EPIC Case Study 2 – Extension of a Tertiary Study

Barbara Kitchenham
O. Pearl Brereton
Mark Turner

EPIC Technical Report
EPIC-2008-07

July 2008

Software Engineering Group
School of Computer Science and Mathematics
Keele University
Keele, Staffs
ST5 5BG, UK

and

Department of Computer Science
University of Durham
Durham, UK

Version 1.3
Abstract

This protocol describes a case study undertaken as one of a series of case studies planned by the EPIC project aimed at investigating different procedures that can be adopted when undertaking a software engineering systematic literature review. In particular, the case study will investigate the value of a broad search process, the importance of grey literature, and the value of an automated rather than a manual search. This will be undertaken as a participant-observer study using a multiple-case embedded case study design.

The protocol outlines the case study roles and responsibilities, the data collection, analysis and interpretation process and the quality assurance process used to ensure that the results are not biased by unexpected events. The main limitations of the case study are that the case is rather atypical of software engineering SLRs and the context of a broad search is limited to that of automated search of many digital libraries.
1. Background

This case study is part of a series of EPIC case studies aimed at investigating the systematic literature review (SLR) process and its adoption by software engineering researchers [1].

The case study will observe the progress of an SLR aimed at extending an existing tertiary study that surveyed the use of 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. The case directly observed in this case study will extend the search to a large number of digital libraries. The extension to the SLR has been requested by reviewers of a journal paper reporting the original SLR.

Medical guidelines for performing SLRs recommend broad search procedures including automated searches and efforts to identify any relevant grey literature ([3], [4]). However, SE researchers have taken somewhat different approaches to the SLR search process in different published SLRs, particularly mapping SLRs. For example, some researchers have restricted their searches to specific digital libraries [7] or a specific set of journals and conference proceedings [9] or have rejected the use of automated searches [6].

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 [8]. This case study, which is based on extending a mapping study that was initially based on a restricted selection of journals and conference proceedings using a broader search process, gives us the opportunity to investigate the following research questions:

- 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?

The case used in this case study is not a typical software engineering SLR. It is:

- A tertiary study.
- The subject of the study is the SLR methodology not a software technology.
- A mapping study.
- A study where, due to the topic, relatively few additional primary studies are expected.

In addition, the case only investigates broadening the search by automated searching of several digital libraries compared with manual search of journal and conference proceedings that are a subset of articles referenced by three digital libraries (ACM, IEEE and Scopus). This differs from the broad manual search proposed by Jørgensen and Shepperd [6]. The results of this case study must, therefore, be interpreted carefully in the light of the specific nature of the case.
2. Design

We regard this case study as a multiple-case case study because although we only plan to observe one SLR, we are using a previous SLR (the original tertiary study) as a baseline to enable us to answer the research questions. We regard this case study as an embedded case study because we will be observing in detail specific SLR tasks (i.e. searching and data analysis).

Each research question leads to several different propositions as specified in Table 1.

Table 1 Research goals and propositions

<table>
<thead>
<tr>
<th>Question</th>
<th>Topic</th>
<th>Proposition Id</th>
<th>Proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ3</td>
<td>Breadth of Literature Search</td>
<td>P3.1</td>
<td>A broad search will identify more relevant primary studies than a restricted search</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P3.2</td>
<td>The additional primary studies found by a broad search will change the conclusions of the study</td>
</tr>
<tr>
<td>RQ2</td>
<td>The importance of grey literature</td>
<td>P2.1</td>
<td>Primary studies not published in journals or conference proceedings are of equivalent quality to other primary studies.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P2.2</td>
<td>Additional primary studies not published in journals or conference proceedings will change the conclusions of the study even if low quality studies are excluded.</td>
</tr>
<tr>
<td>PQ6</td>
<td>Manual versus Automated searches</td>
<td>P6.1</td>
<td>Automated searches will find more relevant primary studies than manual searches</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P6.2</td>
<td>Automated searches require less effort than manual searches</td>
</tr>
</tbody>
</table>

The case study will address the research questions as follows:

- The number of additional primary studies found by this process and the impact of the additional primary studies on the results of the study will address RQ3.
- The type of additional primary studies, the quality of additional primary studies and the impact of studies of different types on the results will address RQ2.
- The number of additional studies and the time taken to complete the research process will address RQ6.

