Conducting Systematic Literature Reviews

Sandro Schulze

based on the guidelines of Kitchenham et al.
Why to do this in a Systematic Way?

• Decades of research have produced tons of papers
• Digitalisation provides easy access too all of them
• even for a specific domain, this results into hundreds of thousands of papers
• Quality and content differs widely

You need to prevent from getting lost in this mess!!
Overview — Stages of an SLR

- Planning
  - Identifying the need for a review (done by your advisor)
  - Specifying research questions
  - Developing a review protocol

- Conducting the Review
  - Searching Relevant Studies
  - Selection of primary studies
  - Study quality assessment
  - Data extraction and monitoring
  - Data synthesis

- Reporting the Review (is specified by the seminar regulations)
Planning
Specifying Research Questions

• RQs drive the whole SLR methodology
  • search process: primary studies must address RQs
  • data extraction: data items must support answering the RQs
  • data analysis/synthesis: in such a way that RQs can be answered
What Makes Up Good RQs?

• The RQ is meaningful and important to practitioners and researchers.
• The RQ will lead to changes in current practice or to increase confidence in the value of current practice.
• The RQ will identify discrepancies between commonly held beliefs and the reality.
• The RQ reveals new insights (challenges, shortcomings, etc.)
How to get there?

- PICOC method has been suggested to get the focus right (Petticrew & Roberts, 2006):
  - **Population** — target group for the investigation (e.g., bug reports)
  - **Intervention** — specifies the investigation aspect(s) or issues of interest
  - **Comparison** — what is the investigation compared to?
  - **Outcomes** — effect of intervention
  - **Context** — setting or environment of investigation
Example

- **Reverse Engineering Variability from Natural Language Documents: A Systematic Literature Review**
- paper by Li et al. (I was heavily involved)

<table>
<thead>
<tr>
<th>PICOC</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Literature in reverse engineering variability from NL documents.</td>
</tr>
<tr>
<td>Intervention</td>
<td>Mechanisms, i.e., techniques, methods, tools, approaches that realize such a reverse engineering process.</td>
</tr>
<tr>
<td>Comparison</td>
<td>Techniques together with their performance, evaluation and tool support proposed by each primary study.</td>
</tr>
<tr>
<td>Outcome</td>
<td>Several observations regarding applicability and quality of current approaches and major gaps and open challenges in this field.</td>
</tr>
<tr>
<td>Context</td>
<td>The SPLE process, in particular the reverse engineering step to enable a systematic product-line development (e.g., in an extractive way).</td>
</tr>
</tbody>
</table>

Table 1: Research questions structured by PICOC criteria.
Example (cont’d)

• RQ1: What approaches of feature & variability extraction from NL documents have been proposed for SPLs?

• RQ2: How are the techniques supported regarding tools and automation?

• RQ3: How reliable are the approaches, proposed for feature extraction in SPLs?
Developing a Review Protocol

Prevents you from cheating yourself
SLR Protocol

• Plan that specifies the basic review procedures
• Components
  • (background)
  • Research questions
  • search terms & resources (i.e., databases)
  • selection criteria
  • (quality checklist and procedures)
  • data extraction —> how obtain data from primary study
  • data synthesis —> how to put them together
  • SLR schedule
Conducting the Review
Searching Relevant Studies

- Involves a comprehensive and exhaustive searching of studies to be included in the review.
  - Define a search strategy
  - Consult the subject librarian(s)
  - Search strategies are usually iterative and benefit from:
    - Preliminary searches (to identify existing review and volume of studies)
    - Trial searches (combination of terms from RQ)
    - Check the search results against list of known studies
    - Consult the experts in the field
Search Strategy — Aspects to be Defined

1. **Which** approach to be used in search process (e.g., manual or automated search)?

2. **Where** (venues or databases) to search, and which part of article (field) should be searched?

3. **What** (subject, evidence type) to be searched, and what are queries (search strings) fed into search engines?

