The Impact of Limited Search Procedures for Systematic Literature Reviews
– A Participant-Observer Case Study

Barbara Kitchenham¹, Pearl Brereton¹, Mark Turner¹, Mahmood Niazi¹, Stephen Linkman¹, Rialette Pretorius² and David Budgen²

¹School of Computing and Mathematics, Keele University, Stoke-on-Trent ST5 5BG, UK
²Dept Computer Science, Durham University, Durham, DH1 3LE, UK
b.a.kitchenham@cs.keele.ac.uk

Abstract

This study aims to compare the use of targeted manual searches with broad automated searches, and to assess the importance of grey literature and breadth of search on the outcomes of SLRs. We used a participant-observer multi-case embedded case study. Our two cases were a tertiary study of systematic literature reviews published between January 2004 and June 2007 based on a manual search of selected journals and conferences and a replication of that study based on a broad automated search. Broad searches find more papers than restricted searches, but the papers may be of poor quality. Researchers undertaking SLRs may be justified in using targeted manual searches if they intend to omit low quality papers; if publication bias is not an issue; or if they are assessing research trends in research methodologies.

1. Introduction

We are currently undertaking a program of case-study based research aimed at better understanding the role of systematic literature reviews (SLRs) in software engineering [1]. This is part of the Evidence-based Practices Informing Computing (EPIC) project which is funded by the UK Engineering and Physical Sciences Research Council. SLRs are a type of secondary study (i.e. a study based on previously published research) used to find, critically evaluate, and aggregate all relevant research papers (referred to as primary studies) on a specific research question or research topic. The method was developed in medicine and has been adopted by many other disciplines including social sciences, economics, and nursing. The method is intended to ensure that the review is unbiased, rigorous and auditable. The basic methodology is similar whatever the discipline using it although medical standards emphasize meta-analysis (a means of statistically aggregating the results from different studies of the same phenomena) more than other disciplines (see for example, [2], [3], [4]).

We are using the participant-observer case study approach as our main research methodology for investigating software engineering SLRs. The cases in each case study are SLRs performed by the EPIC research group who act as SLR participants as well as case study researchers (see Figure 1).

Figure 1 EPIC Case study methodology

This case study reports the progress of an SLR aimed at extending an existing tertiary study (i.e. a systematic review where the review is based on aggregating secondary studies) that surveyed SLRs in the time period 1st Jan 2004 to 30th June 2007. The original tertiary study restricted its search process to a set of 13 journals and conferences [5]. The case directly observed in this case study extends the search to a large number of digital libraries.

Medical guidelines for performing SLRs recommend broad search procedures including automated searches and efforts to identify any relevant grey literature [2]. However, SE researchers have taken somewhat different approaches to the SLR search process in different published SLRs. For example, some researchers have restricted their searches to specific digital libraries [6] or a
specific set of journals and conference proceedings [7]. Other researchers have strongly advocated the use of manual searches as opposed to automated searches [8].

We believe it is important to assess the impact of different procedures, in order to improve the advice given to researchers in our own SLR guidelines [3]. This case study, which is based on extending an existing SLR using a broader search process, gives us the opportunity to investigate the following research questions (note the research question numbering system refers to the set of research questions addressed by our full research program [1]):

- RQ3 (Breadth of Literature Search): To what extent is the adoption of an extended search space vital for answering detailed research questions?
- RQ2 (Importance of Grey Literature): To what extent is the grey literature necessary for SE SLRs?
- RQ6 (Manual versus Automated Search Strategies): Are automated search strategies preferable to manual search strategies in the SE domain?

We describe our research methodology in Section 2, present our results in Section 3 and discuss our results in Section 4.

2. Methodology

We investigated our research questions using participant-observer case studies based primarily on Yin’s methodology [9]. A participant-observer case study has several advantages:

- There is no problem about access to information about the case.
- There is no requirement to liaise with external organisations or individuals.

However, there is the possibility of bias if we are investigating an issue in which we have a vested interest. With respect to vested interests, we are very much in favour of the use of systematic literature reviews and our research questions are based on the assumption that SLRs are useful. We merely seek to determine the most appropriate procedures for performing SLRs not to investigate their value as a scientific methodology, so our personal bias will have limited impact in this case study. In addition, we have specified our methodology clearly in a protocol prior to undertaking the case study [10]. This methodology is described below.