3. Case Study Procedures and Roles

The case study requires the following roles:

- SLR Supervisor: David Budgen (DB)
• SLR Research team including a RA (Rialette Pretorius) responsible for most of the SLR activities and members of the Case Study team: O. Pearl Brereton (OPB), Barbara Kitchenham (BAK), Steve Linkman (SL), Mark Turner MT)
• Case Study leader (BAK)
• Case Study Team (DB, OPB, MT, SL,).

The SLR supervisor will be responsible for organizing the SLR and supervising the RA responsible for performing the SLR. He will be responsible for ensuring that the RA collects information about the SLR process required by the case study.

We see no reason why the RA should be blinded to the goals of this case study, so the RA will be given a copy of the case study document to ensure that she is aware of data collection requirements placed on her by the demands of the case study.

The case study team leader will be responsible for constructing the SLR protocol. Members of the case study team including the case study leader will provide research support for the SLR process (i.e. assisting as required with primary study identification, quality data extraction and SLR data extraction).

The case study team will be responsible for documenting the results of the case study.

4. Data Collection and Analysis

4.1 Data

To address the research questions the following data will be collected:

• The number of new primary studies identified by the automated search. This will be collected by the SLR RA as part of the search process.
• The type of new primary studies (i.e. journal papers, conference papers – in proceedings, book chapters, conference papers – not published, technical reports, other). This information will be collected by the SLR RA 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 determining the nature of the change and which primary studies necessitated the change. This information will be collected by the SLR RA as part of the study aggregation and reporting process.
• Time taken to complete SLR tasks. This will be collected by the SLR RA as part of the SLR procedures. Case study team members undertaking activities to assist the primary study selection process must also record the time spent on those activities.
• The RA will also be asked to keep a log of any problems encountered during the SLR activity and to complete a debriefing questionnaire at the end of each major phase of the SLR (see Appendix 2). This information will be used to assess whether unexpected events have invalidated the planned analyses (See Section 4.6).
4.2 Data Collection
Most data collection will take place as part of the SLR process although the RA and case study team members will also collect effort data using the excel form for effort collection defined in the protocol for the first EPIC case study [2] (see Appendix 2).

The RA will keep a personal log of problems encountered.

4.3 Analysis and Interpretation of Data for RQ3 - Breadth of Search

4.3.1 Proposition P3.1 (Broad searches find more studies) Analysis
In order to assess the importance (or not) of a broad search process we will identify

- The number of primary studies identified in the initial tertiary study.
- The additional number of primary studies identified using the extended search.

The percentage of relevant studies found by the extended search will be reported.

4.3.2 Proposition P3.1 Interpretation
We consider a level of missing primary studies

- Less than 20% to indicate that an extended search may not be important.
- Between 21 and 50% to suggest that an extended search is important.
- 51% or more to suggest that an extended search is critical.

The number of papers found by the initial search was 20, so P3.1 is supported if the number of primary studies is greater than 24, otherwise P3.1 is not supported. Failure to support P3.1 may be due to the nature of the case (i.e. the topic of the case study SLR is not an SE topic it is a research methodology so may not be of general interest to SE researchers).

4.3.3 Proposition P3.2 (More primary studies change conclusions) Analysis
The changes made to the results and conclusions of the original tertiary study that will be assessed as follows:

- Changes do not change the overall trends and conclusions will be rated as “not essential”
- Changes that alter the overall trends and conclusions will be rated as “essential”
- Changes that alter overall trends and conclusions will also be assessed as to their importance (major or minor) depending on whether the case study team assesses the altered trend/conclusion to be a major or minor issue.

