Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering

Claes Wohlin
Blekinge Institute of Technology
SE – 371 79, Karlskrona
Sweden
+46-(0)455-385820
Claes.Wohlin@bth.se

ABSTRACT

Background: Systematic literature studies have become common in software engineering, and hence it is important to understand how to conduct them efficiently and reliably.

Objective: This paper presents guidelines for conducting literature reviews using a snowballing approach, and they are illustrated and evaluated by replicating a published systematic literature review.

Method: The guidelines are based on the experience from conducting several systematic literature reviews and experimenting with different approaches.

Results: The guidelines for using snowballing as a way to search for relevant literature was successfully applied to a systematic literature review.

Conclusions: It is concluded that using snowballing, as a first search strategy, may very well be a good alternative to the use of database searches.

Categories and Subject Descriptors

D.2 [**Software Engineering**]: Management, and G.3 [**Probability and Statistics**]: Experimental design.

General Terms

Experimentation, Measurement

Keywords

Systematic literature review, systematic mapping studies, snowballing, snowball search, replication

1. INTRODUCTION

Systematic literature studies, including both reviews and maps, have emerged as a way of synthesizing evidence and then ultimately allowing researchers to come to a joint understanding of the status of a research area in software engineering in the last decade. Inspired by medicine, the concept of evidence-based software engineering was coined by Kitchenham et al. [1]. Similar ideas have been brought into information systems research, e.g. by Webster and Watson [2].

EASE '14, May 13 - 14 2014, London, England, BC, United Kingdom
Copyright 2014 ACM 978-1-4503-2476-2/14/05…$15.00.
http://dx.doi.org/10.1145/2601248.2601268

However, the need to synthesize research results in software engineering was discussed already in the late 1990s [3, 4, 5]. Pickard et al. [3] discuss combining research results, Miller [4] addresses the issue of combining research results through meta-analysis and Hayes [5] uses the concept of synthesis of research results. They all have in common that they stress the need for a systematic approach to not only conducting individual research studies, but also to building knowledge from combining findings from different studies on a topic. One such early example is the work by Basili et al. [6], where the authors look into combining the research and hence knowledge we have regarding research on software inspections.

Based on the original EBSE ideas [1], research related to systematic literature studies has subsequently evolved. Guidelines for conducting systematic literature reviews have been developed [7]. Systematic mapping studies have been highlighted as a complement to systematic literature reviews [8]. Kitchenham et al. [9] discuss the use of systematic mapping studies as a starting point for further research. Here, we use systematic literature studies as a collective term for systematic literature reviews and systematic mapping studies.

This paper complements previous guidelines for systematic literature reviews in software engineering. It does so by extending and detailing the steps for using snowballing as a search approach for systematic literature studies. Snowballing refers to using the reference list of a paper or the citations to the paper to identify additional papers. However, snowballing could benefit from not only looking at the reference lists and citations, but to complement it with a systematic way of looking at where papers are actually referenced and where papers are cited. Using the references and the citations respectively is referred to as backward and forward snowballing. It builds on ideas presented by for example Webster and Watson [2] in information systems and the procedure outlined by Wohlin and Prikladnicki [10]. The snowballing guidelines are illustrated and evaluated by replicating a published reliability study of systematic literature reviews [11]. In this paper, the authors conducted two systematic literature reviews in parallel to evaluate the reliability of literature reviews. The evaluation here provides a third data point using snowballing as the main approach to identify relevant literature while a database-driven search was applied in the reliability study by MacDonell et al. [11].

Based on the above motivation, this paper has two main research objectives:

1. Formulate a systematic snowballing procedure for systematic literature studies in software engineering,
2. Illustrate and evaluate the snowballing procedure by replicating a published systematic literature review.

The remainder of the paper is outlined as follows. Related work is presented in Section 2. The snowballing procedure is presented in Section 3. The replication of the systematic literature reviews presented in [11] using the snowballing procedure is presented in Section 4. A discussion of the snowballing procedure is provided in Section 5. Finally, a summary is presented in Section 6.

2. RELATED WORK

The guidelines for systematic literature reviews [7] are very important in supporting researchers in conducting systematic literature studies. However, they have also generated discussions. According to the guidelines, the objective is to identify all relevant research. This is fine as an objective, but it is unlikely to work in practice in particular for literature studies of a broader area. The challenge related to population of papers vs. the actual sample identified is discussed in, for example, [12]. A study by Kitchenham et al. [13] may be used to exemplify the challenge. They conducted a systematic literature review using manual search and found 20 relevant papers. When the authors discussed limitations of the study, they mentioned the fact that they used manual search and may have missed some relevant studies. Due to this limitation, Kitchenham et al. [14] repeated the study using an automated database search and found 33 additional studies. This example illustrates that we end up being forced into accepting samples of relevant papers; the challenge is to get the best possible sample from the population. Thus, the search strategy is key to ensure a good starting point for the identification of studies and ultimately for the actual outcome of the study.

The example from Kitchenham et al. [13, 14] gives the impression that automation is better than manual search. However, the point is not really about manual vs. automatic; it is really about being systematic. The claim is based on the fact that even if database searches can be made automatically, the search is not better than the search string used. It is very difficult to formulate good search strings, since all too often the terminology used is not standardized and if using broad search terms then a large number of irrelevant papers will be found in the search. The latter creates substantial manual work that also is error-prone.

The guidelines [7] take database searches using search strings from the area of study as a starting point. However, the guidelines also state that other complementary searches are needed. The latter includes for example: reference lists, grey literature, specific research outlets (journals or conferences) and researchers in the field. Unfortunately, most systematic literature studies stop short of these complementary searches. This may be understandable given the amount of work it is to conduct systematic searches in a number of databases and then identify the relevant papers of sufficiently high quality. The searches in databases are challenging for several reasons, including selection of databases, different interfaces for the databases, different ways of constructing search strings, different search limitations in the databases and identification of synonyms of terms used. This reasoning leads to two conclusions: 1) the choice of the first step in the search strategy often becomes the only step, i.e. search databases (if using the guidelines), and 2) given the challenges with the database searches, we may miss important literature. Thus, other alternative approaches may be considered, for example, the use of a snowballing procedure [2].