2.1. The case and basic design

The “case” in this study is a tertiary SLR. It is also a mapping study. A mapping study is a form of systematic review that asks general questions about research in a topic area (e.g. what do we know about topic x) rather than specific questions about research outcomes (e.g. is method a better than method b). The original tertiary mapping study [5] restricted its search process to a manual search of a set of 13 journals and conferences in the time period 1st Jan 2004 to 30th June 2007. The selected sources included those used by [7]. We replicated the original tertiary study using a broad automated search process searching both digital libraries and indexing systems. Since we have a baseline “case” with which to compare the results of the broad search, our case study design can be categorised as a multi-case case study. Furthermore since we investigated specific SLR tasks (i.e. searching and selection), we regard this as an embedded case study (i.e. a case study that investigates various sub-elements of the case).

2.2. Case Study Propositions

The Research Questions and their related propositions (using Yin’s terminology [9]) are shown below. Propositions in case studies play a similar role to hypotheses in formal experiments.

- RQ3: Breadth of literature search has two propositions:
  - P3.1: A broad search will identify more relevant primary studies than a restricted search.
  - P3.2: The additional primary studies found by a broad search will change the conclusions of the study.

- RQ2: The importance of grey literature has two propositions:
  - P2.1: Primary studies not published in journals or conference proceedings are of equivalent quality to other primary studies.
  - P2.2: Additional primary studies not published in journals or conference proceedings will change the conclusions of the study even if low quality studies are excluded.

- RQ6: Manual versus Automated searches has two propositions:
  - P6.1: Automated searches will find more relevant primary studies than manual searches.
  - P6.2: Automated searches require less effort than manual searches.

2.3. Case study roles

The EPIC team conducted the case study and the SLR (see Figure 1). We assigned roles as follows:

- SLR Supervisor: David Budgen (DB)
- SLR Research team including a Research Assistant (Riallette Pretorius) responsible for most of the SLR activities with support from Pearl Brereton (PB), Barbara Kitchenham (BAK), Stephen Linkman (SL), Mark Turner (MT), Mahmood Niazi (MN)
• Case Study leader (BAK)
• Case Study Team (DB, PB, MT, SL, MN).

The SLR supervisor was responsible for organizing the SLR and supervising the RA. He also ensured that the RA collected information about the SLR process required for the case study.

The case study team leader was responsible for constructing the case study protocol [10] and the SLR protocol [12]. Members of the case study team including the case study leader provided research support for the SLR process (i.e. assisting as required with primary study identification, quality data extraction and SLR data extraction). The data extraction and preliminary selection process for the SLR took much longer than expected, so BAK took over the organization of the SLR after the RA’s internship finished. The RA completed the initial search process, and initial screening of papers to remove obviously inappropriate studies. BAK organized the subsequent SLR processes.

2.4. SLR methodology used for the broad search

The SLR used in this case study replicated the original SLR, in terms of research questions, but extended the search space by undertaking an automated search of four digital libraries and two broad indexing services: IEEE Digital Library; ACM; Citeseer; SpringerLink; SCOPUS; Web of Science. We used the standard SLR methodology [3] for searching and data extraction.

The search strings used for all sources except SCOPUS were a set of 15 simple strings. 14 strings were of the type “Software engineering” AND “string” where the value of the “string” was one of “review of studies”, “structured review”, “systematic review”, “literature review”, “literature analysis”, “in-depth survey”, “literature survey”, “meta analysis”, “past studies”, “subject matter expert”, “analysis of research”, “empirical body of knowledge”, “overview of existing research”, “body of published research”, The final string was “Evidence-based software engineering” OR “evidence based software engineering”.

The SCOPUS search compressed the set of strings into two more complex queries. Searching the other sources involved applying each search string individually and aggregating the output. The search strings were derived by using methodology-related terms found in the SLRs in the original study. The aim of the search was to identify a set of search strings that would find as many of those SLRs as possible. The set of papers found in the original tertiary study acted as a set of known papers against which the results of the automated search could be assessed.

