Search Strategy to Update Systematic Literature Reviews in Software Engineering

Emilia Mendes
Department of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
emilia.mendes@bth.se

Katia Felizardo
Department of Computing
Federal Technological University of Paraná
Cornélio Procópio, Brazil
katiascannavino@utfpr.edu.br

Claes Wohlin
Department of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
claes.wohlin@bth.se

Marcos Kalinowski
Department of Informatics
Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro, Brazil
kalinowski@inf.puc-rio.br

***Abstract*—[*Context*] Systematic Literature Reviews (SLRs) have been adopted within the Software Engineering (SE) domain for more than a decade to provide meaningful summaries of evidence on several topics. Many of these SLRs are now outdated, and there are no standard proposals on how to update SLRs in SE. [*Objective*] The goal of this paper is to provide recommendations on how to best to search for evidence when updating SLRs in SE. [*Method*] To achieve our goal, we compare and discuss outcomes from applying different search strategies to identifying primary studies in a previously published SLR update on effort estimation. [*Results*] The use of a single iteration forward snowballing with Google Scholar, and employing the original SLR and its primary studies as a seed set seems to be the most cost-effective way to search for new evidence when updating SLRs. [*Conclusions*] The recommendations can be used to support decisions on how to update SLRs in SE.**

Keywords— Systematic Literature Review Update, Systematic Literature Reviews, Software Engineering, Snowballing, Searching for evidence

I. INTRODUCTION

In 2004, Kitchenham et al. [12] argued for an Evidence-Based paradigm in Software Engineering (EBSE), to be particularly employed by "researchers interested in empirical software engineering and practitioners faced with decisions about the adoption of new software engineering technologies". EBSE's goals are to: "provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision-making process regarding the development and maintenance of software", and also to encourage the use of Systematic Literature Reviews (SLRs) to obtain such current best evidence.

Such call to arms prompted the SE community to publish SLRs, thus leading to more than 430 SLRs published, within the period from January 2004 to May 2016 [11], [16], [35], [3]. Despite such large number of published SLRs, Mendes et al. [32] identified only 13 updated SLRs, within the period 2006 to 2018; such findings show that many SLRs in SE are potentially outdated, thus influencing our current aggregated understanding of the state of the art in those obsolete SLRs' research topics. If we also take into account outdated primary studies, e.g. studies that used a particular technology that is not suitable any longer, the situation is worsened.

There are a few studies on the topic of SLR updates in SE (see Section II); however, to date there has been no study that has systematically compared different search approaches to make recommendations on "how" to identify new evidence when updating SLRs in SE. Such a systematic approach is the goal and the main contribution of this paper. To do so, we compare and discuss outcomes from applying different search strategies (e.g., database search and forward snowballing processes) to identify the new evidence found by a previously published SLR update on effort estimation. Note that we have also addressed the issue of "when" to update an SLR in SE, which is detailed elsewhere [32].

Note that there are three very important points that we would like to highlight upfront relating to achieving our research goal. First, we wanted to use the results from both an SLR update and its replications, where such replications would have used different search strategies to identify primary studies. Second, we also looked for an SLR update that had replications carried out by sets of authors with as little overlap as possible between them. Such diversity of authors was important to reduce as much as possible any bias when applying different methods to identify primary studies. Out of the 13 SLR updates found [32], only one met our criteria. Third, at the expense of a stronger conclusion validity, our focus in this research was not to carry out a formal experiment comparing different ways to identify primary studies. We wanted to base our suggestions upon already existing evidence from replications of SLR updates in SE. Furthermore, both the updated SLR used herein, and its replications were conducted and reported at different times, which is not optimal. However, if we were to replicate our searches and examine the studies at the same time, we would potentially bring enough bias into the process to make our findings less trustworthy.

The remainder of this paper is organised as follows. In Section 2, we present related work. In Section 3, we provide background information, our research questions, and

present the analysis of different search strategies, with recommendations on how to search for evidence when updating SLRs in SE. Section 4 contains a discussion on threats to validity, followed by our conclusions in Section 5.

II. Related Work

With regard to previous work relating to updating SLRs, there have been a few initiatives, as follows:

Dieste et al. [40] and Ferrari et al. [41] developed processes to support updating SLRs in SE. Nevertheless, their processes do not focus on how to best search for new evidence.

Felizardo et al. [42] proposed an approach based on Visual Text Mining (VTM), which supports the selection of new evidence to update an SLR using as starting point this original SLR's included studies. Our work differs from theirs because it does not urge the use of a specific tool (e.g. VTM); is based on a detailed comparison of different search mechanisms carried out in an SLR update and its replications; and whether different approaches lead to significant differences in the set of included studies or in the SLR conclusions.

Silva et al. [43] evaluated different databases (specific – IEEE Xplore and generic – Google Scholar databases) for supporting secondary studies' updates. IEEE Xplore is judged not sufficient to identify most studies; conversely, Google Scholar seems sufficient. Despite their focus being on the selection of databases, they neither compare different search approaches, nor whether different approaches led to significant differences in the set of included studies or in the SLR conclusions.

