trends in SPI and SPI-related research can be observed?* The third research question aims at investigating the focus points addressed by SPI research so far, and to work out gaps as well as trends. This research question shall thus pave the way to direct future research on SPI.

3.3 Data Collection Procedures

In the following, we describe the query construction process, data source selection, and data storage format.

3.3.1 Query Construction

In a series of workshops, we defined the keywords that we are interested in and defined the general search strings (Table. 1), which were then validated in several test runs before being used in an automated full-text search in several literature databases. The queries were built based on keyword lists given by the common terminology in the area of software processes and SPI.

General Queries. The general search strings[1] S_1—S_8 were defined according to the relevant topics in SPI, e.g., improvement, assessment, measurement, ISO/IEC 15504, CMMI, quality management, and so forth. Due to the expected large number of results, we decided to complement the general search strings with context selectors C_1 and C_2 to limit the search to the domain of interest. Finally, we concluded the search strings shown in Table 1. These search strings were used to conduct a full-text search in the selected literature databases (Sect. 3.3.2).

Filter Queries. Because of the full-text search, we expected a variety of publications including some overhead. Hence, we defined two filter queries (Table 1) to be applied to the initial result set with the purpose of reducing the result set to the key publications. Query F_1 aims at finding all publications in the result set that explicitly present SPI approaches and practices, or that address the management of SPI. F_2 aims at finding all reports in the context of SPI in which feasibility is analyzed or experiences are reported. While the initial search was a full-text search, the filter queries were applied to the abstracts only. However, for technical reasons, ACM and Springer abstracts were partially not available[2] in the initial result set and, thus, the filtering was done manually during the voting procedure (Sect. 3.4).

Table 2: Data collection table (simplified).

Info. Set	Attributes
Meta data	ID, Citation-key,
Content	Authors, Title, Abstract, Year
Voting	Relevance (defined during further analysis and voting by the different authors), Comments
Analysis data	Publication type, Research type classification, etc.

3.3.2 Data Sources and Data Format

The data collection was an automated full-text search in several literature databases. As main data sources, we relied on established literature databases, which we consider

[1]Due to technical limitations of the search engines, we decided to perform multiple requests with simple strings. The request structure can be depicted from Table 7.

[2]Due to technical limitations, publications retrieved from ACM and Springer generated partially incomplete data, which we compensated in the final selection procedure.

Table 3: Inclusion and exclusion criteria (summary).

Crit.	Description
IC_1	Title, keyword list, and abstract make explicit that the paper is related to SPI.
IC_2	Paper presents SPI-related topics, e.g., SPI models, assessments, experiences in adopting and deploying software processes, and reports on improving specific methods/practices.
EC_1	Paper is not in English.
EC_2	Paper is not in the field of software engineering or computer science in general.
EC_3	Paper is a tutorial or workshop summary only.
EC_4	Paper occurred multiple times.
EC_5	Paper full text is not available for download.

Table 4: Research type facets (summary).

Crit.	Description
Evaluation research	implemented in practice, evaluation of implementation conducted; requires more than just one demonstrating case study
Solution proposal	solution for a problem is proposed, benefits/application is demonstrated by example, experiments, or student labs; also includes proposals complemented by one demonstrating case study for which no long-term evaluation/dissemination plan is obvious
Philosophical paper	new way of thinking, structuring a field in form of a taxonomy or a framework, secondary studies like SLR or SMS
Opinion paper	personal opinion, not grounded in related work and research methodology
Experience paper	personal experience, how are things done in practice

most appropriate for a search. In particular, we selected the following databases: ACM Digital Library, SpringerLink, IEEE Digital Library (Xplore), Wiley, Elsevier (Science Direct), and IET Software. If there was a paper listed in one of those databases, but was only referred, we counted it for the database that generated the item, regardless of the actual publication location. To structure the data, we created a spreadsheet that contains the attributes shown in Table 2.

3.4 Analysis Procedures

We describe the analysis preparation as well as the steps conducted to answer the research questions.

3.4.1 Analysis Preparation

We performed an automated search that required us to filter and prepare the result set. The data analysis is prepared by harmonizing the data and performing a 2-staged voting process to prepare the result set analysis.

Harmonization. Due to the query construction, we found a vast amount of multiple occurrences in the result set, and we also found a number of publications that are not in software engineering or computer science. To make the selection of the contributions more efficient, we first cleaned the initial result set (cf. Table 7 for the results per phase).

In the first step, we removed the duplicates, which we identified by title, year, and author list. In the second step,

Table 5: Contribution type facets (summary).

Crit.	Description
Model	representation of observed reality by concepts after conceptualization
Theory	construct of cause-effect relationships
Framework	frameworks/methods related to SPI
Guideline	list of advices
Lessons learnt	set of outcomes from obtained results
Advice	recommendation (from opinion)
Tool	a tool to support SPI

Table 6: Focus type facets from key wording.