The automated search found 1755 candidate papers in the time period 1st Jan 2004 to 30th June 2008 (i.e. one year more than the original study). After removals of duplicate papers and irrelevant papers and papers found in the original search, the RA reduced this set to 161 papers. For each article, one researcher from the set of five researchers excluding BAK and the RA, was assigned at random to each article, BAK was assigned to every paper. The researchers then screened each assigned paper to identify papers that could be rejected based on not being in English, not being a literature study (i.e. being another form of “survey”), not being a software engineering topic, or being a duplicate paper. This screening process considered only the abstract and title and emphasized including any paper we were not sure about. Any disagreements were discussed and resolved leaving 119 papers. These papers were subject to a second screening based on the full paper and using the same process. Our inclusion criteria were:

- That there was a full paper (not a PowerPoint presentation or extended abstract)
- That the paper included a literature review where papers were included based on a defined search process.
- The paper should be related to software engineering rather than IS or computer science.

We finally agreed on 39 papers. We added one more paper which was referenced by a rejected paper. 14 of the papers were published in the same time period as the original tertiary study [5]. These 14 papers are the basis of this study. Note we did not undertake any data extraction of the remaining 26 papers until after this case study was completed.

The selection criteria used in the original tertiary study were stricter but the selection process was less rigorous. Papers were included if they were systematic (i.e. had defined research questions, a search process, and a data extraction and analysis process) or were a meta-analysis. Papers were excluded if they were informal reviews, duplicate reports of the same study, or related to EBSE or SLR methodology. The decision to include was made by a single researcher but if the researcher was unsure about excluding a paper another researcher was consulted.

Quality evaluation for both tertiary studies used the Centre for Reviews and Dissemination DARE criteria [11] which are based on four questions:

- Q1: Are the review’s inclusion and exclusion criteria described and appropriate?
- Q2: Is the literature search likely to have covered all relevant studies?
- Q3: Did the reviewers assess the quality/validity of the included studies?
- Q4: Were the basic data/studies adequately described?

We scored each question on a scale of Yes (1), No(0), Partly (0.5) and summed the scores. The DARE criteria concern the rigour of a systematic review in terms of the extent to which it is repeatable (Q1), complete (Q2), able
to deliver trustworthy conclusions (Q2 and Q3) and auditable (Q4).

The process used to extract the quality information and classification information from each additional primary study was as follows:

- Each of the 5 researchers from the case study team (excluding BAK and Pretorius) was assigned at random to either 4 or 5 of the papers, such that each paper was reviewed by two different researchers.
- BAK extracted the data from every paper.
- The data for each paper were aggregated using the median value for quality scores and the modal values for categories.

The aggregated data was reviewed and agreed by all extractors. In the original tertiary study, BAK extracted the quality data for each papers and the extraction was checked by another researcher.

2.5. Case Limitations

We note that our “case” is not a typical software engineering SLR:

- It is a tertiary study, not a conventional secondary study.
- The subject of the study is the SLR methodology not a software technology.
- It is a mapping study, not a conventional SLR looking at detailed research questions(s).
- It is a study where, due to the topic, relatively few additional primary studies are expected.

In addition, the case only compares broadening the search using an automated search of six digital sources, with manual search of journal and conference proceedings that are a subset of articles referenced by three electronic sources (ACM, IEEE and SCOPUS). This differs from the broad manual search proposed by Jørgensen and Shepperd [7]. The results of this case study must, therefore, be interpreted carefully in the light of the specific case.

2.6. Case study data collection

The RA was given a copy of the case study protocol to ensure that she was aware of data collection requirements placed on her by the demands of the case study.

To address the case study research questions the following data were collected:

- The number of new primary studies identified by the automated search. This was collected by the RA as part of the search process.
- The type of new primary studies (i.e. journal papers, conference papers, book chapters, workshop papers, technical reports). This information was collected by BAK as part of the SLR data extraction process.
- Information about each change to the results and conclusions of the study due to the additional literature. This information was collected by BAK as part of the study aggregation and reporting process.
- Time taken to complete SLR tasks. This was collected by the RA and SLR Team as part of the agreed SLR process.

The data analysis and interpretation procedures for the case study were specified in advance in our protocol, but managed both to be too simplistic (at the detailed propositions level) and too complex (at the interpretation level) to cope with the data we collected. In particular, counting the number of primary studies found by each search strategy was more complicated than we had expected, so simple interpretations based on the percentage of studies missed was inappropriate. We also decided not to attempt to evaluate the importance of changed results since such an evaluation is extremely subjective. In addition, the impact of poor quality studies affected the search strategy propositions as well as the grey literature propositions.

