Experiences Using Systematic Review Guidelines

Mark Staples
National ICT Australia
Australian Technology Park
Garden Street, Eveleigh, NSW 1430 Australia
mark.staples@nicta.com.au

Mahmood Niazi
National ICT Australia
Australian Technology Park
Garden Street, Eveleigh, NSW 1430 Australia
and
School of Computing and Mathematics
Keele University
Staffs ST5 5BG United Kingdom
m.k.niazi@cs.keele.ac.uk

A systematic review is a defined and methodical way to identify, assess and analyse published primary studies in order to investigate a specific research question. Kitchenham has recently published guidelines for software engineering researchers performing systematic reviews. The objective of our paper is to critique Kitchenham's guidelines and to comment on systematic review generally with respect to our experiences conducting our first systematic review. Our perspective as neophytes may be particularly illuminating for other software engineering researchers who are also considering conducting their first systematic review. Overall we can recommend Kitchenham's guidelines to other researchers considering systematic reviews. We caution researchers to clearly and narrowly define the research questions they will investigate by systematic review, to reduce the overall effort and to improve the quality of the selection of papers and extraction of data. In particular we recommend defining complementary research questions that are not within the scope of the systematic review in order to clarify the boundaries of the specific research question of interest. An instance of this recommendation is that researchers should clearly define the unit of study for the systematic review.

Keywords: Systematic review, empirical software engineering.

1. INTRODUCTION

A systematic review is a defined and methodical way of identifying, assessing, and analysing published primary studies in order to investigate a specific research question. A systematic review can also discover the structure and patterns of existing research, and so identify gaps that can be filled by future research. [6] Systematic reviews differ from ordinary literature surveys in being formally planned and methodically executed. A good systematic review should be independently replicable and so will have much greater scientific value than an ordinary literature survey. However, systematic reviews require much more effort than ordinary literature surveys.

The following features differentiate a systematic review from a conventional literature review [6]:

- Definition and documentation of a systematic review protocol in advance of conducting the review, to specify the research questions and the procedures to be used to perform the review.
- Definition and documentation a search strategy as part of the protocol, to find as much of the relevant literature as possible.
- Description of the explicit inclusion and exclusion criteria as part of the protocol, to be used to assess each potential study.
- Description of quality assessment mechanisms as part of the protocol, to evaluate each study.
- Description of review and cross-checking processes as part of the protocol, and involving multiple independent researchers, in order to control researcher bias.

Kitchenham [6] has recently published guidelines for software engineering researchers performing systematic reviews. Although procedures and systems for systematic reviews are well established in other disciplines

(particularly in medicine), software engineering researchers have yet to come to a well understood consensus about the conduct and value of systematic reviews.

The objective of this paper is to comment on Kitchenham's guidelines and comment on systematic review generally with respect to our experiences conducting a systematic review [8] informed by Kitchenham's guidelines. This systematic review was the first that we have conducted, and so our critique is from the perspective of neophytes and may be particularly illuminating for other researchers who are also considering conducting their first systematic review.

This paper is organised as follows. Section 2 sketches an overview of the systematic review process as described by Kitchenham [6]. Section 3 describes our experiences performing our first systematic review. In Section 4 we critique Kitchenham's systematic review guidelines in light of our experiences and discuss some the lessons we have learned. Section 5 concludes and discusses a few suggested improvements to the systematic review approach for software engineering research.

2. SYSTEMATIC REVIEW GUIDELINES

Kitchenham describes the three main phases of a systematic review process: planning the review, conducting the review, and reporting the review. Each of these phases contains a sequence of stages, but the execution of the overall process involves iteration, feedback, and refinement of the defined process. [6] In this section we describe the three phases of a systematic review, and for their constituent phases identify some of the important guidelines described by Kitchenham. [6]

2.1. Planning the Review

The output from this phase is a *systematic review protocol* that defines the purpose and procedures for the review.

2.1.1. Identify the Need for a Systematic Review

The need for a systematic review springs from the need to thoroughly summarise all existing information about a phenomenon. Kitchenham notes that researchers should first find and review any existing systematic reviews related to the phenomenon, which if found may obviate the need for a new systematic review, or at least provide examples to help in the development of a protocol for a new systematic review.