Crit.	Description
Standard SPI models	Application, adaptation, and evaluation of standard SPI models, e.g., CMMI or ISO/IEC 15504
SPI models in SME	SPI models (new models and adaptations of standards) for SME and VSE
Assessment	General assessment and/or measurement approaches and models
General SPI	Work on general SPI initiatives (e.g., company level)
Method-specific SPI	Improvement of specific methods and/or procedures, e.g., testing
Success factors	Investigation of success factors, e.g., survey-based research

we applied the filter queries (Sect. 3.3.1) to sort out those publications not devoted to software processes and SPI. In order to double-check the result set, we used *word clouds* generated from abstracts and keyword lists to validate if the result set meets our requirements[3]. This procedure was performed individually per database and again on the integrated result set. Finally, we added the yet non-filtered ACM and Springer sets to prepare the voting procedure.

Voting. Similar to [18], we performed a multi-staged voting process to classify the papers as relevant or irrelevant and to build a set of publications for further investigation. The integrated result table therefore contains several columns (attribute "relevance" in Table 2).

In the voting, the inclusion and exclusion criteria listed in Table 3 guided the decision-making process. Two researchers performed individual votings (initially: publication title). If both agreed, the paper was initially decided. For those papers that were not immediately decided, a number of workshops were performed in which the decision was made based the papers' title and abstract. After the initial voting, the selection was reviewed by a third researcher and refined. The goal of this stage was to figure out those publications that are relevant for the analyses and classifications.

3.4.2 Analysis and Classification

After the voting, the final set of publications was defined. On this set, the analysis and classification was performed using the abstracts and—where necessary—the complete pub-

[3]We used the word clouds to visually inspect the result set for "intruders", e.g., medicine, chemistry, and cancer therapy. Terms not matching our search criteria were collected and used to identify and remove the misselected papers from the result set.

Table 7: Data collection and filtering results (tentative result sets during selection and final result set).

Step	IEEE	ACM	Springer	Elsevier	Wiley	IET	Total
Step 1: Search (Sect. 3.3.1)							
S_1 **and** (C_1 **or** C_2)	71	543	306	991	1,185	89	3,185
S_2 **and** (C_1 **or** C_2)	68	539	306	989	1,133	89	3,124
S_3 **and** (C_1 **or** C_2)	1,310	2,341	1,032	2,675	16,113	726	24,197
S_4 **and** (C_1 **or** C_2)	130	925	438	945	2,480	479	5,397
S_5 **and** (C_1 **or** C_2)	1,585	2,459	1,038	2,731	17,184	822	25,819
S_6 **and** (C_1 **or** C_2)	535	1,746	762	1,863	9,182	484	14,572
S_7 **and** (C_1 **or** C_2)	168	324	143	242	765	41	1,683
S_8 **and** C_2	114	105	433	1,015	6,341	366	8,374
Step 2: Removing Duplicates (Sect. 3.4.1)							
Duplicates per database	1,486	566	4,388	7,161	1,328	1,714	16,643
Duplicates across all databases	916	551	1,059	2,043	370	376	**5,315**
Step 3: In-depth Filtering (Sect. 3.3.1)							
Applying filters F_1 and F_2	578	–	–	710	221	53	1,562
Unfiltered	–	551	1,059	–	–	–	1,610
Result set (search process)	578	551	1,059	710	221	53	**3,172**
Step 4: Voting (Sect. 3.4.1)							
Final result set	283	65	114	103	67	3	**635**

lication. In the following, we summarize the analysis procedures used to answer our research questions. In particular, we describe the schema construction to perform the map creation.

Research Type Facets. In order to classify the publications, we rely on the classification according to the *research type facet* as proposed by Wieringa et al. [41]. However, during a test classification on a small sample, we found the need to adjust the facet definitions. Table 4 lists the research type facets applied to the result set.

Contribution Type Facets. In order to analyze what and how publications contribute to the body of knowledge, we adopted the *contribution type facets* as proposed by [31]. Table 5 lists the facet types applied to the result set.

Focus Type Facets. Similar to Paternoster et al. [23], we developed a set of *focus type facets* to gain insights into the actual contribution of the respective papers. The focus type facets were created using word clouds generated from the abstracts and keyword lists. Since we found a diversity of different terms and their spelling, we initially integrated all keyword lists into one text file, coded the keyword list, and generated a word cloud from the integrated list, which supported the definition of the focus type facets. Table 6 lists the focus type facets applied to the result set.

Trend Determination. In order to study trends, we use the result set's metadata and the outcomes of the different classifications and put them into a timeline similar to [18].

3.5 Validity Procedures

To increase the validity of our study, we implemented the following procedures: First, to avoid limitations caused by a too specific set of search criteria, we performed a breadth-first full-text search. Therefore, we used a set of keywords to build our search queries and performed several test runs. During the whole study, we performed several quality assurance activities (partially tool-supported), iterated through the single steps, and stepwise analyzed and refined tentative result sets. Furthermore, in different stages of the preparation, selection, and classification processes, we used randomly selected samples to test next steps and to confirm our approach.