Greenhalgh and Peacock [15] applied three different search methods in their research, and concluded that protocol-driven search approaches by themselves are not necessarily the most efficient method regardless of the number of databases used, since some sources may be found through personal knowledge/contacts (e.g. browsing library shelves and asking colleagues), and snowballing is the best approach for identifying sources published in obscure journals according to their study.

Skoglund and Runeson [16] studied a reference-based search approach with the primary objective to reduce the number of initial articles found in systematic literature studies. Despite that the proposed method increased the precision without missing too many relevant papers for the technically focused reviews, its results were not satisfactory when the search area was wide or the searches included general terms. This implies that the choice of approach to searching may be context-dependent.

In summary, too few studies have addressed the reliability of systematic literature studies. As discussed here, they have either compared different systematic literature studies to check whether the same results are achieved [11], [12] and [17], or investigated more efficient approaches of searching [15], [16] and [18]. As a complement to previous studies, Jalali and Wohlin [19] investigated the reliability of systematic literature studies using different search strategies. This was done by comparing the outcome of two studies on the same topic using the guidelines by Kitchenham and Charters [7] and the steps for snowballing outlined by Webster and Watson [2] for finding the relevant literature. Here, it should be noted that the steps for snowballing are only outlined and they cannot be viewed as a guideline in the same way as those presented by Kitchenham and Charters. Thus, there is a need for more detailed guidelines for snowballing to conduct thorough and repeatable systematic literature studies using a snowballing approach as the first step.

When it comes to the reliability of systematic literature reviews, the paper by MacDonell et al. [11] is of particular interest. The reason being that the snowballing procedure presented next is evaluated in Section 4 based on study presented in [11]. MacDonell et al. evaluated the reliability of systematic reviews through comparing the results of two studies with a common research question performed by two independent groups of researchers. In their case, the systematic literature review seemed to be robust to differences in process and people, and it produced stable outcomes.

3. SNOWBALLING PROCEDURE

The basic planning and motivation of a systematic literature study is independent of the search approach, which is the main concern here. Thus, the basic steps for planning a literature study as presented in [7] are still relevant even if applying a different approach to the search.

The snowballing procedure is outlined in steps in Figure 1 and described in the following subsections.

3.1 Start Set

In database searches, the first step is to identify keywords and formulate search strings. When applying a snowballing approach, the first challenge is to identify a start set of papers to use for the snowballing procedure. The start set is shown in the top of Figure 1. Any search for papers to include in the start set generates a tentative start set. The actual start set is only those papers in the tentative start set that at the end will be included in the systematic literature study.

A good start set may be identified by using, for example, Google Scholar. It is a good alternative to avoid bias in favour of any

specific publisher. A good start set has the following characteristics:

- If relevant papers may come from different communities, then it is important to have these covered in the start set. This is particularly crucial if there is a risk that relevant papers may be in independent clusters, i.e. in clusters of papers not referring to each other.
- The number of papers in the start set should not be too small. The actual size of the start set depends on the breadth of the area being studied. A smaller area (more specific focus) requires fewer papers than a broad area.
- If too many papers are found, for example due to having very general search terms in Google Scholar, then identifying a number of relevant and highly cited papers may be an alternative.
- The start set should cover several different publishers, years and authors. The important issue here is diversity.
- The start set ought to be formulated from keywords in the research question, while preferably also taking synonyms into account. The latter is to avoid only capturing papers using a specific terminology and missing papers using a slightly different terminology.

There is no silver bullet for identifying a good start set, which is very similar to the challenges in identifying search strings in database searches. One possibility in snowballing is to identify a seminal and highly cited paper in the area of the systematic literature study. The challenge of identifying a good start set for snowballing is an area for future research. An illustrative example of terminology challenges is provided in [19], where agile practices in global software engineering were studied. In the database search, a paper using the formulation "cross-continent" development was not caught, but it became obvious in the snowballing that the paper should be included. This illustrates the difficulty with inconsistent terminology. The actual results from the systematic literature review are presented in [20]. This example also illustrates the need for a more consistent usage of terminology to enable good systematic literature studies. An attempt to address this in the area of global software engineering is presented in [21].

3.2 Iterations

Once the start set is decided, including only papers that will be included in the final analysis, it is time to start the first iteration conducting backward and forward snowballing. To finally decide to include a paper means that the full paper should be examined before deciding to use it as a paper in the snowballing. If not doing this, a rollback is needed if other papers are included based on a paper that later is excluded. Thus, it is important to be certain on inclusion before using the paper for snowballing at all.

3.2.1 Backward Snowballing

If starting to the left in Figure 1, backward snowballing means using the reference list to identify new papers to include. The first step is to go through the reference list and exclude papers that do not fulfil the basic criteria such as, for example, language, publication year and type of publication (if only considering peer-reviewed papers). The next step is to remove papers from the list that have already been examined based on being found earlier through either backward or forward snowballing in this or a previous iteration. Once these are removed, the remaining papers are real candidates for inclusion.

The first two steps in the backward snowballing is to extract as much information as possible from the paper being examined and not go to the new paper until no more information is available in the paper being examined. The following is examined in the reference list:

- Title – Is it tentatively a paper to include?
- Publication venue – Is it published in a place where relevant papers may be published?
- Authors – Do we know that the authors have published relevant paper in the area studied before?

Papers cannot be excluded based on, for example, that the author is not known for publishing in the area, but a paper may be more likely to be included if the author regularly publishes in the area. Thus, the information found in the reference list must be examined and evaluated carefully. If the paper still is a candidate for inclusion after having looked at it in the reference list, then the next step is to examine where and how the paper is referenced. The place and context of the reference may provide important information about the actual content of the candidate paper, and it is practical to get this information from the paper being examined instead of having to find the candidate paper directly.

If the paper is candidate for inclusion after having examined all information available in the paper being examined, then it is time to find the potentially new paper to include.

Once the paper is found, the abstract is read first and then other parts of the paper are read until a definitive decision can be taken to either include or exclude the paper. It is recommended not to start reading the paper from the beginning to end directly, instead it is recommended to browse through the paper and read the most relevant parts of the paper to be able to make a decision about inclusion of exclusion in an efficient way.