2.1.2. Development of a Review Protocol

A systematic review protocol is a formal and rather concrete plan for the execution of the systematic review. Kitchenham notes that a pre-defined protocol is necessary to reduce the possibility of researcher bias. The contents of a systematic review protocol in many ways foreshadow the structure of the final report – it describes the background context for the research, the specific research questions, the planned search strategy, criteria for publication selection, the treatment of publication quality assessment, the data extraction plan, the data synthesis plan, and a project plan.

Kitchenham discusses research questions at some length, detailing various types of research questions appropriate for systematic review, broader justification for research questions, and the detailed structure (population, intervention, outcomes) of research questions. Kitchenham here also discusses the issue of admitting studies with different types of experimental designs. In particular she discusses whether studies based on expert opinion should be admitted for systematic reviews on software engineering.

Kitchenham recommends in several places that aspects of the protocol should be piloted during its development. In particular, the search terms, selection criteria, and data extraction procedures should all be trialled before finalising the protocol.

2.1.3. Protocol Review

Kitchenham notes that the protocol is critical for the systematic review, and so should be itself reviewed.

2.2. Conducting the Review

This phase ultimately generates final results, but also generates the following intermediate artifacts: the initial search record and archive, the list of selected publications, records of quality assessments, and extracted data for each of the selected publications.

2.2.1. Identification of Research

A formal search strategy (described in the protocol) is used to find the entire population of publications relevant to the research questions. Explicit description of the search strategy helps to make the study replicable and open to external review. Kitchenham notes that the search strategy should attempt to address publication bias by trying to find publications that report “negative” results. The search terms and results should be documented and archived.

2.2.2. Selection of Primary Studies

The selection process is intended to identify the found primary studies that provide direct evidence about the research questions. Again, the selection process should follow the plan described in the protocol. Kitchenham notes that selection is a multistage process: first researchers only exclude clearly irrelevant publications, and then from the resulting short list researchers only include publications that contain extractable data addressing the research questions. Kitchenham also describes the importance of checking the reliability of this selection process, in order to reduce the risk of researcher bias.

2.2.3. Study Quality Assessment

Kitchenham discusses quality assessment with regards to defining the exclusion criteria for the systematic review. After selecting the primary studies, a more detailed quality assessment is needed to allow researchers to assess differences in the implementation of studies. For detailed quality assessment, checklists can be designed using factors that could bias study results. Kitchenham refers to four types of bias: selection bias, performance bias, measurement bias and attrition bias.

2.2.4. Data Extraction and Monitoring

Data extraction should be performed as indicated in the systematic review protocol. The protocol will describe data extraction forms, and will describe procedures for data extraction. Kitchenham recommends performing data extraction by two or more researchers and settling disagreements by consensus or by use of additional researchers. Data monitoring is also performed in this stage – multiple reports of the same study should be identified, and missing or unpublished data should be sought from the publications’ authors.

2.2.5. Data Synthesis

When data has been extracted, it must be grouped and summarised so as to shed light on the research questions for the systematic review. As with other stages, the procedures to be followed should be defined in the protocol. Kitchenham discusses options for combining data from different types of studies, and combining different types of data. Where some studies are of much higher quality, it is possible to perform sensitivity analyses to determine the effects on the synthesis results of ignoring low quality publications.

2.3. Reporting the Review

Reporting the review is a single stage phase. Usually, systematic reviews are reported using two formats: in a technical report and in a journal or conference papers. The structure and contents of reports is presented in [6].

3. EXPERIENCES WITH OUR SYSTEMATIC REVIEW

This section describes our experiences performing a systematic review using Kitchenham’s guidelines.

3.1. Planning the Review

We referred to Kitchenham’s guideline before planning the review.

3.1.1. Identifying the Need for a Systematic Review

As part of a larger research project investigating the adoption and impact of CMMI (Capability Maturity Model Integration) on SMEs (Small-to-Medium-sized Enterprises) we wanted to investigate the applicability of CMMI to SMEs. CMMI is fairly recent and so would have few associated studies, but we thought that studies about earlier CMM and related models would be relevant. Although primarily interested in SMEs, we also wanted to know how motivations for SPI differed between large enterprises and SMEs, and so we were led to consider the following two more general research questions:

1. Why do organisations embark on CMM or CMM-based SPI initiatives?
2. Why do organisations not embark on CMM or CMM-based SPI initiatives?

These questions seemed *prima facie* to be amenable to investigation by systematic review, by aggregating information from surveys and experience reports. We would also attempt to extract data about organisation size and industry and treat them as auxiliary variables in our study.