4. **When** is the search carried out, and what time span to be searched?
Constructing the Search String

- Derive major terms used in the review questions based on the PICOC.
- List the keywords mentioned in the article.
- Search for synonyms and alternative words
- Use the boolean OR to incorporate alternative synonyms.
- Use the boolean AND to link major terms

- In this seminar: discuss with advisor (he/she may have ideas on search terms)
Example (cf.~paper on slide 8)

("feature extraction" OR "feature selection" OR "feature location" OR "feature identification" OR "feature detection" OR "feature mining" OR "feature clustering" OR "feature similarity") AND ("natural language" OR "requirement" OR "textual requirement" OR "description" OR "specification" OR "product review" OR "Natural language processing") AND ("Software Product Lines" OR "product family" OR "software family" OR "Feature-oriented software development")

- Too long search strings may result into false positives
- Simpler search strings might just as effective ;-)
Sources for Your Search

- Digital libraries (ACM, IEEE, dblp)
- Reference lists from relevant primary studies and review articles
  - Journals (including company journals such as the IBM Journal of Research and Development), grey literature (i.e. technical reports, work in progress)
- Conference proceedings
- Research registers
- The Internet (google scholar!)
- Direct contact specific researcher(s)
Complementary Search — Snowballing

- Especially useful when applied to a small number of initial papers
- Snowballing approach requires a starting set of papers, which should be based on identifying a set of papers from leading journals in the area.
- **Forward snowballing** — identify articles that have cited the articles found in the search
- **Backward snowballing** — identify articles from the reference lists.
Snowballing Technique

- References

Set of studies found from database search

Backward Snowball

Forward Snowball

What is “publication bias”?

Refers to the problem that positive results are more likely to publish than negative results (Kitchenham & Charter, 2007).

E.g. formal experiments that was failed to reject null hypothesis are considered less interesting. Publication bias can lead to systematic bias in SLR.

Strategies to address publication bias:

- Scanning grey literature
- Scanning conference proceedings
- Contacting experts working in the area
Managing Your Bibliography

• Tools can help you in managing all your references

Jabref

Mendeley

BibSonomy

LibraryThing

What’s on your bookshelf?
Documenting the Search

- The process of conducting SLR must be transparent and replicable.
- The review should be documented in sufficient detail.
- The search should be documented and changes noted.
- Unfiltered search results should be saved for possible reanalysis.

### Table 2: Search process documentation

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Library</td>
<td>Name of database</td>
</tr>
<tr>
<td></td>
<td>Search strategy for the database</td>
</tr>
<tr>
<td></td>
<td>Date of search</td>
</tr>
<tr>
<td></td>
<td>Years covered by search</td>
</tr>
<tr>
<td>Journal Hand Searches</td>
<td>Name of journal</td>
</tr>
<tr>
<td></td>
<td>Years searched</td>
</tr>
<tr>
<td></td>
<td>Any issues not searched</td>
</tr>
<tr>
<td>Conference proceedings</td>
<td>Title of proceedings</td>
</tr>
<tr>
<td></td>
<td>Name of conference (if different)</td>
</tr>
<tr>
<td></td>
<td>Title translation (if necessary)</td>
</tr>
<tr>
<td></td>
<td>Journal name (if published as part of a journal)</td>
</tr>
<tr>
<td>Efforts to identify unpublished studies</td>
<td>Research groups and researchers contacted (Names and contact details)</td>
</tr>
<tr>
<td></td>
<td>Research web sites searched (Date and URL)</td>
</tr>
<tr>
<td>Other sources</td>
<td>Date Searched/Contacted</td>
</tr>
<tr>
<td></td>
<td>URL</td>
</tr>
<tr>
<td></td>
<td>Any specific conditions pertaining to the search</td>
</tr>
</tbody>
</table>
Selection of Primary Studies
Overview

• Primary studies need to be assessed for their actual relevance
• Set the criteria for including or excluding studies (decided earlier during protocol development, can be refined later)
• Inclusion & exclusion criteria should be based on RQ
• Selection process should be piloted.
• Study selection is a multistage process
  • specify initially during protocol definition
  • refine later (if necessary)
Example (cf. paper on slide 8)

• Inclusion criteria

  IC01 Articles matching the search string mentioned above and within the scope of our analysis, i.e., they propose a technique or mechanism for feature & variability extraction from NL documents.