Rodriguez et al. [44] reported lessons learned considering their experience in updating SLRs. Some of these lessons are: (i) to adopt software tools to support the updating process; (ii) to provide as much information as possible about the SLR being updated; (iii) to involve some of the authors from the SLR being updated; and (iv) to reuse the protocol from the SLR being updated. Nepomuceno and Soares [45] extended Rodriguez et al.'s work [44], by asking researchers about the lessons documented by Rodriguez et al. [44] and reached similar conclusions.

Similar to Rodriguez et al., and Soares et al.'s work [10], in this paper we also provide lessons learned when updating SLRs. However, our main focus is to use evidence from a detailed comparison between different search strategies to provide concrete recommendations on "how" to search for new evidence when updating SLRs in SE. To date there has been no such detailed study in SE, as far as searching for new studies to update an SLR is concerned.

III. How Best to Search for Evidence to Update SLRs in Software Engineering

A. Background

The SLR update and its two replications relate to an SLR on the topic of cross-company (CC) vs. within-company (WC) effort estimation. Hereafter this original SLR is called OSLR; its results were published as a conference paper [14] and later as a journal paper [15]. The list of papers included in OSLR is given in Table I.

TABLE I. OSLR included studies

SID	Authors	Ref.	Year
CC model NOT significantly different from WC model			
S2	L.C. Briand, et al.	[1]	1999
S3	L.C. Briand, et al.	[2]	2000
S6	I. Wieczorek and M. Ruhe	[37]	2002
S10	E. Mendes, et al.	[31]	2005
CC model significantly different from WC model			
S4	R. Jeffery, et al.	[8]	2000
S5	R. Jeffery, et al.	[7]	2001
S8	B. A. Kitchenham and E. Mendes	[13]	2004
S9	E. Mendes and B. A. Kitchenham	[28]	2004
Inconclusive			
S1	K. Maxwell, et al.	[24]	1999
S7	M. Lefley and M. Shepperd	[20]	2003

This OSLR was updated once [27] (U1-OSLR), and later replicated twice by studies investigating different search mechanisms and whether they would retrieve the same primary studies as in U1-OSLR [39][4]. These two replications are named henceforth as R1-U1-OSLR and R2-U1-OSLR, respectively. Together they identified a superset containing 15 studies published until the end of 2013, shown in Table II.

TABLE II. List of 15 studies in the superset

SID	Authors	Ref.	Year
CC model NOT significantly different from WC model			
S13	C. Lokan and E. Mendes	[22]	2008
S14	E. Mendes and C. Lokan	[30]	2008
S15	E. Mendes, et al.	[25]	2008
S16	C. Lokan and E. Mendes	[21]	2009
S18	E. Kocaguneli and T. Menzies	[19]	2011
S20	F. Ferrucci, et al.	[6]	2012
S22	R. Premraj and T. Zimmermann	[34]	2007
S23	E. Kocaguneli, et al.	[18]	2010
CC model significantly different from WC model			
S11	C. Lokan and E. Mendes	[23]	2006
S12	E. Mendes, et al.	[26]	2007
S14	E. Mendes and C. Lokan	[30]	2008
S15	E. Mendes, et al.	[25]	2008
S16	C. Lokan and E. Mendes	[21]	2009
S17	E. Mendes and C. Lokan	[29]	2009
S20	F. Ferrucci, et al.	[6]	2012
S21	L. L. Minku and X. Yao	[33]	2012
Inconclusive			
S19	O. Top, et al.	[36]	2011
S24	E. Kocaguneli, et al.	[17]	2013
S25	F. Ferrucci, et al.	[5]	2009

Note that there were several studies where results contrasted, depending on the prediction models being compared. This is why we have some studies (S14, S15, S16 and S20) shown under different categories in Table II. U1-OSLR should have identified, except for S24, all 14 studies (explanation given later); and both replications should have identified all 15 studies; however, this was not the case, as shown in Fig. 1. It is clear that the results did not fully agree. Such findings motivated us to investigate further the issue

of searching for evidence when updating SLRs; this is reflected via our research questions, detailed next in subsection B. Note that further details on U1-OSLR and its two replications are given in subsection C.

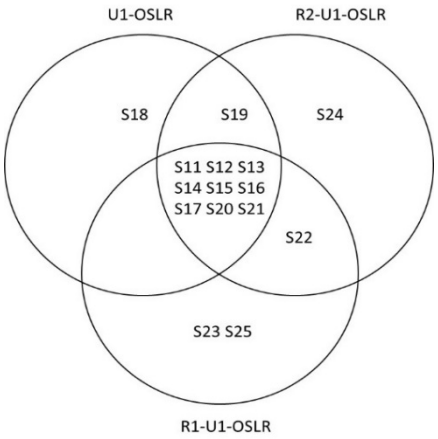

Fig. 1. Superset organised by studies (SLR update and replications)

B. Research Questions

The five main research questions (RQs) relating to how best to search for evidence when updating an SLR follow. RQ1 focuses upon the search strategy employed in the SLR updated and its two replications. We consider the search strategy separately from the selection of studies to investigate the cause(s) for some of the 15 studies not being included in both U1-OSLR and its two replications. Although RQ2 has been addressed in U1-OSLR and its two replications, we included it here because such information is also relevant to understand how to search for new evidence when updating SLRs in SE.