3. Results

We present our case study results and analysis in this section. The data extracted from each additional paper are summarized in Table 1. In Table 1, the “Article” column indicates whether the study was an SLR or a mapping study (MS). The “EBSE” column identifies whether the paper referenced either of the Evidence-Based Software Engineering papers ([27], [28]) or either of the SLR Guidelines technical reports ([3], [29]). We refer to such papers as “EBSE-related” papers. The “PG” column indicates whether the study recommended practitioner guidelines. The Total Quality score is the value obtained using the DARE criteria [11]. “Survey type” indicates whether the study addressed a specific research question (RQ) or was concerned with general research trends (RT). For SLRs the survey type is usually RQ and for mapping studies it is usually RT but there can be exceptions.
Table 1 Data Extraction for Additional Papers

<table>
<thead>
<tr>
<th>Study ref</th>
<th>Article</th>
<th>Total Quality Score</th>
<th>Year</th>
<th>EBSE</th>
<th>Type</th>
<th>Number of Primary Studies</th>
<th>PG</th>
<th>Survey type</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>[15]</td>
<td>SLR</td>
<td>3.5</td>
<td>2006</td>
<td>Y</td>
<td>Conference</td>
<td>26</td>
<td>Y</td>
<td>RQ</td>
<td>Requirements Engineering</td>
</tr>
<tr>
<td>[16]</td>
<td>MS</td>
<td>1.5</td>
<td>2007</td>
<td>Y</td>
<td>Workshop</td>
<td>653</td>
<td>N</td>
<td>RT</td>
<td>Cost estimation</td>
</tr>
<tr>
<td>[17]</td>
<td>SLR</td>
<td>2.5</td>
<td>2005</td>
<td>Y</td>
<td>Workshop</td>
<td>50</td>
<td>N</td>
<td>RT</td>
<td>Cost estimation</td>
</tr>
<tr>
<td>[18]</td>
<td>MS</td>
<td>1.5</td>
<td>2006</td>
<td>N</td>
<td>Workshop</td>
<td>57</td>
<td>N</td>
<td>RT</td>
<td>Outsourcing</td>
</tr>
<tr>
<td>[19]</td>
<td>MS</td>
<td>1.5</td>
<td>2007</td>
<td>N</td>
<td>Journal</td>
<td>80</td>
<td>N</td>
<td>RT</td>
<td>Mining Software Repositories</td>
</tr>
<tr>
<td>[22]</td>
<td>MS</td>
<td>2.5</td>
<td>2005</td>
<td>N</td>
<td>Conference</td>
<td>105</td>
<td>N</td>
<td>RT</td>
<td>Mobile Systems Development</td>
</tr>
<tr>
<td>[24]</td>
<td>MS</td>
<td>1.5</td>
<td>2007</td>
<td>N</td>
<td>Book Chapter</td>
<td>4089</td>
<td>N</td>
<td>RT</td>
<td>Requirements Engineering</td>
</tr>
<tr>
<td>[26]</td>
<td>MS</td>
<td>2</td>
<td>2007</td>
<td>N</td>
<td>Book Chapter</td>
<td>155</td>
<td>N</td>
<td>RT</td>
<td>Open Source Software</td>
</tr>
</tbody>
</table>

3.1. RQ3 - Breadth of Search

3.1.1. P3.1: Broad searches find more relevant studies.
The comparison of the original restricted search and the extended broad search are shown in Table 2. Although overall the proposition that broad searches find more papers than restricted searches is supported, the results are more complicated than the simple proposition suggests.

The broad search missed 5 studies included in the original SLR. Two of those studies were found by contacting specific researchers, and so were not discovered by the manual search process ([30], [31]). Of the other three papers, one had an embedded review that was borderline for inclusion [32]; one used the term “review” but not “literature review” [33]; and the final paper was a literature review of computer science papers and should probably have been omitted from the initial study [34].

The original manual search missed three papers from its set of 13 sources: one journal paper ([14]) and two conference papers ([13], [21]). One of the papers would have been excluded from the original study because it did not have a defined data extraction process [14]. Nonetheless missing papers suggests that the process of having only one researcher search each source was not as effective as it should have been.