  IC02 Articles published between January 1st 2000 to May 30th 2017, since research on automatic feature & variability extraction from NL documents in SPLs began in 21st century.

• Exclusion criteria

  EC01 Articles not focusing on feature and variability extraction from NL documents in SPL, i.e., feature extraction from legacy code, approaches improving feature modelling, functional requirements extraction, etc;

  EC02 Articles not written in English;

  EC03 Articles not belonging to research papers, i.e., proposals, summary of conference, lecture notes, etc;

  EC04 Articles not pertaining to the firsthand researches, namely, related literature review or survey papers.
Study Quality Assessment
Why is this important?

• To provide more detailed **Inclusion/Exclusion criteria**
• To check whether quality differences provide an **explanation for differences** in study results
• As a means of **weighting the importance** of individual studies when results are being synthesized.
• To guide the **interpretation of findings** and determine the **strength of inferences**.
• To **guide recommendations** for further research.
Assessing Studies’ Quality

• “Quality relates to the extent to which the study minimizes bias and maximizes internal and external validity” (Khan et al, 2001)

• Quality Concepts Definition (Kitchenham & Charter, 2007)

<table>
<thead>
<tr>
<th>Term</th>
<th>Synonyms</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>Systematic error</td>
<td>A tendency to produce results that depart systematically from the ‘true’ results. Unbiased results are internally valid</td>
</tr>
<tr>
<td>Internal validity</td>
<td>Validity</td>
<td>The extent to which the design and conduct of the study are likely to prevent systematic error. Internal validity is a prerequisite for external validity.</td>
</tr>
<tr>
<td>External validity</td>
<td>Generalisability, Applicability</td>
<td>The extent to which the effects observed in the study are applicable outside of the study.</td>
</tr>
</tbody>
</table>
Assessing Studies’ Quality

• Assessing quality of studies:
  • Methodology or design of the study
  • Analysis of studies’ findings.
• Quality checklist or instrument need to be designed to facilitate quality assessment.
• Most quality checklists include questions aimed at assessing the extent to which articles have addressed bias and validity.
Example

- Kitchenham et al. (2007) constructed a quality questionnaire based on 5 issues affecting the quality of the study:
  1. Is the data analysis process appropriate?
  2. Did studies carry out a sensitivity or residual analysis?
  3. Were accuracy statistics based on the raw data scale?
  4. How good was the study comparison method?
  5. The size of the within-company data set (e.g. < 10 projects considered poor quality)
Example (2)

- Salleh et al., 2011

<table>
<thead>
<tr>
<th>Item</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the article referred? [30]</td>
<td>Yes/No</td>
</tr>
<tr>
<td>2. Were the aim(s) of the study clearly stated? [16], [67]</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>3. Were the study participants or observational units adequately described? For example, students’ programming experience, year of study etc. [44], [68]</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>4. Were the data collections carried out very well? For example, discussion of procedures used for collection, and how the study setting may have influenced the data collected [44], [48], [67], [68]</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>5. Were potential confounders adequately controlled for in the analysis? [67]</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>6. Were the approach to and formulation of the analysis well conveyed? For example, description of the form of the original data, rationale for choice of method/tool/package [48], [67], [68]</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>7. Were the findings credible? For example, the study was methodologically explained so that we can trust the findings; findings/conclusions are resonant with other knowledge and experience [48], [44], [68]</td>
<td>Yes/No/Partially</td>
</tr>
</tbody>
</table>
Data Extraction
Overview