3.2.2 Forward Snowballing

Forward snowballing refers to identifying new papers based on those papers citing the paper being examined, and it is displayed to the right in Figure 1. The citations to the paper being examined are studied using Google Scholar. Quotes are removed in Google Scholar, and only citations are used.

Each candidate citing the paper is examined. The first screening is done based on the information provided in Google Scholar. If this information is insufficient for a decision, the citing paper is studied in more detail. First, the abstract is studied, and if this is insufficient, the place citing the paper already included is examined. If this is insufficient, the full text is studied to make a decision regarding the new paper. The approach to go through the papers is similar as to papers identified using backward snowballing.

3.2.3 Inclusion and Exclusion

As shown in Figure 1, it is important to decide on either inclusion or exclusion before starting to use a new paper for snowballing. If moving too quickly into using a paper for snowballing and then later realizing that the paper should not have been included there is a problem, and the process has to be rolled back and papers removed if being wrongly included. Only papers found through included papers should be used in the analysis.

After backward and forward snowballing, new papers identified in the iteration are put into a pile to go into the next iteration. It is important to do one iteration at the time to get traceability.

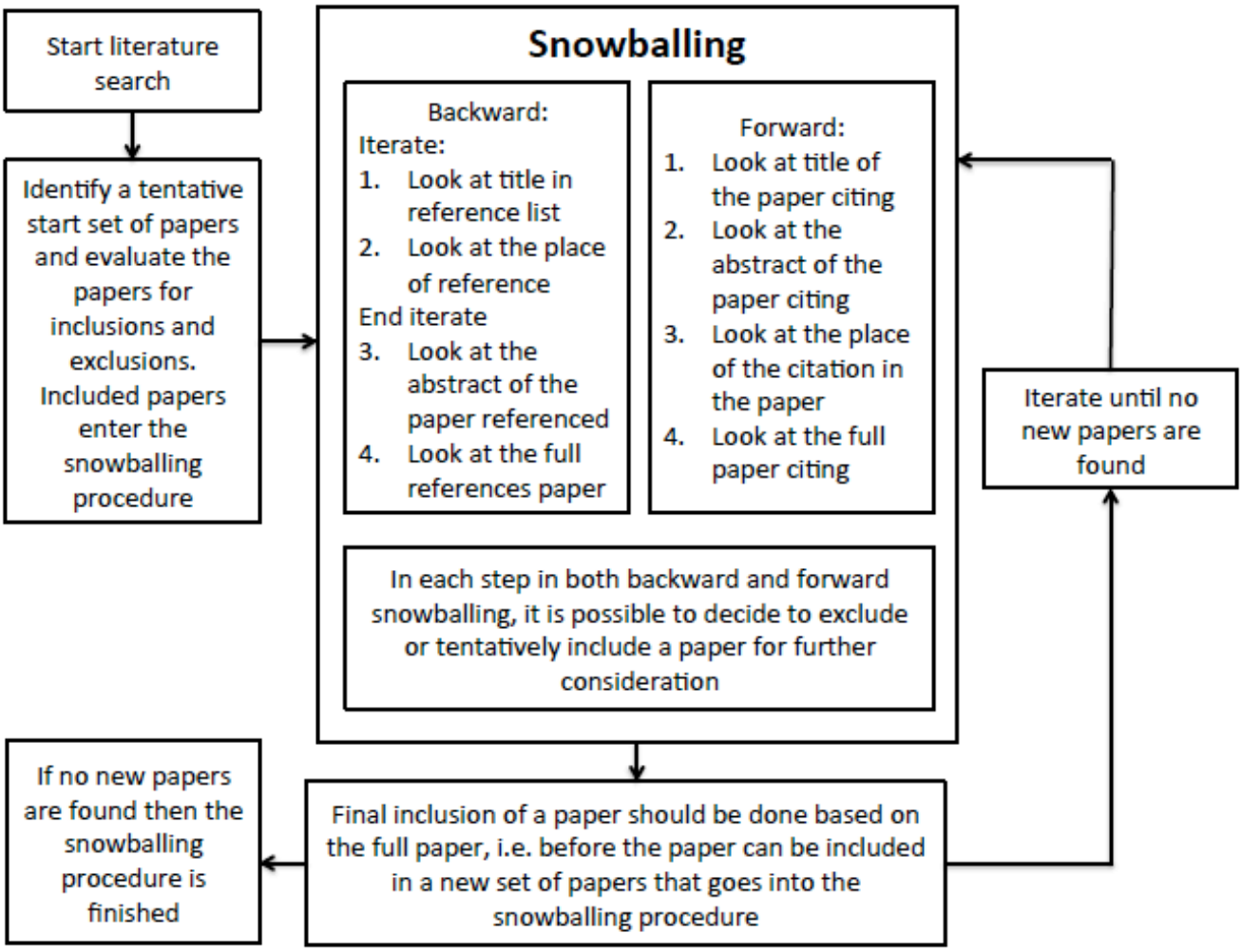

Figure 1. Snowballing procedure.

3.3 Authors

Once no new papers are found in the iterations using both backward and forward snowballing, the loop is ended. To complement the snowballing, it is recommended to contact the authors of included papers to potentially identify some additional papers. It is most important to contact the most active researchers in the area. Based on any new papers identified, the snowballing procedure outlined in Figure 1 must be re-started.

Other alternative methods to identify additional papers may also be considered, for example, searching in specific journals or conferences that are likely to include more papers on the topic. The journals and conferences may be identified through looking at where included papers are published.

3.4 Data Extraction

All papers identified go into data extraction, which should be conducted in accordance with the research questions posed in the systematic literature study. Given that the full papers have to be investigated before they go into the snowballing procedure, it may be considered to conduct the data extraction at the same time as deciding whether the paper should be included or not.