Overall the broad search identified 11 papers that were published in the sources not searched in the original study. However, this is slightly misleading:

- Two of the studies selected ([18] and [23]) would not have been included in the original tertiary, because they did not have a clear data extraction and aggregation process although they did have a defined search process.
- Two of the studies that investigated empirical software engineering ([20] and [25]) considered only one source (i.e. the Empirical Software Engineering journal). This was because previous literature surveys related to empirical studies in software engineering had omitted ESE from their list of sources ([35], [36]). Thus, whether these studies count as ancillary studies or mapping studies in their own right is problematic.

Table 2 Relevant studies found

<table>
<thead>
<tr>
<th>Paper counts</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies used in original tertiary study found by manual search</td>
<td>18</td>
</tr>
<tr>
<td>Studies in original study not found by manual search</td>
<td>2</td>
</tr>
<tr>
<td>Total Studies found by this case</td>
<td>29</td>
</tr>
<tr>
<td>Studies found in both cases</td>
<td>15</td>
</tr>
<tr>
<td>Extra studies found by this case</td>
<td>14</td>
</tr>
<tr>
<td>Studies found by original study but not this case</td>
<td>5</td>
</tr>
<tr>
<td>Studies found by this search that should have been found by original study</td>
<td>3</td>
</tr>
<tr>
<td>Extra studies found by this case but not directly by the broad search</td>
<td>1</td>
</tr>
</tbody>
</table>

3.1.2. P3.2: Additional primary studies will change the conclusions of the study. The results of the original study (i.e. individual trends and observations) are shown in Table 3. In 11 of 17 cases, the original results were confirmed or strengthened. They were contradicted in 6 cases. We have based our comparison on results rather than on conclusions...
because it is easier and more objective to map individual results to one another rather than to try to compare wider conclusions. Furthermore, changed results are the underlying reason for changed conclusions.

### Table 3 Comparisons of Results

<table>
<thead>
<tr>
<th>Research question in original study</th>
<th>Results in original study</th>
<th>Results including all relevant papers</th>
<th>Impact on results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result not related to a specific research question</td>
<td>Quality is improving over time</td>
<td>No relationship between quality and time</td>
<td>Results contradicted</td>
</tr>
<tr>
<td>Result not related to a specific research question</td>
<td>Quality not better for EBSE-related papers.</td>
<td>No relationship between quality and citing EBSE-related papers.</td>
<td>No change</td>
</tr>
<tr>
<td>4.1 How much EBSE activity has there been since 2004</td>
<td>Of 20 studies, 10 cited Guidelines or EBSE paper</td>
<td>Of 34 studies 15 cited Guidelines or EBSE paper</td>
<td>Less EBSE-positioned research</td>
</tr>
<tr>
<td>Stable numbers per year: 2004 (6); 2004 (5); 2006 (6); 2007 (3) where 2007 covers only half year</td>
<td>Probably increase in 2007: 2004 (6); 2005 (11); 2006 (9); 2007 (8)</td>
<td>Recent activity underestimated</td>
<td></td>
</tr>
<tr>
<td>Main publication source: IEEE SW (4 studies), TSE (4), JSS (3), IST (2)</td>
<td>Main publication source changes: TSE (5), IEEE SW (4), JSS (3), IST (2), Metrics (2), ICSE (2)</td>
<td>No major change. Changed counts due to papers missed in original search.</td>
<td></td>
</tr>
<tr>
<td>4.2 What topics are being addressed</td>
<td>Topics addressed at least twice: Cost estimation (7 papers); SE Empirical Methods (4); Testing (3)</td>
<td>Cost estimation papers (11 papers); SE Empirical Methods (7); Testing (4); Requirement Engineering (2); Architecture (2)</td>
<td>Results changed. More work on conventional SE topics</td>
</tr>
<tr>
<td>4.3 Who is leading research</td>
<td>Studies per person: Jørgensen (5);</td>
<td>Studies per person: Jørgensen (8);</td>
<td>No change</td>
</tr>
<tr>
<td>After Jørgensen, Sjöberg (3);</td>
<td>After Jørgensen, Shepperd (4);</td>
<td>Results change</td>
<td></td>
</tr>
<tr>
<td>Most activity organization: Simula Laboratory (8)</td>
<td>Most activity organization: Simula Laboratory (11)</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>Most studies have European authors (14 of 20)</td>
<td>Most studies have European authors (26 of 34)</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>Few studies have US authors (3 of 20)</td>
<td>Few studies have US authors (7 of 34)</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>4.4 Current Limitations of SLRs</td>
<td>Many research trends papers (8 of 20)</td>
<td>Many research trends papers (19 of 34)</td>
<td>Original result strengthened</td>
</tr>
<tr>
<td>Limited number of conventional SE topics addressed</td>
<td>Requirements and software architecture addressed</td>
<td>More work on conventional SE topics, but general conclusion holds</td>
<td></td>
</tr>
<tr>
<td>Sample sizes for conventional SLRs relatively small (range 6-54) compared with mapping studies (range 63-1485)</td>
<td>Sample sizes for conventional SLRs relatively small median (21) compared with mapping studies (median 105)</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>Quality relatively good: Only 3 studies scored less than 2 on the DARE scale</td>
<td>10 of 34 papers scored less than 2 on the DARE scale</td>
<td>Original results contradicted</td>
<td></td>
</tr>
<tr>
<td>Few studies addressed the quality of primary studies. 3 fully, 5 partially, 12 not at all</td>
<td>4 fully, 6 partially, 24 not at all</td>
<td>No change</td>
<td></td>
</tr>
<tr>
<td>Few studies provided practitioner guidelines (4 of 20)</td>
<td>6 of 34</td>
<td>No change</td>
<td></td>
</tr>
</tbody>
</table>