• Involve reading the full text article.
• Data extracted from primary studies should be recorded using data extraction form.
• The form should be designed and piloted when the protocol is defined.
• Collect all the information that can be used to answer the RQ and the study’s quality criteria.
• Both quality checklist and review data can be included in the same form.
• In case of duplicates publication (reporting the same data), refer the most complete one (e.g., journals over conferences)
• For validation, a set of papers should be reviewed by 2 or more researchers. Compare results and resolve any conflicts. (skipped here)
4) Data Extraction

- Involve reading the full text article.
- Data extracted from primary studies should be recorded using a data extraction form.
- The form should be designed and piloted when the protocol is defined.
- Collect all the information that can be used to answer the RQ and the study’s quality criteria.
- Both quality checklist and review data can be included in the same form.
- In case of duplicates publications (reporting the same data), refer to the most complete one.
- For validation, a set of papers should be reviewed by 2 or more researchers. Compare results and resolve any conflicts.
Synthesis of Evidence
Overview

• Involves collating and summarizing the results of the included primary studies.

• Key objectives of data synthesis (Cruzes & Dyba, 2011):
  • to analyze and evaluate multiple studies
  • to select appropriate methods for integrating or providing new interpretive explanations about them

• Types of Synthesis:
  • Descriptive (narrative/qualitative, thematic analysis)
  • Numerical synthesis (e.g. meta-analysis)
Descriptive Synthesis

• “An approach to the synthesis of findings from multiple studies that relies primarily on the use of words and text to summarize and explain the findings of the synthesis. It adopts a textual approach to the process of synthesis to ‘tell the story’ of the findings from the included studies.”

• Methods proposed for synthesis of qualitative studies:
  • Meta-ethnography
  • thematic analysis
  • content analysis
  • narrative synthesis
  • etc.
Meta-Ethnography

• It is a method that involves induction and interpretation.
• The outcome is the translation of studies into one another, which encourages the researcher to understand and transfer ideas, concepts and metaphors across different studies.
• Studies can relate to one another in one of three ways:
  • they may be directly comparable as reciprocal translations
  • they may stand in opposition to one another as refutational translations, or
  • taken together they may represent a line of argument.
Numerical Analysis (Meta-Analysis)

- Meta-analysis can be used to aggregate results or to pool data from different studies.
- The aim is to resolve uncertainty when the results of studies disagree; and to increase confidence in the results obtained from individual studies.
- The outcome of a meta-analysis is an average effect size (ES) with an indication of how variable that effect size is between studies.
- ES can be calculated based on Hedges or Cohen’s approach; i.e. the standardized mean difference between the two groups.
Meta-Analysis

• Meta-analysis involves three main steps:
  1. Decide which studies to be included in the meta-analysis.
  2. Estimate an effect size for each individual study.
  3. Combine the effect sizes from the individual studies to estimate and test the combined effect.

• Results of the meta-analysis can be presented in a forest plot.

• Software tools: Comprehensive Meta-Analysis (CMA), MIX, OpenMeta [Analyst], Review Manager (RevMan) from Cochrane Collaboration, Stata, R.
Reporting the Results
Overview

- Usually, conferences & journals have dedicated tracks or editors for that

- Examples:
  - Information & Software Technology (IST) has an editor specializing in systematic reviews.
  - Int’l Symposium on Empirical Software Engineering & Measurement (ESEM)
  - You have a dedicated Seminar for writing a (small) SLR :-)

How to Conduct a Systematic Literature Review — Sandro Schulze
General Reporting Structure

- **Introduction**
  - General introduction about the research. State the purpose of the review. Emphasize the reason(s) why the RQ is important. State the significance of the review work and how the project contributes to the body of knowledge of the field.

- **Main Body**
  - Review method – briefly describe steps taken to conduct the review
  - Results – Findings from the review, i.e. the synthesis
  - Discussion – implication of review for research & practice, Threats to Validity

- **Conclusions**

- More details in next lecture on Academic Writing
References


• Cohen, J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. (70) 1968, pp. 213-220.


