Systematic secondary studies

TIEJ601 Postgraduate Seminar in Mathematical Information Technology

Antti-Juhani Kaijanaho

Department of Mathematical Information Technology
University of Jyväskylä

September 30, 2014

Arity of studies

Definition

A *secondary study* is a study about some topic that uses published scientific literature (called *primary studies*) about the same topic as its source of data.

Note

The term *tertiary study* is sometimes used of secondary studies using other secondary studies as their source of data.

Systematic secondary studies

- deliberately designed
- best current practice followed
- audit trail preserved
- meticulously documented
- threats to validity assessed
- time and effort intensive

Main types

- Systematic Literature Review (SLR)
  - very focused research questions
  - of practical relevance
  - synthesizes a result from the primary studies
  - very common in medicine[1]
  - subtype: meta-analysis
    - a set of statistical methods for pooling quantitative primary studies' data
    - often used to designate SLRs that use meta-analysis methods
- Systematic Mapping Study (SMS)
  - broad research questions
  - of relevance primarily to research
  - generates a "map" of the literature
  - rarely attempts a synthesis
- (not exhaustive)

[1] See e. g. http://summaries.cochrane.org/

Recommended methodological sources

Software Engineering

- Kitchenham & Charters: *Guidelines for performing Systematic Literature Reviews in Software Engineering. Version 2.3.* EBSE Technical Report EBSE-2007-01, 2007. [2]
- Kitchenham & Brereton: *A systematic review of systematic review process research in software engineering.* Information and Software Technology 55 (12), 2013. doi:10.1016/j.infsof.2013.07.010
- See also the CSE recommendations on the next slide

[2] https://community.dur.ac.uk/ebse/resources/guidelines/Systematic-reviews-5-8.pdf

Recommended methodological sources

Computer Science Education

- Petticrew & Roberts: *Systematic Reviews in the Social Sciences. A Practical Guide.* Malden, MA: Blackwell, 2006.
- *Campbell Collaboration Systematic Reviews. Policies and Guidelines. Version 1.0.* Campbell Systematic Reviews supplement 1, 2014.[3]
- *What Works Clearinghouse Procedures and Standards Handbook. Version 3.0.* 2014.[4]
- see also the SE recommendations on the previous slide

[3] http://www.campbellcollaboration.org/lib/download/3308/C2_Policies_Guidelines_Version_1_0.pdf

[4] http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Conducting a systematic secondary study

1. Determine whether a new systematic secondary study is needed.
2. Design the study.
3. Review (and optionally publish) the design.
4. Conduct literature searches.
5. Select primary studies from the literature found using predefined criteria.
6. Assess the quality of the selected studies (optional for mapping studies).
7. Collect and synthesize data.
8. Assess the limitations of the study.
9. Write and publish one or more reports.

Crosscutting concerns

Steps 4–7 should be

- designed
- piloted

beforehand,

- executed carefully
- documented

while in progress and

- assessed for reliability

afterward.