4. REPLICATION

4.1 Introduction

The paper by MacDonell et al. [11] was read 6-12 months before deciding to replicate it using the snowballing procedure formulated in this paper. Once it was decided to replicate the systematic literature review, Sections 1-3 were read in detail to ensure that the replication was conducted in a fair way in relation to the original study. In particular, the research question in the paper is carefully studied to enable replication. The research question in MacDonell et al. [11] is formulated as follows: *What evidence is there that cross-company estimation models are at least as good as within-company estimation models for predicting effort for software projects?*

Thus, the objective of the replication is primarily to illustrate and evaluate the snowballing procedure proposed in Section 3. While doing this, the intention is to answer the research question posed above, and then to reflect on the results in comparison with the two systematic literature reviews presented in [11].

4.2 Start Set

As mentioned above a key challenge is to identify a good tentative start set. In this particular case, the research question posed in [11] was a good starting point. It was decided to avoid publisher bias (e.g. searching in one publisher's database) and do the search in Google Scholar. The following string of words was put into Google Scholar: cross-company within-company software effort estimation, and the time frame chosen was 1995-2005. The latter is for replication purposes. The actual search was conducted August 20, 2013. But given the time frame of the search, the actual search date is of less importance in this case. However, under other circumstances the date could be of importance since the content of the databases changes and what Google Scholar indexes may also change over time.

In total, 13 candidates for inclusion were identified; they are denoted C1, C2 and so forth to indicate that they are candidates for inclusion. The 13 papers are:

C1. Mendes, E. and B.A. Kitchenham, Further Comparison of Cross-Company and Within Company Effort Estimation Models for Web Applications. Proceedings Metrics'04, Chicago, Illinois September 11-17th 2004, IEEE Computer Society, pp 348-357, 2004.

C2. Kitchenham, B.A., and E. Mendes. A Comparison of Cross-company and Within-company Effort Estimation Models for Web Applications, Proceedings 8th International Conference on Empirical Assessment in Software Engineering EASE 2004, Computer Society Press, 2004, pp. 47-55.

C3. Mendes, E., Lokan, C., Harrison, R., and Triggs, C. A Replicated Comparison of Cross-company and Within-company Effort Estimation Models using the ISBSG Database, in Proceedings of Metrics'05, Como, 2005.

C4. Mair, C. and Shepperd, M., "The Consistency of Empirical Comparisons of Regression and Analogy-based Software Project Cost Prediction", Proc. Int. Symp. On Empirical Software Engineering, 2005.

C5. Premraj, R., Shepperd, M., Kitchenham, B. and Forselius, P., "An Empirical Analysis of Software Productivity over Time", Proc. Int. Symp. On Software Metrics, 2005.

C6. B. Kitchenham and E. Mendes, "Software Productivity Measurement using Multiple Size Measures", IEEE Trans. On Softw. Eng., vol. 30, pp. 1023-1035, 2004.

C7. Mendes, E., N. Mosley, and S. Counsell, "Investigating Web Size Metrics for Early Web Cost Estimation", Journal of Systems and Software, Vol. 77, No. 2, pp. 157-172, 2005.

C8. Mendes, E., N. Mosley, and S. Counsell, Investigating Early Web Size Measures for Web Cost Estimation, Proceedings of EASE'2003 Conference, Keele, April, 2003, pp. 1-22.

C9. Mendes, E., Dinakaran, G. and Mosley, N., "How Valuable is it for a Web Company to Use a Cross-company Cost Model, Compared to Using Its Own Single-company Model?", Technical report, 2005.

C10. Diaz, L.M. and Buxmann, P., "The Value of Cooperative Planning in Supply Chains – A Simulative Approach –", Proceedings European Conference on Information Systems, 2003.

C11. Bishop, L. and Levine, D. I., "Computer-Mediated Communication as Employee Voice: A Case Study", Industrial and Labor Relations Review, Vol. 52, No. 2, pp. 213-233, 1999.

C12. Sin, C. C-S., "An Exploratory Empirical study of the Role of Manufacturing in Product Formulation", PhD thesis, 1997.

C13. Thompson, M., Zimbardo, P. and Hutchinson, G., "Consumers are Having Second Thoughts about Online Dating – Are the Real Benefits Getting Lost in Over Promises?", Technical report weAttract.com, 2005.

First non-peer reviewed candidates were excluded and then candidates covering the same study were excluded. Candidates 9, 12 and 13 were excluded based on not being a peer-reviewed journal of conference/workshop paper. Candidate 8 was excluded since it was judged that Candidate 7 is an extension of Candidate 8. The other candidates were reviewed more in-depth. After which, it was decided to exclude Candidates 4, 5, 6, 7, 10 and 11 due to being out of scope. Based on the inclusion/exclusion criteria, it was decided to include candidates 1, 2 and 3. The latter three are denoted P1, P2 and P3 respectively and these papers form the start set for the snowballing.

The identified start set is far from perfect, since the three papers identified have one author in common. It would have been better to have at least one paper from someone else to mitigate the risk of missing papers not being linked to these three papers. However, given that it was decided to use the research question in [11] as a starting point no action was taken.

4.3 Iteration 1

From the start set of three papers, both backward and forward snowballing were conducted.

4.3.1 Backward Snowballing

In backward snowballing, the references of the three included papers are studied to identify more papers to include in the study. Only references in the time frame studied are considered. The three papers are evaluated one at the time.

P1 includes 17 references where one reference is already included and one reference is already excluded. This leaves 15 references to evaluate. Four references are excluded based on publication year, and four references are excluded based on title or type of publication. Two references identified as candidates based on their title (denoted P6 and P8 below). Three papers were identified based on how they were used when referring to them (denoted P4, P5 and P7 below), and two papers were excluded based on the place and context of the reference. The full text of the five candidate papers were evaluated to avoid using a paper in the snowballing procedure that later may be excluded, since final inclusion must be based on the full paper. All five papers were judged as relevant and hence included in the study. Thus, the following five papers were added to the list of papers to be included:

P4. Briand, L.C., K. El-Emam, K. Maxwell, D. Surmann, I. Wieczorek. An Assessment and Comparison of Common Cost Estimation Models. Proceedings of the 21st International Conference on Software Engineering, ICSE 99, 1999, pp. 313-322.