4.3.4 Proposition P3.2 Interpretation
If there are no essential changes to overall trends and conclusions, then in this case we can conclude that a broad search was not necessary and P3.2 is not supported. If there were essential changes to overall trends and conclusions then we can conclude that P3.2 is supported.
4.3.5 RQ3 Overall Interpretation

The overall interpretation of the results is shown in Table 2. It shows how the results of the individual propositions lead to an assessment of overall support for the view that a broad search should be performed for small research method-based mapping studies.

Table 2 Interpretation of RQ3 (Impact of Broad Searches)

<table>
<thead>
<tr>
<th>P3.1</th>
<th>P3.2</th>
<th>RQ3</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported</td>
<td>Supported</td>
<td>Supported</td>
<td>Small mapping studies of new research methods should undertake a broad search process. No generalization can be made to mapping studies of new software engineering methods because they are likely to be concentrated in specialized conferences and journals, making a broad search less necessary. Furthermore, results only concern the impact of broadening the search by increasing the number of digital libraries searched. No conclusions can be drawn about the implication of increasing the number of journals searched as suggested by Jørgensen and Shepperd [6].</td>
</tr>
<tr>
<td>Supported</td>
<td>Not Supported</td>
<td>Partly supported</td>
<td>In this case, the broader search confirms existing trends and gives a better assessment of the strength of existing trends, but does not add additional information.</td>
</tr>
<tr>
<td>Not Supported</td>
<td>Supported</td>
<td>Invalid</td>
<td>This case is contradictory and should not occur</td>
</tr>
<tr>
<td>Not Supported</td>
<td>Not Supported</td>
<td>Not Supported</td>
<td>In this case, an automated search was unnecessary. A focused manual search may be preferable to an automated search for small mapping studies.</td>
</tr>
</tbody>
</table>

4.4 Analysis and Interpretation of Data for RQ2 – Importance of Grey Literature

If there are no primary sources other than journal papers and published conference papers, there is nothing that can be learnt about the value of grey literature from this case. Otherwise the two propositions will be investigated as described below.

4.4.1 Proposition P2.1 (Quality of Grey Literature) Analysis

The number, percentage and quality of new primary studies in each major category (journal, conference paper – published proceedings, book chapter, technical report, conference paper- not formally published other) will be tabulated.

When a primary study is reported in two sources or more sources, it will be counted against the most formal publication type using the following ordering
• Journal (most formal)
• Conference paper – published in proceedings
• Book chapter
• Technical report
• Conference/workshop paper – not published in proceedings
• Other (least formal).

The quality of primary studies reported in journals and conferences will be compared with the quality of other primary studies.

4.4.2 Proposition P2.1 Interpretation
Proposition P2.1 will be supported if the quality of primary studies obtained from the grey literature are no worse than the quality of studies reported in journals and conference proceedings.

4.4.3 Proposition P2.2 (Grey Literature change SLR results) Analysis
Any changes made to the original SLR conclusions identified in the analysis of P2.1 will be tracked to the primary studies that prompted the change. The type of primary study and the quality of the primary study will be tabulated.

4.4.4 Proposition P2.2 Interpretation
P2.2 will be supported if any essential change was caused only by inclusion of grey literature studies of which at least one is high quality. P2 will be partly supported if any essential change was caused by results from primary studies from both high quality grey literature and journals or conference proceedings.

4.4.5 Overall Interpretation
The overall interpretation of RQ2 is shown in Table 3. It shows how the results of the individual propositions lead to an assessment of overall support for the view that primary studies should be sought from the grey literature.