The quality results presented in Table 1 and Table 3 suggested that additional studies found in the broad search were of relatively poor quality. The average quality scores for the studies are shown in Table 4. Using the Mann-Whitney rank sum test, the quality score for the additional 14 studies was significantly less than the quality scores of the initial 20 studies (p<0.001). Excluding the quality scores from the three studies that should have been found in the original search process makes no difference to the results. It is therefore relevant to consider what would have happened if poor quality papers were omitted from the aggregated data. Excluding the two results related to quality (since removing low quality papers renders any results related to overall quality issues invalid) four of the original results were contradicted by the additional studies. Table 5 shows that when the poor quality studies are removed, only the
researcher who was involved in most studies, after Jørgensen, still contradicts the results of the original study.

Table 4 Quality scores for studies

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Studies</th>
<th>Median quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies found in original tertiary study</td>
<td>20</td>
<td>2.5</td>
</tr>
<tr>
<td>Studies found in broad search in sources used in original search</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Studies found in broad search but in sources other than original ones</td>
<td>11</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 5 Effect on results of removing low quality papers

<table>
<thead>
<tr>
<th>Results from original SLR</th>
<th>Evidence from original SLR</th>
<th>Evidence from all studies with a quality score of 2 or more</th>
<th>Impact of new evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable number of papers per year</td>
<td>2004 (3); 2004 (5); 2006 (6); 2007 (3)</td>
<td>2004 (3); 2004 (10); 2006 (7); 2007 (4)</td>
<td>Original result confirmed</td>
</tr>
<tr>
<td>Many studies were evidence-based SE articles</td>
<td>10 of 20 referenced evidence based SE articles or SLR guidelines</td>
<td>13 out of 24 papers referenced evidence based SE articles or SLR guidelines</td>
<td>Original result confirmed</td>
</tr>
<tr>
<td>Topics addressed at least twice</td>
<td>Cost Estimation (7); Empirical SE (4); Testing (2)</td>
<td>Cost Estimation (10); Empirical SE (5); Testing (3)</td>
<td>Original result confirmed</td>
</tr>
<tr>
<td>Other active researchers</td>
<td>Sjøberg (3)</td>
<td>Shepperd (4)</td>
<td>Original result changed</td>
</tr>
</tbody>
</table>

3.2.2 P2.2: Grey literature changes results even if low quality studies are excluded. This is a poorly worded proposition. It would be better expressed as “good quality grey literature changes results”. Only two of the eight grey literature studies found by the broad search had quality scores of 2 or more, so the analysis of grey literature overlaps with that of the impact of removing low quality papers. Thus, in this case, the results of removing low quality grey literature are essentially the same as the results of removing low quality studies.