- **RQ1:** Were all the 15 studies in the superset retrieved by the different searches employed by U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR?
- **RQ2:** Which were the studies selected/included by U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR?
- **RQ3:** With respect to forward snowballing,
  - **RQ3.1:** Do different seed sets lead to significant differences in the set of included studies, and in the SLR conclusions?
  - **RQ3.2:** Do different processes lead to significant differences in the set of included studies, and in the SLR conclusions?
- **RQ4:** Were there differences between the conclusions from U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR, when compared to the conclusions that would have been obtained using the superset of 15 studies?
- **RQ5:** What are the differences, if any, between the conclusions using the superset of 15 studies, and those in OSLR?

C. Further Details on U1-OSLR and its replications

The OSLR's update – U1-OSLR [27] was published in 2014 and had the participation of two of the four co-authors of this paper – Mendes and Kalinowski. It used the same search string and protocol as in the OSLR. In total, U1-OSLR identified 11 primary studies, and missed studies S22 to S25. The first replication of R1-U1-OSLR [39] was published in the first semester of 2016, and was authored by one of the four co-authors of this paper – Wohlin. The seed set Wohlin used to replicate the U1-OSLR's results contained 12 sources: the two papers that detailed the OSLR [14] and [15] plus the 10 primary studies originally included in the OSLR. Wohlin analysised the citations provided by Google Scholar to each of the 12 sources as means to identify possible additional studies. He followed his own method, previously proposed in [38]. A total of 12 studies were selected. In comparison to U1-OSLR, it missed S18, S19 and S24, however included three additional studies (S22, S23 and S25), missed by U1-OSLR. Finally, the second replication of U1-OSLR's results – R2-U1-OSLR [4], was published in the second semester 2016, and was co-authored by three of the four co-authors of this paper – Felizardo, Mendes and Kalinowski. They performed forward snowballing on a seed set containing the 10 primary studies included in the OSLR. Citations were identified via search engines, such as, IEEEXplore and ACM, instead of using Google Scholar. Four iterations were carried out, until reaching a saturation point. A total of 172 studies were found, of which 12 were selected for inclusion. The approach identified all the studies included in U1-OSLR, except for one – S18; and in addition identified two studies (S22 and S24) not included in U1-OSLR. In relation to R1-U1-OSLR, it missed two studies (S23 and S25) and included two studies missed by R1-U1-OSLR (S19 and S24). Note that R2-U1-OSLR had originally included 13 papers. However, during the writing of this paper, we noticed that one of the papers (called N3 in R2-U1-OSLR) was incorrectly included, due to a miscommunication between two of its authors.

In summary, the two replications of U1-OSLR jointly found another four studies that were missed by U1-OSLR (S22, S23, S24 and S25). Furthermore, Table III provides an overview of R1-U1-OSLR and R2-U1-OSLR, showing the number of iterations, studies found, studies included, unique studies identified by each review and studies that are common with U1-OSLR.

TABLE III. SUMMARY OF RESULTS FOR THE TWO REPLICATIONS

Characteristics	R1-U1-OSLR [39]	R2-U1-OSLR [4]
Seed set	12 (10 + 2 SLRs)	10
Iterations	1	4
Citations analysis of studies found	1018	172
Studies included	12	12
In common with U-OSLR	9	9
Unique studies	2	1

D. Addressing the Research Questions

RQ1: Were all the 15 studies in the superset retrieved by the different searches employed by U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR?

Table IV shows the studies retrieved by each of the three studies. Note that these are **not** the studies that were ultimately selected. The only searches that retrieved all 15 studies were those carried out in R1-U1-OLSR, using as seed set the two references to OSLR, and all the 10 studies included in OSLR.

When it comes to U1-OSLR, paper S25 was not indexed by any of the search engines, so it would not be possible to find it either via a normal database search or using the ACM and IEEE citation search mechanisms. Paper S24 was presented at a conference in October 2013, which probably explains why it was not found during the database searches carried out in U1-OSLR early November 2013.

TABLE IV. SUPERSET RETRIEVED BY THE DIFFERENT SEARCHES EMPLOYED BY U1-OSLR, R1-U1-OSLR, AND R2-U1-OSLR

	Retrieved By		
Superset	**U1-OSLR**	**R1-U1-OSLR**	**R2-U1-OSLR**
S11	√	√	√
S12	√	√	√
S13	√	√	√
S14	√	√	√
S15	√	√	√
S16	√	√	√
S17	√	√	√
S18	√	√	×
S19	√	√	√
S20	√	√	√
S21	√	√	√
S22	√	√	√
S23	√	√	×
S24	×	√	√
S25	×	√	×
Type of search: U1-OSLR (Search string); R1-U1-OSLR (Forward Snowballing – Google Scholar); R2-U1-OSLR (Forward Snowballing – IEEEXplore and ACM). **Legend**: √ – Yes; × – No			

Finally, in relation to R2-U1-OSLR, papers S18 and S23 were not retrieved by neither the ACM nor IEEE citation searches, despite being indexed by both search engines.

RQ2: Which were the studies selected/included by U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR?

Except for R2-U1-OSLR, there were studies that were retrieved however not included in both U1-OSLR and R1-U1-OSLR (Table V). Both U1-OSLR and R1-U1-OSLR had to look over a large number of titles and abstracts, which we hypothesize as a likely reason for the false negative results. Moreover, R1-U1-OSLR was conducted by a sole researcher, which may also affect the results. Note that we also checked (using Google Scholar) how many studies would have been retrieved using only OSLR as a seed set; it would have retrieved all studies, except for S19. This shows that using solely OSLR as a seed set would not have been sufficient to retrieve all the 15 studies in the superset.