During the publication selection and classification, we relied on researcher triangulation, e.g., within a rigorous multi-staged voting procedure, and calling in further researchers to confirm the classification based on randomly selected 5% samples. This procedure was partially complemented by an in-depth analysis of the contents of the papers going beyond the abstracts.

For the development of the classification schemas, we either ground the developed schemas in external proposals or rely on tool-supported techniques to generate the schemas to a large extent.

4. STUDY RESULTS

In this section, we present the initial results of our study. Table 7 summarizes the set of publications resulting from the collection and preparation phases. We summarize the numbers per database, the total number of results, the cleaned number of results after the first harmonization (removing duplicates), after applying the filter queries, and after the multi-staged voting of the papers for their relevance. Finally, 635 papers[4] remained in the final result set for further investigation.

In the Sections 4.1—4.3, we present the detailed study results according to the research questions (Section 3.2). In Sect. 4.4, we provide a discussion on the study results.

[4]Raw data for download: `https://goo.gl/TQGTlI`

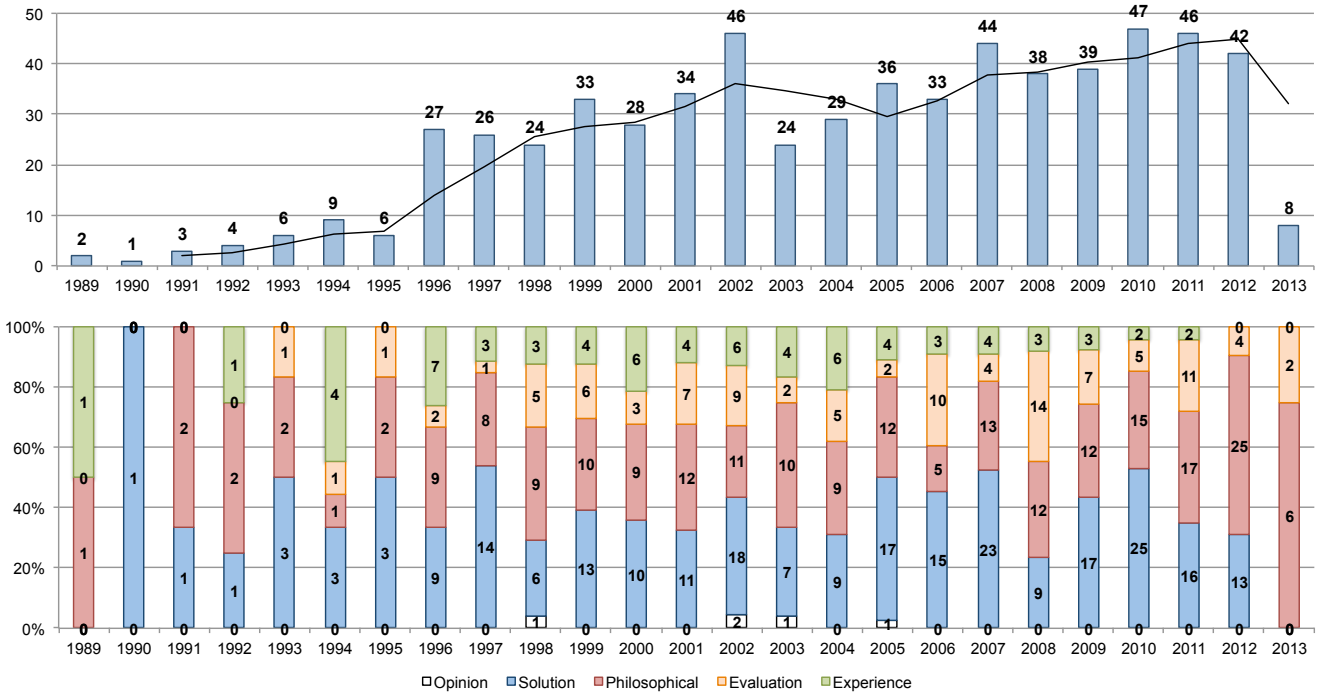

Figure 2: Number of published papers per year (incl. trend) and distribution over the research type facets.

4.1 RQ1: General Publication Flora

Figure 2 illustrates the distribution of the 635 selected publications over time beginning in 1989 and ending in 2013[5]. The figure provides two perspectives: The upper part shows the number of publications over time including a trend line (trend calculation basis: mean, 3-year period). In 1996, the numbers show a growing interest in SPI. From this point on, SPI became an inherent part of software engineering research. Figure 2 shows periodical waves over the years starting three to five years. Within these waves the largest gap/decrease is between 2002 and 2003. In the lower part, Figure 2 shows the detailed numbers of the publications and their relative distribution regarding the research type facets (cf. Table 4): The chart shows that the majority (approx. 2/3 of the result set) of the contributions was categorized as solution proposal ($n = 244$) or philosophical paper ($n = 214$). However, the result set also contains a number of evaluation research papers ($n = 102$) and experience papers ($n = 70$), which shows that the field of SPI is still moving, but accumulates evidence (solutions are proposed that are followed by evaluation papers or papers reporting experiences). The least published papers are opinion papers ($n = 5$).