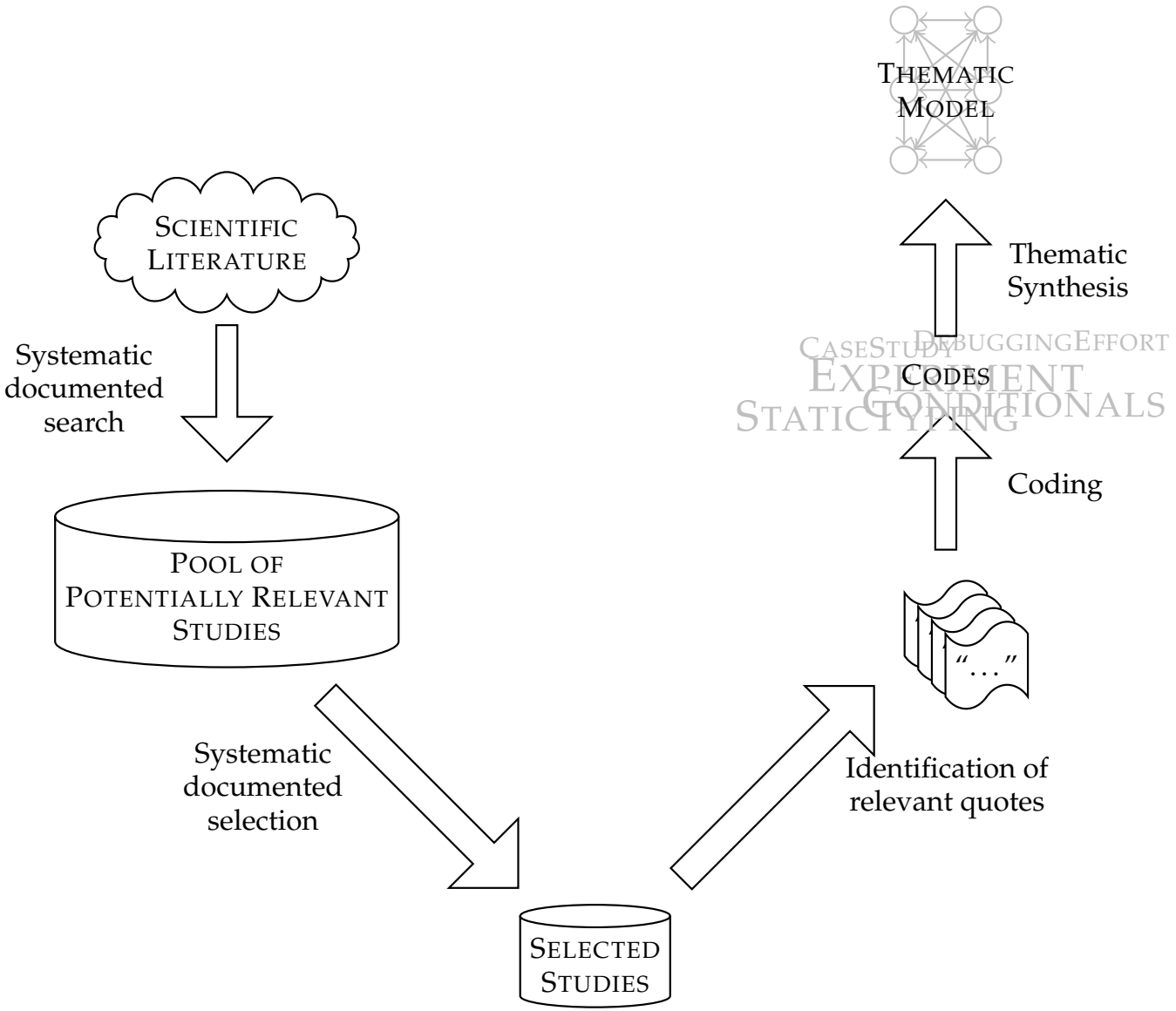
Case

Antti-Juhani Kaijanaho: *The extent of empirical evidence that could inform evidence-based design of programming languages. A systematic mapping study.* Jyväskylä: University of Jyväskylä, 2014. Jyväskylä licentiate theses in computing 18.[5]

- Chapter 3 contains an extensive summary of methodological literature (particularly for software engineering).
- Designed in November 2010.
- Completed in May 2014.
- Approved and published in August 2014.

[5] http://urn.fi/URN:ISBN:978-951-39-5791-9

Overall design



Research questions

What scientific evidence is there about the efficacy of particular decisions in programming language design?

1. How much has the efficacy of particular programming language design decisions been empirically studied?
2. Which programming language design decisions have been studied empirically for efficacy?
3. Which facets of efficacy regarding programming language design decisions have been studied empirically?
4. Which empirical research methods have been used in studying the efficacy of particular programming language design decisions?
5. How common are follow-up or replication studies, either by the original researchers or by others?

Searches

Journals

Journal	Vols.	Years	Date of search	Yield
ESE	1–17	1997-2012	Dec. 10, 2010, Jan. 4, 2013	9
CACM	1–33	1958–1990	Dec. 13–21, 2010, Jan. 17–19, 2011	280
TOPLAS	1–34	1979–2012	Dec. 17–20, 2010, Jan. 7–10, 2013	182
LOPLAS	1–2	1992–1993	Dec. 21, 2010	7
IJMMS & IJHCS	1–70	1969–2012	Dec. 20–21, 2012, Jan. 4, 2013	109

Proceedings

Proc. of	Years	Date of search	Yield
PPIG	1989-2012	Dec. 9, 2010, Jan. 4, 2013	63
ISESE & ESEM	2002–2012	Dec. 10, 2010, Jan. 4, 2013	9
OOPSLA & SPLASH	1986–2012	Jan. 19–28, 2011, Jan. 7, 2013	207
ECOOP	1987–2012	Jan. 28, Feb. 1–7, Jun. 1–17, Aug. 4–19, 2011, Jan. 4, 2013	286
POPL	1973–2012	Aug. 19–22, Sep. 1, 2011, Jan. 4, 2013	219

Keyword search

Keyword search in Google Scholar, IEEE Xplore, ISI Web of Science, and ScienceDirect performed in 2011 and 2013 yielded 420 candidate publications.

Snowball search

Searching in the references and in the set of citing articles of already selected articles was performed in Spring 2013 and yielded 293 candidate publications.