Table 3 Interpretation of RQ2 (Importance of Grey Literature)

<table>
<thead>
<tr>
<th>P2.1</th>
<th>P2.2</th>
<th>RQ2</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported</td>
<td>Supported</td>
<td>Supported</td>
<td>Grey literature is important. Furthermore if it is important for a new research methodology it is even more likely to be important for a new software engineering technology.</td>
</tr>
<tr>
<td>Supported</td>
<td>Not Supported</td>
<td>Partly Supported</td>
<td>In this case, grey literature supports existing trends but does not add new information. This suggests grey literature is important to assess the strength of trends.</td>
</tr>
<tr>
<td>Not Supported</td>
<td>Supported</td>
<td>Invalid</td>
<td>This result is contradictory and should not occur.</td>
</tr>
<tr>
<td>Not Supported</td>
<td>Not Supported</td>
<td>Not Supported</td>
<td>Grey literature is not useful in this case, and may be unnecessary for research method-based</td>
</tr>
</tbody>
</table>
4.5  **Analysis and Interpretation of Data for RQ6 – Manual versus Automated Search Strategies**

4.5.1 Proposition P6.1 (Effectiveness of Automated Searches) Analysis

The number of papers found by the manual search will be compared with the number of papers found by the automated search. We will record:

- **NM**: The number of papers found by the manual search that were not identified by the automated search.
- **NA**: The number of papers found by the automated search that were not found by the manual search.

4.5.2 Proposition P6.1 Interpretation

If NM > 0 and NA = 0, then the manual search was more effective than the automated search and P6.1 is not supported. If NM = 0 and NA > 0, then the automated search was more effective than the manual search and P6.1 is supported.

If NM & NA are both greater than 0, both types of search add to the number of primary studies but if NM > NA P6.1 is not supported and if NA > NM P6.1 is supported.

4.5.3 Proposition P6.2 (Efficiency of Automated Searches) Analysis

The researchers responsible for the preliminary tertiary study estimate that it took on average 3.5 hours to search each source (journal, conferences proceeding for all relevant years) and 15 minutes for each extra check on papers that included surveys or evidence-based assessments but which were candidates for exclusion because of lack of methodological rigour. This suggests that the time spent on the original search was approximately:

\[
\text{Search time (hours)} = 13 \times 3.5 + 14 \times 0.25 = 45.5 + 3.5 = 49 \text{ hours.}
\]

As an estimate this cannot be considered accurate to more than about 20%, so the search was probably between 49 +/- 9.8, i.e. 39.2 to 58.8 hours.

This estimate will be compared with the time taken to complete the automated search. This includes the time taken by the RA and the time taken by the SLR supervisor and the case study team to assist with the construction of search strings and selection of primary studies.

4.5.4 Proposition P6.2 Interpretation

Proposition P6.2 will be supported if the automated search took less effort than was expended on the manual search.
The main concerns with this comparison are that:

- The time taken for the manual search is a post-hoc estimate not an actual. For this reason we are allowing a 20% inaccuracy in our estimate.
- The RA (who will contribute most of the effort to the search process) will not be as experienced at performing automated searches as the original researchers were at performing a manual search process. We will reduce the impact of this issue by ensuring that the RA is not a complete novice.
- The RA will find the identification of primary studies more difficult than the original researchers. This problem might lead to the RA taking more time to perform the selection process and perhaps misclassify primary studies. To address the second issue we plan to have a rather more formal selection process so that papers that include a literature survey will be checked by two researchers for selection. To address the fact that the search processes in the two cases will not be exactly equivalent, we will add 15 minutes per selected primary study to our estimate of the time taken for the manual search (i.e. an extra 5 hours). This leads to a revised estimate of the time taken for the first of 44.2 to 63.8 hours.

4.5.5 Overall Interpretation

The interpretation of this research question depends on the answer to research question 3 (breadth of search) as shown in Table 4. It shows how the results of the individual propositions and the support for a broad search process lead to an assessment of overall support for the view that automated searches are preferable to manual searches at least for small research-method based SLRs.