Please note that Fig. 1 shows the primary studies included, whereas Tables IV and V detail respectively studies retrieved by the searches and whether they were included or not in the SLR update or its replications.

TABLE V. STUDIES INCLUDED IN U1-OSLR, R1-U1-OSLR, AND R2-U1-OSLR

Study		**U1-OSLR**	**R1-U1-OSLR**	**R2-U1-OSLR**
Study included	**S11**	√	√	√
	S12	√	√	√
	S13	√	√	√
	S14	√	√	√
	S15	√	√	√
	S16	√	√	√
	S17	√	√	√
	S18	√	**?**	×
	S19	√	**?**	√
	S20	√	√	√
	S21	√	√	√
	S22	**?**	√	√
	S23	**?**	√	×
	S24	×	**?**	√
	S25	×	√	×
Legend: √ – Included; ? - retrieved but not included; × – Not retrieved				

RQ3: With respect to forward snowballing. There were two specific questions concerning forward snowballing. The answers to these questions follow.

RQ3.1: Do different seed sets lead to significant differences in the set of included studies, and in the SLR conclusions?

Here we compare three different choices of seed sets (SLRs only; SLRs + primary studies; and primary studies only). Hence, we added a third possible seed set using solely the two papers that described the OSLR. This third seed set is searched using IEEE Xplore + ACM (to complement the seed set used in R2-U1-OSLR, so to be equivalent to the seed set employed in R1-U1-OSLR), and also searched using Google Scholar (so to assess whether R1-U1-OSLR would have the same studies using only the OSLR as seed set). Table VI shows the total number of studies (titles and abstracts) that had to be checked, and the list of included studies. If R2-U1-OSLR had also included in its seed set the two papers describing OSLR, it would have identified 13 new studies, rather than 12 (S23 would be the 13th paper).

Concerning the differences in the SLR conclusions, we have the following:

- S14: Cross-company model NOT significantly different from within-company model (missed by SS1)
- S18: Cross-company model NOT significantly different from within-company model (missed by all)
- S23: Cross-company model NOT significantly different from within-company model (missed by SS4)
- S22: Cross-company model NOT significantly different from within-company model (missed by SS1)
- S19: Inconclusive results (missed by SS1, SS2 and SS3)
- S24: Inconclusive results (missed by SS2 and SS3)
- S25: Inconclusive results (missed by SS1 and SS4)

TABLE VI. SEED SET AND THEIR RESPECTIVE SET OF INCLUDED STUDIES

SS#	SS1	SS2	SS3	SS4
Seed set	OSLR only – 2 papers [14, 15] IEEE Xplore + ACM)	OSLR only – 2 papers [14, 15] Google Scholar	OSLR + Primary studies Google Scholar	Primary studies only IEEE Xplore + ACM
# Studies Retrieved	114	224	1018	172
Studies Included (Super set = 15 studies)	(10) S11; S12; S13; S15; S16; S17; S20; S21; S23; S24	(12) S11; S12; S13; S14; S15; S16; S17; S20; S21; S22; S23; S25	(12) S11; S12; S13; S14; S15; S16; S17; S20; S21; S22; S23; S25	(12) S11; S12; S13; S14; S15; S16; S17; S19; S20; S21; S22; S24

There were four conclusive studies (S14, S18, S22 and S23) missed by at least one of the seed sets, all providing evidence supporting cross-company models NOT being significantly different from within-company models. However, all these seed sets included at least four studies (S13, S15, S16, and S20) that also provided evidence supporting cross-company models NOT being significantly different from within-company models. Therefore, the overall SLR conclusions, based on each of the four seed sets, are well aligned and almost similar.

RQ3.2: Do different processes lead to significant differences in the set of included studies, and in the SLR conclusions?

Here we compare two different processes: (a) no iteration, which assumes that SLRs and/or primary studies are cited; and (b) iteration until saturation is reached.

To answer this question, we merged the results from the first and fourth seed sets (see Table VI) into Choice 2 (Table VII).

In doing so, we acknowledge that any update to an existing SLR that uses forward snowballing should include in its seed set not only the primary studies included in that SLR, but also the reference(s) to the published SLR. Furthermore, to provide a genuine discussion, it is also important to consider the ideal results that could have been achieved using each of the two processes if all the papers that should have been selected were selected by the researchers carrying out those tasks. In other words, if there had been no human error during the filtering process in Choice 1, all the 15 studies would have been retrieved (see Table VII); this does not apply to Choice 2, as it did not retrieve S18 and S25 in the first place.

Concerning the differences in the SLR conclusions (considering only the studies selected by the researchers, as per Table VII), we have the following:

- S18: Cross-company model NOT significantly different from within-company model (missed by both Choice 1 and Choice 2)
- S19 and S24: Inconclusive results (missed by Choice 1)
- S25: Inconclusive results (missed by Choice 2)

There was only one conclusive study (S18) missed by both Choice 1 and Choice 2, so suggesting that, at least within this context, there would not have been any significant changes to the SLR results with using either Choice 1 or Choice 2.