P5. Briand, L.C., T. Langley, I. Wieczorek. A Replicated Assessment of Common Software Cost Estimation Techniques. Proceedings of the 22nd International Conference on Software Engineering, ICSE 20, 2000, pp. 377-386.

P6. Jeffery, R., .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, 42, 2000, pp. 1009-1016.

P7. Jeffery, R., M. Ruhe and I. Wieczorek. Using Public Domain Metrics to Estimate Software Development Effort. Proceedings 7th International Software Metrics Symposium, London, IEEE Computer Society Press, 2001, pp. 16-27.

P8. Wieczorek, I. and M. Ruhe. How Valuable is Company-specific Data Compared to Multi-company Data for Software Cost Estimation? . Proceedings 8th International Software Metrics Symposium, Ottawa, IEEE Computer Society Press, June 2002, pp. 237-246.

P2 has 18 references, but a majority of them are the same as for P1. Only four new references were identified. Neither the title nor the place and context of the references gave definitive information about whether or not to include or exclude the four papers. Thus, the abstract was first studied and it was still inconclusive, and hence the full papers were evaluated. After having gone through the full papers, it was decided that all four papers should be excluded.

P3 includes 21 references with many references in common with P1 and P2. In total six new references were identified. One was excluded based on the publication year and three were excluded based on the titles. The other two papers were candidates for inclusion. One paper was identified based on the title (P10) and one was identified from the place and context of the reference (P9). The full text of the papers was evaluated and it was decided to include both papers. The two papers included are:

P9. Lefley, M., and M.J. Shepperd, Using Genetic Programming to Improve Software Effort Estimation Based on General Data Sets, Proceedings of GECCO 2003, LNCS 2724, Springer-Verlag, pp. 2477-2487, 2003.

P10. Maxwell, K., L.V. Wassenhove, and S. Dutta, Performance Evaluation of General and Company Specific Models in Software Development Effort Estimation, Management Science, 45(6), June, pp. 787-803, 1999.

4.3.2 Forward Snowballing

In forward snowballing, the papers citing the three papers in the start set are evaluated. The citation analysis is conducted using Google Scholar. Given that the three papers in the start set are published in the end of the time frame considered, it is no surprise that there are few citations. The time frame studied is 1995-2005, and the three papers are published in 2004, 2004 and 2005 respectively. Five papers cite P1, but all of them are in the tentative start set (C1-C13). The situation is similar for P2 and P3. Four papers cite P2 and one paper cites P3 in the time frame studied, and once again the papers are already in the tentative start set. Thus, no new papers were identified from forward snowballing from the start set (P1-P3).

4.3.3 Summary of Status

The tentative start set included 13 candidates, which were evaluated. From these 13 candidates, three papers were included in the study. From these three papers, 25 candidates were evaluated (15 from P1, 4 from P2 and 6 from P3) from the backward snowballing. Seven new papers were included, denoted P4-P10. No new candidates needed to be evaluated based on the forward snowballing. In total 38 papers have been evaluated so far and 10 papers have been included in the study.

4.4 Iteration 2

The seven new papers identified (P4-P10) go into the first iteration. Thus, first backward snowballing is conducted for these seven papers and then a forward snowballing is done too.

4.4.1 Backward Snowballing

P4 includes 33 references, and 28 of them are perceived as new. Here it should be noted that given the number of references studied, there is always a risk that the same reference is studied more than once. The reason being that it is not deemed efficient to put a number of references into a tool, which later are excluded immediately and hence there may be some random errors in the actual numbers. For example, among the perceived new 28 references, a paper already evaluated maybe hiding but the researcher may not remember.

Out of the 28 references, 14 of them are excluded based on the publication year and 12 references are excluded based on the title. Only two references call for a closer examination. For these two papers, the places and context of the reference were identified in the paper, and it was concluded that none of them should be included. Thus, no new paper was identified from P4.

P5 has 32 references, but a large number of them have already been examined. In total, 12 references are perceived as new. Out of these 12 references, three references are excluded based on the publication year and the others are excluded based on the title. Thus, no new paper was identified from P5.

P6 includes 19 references, and the situation is quite similar as for P5. Only eight new references are identified. Two references are excluded based on publication year and the other six references are excluded based on title. Once again no new papers are identified.

P7 has 31 references, and nine of them are perceived as new. These nine are examined. It is concluded that three references can be excluded based on the publication year and five are excluded based on the title. It leaves one paper for further examination. The place and context of the reference to this paper is investigated, and it is concluded to exclude the paper. No new papers are included from P7.

P8 includes 28 references, and only one of them is new. The new paper is excluded based on title.

P9 has also 28 references, but in this case 18 references are perceived as new. Although having many new references, no new paper is identified for inclusion. Six references are excluded based on the year and the remaining 12 references are excluded based on their title.

P10 is the oldest paper of those included, and hence the following outcome is not so surprising. P10 includes 28 references with 21 of them being perceived as new. Unfortunately, all 21 references left are excluded based on the publication year.

In summary, 97 (28+12+8+9+1+18+21) references were examined after having removed those that had already been investigated. A large majority of the 97 references were excluded based on publication year or title. For only three references there was a need to look in the paper for the place and context of the reference. In all other cases, the references could be excluded based on either publication year or the title of the reference.

4.4.2 Forward Snowballing

The next step is to examine the seven papers from a forward snowballing point of view. This includes examining the citation to the seven papers within the studied time frame (1995-2005). The outcome of the examination of these seven papers is as follows:

62 candidates for inclusion cite **P4**. 11 of these have already been examined. 40 candidates are excluded based on the information available in Google Scholar, which includes publication year, type of publication, language and title. This leaves 11 candidates. The abstract is examined for these 11 candidates, and it is concluded to exclude 10 of them. One paper is viewed as a candidate for inclusion, and it is decided to include the paper after having investigated the full paper. The new paper is:

P11. E. Mendes, N. Mosley and S. Counsell, "Early Web Size Measures and Effort Prediction for Web Costimation", Proceedings Ninth International Software Metrics Symposium, pp. 18-39, 2003.

P5 has 74 relevant citations in Google Scholar. 37 of these citations have already been examined and 36 of the citations were excluded based on the information available in Google Scholar. Thus, only one abstract was investigated and it was decided to exclude the paper.