Table 4 Interpretation of RQ6 - Manual v. Automated search

<table>
<thead>
<tr>
<th>P6.1</th>
<th>P6.2</th>
<th>RQ3 Supported</th>
<th>RQ3 Not supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported</td>
<td>Supported</td>
<td>An automated search is appropriate and necessary for a small mapping study.</td>
<td>An automated search is more effective and efficient than a manual search but may not lead to more reliable conclusions</td>
</tr>
<tr>
<td>Supported</td>
<td>Not supported</td>
<td>An automated search is to be preferred to ensure that all relevant papers are found.</td>
<td>A focused manual search will lead to similar conclusions to a broader automatic search.</td>
</tr>
<tr>
<td>Not supported</td>
<td>Supported</td>
<td>These situations are contradictory – RQ3 suggests a broader search is necessary but RQ6 suggests an automated search does not find more high quality primary studies than a manual search. Further</td>
<td>Although an automated search is more efficient, manual search is preferable to an automated search in order to ensure the most relevant primary studies are obtained.</td>
</tr>
<tr>
<td>Not supported</td>
<td>Not supported</td>
<td>An automated search</td>
<td></td>
</tr>
</tbody>
</table>
supported | investigation of individual primary studies and their impact is necessary to discover why this has happened. | should be avoided. A manual search is much to be preferred.

### 4.6 Quality Assurance Data

Data recorded by the RA concerning process problems and individual experiences while undertaking the SLR will be reviewed to assess whether any unexpected events have invalidated the planned analyses or study conclusions.

### 5. Plan Validity

#### 5.1 General

The extent to which the case study addresses various quality concerns is shown in Appendix 4. We have addressed all the questions raised in the checklist.

#### 5.2 Construct validity

We will use several sources of evidence that arise naturally due to the different SLR procedures:

- The results of the manual and automated searches.
- The quality assessment of the selected primary studies.
- The nature of the selected primary studies.
- Time sheets related to SLR activities.

#### 5.3 Internal Validity

The relationship between the collected data and conclusion is specified in section 4. The collection of additional quality assurance data will ensure that unexpected events that could bias conclusions are recorded (see Section 4.6).

#### 5.4 External validity

The results of this study must be interpreted carefully because of the nature of the SLR which is the subject of the case study. This is discussed in section 4.

### 6 Study Limitations

The case used in this study is a mapping study of a research methodology not a mapping study of a new software engineering methodology. This fact will impact the generality of any results. In addition, the definition of a “broader search” in the context of this case is restricted to an extended automated search. Nonetheless, the case study should be useful as one of several related case studies undertaken by the EPIC project.
7 Reporting
The results of the case study will be reported in an EPIC technical report which will be published on the EPIC web site, and will be made more widely available in relevant journal/conference papers.

8 Schedule

<table>
<thead>
<tr>
<th>Task</th>
<th>Role/Person Responsible</th>
<th>Required Completion Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyze data to address RQ3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analyze data to address RQ10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analyze data to address RQ7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document results in a TR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References
Appendix 1 Coding Scheme and Data Recording for the Case Study

Daily time sheet should be a six column data table in an EXCEL spread sheet with the following columns:
- Date
- Researcher initials
- Task
- Subtask
- Hours
- Comments.

Task should be one of Planning (P), Conduct, (C) Reporting (R), Case Study Activities (CA), Other (O).

The case study activities field should be used to record time spent on task such as data collection and reporting required for the Case study.

Subtask for Planning should be one of:
- Reviewing background literature (BL)
- Research question formulation (RQ)
- Search strategy definition (SSD)
- Search String specification (SSP)
- Search String trial (SST)
- Constructing Inclusion and exclusion criteria (IEC)
- Constructing Quality Checklists (QC)
- Data Extraction form construction (DEF)
- Data Extraction Form Trial (DEFT)
- Data Aggregation Specification (DAS)
- Data Aggregation Specification Trial (DAT)
- Protocol documentation (PD)
- Response to Protocol review queries (PRR)
- Other

The subtasks for Conduct should be:
- Protocol revision (PR)
- Search process (SP)
- Primary study selection (PSS)
- Quality Evaluation (QE)
- Data Extraction (DE)
- Data Aggregation (DA)
- Other (O)

The Reporting subtasks should be:
- Specifying the results & conclusions of the study (RC)
- Preparing the final report (RP)
• Response to final report review (RRR)
• Other (O)

When the “Other” category is used a textual description of the task should be provided in the comments field. Time related to keeping records of the process should also be recorded.