TABLE VII. SEED SET AND ITS RESPECTIVE SET OF INCLUDED STUDIES

Seed set	Studies Included (Super set = 15 studies)
Choice 1 - OSLR + Primary studies (Google Scholar) + no iteration	(12) – S11; S12; S13; S14; S15; S16; S17; S20; S21; S22; S23; S25 (retrieved but did not include S18, S19 and S24)
Choice 2 - OSLR + Primary studies (IEEEXplore + ACM) + saturation	(13) – S11; S12; S13; S14; S15; S16; S17; S19; S20; S21; S22; S23; S24 (did not retrieve S18 and S25)

RQ4: Were there differences between the conclusions from U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR, when compared to the conclusions that would have been obtained using the superset of 15 studies?

Fig. 2 shows the results arranged by the SLR update and replications, and also by the four different study outcomes (Within-Company (WC) superior, Cross-company (CC) and WC not significantly different, CC superior, and Inconclusive results). When we discard the Inconclusive results, we can see that there are only three studies that were not included by the SLR update and replications, namely S18, S22 and S23. S18 and S23 show results that are also recurrent in many other studies (WC superior and/or CC and WC not significantly different). However, S22 presents one of the only two results to date where CC showed superior accuracy to WC. R1-U1-OSLR and R2-U1-OSLR both included S22; however, despite being retrieved via database search, it was not included by U1-OSLR. Similarly to RQ3, if there were no false negative results due to human judgement, our answer to this research question would be that differences between conclusions from the SLR update and its replications, and conclusions using the superset of 15 studies did not differ in a significant way.

RQ5: What are the differences, if any, between the conclusions based on the superset of 15 studies, and those in OSLR?

In relation to OSLR, we have 10 results (See Fig. 3): four results for WC showing superior accuracy (40%); four for CC and WC presenting similar accuracy (40%); and two with inconclusive results (20%). In relation to the OSLR aggregated revisions we have 23 results, arranged as follows: nine for WC showing superior accuracy (39%); nine for CC and WC presenting similar accuracy (39%); two for CC showing superior accuracy (9%); and three for inconclusive results (13%). The only representative change observed is given by two results showing CC accuracy to be superior to WC accuracy; other than that, percentages for the other three types of results are very similar

.

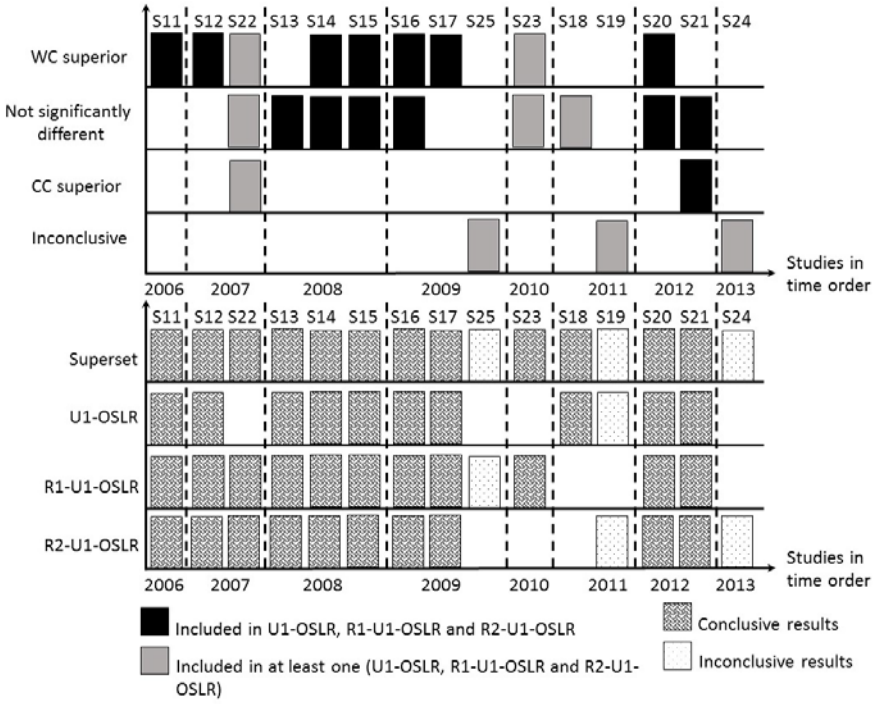

Fig. 2. Results for U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR, arranged by study along time axis