Among all publications, one trend can be observed over time: solution proposals and philosophical papers continually represent the majority of the publications. Furthermore, the different paper types follow cycles. For instance, in the categories solution proposals and philosophical papers, the papers dip sometimes to one type and then again to the other. Papers in the category of evaluation papers also roughly follow a 3-year cycle. The category of experience papers is the most constant one, even if it contributes only a small number of papers to the result set.

[5]Due to the search date, 2013 is not fully covered. The overall time frame results from the paper selection, i.e., the first papers voted in during the selection were published in 1989.

4.2 RQ2: Contribution and Focus Types

In order to get an overview of the harvested papers, we performed a categorization to define contribution type facets (Table 5) and focus type facets (Table 6) of the publications. In the following, we provide two perspectives. In Figure 3, we provide a comprehensive map in which we relate the contribution- and focus type facets to the research type facets. In a second perspective, Figure 4 relates the contribution- to the focus type facets.

4.2.1 *The Big Picture*

Figure 3 illustrates the big picture of the publication classification. On the right hand side, the figure lists the contribution type facets and shows that lessons learned ($n = 290$, 46%) and frameworks ($n = 235$, 37%) make the majority of the contributions. The other categories are barely represented: tools ($n = 36$), guidelines ($n = 27$), models ($n = 24$), theories ($n = 12$), and advice ($n = 11$). Notably, the chart shows two striking results: the majority of the solution proposals focuses on frameworks ($n = 167$), i.e., 26% of all considered publications propose a new SPI-related framework. Second, the largest share of the philosophical papers is devoted to lessons learned ($n = 155$), i.e., 24% of all publications report on learnings from SPI-related activities (e.g., from conducted SPI endeavors or from survey-based research). However, looking at the research type facet "Evaluation Research", we find only 44 publications on the evaluation of frameworks and, respectively, 45 publications

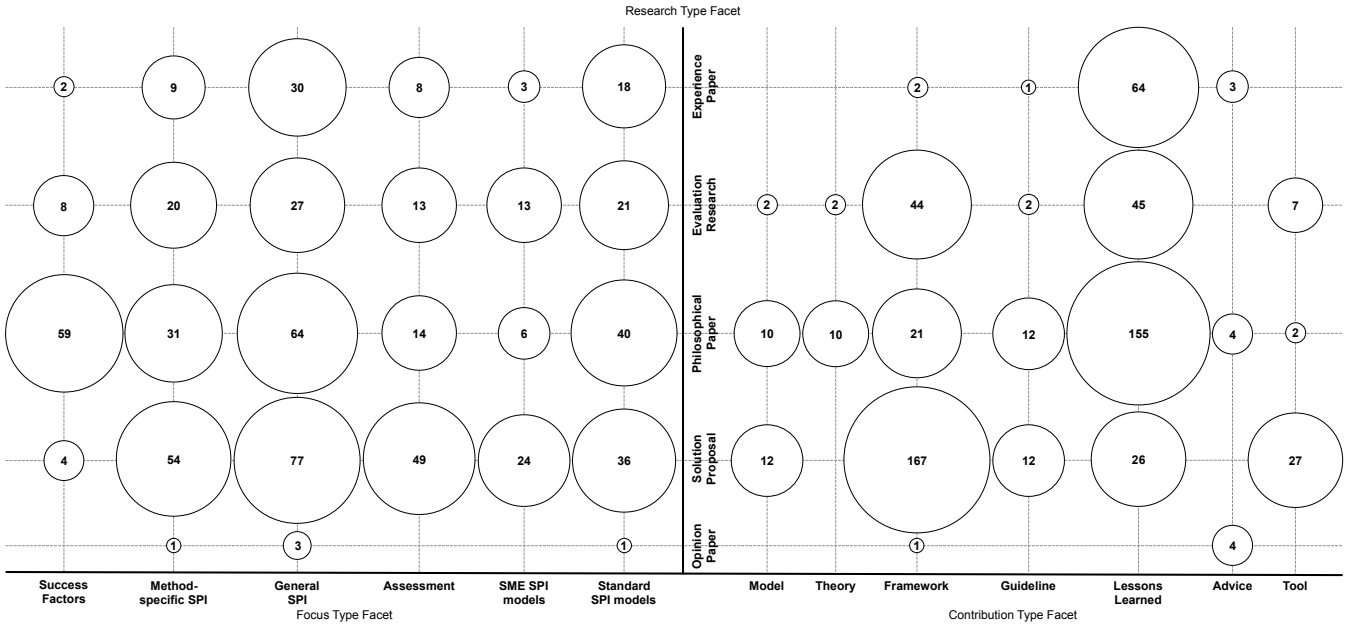

Figure 3: Systematic map: focus-, contribution-, and research type facets.

on lessons learned[6]. That is, the chart shows an imbalance between proposals and philosophical discussion, and respective evaluating research. However, this situation is a fundamental observation in the result set, as evaluation research accounts for only 16% ($n = 102$) of all publications.