As suggested in Kitchenham's guidelines, we did identify a previous systematic review related to CMM [4]. However, that study did not address our research question, and relied on a private database of CMM assessment results. So it did not make a useful contribution to the development of our systematic review protocol.

3.1.2. *Development of a Review Protocol*

Here we discuss how our initial protocol addressed our research questions, search strategy, selection criteria and process, quality assessment criteria, data extraction model and process, and data analysis plan.

Our research questions were similar to one of the general kinds of question suitable for investigation by systematic review listed by Kitchenham: "assessing the frequency or rate of a project development factor such as the adoption of a technology..." We were trying to discover the frequency of (de-)motivations for the adoption of CMM or CMM-based SPI. Again as suggested by Kitchenham we considered our research questions to have broader justification because they could help us understand factors affecting the adoption of SPI, which could lead to practical improvements in either SPI approaches or in their dissemination, and might reveal discrepancies between beliefs and reality about adoption of CMM or CMM-based SPI.

Kitchenham describes a general structure for research questions based on the population, intervention, and outcomes of interest. The structure of our research question did not fit entirely comfortably with all of these viewpoints as suggested by Kitchenham. If we were to force our question into this structure, the "populations" are organisations who have made a decision about adopting CMM or CMM-based SPI, and the "intervention" is best seen as the organisation's motivation or key driver for our "outcome": adopting CMM or CMM-based SPI.

We decided to admit surveys, case studies, and self-reported experience reports, but decided to exclude expert opinions. We felt that there would be enough data without resorting to expert opinion, and we were concerned about the possibility that "received opinion" within the SPI community may not reflect reality, especially as experienced by SMEs.

The definition of our search strategy was unexpectedly difficult. We used an initial list of resources, and an initial uniform search term. The initial search term was intended to be logically similar to: ("CMM" OR "CMMI") AND ("motivation" OR "reason"). During an initial trial period we tried to communicate the search instructions by email to each other and replicate each other's search results. We discovered that each of our searchable resources had different search syntaxes and form interfaces. Moreover, some resources would return different results for the "same" term depending on whether their "Basic" or "Advanced" search form was used.

We found our initial search term was too restrictive, and realised that we could not invent a search term that restricted results to only those that discussed the organisational motivation for SPI. We broadened our search term to be logically similar to:

"CMM" OR "CMMI"

However, the resources we searched covered a different variety of fields, each using different terminology. In some resources the search term above would return papers in other fields using the acronym "CMM" (e.g. "cutaneous malignant melanoma"). So for those resources we used a more specific search term:

("CMM" OR "CMMI") AND "capability maturity"

We broadened our list of resources from journal and conference databases to also include the SEI website, and recorded the specific search string that we used for each resource.

Following Kitchenham's guidelines, and because we had such broad search terms, we planned to have a two-stage selection process: first to exclude any obviously irrelevant publications, and second to include only those publications that contained data relevant to our research question. That is, we only ultimately selected those publications that gave direct evidence about actual organisations' explicitly stated reasons for adopting or not adopting CMM-based or CMMI-based SPI initiatives. This evidence could include publications that described why organisations had chosen one SPI approach over another.

The selection process was initially planned to be performed by one researcher as opposed to two [6], in order to reduce the major effort associated with this task. We initially planned that a second researcher would independently select publications from a random sample of the archived search results, and perform an inter-rater reliability test to confirm the accuracy of the selection process.

Kitchenham's guidelines suggest performing a quality assessment of each selected publication. We did not feel it would be possible for us (or perhaps any other individuals) to assess the extent to which other authors were able to identify and actually control threats to the validity of their studies. So, instead of trying to gauge the actual quality of publications, we only extracted a "YES" or "NO" for attributes for each of publication bias, internal validity, and external validity solely on the basis of whether the publication mentioned methodological issues related to these threats. That is, we did not make any judgements about the publications' effective treatment of these threats, but rather only if the publication discussed the possibility of these threats. Kitchenham's guidelines suggest performing the quality assessment in a separate phase immediately prior to data extraction. However, we treated our publication quality attributes like another piece of data, to be extracted at the same time as data extraction.

In order to extract data, we constructed a data model to suit our systematic review, as shown below in Figure 1.