The RA and the case study researchers acting in their roles as an SLR research team should maintain their own spreadsheets.
Appendix 2 Incident Log

An incident log should specify any “breakdown” that occurs during the SLR process. The format of the log should be a word document with a separate page for each incident. The RA should maintain his/her own incident log.

The incident form should comprise:

1. Date
2. Name (initials) of researcher
3. Task (as for the timesheet)
4. Subtask (as for the timesheet)
5. Problem description (text)
6. Problem Solution (text)
7. Problem Code

The Problem Code should be:

- NA: No Advice on the problem was available in the Guidelines
- IA: Incorrect Advice was provided in the Guidelines
- AA: Ambiguous advice or incomplete advice was provided in the Guidelines
- Other: The problem was outside the scope of the Guidelines.
Appendix 3 RA Debriefing Questionnaire

Name
Date

1. Which SLR Phases have you completed?
   Planning: Yes/No
   Conduct: Yes/No
   Reporting: Yes/No

2. Which tasks were the most difficult? Please identify up to 5 tasks (you do not have to identify 5 tasks). For each task record (use as many extra sheets of paper as necessary):
   a. Task Name

   b. Was the difficulty due to:
      1. Lack of experience applying the Systematic Literature Review Process
         Yes/No/Partly
      2. Lack of Testing Experience: Yes/No/Partly
      3. Lack of experience with empirical studies/statistics: Yes/No/Partly
      4. The constraints of undertaking the SLR within the context of a case study:
         Yes/No/Partly
      5. Working as a single researcher with a supervisor as opposed to in team of researchers: Yes/No/Partly
      6. Other: Yes/No/Partly
         If Other please specify

c. Could the difficulty be reduced if the guidelines were revised: Yes/No/Partly
   If Yes/Partly, what changes would you recommend?

d. Could the difficulty be addressed by initiatives other than amending the guidelines (e.g. statistical training, other forms of training) Yes/No/Partly
   If Yes/Partly, what changes would you recommend?
## Appendix 4 – Quality Evaluation of Case Study Plan

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is the object of study?</td>
<td>To investigate the impact of different SLR search procedures.</td>
</tr>
<tr>
<td>2</td>
<td>Is a clear purpose/objective/research question/ hypothesis/ proposition defined upfront?</td>
<td>See Table 1</td>
</tr>
<tr>
<td>3</td>
<td>Is the theoretical basis – relation to existing literature and other cases – defined?</td>
<td>The difference between medical guidelines and procedures adopted by SE researchers is documented in Section 1.</td>
</tr>
<tr>
<td>4</td>
<td>Are the authors’ intentions with the research made clear?</td>
<td>Section 4 indicates how we will interpret the data we collected to answer our research questions.</td>
</tr>
<tr>
<td>5</td>
<td>Is the case adequately defined (size, domain, process).</td>
<td>The case is described in Section 1.</td>
</tr>
<tr>
<td>6</td>
<td>Is a cause-effect relation under study? If yes, is the cause distinguished from other factors?</td>
<td>Several cause-effect relations are being studied. We attempt to identify the implications of each and how they interact.</td>
</tr>
<tr>
<td>7</td>
<td>Will data be collected from multiple sources? Using multiple methods?</td>
<td>Data will be collected from the SLR search process results and from the quality evaluation of primary studies.</td>
</tr>
<tr>
<td>8</td>
<td>Is there a rationale behind the selection of roles, artefacts, viewpoints etc.</td>
<td>The rational for the case study is two-fold: it fits goals of the EPIC study and with work required for a journal paper.</td>
</tr>
<tr>
<td>9</td>
<td>Are the case study settings relevant to validly address the research question?</td>
<td>The case is not a typical SLR, so results have to be interpreted carefully.</td>
</tr>
<tr>
<td>10</td>
<td>Is the integrity of the individuals/organisations taken into account?</td>
<td>The main issue is the participant-observer issue. We have identified the case study and SLR roles of team members and will ensure that the RA performing the SLR is not aware of the case study research questions</td>
</tr>
</tbody>
</table>