36 citations are found for **P6**. A majority of these have already been examined (26 candidates citing P6), and the other 10 candidates for inclusion are excluded based on the information available in Google Scholar.

P7 has 47 citations in Google Scholar, and 31 of them have already been examined. The remaining 16 candidates are excluded based on the information available in Google Scholar.

12 candidates for inclusion cite **P8**. 10 of these citations have already been examined and the other two citations can be excluded based on the information provided in Google Scholar.

Only three citations are identified for **P9**. Two of them have already been examined and the third can be excluded based on information available in Google Scholar.

Finally, **P10** has 17 citations in Google Scholar. Eight of these citations have already been examined and the other nine citations can be excluded based on the information available in Google Scholar.

In summary, 126 citations have been examined for these seven papers (51+37+10+16+2+1+9) after having removed those that have already been examined. Most of the papers were excluded based on information available in Google Scholar and only 12 abstracts were read to make the decisions. At the end, only one paper was included based on the forward snowballing in Iteration 2.

4.5 Iteration 3

Given that Iteration 2 only identified one paper, the backward and forward snowballing become easy in the third iteration. In the backward snowballing, it is noted that **P11** has 39 references. Four of them are excluded based on the publication year, and 13 references have already been examined. The remaining 22 references are excluded based on the title. Moving on to forward snowballing, P11 is cited by seven candidates for inclusion. All seven of them are excluded based on information available in Google Scholar.

In summary, the second iteration resulted in examining 33 (4+22+7) additional candidates for inclusion.

4.6 Efficiency

One important efficiency measure for systematic literature studies is the number of included papers in relation to the total number of candidate papers examined. It is well known that there is a large risk for noise, i.e. papers that preferably should never have been examined since they were not included at the end. If looking at the efficiency in the different steps:

Number of investigated papers:

- Start set: 13 candidates for start set and 3 papers were included, i.e. efficiency = 3/13 =23%.
- Iteration 1: 25 candidates from snowballing from start set, and 7 papers were included, i.e. efficiency = 7/25 = 28%.
- Iteration 2: 223 candidates for inclusion were generated in backward and forward snowballing, only one paper was included, i.e. efficiency 1/223 = 0.4%.
- Iteration 3: 33 candidates were examined and no paper was included, and hence the efficiency becomes 0%.

The overall efficiency becomes (3+7+1+0)/(13+25+223+33) = 3.7%. It is important to note that the efficiency is calculated on all candidates. If removing those in backward snowballing where the decision is taken either on publication year (trivial) or title in reference list the efficiency increases. The papers in forward snowballing are handled in the same way, due to that they require either that the information in Google Scholar is read or going to the actual paper. Then the overall efficiency increases substantially. In the backward and forward snowballing from the three papers in the start set, the abstract was examined for 12 papers and five of these were included. Here, four full papers were read that were not included. In Iteration 2, 12 abstracts were read and one paper was included. The total efficiency with this calculation becomes:

Start set:

- 3 of 13 from Start set (as before)

Backward:

- 7 of 12 from Iteration 1 (instead of 7 of 25)
- 0 of 3 from Iteration 2 (instead of 3 of 97)
- 0 of 0 from Iteration 3 (instead of 0 of 26)

Forward:

- 1 of 133 (0+126+7) from the three different sets after the start set.

If removing the candidates where they were removed either on publication year in the reference list or on title, the efficiency becomes: (10+1)/(28+133) = 6.8% with the backward snowballing being very effective.

However, it must be noted that it is a delicate balance to not be too restrictive on titles. If being very restrictive on titles, the following papers would not have been included: P4, P5, P7, P9 and P11.

4.7 Authors

If following the snowballing procedure as described in Section 3 and illustrated in Figure 1, authors of included papers should be contacted. Thus, snowballing should not only be on papers, but also on authors. This has not been done since several of the

authors of the included papers also authored the systematic literature review being replicated here. Furthermore, any responses from the authors may be biased by already having seen the published systematic literature review. Otherwise, the following authors would have been candidates to contact:

- Authored more than one paper: Briand, Jeffrey, Kitchenham, Maxwell, Mendes, Ruhe and Wieczorek.
- Authored one paper: Counsell, Dutta, El Emam, Harrison, Lefley, Lokan, Mosely, Shepperd, Surmann, Triggs and Wassenhove.

Other complementary searches were not done either since the objective is to compare snowballing with a database-first-approach in a similar way as in [19], although with a better specified procedure for using snowballing.

4.8 Citation Matrix and Timeline

To understand the referencing and citing between the papers, a citation matrix is created for illustration purposes. Table 1 shows how the eleven papers refer to each other, denoted with "X". For example, it can be seen how P2 refer to P4-P8 (row 2), and how P10 is cited by P3, P4, P6 and P9 (column 10). For reason of space in the table, the paper numbers are given without the preceding "P".

In Table 1, it can be seen how P10 does not refer to any of the other papers, this is no surprise if looking at the timeline of the publications. The timeline is as follows:

1999: 10 and 4; 2000: 5 and 6; 2001: 7; 2002: 8;

2003: 9 and 11; 2004: 2 and 1 and 2005: 3

The timeline means that Table 1 can be complemented with information about possibility to cite. Table 1 is complemented with this information by introducing "-" when a paper cannot be referenced due to it not being published yet. For example, for P10 the row is filled with "-", since P10 could not refer to any of the other papers, and the column for P10 is left with empty cells since all other papers could have cited this paper given that it was published first. This may not be entirely true for two reasons: 1) it takes time for a paper to be published so although it looks like it should be available it may not have been at the point in time when another paper was written, and 2) authors are aware of their own papers and can cite them even if they are not officially published yet (as P2 is cited by P11). Independently, Table 1 provides some additional information by introducing "-" as a sign for most likely not being able to cite another paper.

A closer look at Table 1 provides some interesting observations:

- The timeline together with Table 1 show that the three papers from the start set (P1-P3) are relatively new in the studied time frame. Thus, most other papers are found through backward snowballing.
- P10 may not have received the citations it deserves despite it being the first study published, only four out of ten papers cite it.
- P11 is not cited by any of the other papers, despite three papers being published after its publication. This is surprising in particular since P1-P3 have one author in common with P11.