Assessing manual search

	QGS		
	total	contrib.	q.-s.
ESE	3	1	33 %
CACM	4	4	100 %
TOPLAS	9	9	100 %
LOPLAS	0		
IJMMS	14	12	86 %
IJHCS	2	2	100 %
PPIG	2	1	50 %
ISESE	0		
ESEM	2	2	100 %
OOPSLA	14	9	64 %
SPLASH	0		
ECOOP	13	13	100 %
POPL	3	3	100 %
	66	56	85 %

QGS = quasi-gold standard: (in this case) the set of relevant publications published in the forum found by any search

contr. = contribution (in the QGS): (in this case) the set of relevant publications published in the forum found by manual search in the forum

q.-s. = quasi-sensitivity: (in this case) the ratio of relevant publications found by manual search in the forum to the number of relevant publications found by any search in the forum

Assessing automatic searches

	yield	contrib. oa.	contrib. QGS	q.-s.	sp.
Google Scholar	8 455	67	16	24 %	1 %
IEEE Xplore	995	18	2	3 %	18 %
ScienceDirect	1 022	8	7	11 %	1 %
Web of Science	20	3	0	0 %	15 %
	10 492	69	18	27 %	1 %

oa. contrib. = overall contribution: the number of relevant publications found by this search engine in any forum

QGS contrib. = contribution in the quasi-gold standard: the number of relevant publications found by this search engine within the manually searched forums

q.-s. = quasi-sensitivity: the ratio of the size of the QGS contribution to the size of the full QGS (the set of relevant publications found by any search in any manually-searched forum)

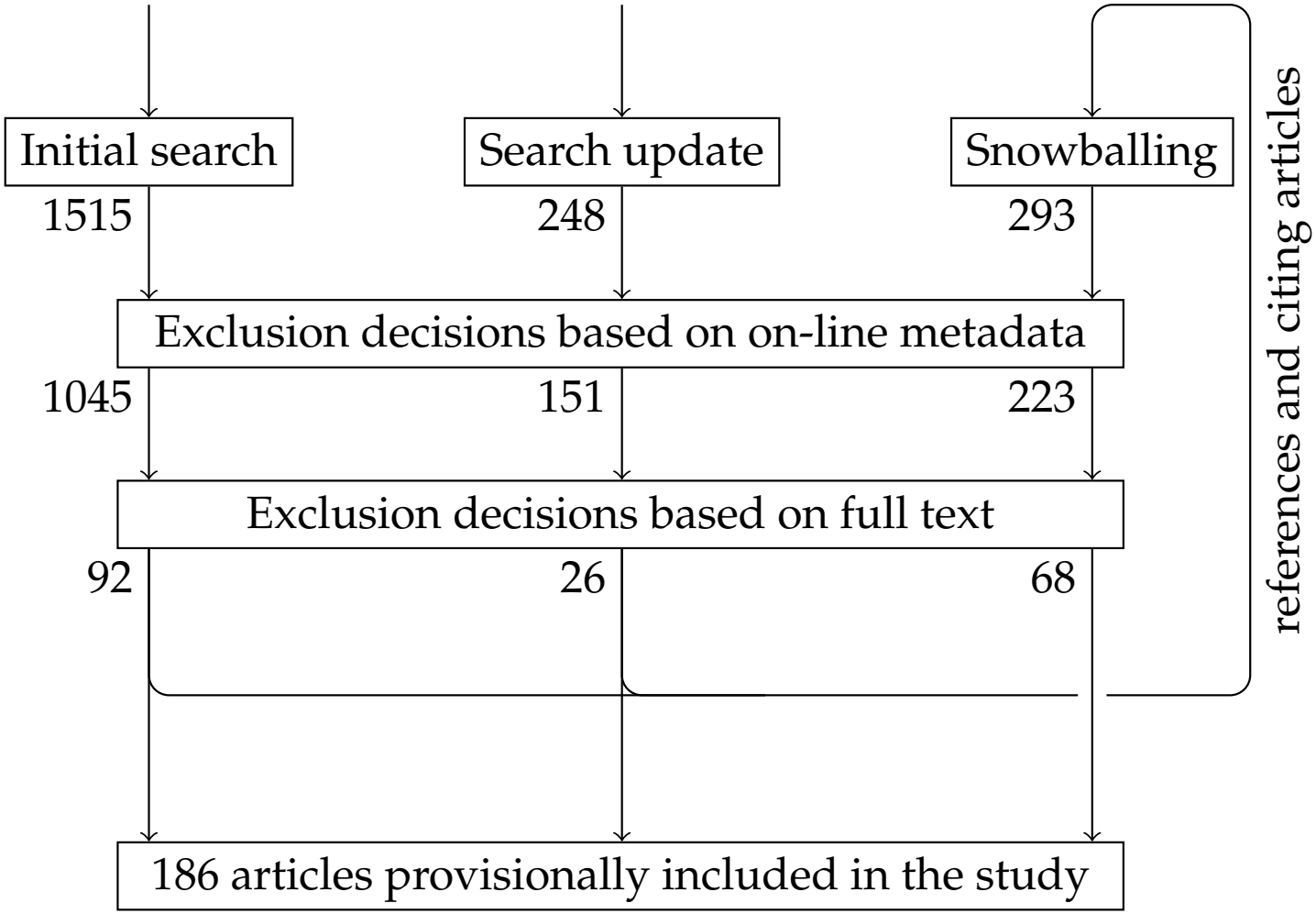
sp. = specificity: ratio of overall contribution to the yield