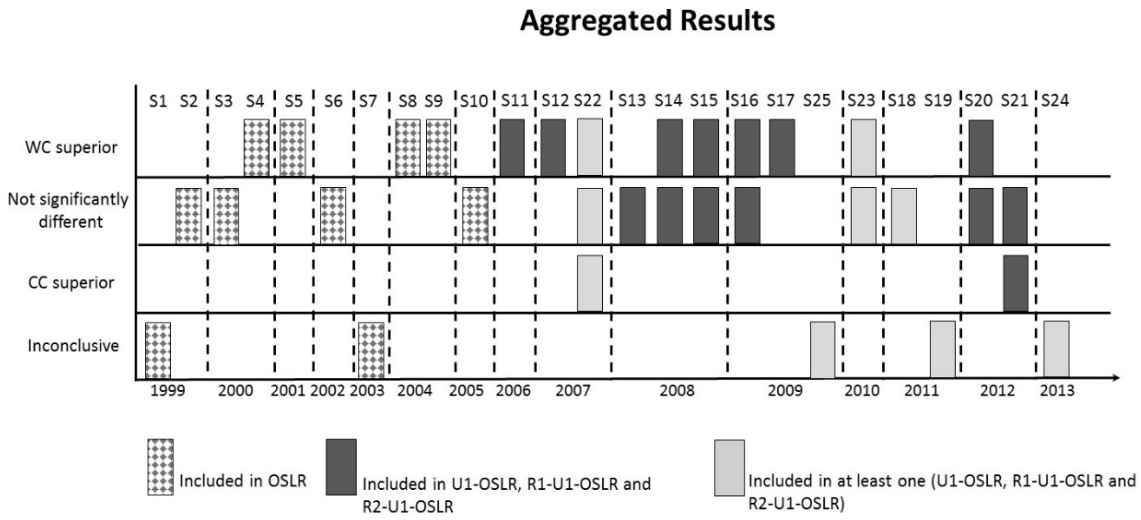

Fig. 3. Results for the OSLR, and also for aggregated results for U1-OSLR, R1-U1-OSLR, and R2-U1-OSLR, arranged by study along time axis

E. Recommendations based on Comparison Results

Table IV shows very clearly that the only approach where all 15 studies were retrieved used: i) forward snowballing with a single iteration; ii) employs Google Scholar to find citations to a seed set containing all the primary studies included in OSLR; and iii) also used the two papers describing OSLR. As pointed out in Section 3, under RQ2, such seed set is needed to retrieve all 15 studies. However, as shown in Table V, it is also important to highlight that the volume of citations returned from Google Scholar (used by R1-U1-OSLR), and the volume of references returned from several database searches employed by U1-OSLR, may increase the likelihood of false negatives, although many of the citations/references are easily discarded as non-relevant.

Further, the volume of citations returned from Google Scholar is primarily related to the high number of citations to the two papers describing the OSLR, and we would argue that citations to the papers describing OSLR could not be ignored. Furthermore, there is a significant overlap in references and hence the number of unique references is substantially lower.

With regard to group-work, R1-U1-OSLR was carried out by a single person; conversely, although U1-OSLR had

a team of people, the initial filtering of titles and abstracts was done separately and individually, and only the titles and abstracts chosen separately were combined and discussed during a joint meeting attended by most of the authors. Therefore, we argue that, whenever there are the resources available, the entire selection of studies should be done independently by two people, and then compared. We also argue that the best choice would be for these two people to be experienced in carrying out SLRs in SE. Although these two recommendations to group-work configuration when carrying out SLRs in SE are not new, and have been previously documented elsewhere (e.g. [9][10][14]), we are reiterating the importance herein. In summary, our recommendation is that any SLR updates in SE use:

1. A seed set containing the original SLR + primary studies
2. Google Scholar
3. Forward snowballing, and one iteration ought to be sufficient since any paper published on the topic of an SLR should refer to either the SLR or at least one of the primary studies
4. More than one researcher in the initial screening to minimize the risk for removing studies that should be included (false negatives).

It is important to note that our recommendation is based upon the approach that was found to be the most suitable when based on the evidence gathered and using as basis the combination of the original SLR's update, its two replications, all done with different researchers and where different search methods were employed to identify primary studies. As other SLR updates are carried out in SE, by different groups of authors, and using different search mechanisms, further evidence can be gathered and used to support (or not) our recommendations.

IV. Threats to Validity

As previously stated, our recommendations relating to how best to search for evidence when updating SLRs in SE are based on evidence gathered from investigating one combination of an original SLR + one update + two replications of the SLR update, which can be seen as a threat to conclusion validity. Our goal was to inform our recommendations by using evidence from an existing SLR, its update and corresponding replications, rather than to carry out a formal experiment with a simulated scenario, or to re-do a few SLR updates ourselves. The chosen combination was the only one that had a range of different authors for the original SLR, its update and replications, and where different search mechanisms were employed to identify primary studies.

The diversity of authors helped reduce bias when applying the different search methods to identify primary studies, and the use of different search methods provided an opportunity for these to be compared; such comparison informed our recommendations. One of the authors in this paper (Mendes) has taken part in and knows the OSLR very well and has frequently cited OSLR in all the following studies. Furthermore, many recent studies in the SLR topic were conducted with the participation of this paper's authors (not including Wohlin), who are well-aware of (and cited) OSLR. Therefore, it would seem that such knowledge could have influenced the effectiveness of the snowballing search in updating SLRs. However, studies from the other authors, retrieved using database searches, were also successfully retrieved using forward snowballing, which contradicts the "higher effectiveness" argument.

V. Conclusions

This paper investigates and makes recommendations towards an important aspect relating to the update of SLRs in SE – "how" best to search for evidence to update an SLR.

This aspect is addressed by using the results from a comparison of different search strategies (e.g., database search and forward snowballing processes) used by an SLR update and two replications, in the topic of effort estimation. Our results suggest that SLRs should be updated using:

1) A seed set containing the original SLR and its primary studies;
2) Google Scholar;
3) Forward snowballing. Note that one iteration should be sufficient since any paper published on the topic of an SLR ought to refer to either the SLR or at least one of the primary studies; and
4) More than one researcher in the initial screening to minimize the risk for removing studies that should be included (false negatives).