Table 6 Quality scores for study types

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Studies</th>
<th>Median quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journals in original set of sources</td>
<td>15</td>
<td>2.5</td>
</tr>
<tr>
<td>Conferences in original set of sources</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>Journals not in original set of sources</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>Conferences not in original set of sources</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Workshop studies (including one from original study)</td>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>Book Chapters</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>Technical report</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

3.3 RQ6: Manual versus Automated searches

3.3.1 P6.1: Automated searches will find more relevant primary studies than manual searches. Our study has confirmed that broad automated searches will find more relevant studies than restricted manual searches (see Section 3.1.1). However, our results do not imply that an automated search would necessarily be better than a broad manual search. The problem with a broad manual search is identifying the relevant journals and conferences. It may require searching many different sources to find relatively few additional papers.

3.3.2. P6.2: Automated searches require less effort than manual searches. The original restricted manual search had a search space of 2506 papers. Each journal and conference was searched by a single researcher who identified candidate papers. The candidate papers that the researcher thought were relevant were included, papers that the researcher thought were not relevant (i.e. were informal literature reviews) were passed to a second researcher for a second opinion. We identified 33 candidate papers that included a literature survey and finally agreed on 19 papers (related to 18 separate studies). We excluded papers that included an informal literature review.

We do not have timesheets for the original search, but we estimate that it took about 4 hours to review a specific source and about 15 minutes per paper to look over the 12 disputed papers. This gives a total of 56 hours to perform the search (although we missed 3 papers). The effort for the automated search is itemised in Table 7. Note only the RA
kept detailed timesheets, effort values for the other team members are based on post-hoc estimates.

Table 7 Effort for case study selection and screening papers

<table>
<thead>
<tr>
<th>Activity</th>
<th>Effort (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search String specification and testing (RA)</td>
<td>46.5</td>
</tr>
<tr>
<td>Search &amp; collating papers found by searches (RA)</td>
<td>117</td>
</tr>
<tr>
<td>Initial candidate selection (by RA)</td>
<td>161</td>
</tr>
<tr>
<td>Organising the screening process (including assigning papers to researchers and collating results) (BAK)</td>
<td>13</td>
</tr>
<tr>
<td>Finding papers (BAK, PB, DB)</td>
<td>11.25</td>
</tr>
<tr>
<td>1st screening</td>
<td>14</td>
</tr>
<tr>
<td>2nd screening</td>
<td>8.5</td>
</tr>
<tr>
<td>Total</td>
<td>357.25</td>
</tr>
</tbody>
</table>

Additional costs and time accrued because we were unable to find 12 papers online. Thus to complete our second round of screening we had to obtain the papers via inter-library loans. This took about four elapsed weeks in all (although the time period included Christmas).

Overall, our results suggest that a broad automatic search requires much more effort than a restricted manual search. This result would still hold if the time for the manual search was doubled to allow two researchers to check each source. We discuss the manual and automated search in more detail below to explain this rather unexpected result.

The manual search involved individual researchers looking through all the papers in each of 13 specific journal and conference proceedings published between 1st January 2004 and 30th June 2007. In all case the same researcher reviewed papers published in a specific source. Sources were searched on-line with the exception of IET Software which was searched using the printed journals. Thus, in all cases, it was simple to view the abstract and title and if necessary consult the full version of the paper. Since for each publication the search was a simple sequential task it could be stopped and started at any point without requiring any iteration. Also, since full versions of the papers were accessible, the initial inclusion/exclusion process was integrated with the search process.

In contrast, the automated search required several different stages as follows:

1. The papers used in the original tertiary [5] study were associated with the digital library that indexed the journal/conference in which they appeared.
2. Various different search strings were developed and tested on each library to find the maximum possible number of known studies. This involved many different searches and manually checking all outcomes against the relevant known papers.
3. The set of 15 search strings were applied to each of the six digital libraries and indexing systems.

4. The outcomes of all the individual searches were collated and duplicates removed.

Apart from the actual searches, none of the above tasks were automated. Furthermore, although the comparison with the set of studies found in the original tertiary study was effort intensive, it is a normal method of validating search strings, so should be considered an integral part of an automated search process.