The left side of Figure 3 shows the focus type facets. This part of the chart shows a more balanced distribution across the different categories. 201 out of 635 publications (32%) are categorized "General SPI", i.e., addressing general SPI activities, such as SPI on the company level, lessons learned in general, or non-standardized framework proposals. The second largest share ($n = 116$, 18.3%) is devoted to publications addressing standard SPI- and maturity models, e.g., CMMI and ISO/IEC 15504; followed by publications of the category "Method-specific SPI" ($n = 115$, 18.1%). The remaining categories (Assessment: $n = 84$, Success Factors: $n = 73$, and SME-specific SPI models: $n = 46$) together make 32% of the result set. Except for publications addressing success factors, all other focus type facets show a balanced distribution regarding the research type facets, i.e., SPI-related topics are studied from different perspectives.

4.2.2 *Contribution & Focus*

In order to get more insight into the result set, we also provide a map in which we relate the contribution type facet with the focus type facet.

On the one hand, the map in Figure 4 shows the previously found focus on frameworks and lessons learned but, on the other hand, provides a more differentiated picture. Especially three combinations stand out among the others: general SPI and lessons learned ($n = 96$, 15%), general SPI and frameworks ($n = 64$, 10.1%), and standard SPI models and lessons learned ($n = 67$, 10.6%). Figure 4 also shows

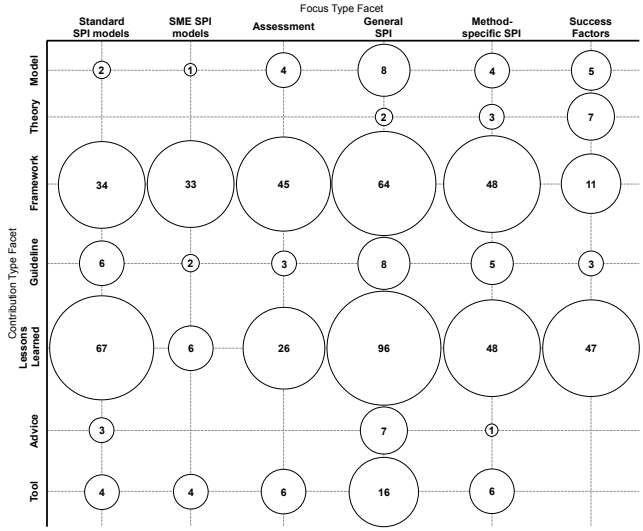

Figure 4: Systematic map: focus- and contribution type facets.

that new frameworks are proposed for standard SPI models as well as for SME's. However, the field of standard SPI models, apparently, provides a large body of knowledge, whereas lessons learned from SME-related SPI are scarcely available.

4.3 RQ3: Trends in SPI-related Research

The third research question aims at analyzing the development of SPI over time and at identifying trends. Figure 2 illustrates the general publication flora showing a growing interest in SPI and related research starting in the mid 1990's. Moreover, Figure 2 shows that—for years—solution proposals and philosophical papers make the majority of all publi-

[6]Lessons learned in evaluation research refer to papers deriving lessons learned from practically conducted SPI projects rather than from survey-based research, i.e., lessons learned are the emphasized (major) outcome of a paper.

cations. Taking into account the maps in Figure 3 and 4, we see that proposing new frameworks and reporting on lessons learned shape the field of SPI so far. However, at this level only few trends become obvious:

- The community is still on the quest for appropriate SPI frameworks (cf. Figure 3).
- The investigation of success factors becomes a research topic to develop theories (cf. Figure 4).

Further Trends. In order to analyze further trends, we enriched the set of collected metadata[7] of the result set. This extended metadata set delivered the following findings:

- There is a growing interest in agile methods and adopting agile principles for SPI ($n = 27$).
- There is a growing number of secondary studies analyzing and structuring reported knowledge ($n = 31$). These studies are mainly used to derive/propose models and theories.
- A considerable share of research is based on surveys ($n = 73$) thus grounding SPI-related research in personal experience/perception of the interviewees. These studies are also used to derive/propose models and theories.
- There is a lack of replication research ($n = 2$), and SPI-related research is oftentimes only confirmed in student labs (cf. research type facet "solution proposal") or by a single (demonstrating) case in industry ($n = 19$).

Major Research Directions. In summary, we find some major directions of SPI research: 1) SPI becomes more relevant for SME's thus leading to more proposals for SME-specific SPI- and assessment models, e.g., [29, 33, 37]. 2) As companies adopting to agile methods and principles, agility becomes more relevant in the context of SPI, which causes a general discussion on various SPI facets, e.g., [20]. 3) The SPI-related body of knowledge comprises various models, experiences, and so forth calling for structuring and consolidation. Therefore, the interest in secondary studies increases, e.g., [36, 38].