We do not claim that our recommendations be set in stone; rather, we suggest that further investigations be carried out on how to update SLRs in SE, based on evidence from SLRs + updates, or even via carrying out formal experiments, and such findings will broaden our understanding in this area, and may support (or not) our recommendations. Thus, the findings and recommendations should be seen as a first stepping stone towards identifying a suitable process for updating SLRs.

References

[1] L.C. Briand, K. El Emam, D. Surmann, I. Wieczorek, and K. Maxwell, "An assessment and comparison of common software cost estimation modeling techniques". In ICSE' 99, 1999, pp. 313–323.

[2] L.C. Briand, T. Langley, and I. Wieczorek, "A replicated assessment and comparison of common software cost modeling techniques". In ICSE' 00. ACM, pp. 377–386, 2000.

[3] D. Budgen, P. Brereton, S. Drummond, and N. Williams, "Reporting Systematic Reviews: Some Lessons from a Tertiary Study", draft report, 2017.

[4] K.R. Felizardo, E. Mendes, M. Kalinowski, E.F. Souza, N. Vijaykumar, "Using Forward Snowballing to update Systematic Reviews in Software Engineering". In ESEM' 16, pp. 1–6, 2016.

[5] F. Ferrucci, C. Gravino, S. Di Martino, L. Buglione, "Estimation web application development effort employing COSMIC: A comparison between the use of a cross-company and a single-company dataset". In 6th Software Measurement European Forum, pp. 77–89, 2009.

[6] F. Ferrucci, F. Sarro, and E. Mendes, "Web effort estimation: The value of cross-company data set compared to single-company data set". In PROMISE' 12, ACM, pp. 29–38, 2012.

[7] R. Jeffery, M. Ruhe, and I. Wieczorek, "Using public domain metrics to estimate software development effort". In 7th IEEE Symposium on Software Metrics (METRICS' 01), pp. 16–27, 2001.

[8] R. Jeffery, M. Ruhe, and I. Wieczorek, "A comparative study of two software development cost modeling techniques using multi-

organizational and company-specific data". Information and Software Technology, vol. 42, nº. 14, pp. 1009–1016, 2000.

[9] B.A. Kitchenham, "Procedures for Performing Systematic Reviews", Technical Report TR/SE-0401 ISSN:1353-7776, School of Computer Science and Mathematics, Keele University, 2004.

[10] B.A. Kitchenham, and S. Charters "Guidelines for performing systematic literature reviews in software engineering", Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University, 2007.

[11] B.A. Kitchenham, O. P. Brereton, D. Budgen, M. Turner and S. Linkman, "Systematic literature reviews in software engineering – a systematic literature review". Information and Software Technology, vol. 51, nº. 1, pp. 7–15, 2009.

[12] B.A. Kitchenham, T. Dybå and M. Jørgensen, Evidence-Based Software Engineering. Proceedings ICSE 2004, pp: 273–281 B.

[13] B.A. Kitchenham and E. Mendes, "A Comparison of Cross-company and Within-company Effort Estimation Models for Web Applications", In: EASE'04, 2004, pp. 47–56.

[14] B.A. Kitchenham, E. Mendes, and G.H. Travassos, "A Systematic Review of Cross- vs. Within-company Cost Estimation Studies". Proceedings of EASE, pp: 1–10, 2006.

[15] B.A. Kitchenham, E. Mendes, and G.H. Travassos, "Cross versus within-company cost estimation studies: A systematic review". IEEE TSE, vol. 33, nº. 5, pp. 316–329, 2007.

[16] B.A Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, Stephen Linkman, "Literature reviews in software engineering – a tertiary study", Information and Software Technology, vol. 52, nº. 8, 2010, p. 792–805.

[17] E. Kocaguneli, B. Cukic, T. Menzies and H. Lu, "Building a second opinion: learning cross-company data", In PROMISE' 13). ACM, pp. 1–10, 2013.

[18] E. Kocaguneli, G. Gray, T. Menzies, Y. Yang and J.W. Kung, "When to use data from other projects for effort estimation". In ASE' 10, pp. 321–324, 2010.

[19] E. Kocaguneli and T. Menzies, "How to find relevant data for effort estimation?" In 5th International Symposium on Empirical Software Engineering and Measurement (ESEM' 11), pp. 255–264, 2011.

[20] M. Lefley and M. Shepperd, "Using genetic programming to improve software effort estimation based on general data sets". In Genetic and Evolutionary Computation (GECCO' 03), ser. LNCS, vol. 2724. Springer-Verlag, pp. 2477–2487, 2003.

[21] C. Lokan and E. Mendes, "Using chronological splitting to compare cross- and single-company effort models: Further investigation". In 32nd Australasian Computer Science Conference (ACSC' 09), ser. CRPIT, vol. 91. ACS, pp. 35–42, 2009.

[22] C. Lokan and E. Mendes, "Investigating the use of chronological splitting to compare software cross-company and single-company effort predictions". In 12th International Engineering (EASE' 08). British Computer Society, pp. 136–145, 2008.