Selection criteria

1. Is this a primary study that attempts to determine the efficacy of a programming language design decision? (If not, skip question 5.)
2. Is this a literature review that attempts to summarize or consolidate research on the efficacy of a programming language design decision? (If not, skip questions 6 and 7.)
3. Can you find a complete written and published report about this study?
4. Is the study reported in English, Finnish or Swedish?
5. Does this primary study present scientific empirical evidence about their claims?
6. Does this secondary study include any primary studies that present scientific empirical evidence?
7. Does this secondary study discuss scientific empirical evidence in the primary studies under review?

EXCLUDE if Q1 and Q2 are both NO or any of Q3–Q7 is NO.

Selection process



Selection validation

Pairwise Cohen κ

κ	AJK-1	AJK-2	VT	VL
AJK-2	0.82 (+0.65 to 0.99)			
VT	0.78 (+0.36 to 1.00)	1.00 (+1.00 to 1.00)		
VL	0.62 (−0.10 to 1.00)	0.62 (−0.10 to 1.00)	0.62 (−0.10 to 1.00)	
TK	0.29 (−0.26 to 0.83)	0.38 (−0.15 to 0.92)	0.22 (−0.45 to 0.90)	0.38 (−0.40 to 1.00)

AJK-1 is my original set of decisions ($n = 2056$), AJK-2 is my set of re-examinations ($n = 100$), and VT ($n = 28$), TK ($n = 20$), and VL ($n = 10$) are my three supervisors; the pairwise comparisons use the smaller n of the pair, except between TK and VT ($n = 19$).

Multi-way Fleiss κ

For all ratings, $\kappa = 0.42$ (95 % CI -0.19 to 1.00, $n = 10$).
For all except TK, $\kappa = 0.77$ (95 % CI 0.02 to 1.00, $n = 10$).

Note

Some sources recommend avoiding the Cohen κ.
Krippendorff α is often recommended as the best choice (not done in my licentiate thesis).[6]

[6] See e. g. Hayes & Krippendorff: *Answering the Call for a Standard Reliability Measure for Coding Data.* Communication Methods and Measures, 1 (1) 77–89. doi:10.1080/19312450709336664

Synthesis and mapping

Adapted from Daniela S. Cruzes & Tore Dybå: Recommended Steps for Thematic Synthesis in Software Engineering. In Proc. ESEM 2011, pp. 275–284. doi:10.1109/ESEM.2011.36

1. I read all included studies at least once, in order to "get immersed with the data" (p. 276).
2. I extracted certain categories of information from each paper, as direct quotes (with page references)
3. I abstracted and created a code book for those categories of information as well as any emergent concepts and categories, while simultaneously applying it to the quotes.
4. I explored the resulting codings to determine interesting themes and patterns.

For results, see the licentiate thesis and

https://yousource.it.jyu.fi/antti-juhani-kaijanaho-s-licentiate-thesis-materials/collected-data.

Threats to validity

- The searches may not have find all relevant publications.
- The selection process may not have reliably determined which publications are relevant and which are not.
- The coding process may not have reliably encoded the relevant information in the primary studies.
- The thematic synthesis process may have created misleading themes and patterns.

Dissemination plan

Recommended at least (done in this case)

- a peer-reviewed detailed report, either a thesis or as a technical report (thesis finished)
- academic article (jounal article in preparation)
- trade article if relevant (no plans at this time)
- reports for other stakeholders if relevant (not relevant)

Personal observations

- properly done a multi-person-year undertaking
  - good idea to have a team of 3–6 researchers
- most abstracts are almost useless
  - I recommend adopting structured abstracts[7]

[7] See e. g. https://community.dur.ac.uk/ebse/abstracts.php

DISCUSSION TIME