Nonetheless, *we still lack evidence*: Over the years, the field accumulated a considerable number of solution proposals, yet, missing evidence of their feasibility. Only the well-established standardized approaches like ISO/IEC 15504 and CMMI appear to be well understood. However, a number of publications explicitly aim at analyzing these standards to outline difficulties, e.g., [34], or to motivate the need for new models, e.g., [24, 30]. Our data shows the field of SPI being shaped by new solution proposals published on a per-year basis, but not providing hard evidence on feasibility, advantages, and disadvantages. So far, our data does not indicate a change in this trend.

4.4 Discussion

In this section, we discuss our results and provide a (tentative) interpretation of the results. Our data shows a diverse picture and, furthermore, shows SPI a frequently researched topic (Figure 2). Moreover, research on SPI addresses a variety of aspects. However, the systematic maps (Figure 3 and Figure 4) show certain focus points: The majority of the investigated publications (approx. $^2/_3$) focus on proposing new frameworks and on reporting lessons learned. Furthermore, our results show a significant imbalance between proposing new solutions and evaluating their feasibility. Among the different categories, the majority of evaluation research is conducted in the context of standardized SPI- and maturity models (Figure 4, standard SPI models and lessons learned). For newly proposed models, we often find—if at all—only single-case validation (in industry or university-hosted labs); only few, e.g., [29] provide a comprehensive evaluation. Another finding is the lack of theorizing approaches, which are often performed for specific domains (e.g., SME's) or grounded in secondary studies only.

In summary, although SPI is around for decades, we still miss a sound theory about SPI. We have a number of standardized and specific SPI models and frameworks. However, we lack evidence. One reason could be that SPI always involves change in behavior of individual persons and changes in the culture of an organization and, due to the varying contexts, SPI cannot be too descriptive. Therefore, frameworks and tools are proposed, which long for adaptation to the respective context. Yet, the constant change or evolution of the context could be considered a continuous stimulus to provide new frameworks that only have a short life cycle and are quickly replaced by other frameworks that aim to "better" solve this issue.

This assumption is supported by the missing long-term and replication studies (the result set only contains 2 explicitly mentioned replication studies). Furthermore, missing is a critical discussion and comparison of available approaches, and their use and feasibility in practice. Although we found 31 secondary studies, these studies lay their focus on investigating success factors rather than providing structure and trying to generalize available knowledge, as for instance done in [38]. In a nutshell, our results show that SPI is a still emerging field characterized by solution proposals and experiences awaiting more effort to systematization.

4.5 Threats to Validity

In this section, we evaluate our findings and critically review our study regarding the threats to validity. As a literature study, this study suffers from potential incompleteness of the search results and a general publication bias, i.e., positive results are more likely published than failed attempts. For instance, the result set does not contain studies that explicitly report on failure and draw their conclusions from respective lessons learned, and we thus cannot analyze proposals to answer the question for: What works and what does *not*? That is, our study encounters the risk to draw an incomplete and potentially too positive picture.

Beyond that general threat, the *internal validity* of the study could be biased by personal ratings of the participating researchers. To address this risk, we relied on a proven procedure [18] that utilizes different supporting tools and researcher triangulation to support dataset cleaning, study selection, and classification. Calling in extra researchers to analyze and/or confirm decisions increases internal validity. However, especially in the study classification for the focus type facets (which were derived from keyword lists),

[7] While performing the initial data analyses, we realized that we observed certain clusters of publications and, thus, decided to make our observation metadata categories. In particular: secondary studies, i.e., systematic mapping studies and literature reviews, survey-based research, replication studies, publications addressing agile methods and software development, and publications claiming to present proven approaches but delivering only one (demonstrating) case.

we realized some deviation in the rating. So far, we elaborated the reasons finding two major problem candidates to be addressed in future stages of the study: 1) general disagreement in the ratings—in this case, we need to revise the rating procedures; 2) inappropriateness of the focus type facets—in this case we either need to revise the focus type facets or we need to adjust the classification rules, i.e., one paper may have more than one focus.

The *external validity* is threatened by missing knowledge about the generalizability of the results. However, as we focused on a broadband analysis accepting a large number of publications, we assume to have created a generalizable result set. Yet, this assumption needs to be confirmed by further independently conducted studies. But here lays a major problem: So far, we miss actionable methods to replicate such secondary studies—especially such comprehensive studies. Nevertheless, the present study provides an ex-post analysis and a snapshot. Considering the average publication frequency over the years (Figure 2), a replication is strongly required to confirm our findings.

5. CONCLUSION & FUTURE WORK

In this paper, we provided a first set of systematic maps drawing the big picture of SPI-related research. We conducted a comprehensive systematic mapping study in which we analyzed 635 publications of the past 25 years. Our results show that SPI is a still emerging field, which is shaped by a considerable number of solution proposals and experience reports, which make about $^2/_3$ of all publications.