[23] C. Lokan and E. Mendes, "Cross-company and single-company effort models using the isbsg database: A further replicated study". In 5th International Symposium on Empirical Software Engineering (ISESE' 06). ACM, pp. 75–84, 2006.

[24] K. Maxwell, L. V. Wassenhove, and S. Dutta, "Performance evaluation of general and company specific models in software development effort estimation". Management Science, vol. 45, nº. 6, pp. 787–803, 1999.

[25] E. Mendes, S. Di Martino, F. Ferrucci, and C. Gravino, "Cross-company vs. single-company web effort models using the tukutuku database: Na extended study". JSS, vol. 81, nº. 5, pp. 673–690, 2008.

[26] E. Mendes, S. Di Martino, F. Ferrucci, and C. Gravino, "Effort estimation: How valuable is it for a web company to use a cross-company data set, compared to using its own single-company data set?" In WWW' 07,. ACM, pp. 963–972, 2007.

[27] E. Mendes, M. Kalinowski, D. Martins, F. Ferrucci, and F. Sarro, "Cross vs. within-company cost estimation studies revisited: An extended systematic review". In EASE' 14, ACM, pp. 12:1–12:10, 2014.

[28] E. Mendes and B.A. Kitchenham, "Further comparison of cross-company and within-company effort estimation models for web applications". In 10th IEEE Symposium on Software Metrics (METRICS' 04), pp. 348–357, 2004.

[29] E. Mendes and C. Lokan, "Investigating the use of chronological splitting to compare software cross-company and single-company effort predictions: A replicated study". In 13th International Conference on Evaluation and Assessment in Software Engineering (EASE' 09). British Computer Society, pp. 11–20, 2009.

[30] E. Mendes and C. Lokan, "Replicating studies on cross- vs single-company effort models using the isbsg database". Empirical Software Engineering, vol. 13, nº. 1, pp. 3–37, 2008.

[31] E. Mendes, C. Lokan, R. Harrison, and C. Triggs, "A replicated comparison of cross-company and within-company effort estimation models using the isbsg database". In 11st IEEE Symposium on Software Metrics (METRICS' 05), pp. 1–10, 2005.

[32] E. Mendes, C. Wohlin, K. Felizardo, and M. Kalinowski, When to update Systematic Literature Reviews in Software Engineering? submitted manuscript, 2019.

[33] L. L. Minku and X. Yao, "Can cross-company data improve performance in software effort estimation?" In 8th International Conference on Predictive Models in Software Engineering (PROMISE' 12). ACM, pp. 69–78, 2012.

[34] R. Premraj and T. Zimmermann, "Building software cost estimation models using homogeneous data". In 1st International Symposium on Empirical Software Engineering and Measurement (ESEM' 07), pp. 393–400, 2007.

[35] F. Q. da Silva, A. L. Santos, S. Soares, A. C. C. Franc¸a, C. V. Monteiro, F. F. Maciel, Six years of systematic literature reviews in software engineering: An updated tertiary study, Information and Software Technology vol. 53, nº 9, 2011, 899–913.

[36] O. Top, B. Ozkan, M. Nabi, and O. Demirors, "Internal and external software benchmark repository utilization for effort estimation". In Software Measurement, 2011 Joint Conference of the 21st International Workshop on and 6th International Conference on Software Process and Product Measurement (IWSM-MENSURA' 11), pp. 302–307, 2011.

[37] I. Wieczorek and M. Ruhe, "How valuable is company-specific data compared to multi-company data for software cost estimation?" In 8th IEEE Symposium on Software Metrics (METRICS' 02), pp. 237–246, 2002.

[38] C. Wohlin. "Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering". In 18th International Conference on Evaluation and Assessment in Software Engineering, 321–330, 2014.

[39] C. Wohlin, "Second-generation Systematic Literature Studies Using Snowballing". In 20th International Conference on Evaluation and Assessment in Software Engineering (EASE' 16). ACM, pp. 15:1–15:6, 2016.

[40] O. Dieste, López, M. and Ramos, F. "Formalizing a Systematic Review Updating Process". In: 6th International Conference on Software Engineering Research, Management and Applications (SERA'08), pp. 143–150, 2008.

[41] F.C. Ferrari and J.C. Maldonado. "Experimenting with a Multi-Iteration Systematic Review in Software Engineering". In 5th Experimental Software Engineering Latin America Workshop (ESELAW'08), pp. 1–10, 2008.

[42] K.R. Felizardo, E.Y. Nakagawa, S.G. MacDonell, J.C. Maldonado. "A Visual Analysis Approach to Update Systematic Reviews". In 18th International Conference on Evaluation and Assessment in Software Engineering (EASE'14), pp. 1–10, 2014.

[43] A.Y.I. da Silva, K.R. Felizardo, E.F. de Souza, N.L. Vijaykumar and E.Y. Nakagawa. "Evaluating electronic databases for forward snowballing application to support secondary studies updates – Emergent Results". In 32nd Brazilian Symposium on Software Engineering (SBES' 18), pp. 1–10, 2018.

[44] L.M.G. Rodriguez, K.R. Felizardo, L.B.R. Oliveira, E.Y. Nakagawa. "An experience report on update of systematic literature reviews". In 29th International Conference on Software Engineering and Knowledge Engineering (SEKE' 17), pp. 1–10, 2017.

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