It is worth noting that finding one of the eleven papers in Table 1 means that the other papers can be found with snowballing. It does not matter which paper is identified; the other papers will be found. This illustrates one of the strengths with snowballing, i.e. papers may be found even if they use different terminology but the authors within the area refer to each other despite these differences.

Table 1. Citation matrix.

	Cited										
Ref.	1	2	3	4	5	6	7	8	9	10	11
1		X	-	X	X	X	X	X			
2	-		-	X	X	X	X	X			
3	X	X		X	X	X	X	X	X	X	-
4	-	-	-		-	-	-	-	-	X	-
5	-	-	-	X		-	-	-	-		-
6	-	-	-	X	X		-	-	-	X	-
7	-	-	-	X	X	X		-	-		-
8	-	-	-	X	X	X	X		-		-
9	-	-	-				X			X	-
10	-	-	-	-	-	-	-	-	-		-
11	-	X	-	X	X	X	X	X			

4.9 Reflections

4.9.1 Ten Lessons Learned from Snowballing

The actual use of the snowballing procedure presented in Section 3 comes with ten lessons learned:

1. Direct exclusion in reference lists are very quickly done when it comes to basic criteria such as, for example, language, publication year and type of publication. The extra work from these is negligible.
2. A large number of references are found in several papers. It is particularly obvious when it comes to papers by the same authors. This has some implication, either it means that a good portion of the papers in the area has been captured or a cluster of papers has been found (e.g. overlap in authors), and other clusters may exist. This comes back to the necessity to identify a good start set, i.e. to avoid bias.
3. It is very difficult to decide whether or not to exclude a paper or evaluate it on the next level, for example, to exclude a paper based on its title, or read the abstract or even read the full paper. This is an important balance. On the one hand, it generates a lot of work to read the full text of papers that are excluded, but on the other hand, it is important to not exclude papers early that actually should be included.
4. In backward snowballing, it is recommended to iterate between the reference list and the place and context for the reference in the paper. Once a paper is found for inclusion, then look at other papers referred to in a similar way. They may be strong candidates for inclusion too.
5. In forward snowballing, for papers included, look where the paper leading to the new paper is referenced and identify papers referenced in a similar way. This is easily missed in backward snowballing since the paper leading to the new paper (through forward snowballing) is already found and hence it is typically not looked at in the backward snowballing.

6. The papers to be examined for a specific paper may very depending on the order in which the papers are investigated. In other words, a paper is examined the first time it is found, which means that the number of papers to be investigated varies. However, the number of papers in each step in the snowballing procedure should be stable.
7. The frequency of papers identified in each step (Start set, Iteration 1, Iteration 2 and so forth) should be tracked. If having a good start set, the number of new papers to be included ought to decrease for each step.
8. If the frequency of detecting new papers is not decreasing substantially consider doing a new search with synonyms to the words used from the research question, since it may indicate that a cluster of papers has been missed, see also item 2.
9. It is recommended to create a citation matrix, since it provides information about citation patterns. A citation matrix with many blank cells is an indication that other papers may have been missed. This analysis does not help if an independent cluster of papers exists.
10. A timeline helps in establishing possibilities for citations and it also gives some indications on the activities in relation to the area studied in the systematic literature study.

In summary, the replication was straightforward, and the actual case was very suited for snowballing given that finding one of the papers finally included papers meant that all the other papers would be found through snowballing.

4.9.2 Validity Threats

The main threat is that the researcher read the original study before deciding to conduct the replication. However, this is very hard to avoid since the main reason to replicate the study was driven by having read it and being convinced that it would be an interesting case to both illustrate and evaluate the snowballing approach to conducting systematic literature studies.

Having said this, the replication was run 6-12 months after reading the paper, which means that the researcher does not remember all details of the original study. However, the researcher did remember approximately the expected number of papers to find. The researcher did not remember the exact number, but remembered that the number of papers found in the original studies were in the interval 10-19 papers. This may have affected the decisions on inclusion and exclusion. Independently, if it has affected the outcome it has only affected where the decisions are taken and not the actual number of papers examined in the replication.

4.10 Comparison

The same papers are identified. Nine studies are in common with both previous systematic literature reviews presented in [11]. In [11], it is noted that one study should have been excluded due to the type of analysis conducted. The analysis was not conducted to this detail here, since the main objective is to evaluate the snowballing procedure. However, it is interesting to note that the paper to exclude is P11, which somewhat surprisingly was not cited by P1-P3. Thus, the citation matrix may indicate some issues that need a more detailed investigation.

It is difficult to compare efficiency numbers. What does it mean to look at a paper in the reference list? Should papers denoted "retrieved" in [11] be compared with all papers in the reference list, or should those being removed based on publication year, non-peer reviewed papers or papers that have already been included not be counted? The papers denoted "Detailed reviewed" in [11] are compared with those where either place of reference, abstract or full paper were evaluated here. This gives the results in Table 2.

Table 2. Efficiency comparison.

	Review 1 [11]	Review 2 [11]	Snowballing
Detailed	24	38	38

If using the papers studied in detail as a efficiency measure, then the different approaches requires about the same effort. Unfortunately, the actual effort for conducting the different reviews is not available for comparison. However, further studies are needed to better understand the advantages and disadvantages with different approaches and how they may complement each other in the best possible way. One of the main benefits with snowballing is its focus on papers actually referenced or papers citing papers included, and hence there is a possibility that the noise is less than using a database approach.

5. DISCUSSION SNOWBALLING

One of the main advantages of snowballing is that it starts from relevant papers and then uses these to drive the further study. Reference lists are quite easily examined and when combined with the place and context of the reference, it becomes in most cases quite straightforward to identify relevant papers. The citation analysis may result in examining a large number of papers (when a paper is highly cited), but the information in Google Scholar is in most cases quite helpful to make a decision about tentative inclusion or exclusion.