Yet, the field of SPI suffers from missing evidence: Proposed solutions are barely evaluated for their feasibility, studies comparing and analyzing proposed solutions for their advantages and disadvantages are missing, and testable theories are—if at all—in the construction phase awaiting their confirmation. Furthermore, our study reveals some trends in SPI-related research: We found growing interest in SPI for SME's and adopting agile principles for SPI. Also, we found an increasing number of secondary studies of which some already started to collect, structure, and generalize knowledge.

5.1 Limitations

The major limitation of the study at hand is its granularity. As the main objective is to provide an overall picture, the present study lacks in details, e.g., which models are proposed in detail, what is the exact degree of validation/evaluation of the respective models, what is the dissemination of certain models, to which extent do large, medium, and small companies implement SPI (and what is the success rate), and so forth. However, this information is available by the result set, but requires the application of the systematic literature review instrument rather than the mapping study instrument.

5.2 Future Work

In order to address the aforementioned limitation, future work includes an update of the study and an in-depth analysis of the result set. This work, inter alia, includes a critical discussion and revision of the different classification schemas (cf. Sect. 4.5), and extension of the mapping study including further categories such as rigorousness (cf. [23]), in-depth analysis of study results, e.g., creation of SPI model catalogs, investigating the published lessons learned in detail, and harvesting published studies to conduct comparative analyses and, finally, to develop testable hypotheses to foster the theory development for SPI.

ACKNOWLEDGMENTS

Conducting this study was a long-term endeavor that involved many people. We thank Michaela Tießler, Ragna Steenweg, Daniel Méndez Fernández for their support in testing the instruments for paper selection and dataset cleaning. Furthermore, we owe special thanks to the students of the "Hiwi-Pool" of the Technische Universität München, who supported us during the data collection and dataset completion and cleaning processes.

6. REFERENCES

[1] I. Allison. Organizational factors shaping software process improvement in small-medium sized software teams: A multi-case analysis. In *International Conference on the Quality of Information and Communications Technology*, pages 418–423. IEEE, 2010.

[2] S. Bayona-Oré, J. Calvo-Manzano, G. Cuevas, and T. San-Feliu. Critical success factors taxonomy for software process deployment. *Software Quality Journal*, 22(1):21–48, 2014.

[3] J. G. Brodman and D. L. Johnson. What small businesses and small organizations say about the cmm. In *International Conference on Software Engineering*, pages 331–340. IEEE, 1994.

[4] CMMI Product Team. CMMI for Development, Version 1.3. Technical Report CMU/SEI-2010-TR-033, Software Engineering Institute, CMU, 2010.

[5] G. Coleman and R. O'Connor. Investigating software process in practice: A grounded theory perspective. *Journal of Systems and Software*, 81(5):772–784, 2008.

[6] P. Diebold, J.-P. Ostberg, S. Wagner, and U. Zendler. What do practitioners vary in using scrum? In *International Conference XP*, pages 40–51. Springer, 2015.

[7] T. Dybå. An instrument for measuring the key factors of success in software process improvement. *Empirical Software Engineering*, 5(4):357–390, 2000.

[8] K. El-Emam and D. R. Goldenson. An empirical review of software process assessments. *Advances in Computers*, 53:319–423, 2000.

[9] Y. Y. L. Helgesson, M. Höst, and K. Weyns. A review of methods for evaluation of maturity models for process improvement. *Journal of Software: Evolution and Process*, 24(4):436–454, 2012.

[10] R. V. Horvat, I. Rozman, and J. Györkös. Managing the complexity of spi in small companies. *Software Process: Improvement and Practice*, 5(1):45–54, 2000.

[11] M. Hull, P. Taylor, J. Hanna, and R. Millar. Software development processes - an assessment. *Information and Software Technology*, 44(1):1 – 12, 2002.

[12] W. S. Humphrey. *Managing the Software Process*. Addison-Wesley Professional, January 1989.

[13] ISO. Software Process Assessment - Part 4: Guidance on use for process improvement and process capability determination. Technical Report ISO/IEC 15504-4:2004, International Organization for Standardization, 2004.

[14] ISO. Systems and Software Life Cycle Profiles and Guidelines for Very Small Entities (VSEs). Technical Report ISO/IEC 29110:2011, International Organization for Standardization, 2011.

[15] Jürgen Münch, Ove Armbrust, Martin Kowalczyk, and Martin Sotó. *Software Process Definition and Management*. Springer, 2012.

[16] B. Kitchenham. Procedures for Performing Systematic Reviews. Technical Report TR/SE-0401, Keele University, 2004.

[17] M. Kuhrmann. Crafting a software process improvement approach - a retrospective systematization. *Journal of Software: Evolution and Process*, 27(2):114–145, 2015.