The difficulty collating papers from different digital libraries raises the issue of whether it is better to search individual digital libraries or broad indexing systems. In principle, automated searches of a single broad-scope indexing system such as ISI Web of Science or SCOPUS would reduce the collation problem significantly. However, the SCOPUS search found only 9 of the 20 papers included in the original tertiary study and two of the additional 14 papers found in broad automated search. One problem is that general indexing systems allow searches based on title, abstract and keywords only, whereas the individual digital libraries can base searches on the full paper.

4. Discussion and conclusions

Overall our case study indicates that:

- A broad automated search finds more relevant studies than a restricted manual search.
- Additional papers will cause some results to be revised. In this case, 6 of 17 results were revised.
- Removing poor quality papers may reduce the number of revised results. In this case, three fewer results were revised (i.e. only 1 of 15 non-quality related results).
- Grey literature studies may be of relatively poor quality, so excluding them will be equivalent to excluding low quality papers. The additional 14 papers found by the broad search, included 8 grey literature studies of which only two scored two or more on the DARE quality scale.
- Broad automated searches take more time and effort than restricted manual searches.

The broad search found seven good quality studies that were not detected by the manual search (although two of those studies should have been found by the original search). However, with respect to this case study, the impact of the broad search on the study results (other than completeness) was rather limited once low quality papers were removed.

Overall these results suggest that researchers would be justified in adopting a restricted manual search if they are intending to exclude low quality studies from their results. We note that any restricted search must be targeted to an appropriate set of sources.

Clearly this conclusion has limitations related to the nature of the case used in this case study. Our case study is based on a tertiary study investigating research trends of a general research methodology. In such a study, “publication
bias” (i.e. the problem that papers that find no statistically significant results are less likely to be published than papers that find significant results) is unlikely to be a problem. In contrast, for a conventional SLR looking at a specific research question such as whether one technology is better than another, publication bias is a potential problem. In such a case, the grey literature is likely to be of much more importance. Thus, restricted manual searches are more justifiable for studies of research trends than for studies of competing SE methodologies.

Research trend studies are usually mapping studies. Jørgensen and Shepperd [8] report results from a broad manual search used for a mapping study of a specific software engineering topic (i.e. cost estimation) and warn against restricted searches because important studies may be omitted. Thus, another issue when considering the use of restricted manual searches relates to the importance of completeness. For studies investigating general research trends (e.g. the extent of empirical validation), or research trends related to a research methodology (such as the use of formal experiments) a restricted manual search may be appropriate, but in order to identify all relevant research on a specific SE topic a broad search strategy is likely to be required.

Our original tertiary study found two additional papers by approaching researchers and this study found an extra paper by reviewing the references of an excluded paper. This suggests any basic search strategy aiming at completeness, whether manual or automated, should also include searching primary study references and contacting individual researchers.

Our case study has three major limitations. Firstly, many of our results rely on being able to assess the quality of the primary studies. We used the DARE criteria because they are relatively straightforward (having only four main questions). Nonetheless there are other suggestions for evaluating SLRs (e.g. [37]) and we cannot be sure that the results would be the same if we had used other quality criteria. In addition, there was considerable disagreement among researchers with respect to answering the individual quality questions; there was only one case in which all three researchers assessed each of the four quality questions identically.

Secondly, we cannot make any excessive claims for the completeness of this study. We have excluded non-English papers and made no attempt to look for PhD theses including literature reviews. However, the search process was comparable with the most extensive automated search processes found in our original study of SLRs [5].

Thirdly, there were subtle differences between the original tertiary study and the broad search tertiary study. For example, the inclusion criteria were more stringent in the first study, but other aspects of the search process were more rigorous in the second study, particularly the screening method and data extraction processes. We have pointed out specific issues in Section 2 and believe that the effects on the case study propositions, with the exception of primary study counts, are relatively minor, but when undertaking a participant-observer case study there is always a danger that personal opinions and preferences might cloud our judgment.

We are currently undertaking another case study, replicating a SLR that used a restricted search, using a broader search process. This will provide a replication of this case study using a more representative “case”. In addition, we are also planning a series of studies comparing formal and informal literature reviews.

Acknowledgment

This study was funded by the UK Engineering and Physical Sciences Research Council project EPIC/E046983/1.

References