Snowballing should not necessarily be seen as an alternative to database searches. Different approaches to identifying relevant literature should preferably used to ensure the best possible coverage of the literature. Future research is needed when it comes to several areas: 1) Identification of a good start set of papers for snowballing, 2) Evaluation of the efficiency for different approaches to systematic literature searches, and 3) Determination of advantages and disadvantages of different approach, in particular in different type of literature searches (e.g. broad area or very focused area), and 4) Formulation of a good hybrid approach where different approaches to identifying the relevant research literature complement each other.

In particular, it should be noted that snowballing is particularly useful for extending a systematic literature study, since new studies almost certainly must cite at least one paper among the previously relevant studies or the systematic study already conducted in the area. Thus, snowballing is by deduction a better approach than a database search for extending systematic literature studies. The actual evidence for this assertion is left for further research.

6. SUMMARY

The two objectives stated in Section 1 are both fulfilled. A procedure for snowballing has been formulated, and it has successfully been illustrated and evaluated. The snowballing procedure is detailed in several steps including both backward and forward snowballing. Ten lessons learned from using the snowballing procedure have been reported, which hopefully will

help others using snowballing in their systematic literature studies.

The replication illustrated the usefulness of the snowballing procedure, and the actual outcome from the replication was similar to the outcome of the original systematic literature reviews. The snowballing procedure was particularly suitable for this case, since it turned out that it was sufficient to find one of the 10-11 papers to be able to find the other papers.

To conclude, a systematic approach to snowballing as the procedure formulated here is definitively an alternative to use as a starting point for a systematic literature study instead of always start by searching different databases. The next challenge is to figure out under which circumstances the snowballing procedure is to prefer over the database approach.

Finally, it should be noted that a key to success for using the snowballing procedure for systematic literature studies is the place and the context of the references in both backward and forward snowballing. In addition, a citation matrix and a timeline have been proposed to get a better overview of papers in the area of the systematic literature study.

7. ACKNOWLEDGMENTS

The author wishes to express his sincere thanks to Samireh Jalali, Deepika Badampudi and Rafael Prikladnicki for inspiring and valuable discussion regarding systematic literature studies in general and the use of snowballing in particular.

The Industrial Excellence Center EASE – Embedded Applications Software Engineering, (http://ease.cs.lth.se) partially funded this research.

8. REFERENCES

[1] Kitchenham, B. A., Dybå, T. and Jørgensen, M. 2004. Evidence-based software engineering. *In Proceedings of 27th IEEE International Software Engineering Conference*, 273-281, IEEE Computer Society, 2004.

[2] Webster, J. and Watson, R. T. 2002. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly* 26, 2, xiii-xxiii.

[3] Pickard, L., Kitchenham B. A. and Jones P. 1998. Combining empirical results in software engineering. *Information & Software Technology* 40, 14, 811–821.

[4] Miller, J. 1999. Can results from software engineering experiments be safely combined?. In *Proceedings IEEE 6th International Symposium on Software Metrics*, 152-158.

[5] Hayes, W. 1999. Research synthesis in software engineering: A case for meta-analysis. In *Proceedings 6th IEEE International Software Metrics Symposium*, 143–151.

[6] Basili V. R., Shull F. and Lanubile F. 1999. Building knowledge through families of experiments. *IEEE Transactions on Software Engineering* 25, 4, 456–473.

[7] Kitchenham B. A. and Charters S. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Version 2.3, EBSE Technical Report, EBSE-2007-01, Keele University.

[8] Petersen K., Feldt R., Mujtaba S. and Mattsson M. 2008. Systematic mapping studies in software engineering. In *Proceedings 12th International Conference on Evaluation and Assessment in Software Engineering*.

[9] Kitchenham B. A., Budgen, D. and Brereton, O. P. 2011. Using mapping studies as the basis for further research – a participant-observer case study. *Information and Software Technology* 53, 6, 638-651.

[10] Wohlin C. and Prikladnicki R. 2013. Systematic literature reviews in software engineering. *Information and Software Technology* 55, 6, 919-920.

[11] MacDonell, S., Shepperd, M., Kitchenham, B. A. and Mendes, E. 2010. How reliable are systematic reviews in empirical software engineering?. *IEEE Transactions on Software Engineering* 36, 5, 676–687.

[12] Wohlin, C., Runeson, P., da Mota Silveira Neto, P. A., Engström, E., do Carmo Machado, I. and de Almeida, E. S. 2013. On the reliability of mapping studies in software engineering. *Journal of Systems and Software* 86, 10, 2594-2610.

[13] Kitchenham, B. A., Brereton, O. P., Budgen, D., Turner, M., Bailey, J. and Linkman, S. 2009. Systematic literature reviews in software engineering: A systematic literature review. *Information and Software Technology* 51, 1, 7-15.

[14] Kitchenham, B. A., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M. and Linkman, S. 2010. Systematic literature reviews in software engineering: A tertiary study. *Information and Software Technology* 52, 8, 792-805.

[15] Greenhalgh T. and Peacock, R. 2005. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ* 331, 7524, 1064–1065.

[16] Skoglund, M. and Runeson, P. 2009. Reference-based search strategies in systematic reviews. In *Proceedings 13th Evaluation and Assessment in Software Engineering*, 31-40.

[17] Kitchenham, B. A., Brereton, O. P., Li, Z., Budgen, D. and Burn, A. 2011. Repeatability of systematic literature reviews. In *Proceedings of the 15th International Conference on Evaluation and Assessment in Software Engineering*, 46–55.

[18] Zhang, H., Babar, M. A. and Tell, P. 2011. Identifying relevant studies in software engineering. *Information and Software Technology* 53, 6, 625-637.

[19] Jalali, S. and Wohlin, C. 2012. Systematic literature studies: Database searches vs. backward snowballing. In *Proceedings 6th International Symposium on Empirical Software Engineering and Measurement*, 29-38.

[20] Jalali, S. and Wohlin, C. 2012. Global software engineering and agile practices: A systematic review. *Journal of Software: Evolution and Process* 24, 6, 643-659.

[21] Smite, D., Wohlin, C., Galvina, Z. and Prikladnicki, R. 2012. An empirically based terminology and taxonomy for global software engineering. *Empirical Software Engineering: An International Journal* 19, 1, 105-153.