[18] M. Kuhrmann, D. M. Fernández, and M. Tiessler. A mapping study on the feasibility of method engineering. *Journal of Software: Evolution and Process*, 26(12):1053–1073, 2014.

[19] M. Kuhrmann and O. Linssen. Welche Vorgehensmodelle nutzt Deutschland? In *PMV 2014*, LNI. Gesellschaft für Informatik (GI) e.V., 2014.

[20] S. C. Misra, V. Kumar, and U. Kumar. Identifying some important success factors in adopting agile software development practices. *Journal of Systems and Software*, 82(11):1869 – 1890, 2009.

[21] L. F. S. Monteiro and K. M. de Oliveira. Defining a catalog of indicators to support process performance analysis. *Journal of Software Maintenance and Evolution: Research and Practice*, 23(6):395–422, 2011.

[22] S. D. Müller, L. Mathiassen, and H. H. Balshøj. Software process improvement as organizational change: A metaphorical analysis of the literature. *Journal of Systems and Software*, 83(11):2128–2146, 2010.

[23] N. Paternoster, C. Giardino, M. Unterkalmsteiner, T. Gorschek, and P. Abrahamsson. Software development in startup companies: A systematic mapping study. *Information and Software Technology*, 56(10):1200 – 1218, 2014.

[24] M. C. Paulk. Comparing iso 9001 and the capability maturity model for software. *Software Quality Journal*, 2(4):245–256, 1993.

[25] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattson. Systematic mapping studies in software engineering. In *International Conference on Evaluation & Assessment in Software Engineering*, pages 68–77. ACM, 2008.

[26] F. Pino, F. García, and M. Piattini. Software process improvement in small and medium software enterprises: a systematic review. *Software Quality Journal*, 16(2):237–261, 2008.

[27] F. J. Pino, J. A. H. Alegría, F. García, and M. Piattini. A process for driving process improvement in VSEs. In *International Conference on Software Process*, volume 5543 of *Lecture Notes in Computer Science*, pages 342–353. Springer Berlin Heidelberg, 2009.

[28] A. Rainer and T. Hall. An analysis of some 'core studies' of software process improvement. *Software Process: Imprmnt and Practice*, 6(4):169–187, 2001.

[29] A. Raninen, J. J. Ahonen, H.-M. Sihvonen, P. Savolainen, and S. Beecham. LAPPI: A light-weight technique to practical process modeling and improvement target identification. *Journal of Software: Evolution and Process*, 25(9):915–933, 2012.

[30] I. Rozman, R. Vajde Horvat, J. Györkós, and M. Hericùko. Processus – integration of sei cmm and iso quality models. *Software Quality Journal*, 6(1):37–63, 1997.

[31] M. Shaw. Writing good software engineering research papers: Minitutorial. In *International Conference on Software Engineering*, pages 726–736. IEEE, 2003.

[32] M. Staples and M. Niazi. Systematic review of organizational motivations for adopting cmm-based spi. *Information and Software Technology*, 50(7-8):605–620, 2008.

[33] M. Staples and M. Niazi. Two case studies on small enterprise motivation and readiness for cmmi. In *International Conference on Product Focused Software Development and Process Improvement*, pages 63–66. ACM, 2010.

[34] M. Staples, M. Niazi, R. Jeffery, A. Abrahams, P. Byatt, and R. Murphy. An exploratory study of why organizations do not adopt CMMI. *Journal of Systems and Software*, 80(6):883–895, 2007.

[35] D. Stelzer and W. Mellis. Success factors of organizational change in software process improvement. *Software Process: Improvement and Practice*, 4(4):227–250, December 1998.

[36] M. Sulayman and E. Mendes. An extended systematic review of software process improvement in small and medium web companies. In *International Conference on Evaluation & Assessment in Software Engineering*, pages 134–143. IET, 2011.

[37] P. S. Taylor, D. Greer, G. Coleman, K. McDaid, and F. Keenan. Preparing small software companies for tailored agile method adoption: Minimally intrusive risk assessment. *Software Process: Improvement and Practice*, 13(5):421–437, 2008.

[38] M. Unterkalmsteiner, T. Gorschek, A. Islam, C. K. Cheng, R. Permadi, and R. Feldt. Evaluation and measurement of software process improvement - a systematic literature review. *IEEE Transactions on Software Engineering*, 38(2):398–424, 2012.

[39] D. Viana, T. Conte, D. Vilela, C. R. B. de Souza, G. Santos, and R. Prikladnicki. The influence of human aspects on software process improvement: Qualitative research findings and comparison to previous studies. In *International Conference on Evaluation & Assessment in Software Engineering*, pages 121–125. IET, 2012.

[40] L. Vijayasarathy and C. Butler. Choice of software development methodologies - do project, team and organizational characteristics matter? *IEEE Software*, 2015.

[41] R. Wieringa, N. Maiden, N. Mead, and C. Rolland. Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. *Requirements Engineering*, 11(1):102–107, Dec. 2005.