FIGURE 1: Initial data model for our systematic review

A publication has attributes describing its publication details, and can contain a number of studies (e.g. a single paper might report both a survey and a case study each addressing a single research question). We recognised that "multiple case study" can be seen as a distinct methodology in its own right [10]. However for our purposes, where a publication contained multiple case studies we treated them each as separate studies. Each study within a publication has attributes that would be determined during the data extraction phase. The reasons for adopting CMM-based SPI were to be recorded using quoted text from the publications.

One researcher was to initially extract information from all selected publications. A second researcher was to independently extract information from a random sample of all selected publications, and the results were to be compared in an inter-rater reliability check.

In order to have a more abstract view of the data, we planned in the data synthesis phase to group together similar reasons into categories. We planned to start with no pre-defined categories and to aggregate reasons into categories incrementally. A second researcher was to independently perform this analysis, and the results were to be compared with an inter-rater reliability check. Then the results of the systematic review would be determined using a frequency analysis of these categories and a statistical analysis of their relationship to the attributes of organizations.

Our protocol went through many review and short trial iterations, principally to improve search terms and our plans for inter-rater reliability checking between researchers. Eventually we declared that we had a "final" protocol, despite being unsure about when "enough was enough" in our review process. At this stage there were still unresolved issues related to inter-rater reliability checking. For example, when checking publication selection on random sample, what action would we take if we did not get good agreement? Could we fix those problems, and any systematic errors those problems revealed, and then take another random sample to check again? If so, would we throw the already checked papers back into the random sampling?

3.1.3. Protocol Review

Our "final" protocol was reviewed by another researcher, but no significant comments were made. The reviewer was an experienced empirical software engineering researcher, but had not conducted any systematic reviews.

3.2. Conducting the Review

After completing and agreeing on the review protocol, we commenced the systematic review proper.

3.2.1. Identification of Research

Our search terms identified 591 publications. The results of the searches were archived to a local computer in a tabular word processing document using a format shown in Table 1. The electronic versions of publications were also stored in a filesystem directory for easy access during the systematic review.

ScienceDirect

Search Term = (CMM OR CMMI) AND 'capability maturity'

ID	Publication	Initial selection decision	Final selection decision
1	Title <i>Journal Title</i> , Volume X, Issue Y1, Date, Pages N-M Authors		

TABLE 1: Example of tabular format for search archive

3.2.2. Selection of Primary Studies

To reduce the time and cost of the systematic review, we had initially planned for one researcher to perform the selection and have a second researcher independently select publications from a random sample of the archived search results, and perform an inter-rater reliability test was to confirm the accuracy of the selection process. We did attempt this, and on a sample of 14 (from 591) had a result that indicated good agreement. However, we were too unsure about how many publications we would have to sample for the agreement to be significant.

In the end, the two researchers independently selected from among all the search results. In our first attempt at the two phases, we did not achieve a result indicating reliable agreement in either the initial or final selection. We took the union of our initially-selected shortlists and from this combined list of 73 publications each independently re-performed the final selection. Again, we did not achieve a result indicating reliable agreement. We resolved this by discussing selection criteria, and by again performing an independent selection on the newly combined list of 62 publications, this time by physically highlighting quotes within each paper to justify its inclusion. In a joint meeting we considered each point of difference in turn and came to agreement about its selection. This resulted in a list of 46 publications. An illustration of the selection process is shown in Figure 2.

FIGURE 2: Flow of selection and agreement in our systematic review

During the selection of primary studies, we became aware of possible problems of interpretation with our selection criteria. We found several studies discussing individual practitioners' motivations for SPI. Should these studies be

included to investigate organisational motivations for SPI? (After all, these practitioners were part of organisations involved with SPI.) We decided initially to select these studies only to include in a sensitivity analysis.

3.3. Study Quality Assessment, Data Extraction and Monitoring

As we had highlighted relevant quotes during the selection process, raw data extraction was straightforward – it just involved copying those quotes into a spreadsheet. One researcher performed quality assessment and data extraction at the same time, largely as described in the protocol. Although the data extraction broadly ran as planned, our data model changed significantly during the execution of the systematic review. Initially the protocol had only described a plan to extract data about organisation industry and size (a category defined in terms of number of employees). However, we could not reliably determine and categorise the industries served by organizations reported in the literature, so we dropped that from our model. However, during data extraction it became apparent that it would be not unduly difficult and perhaps scientifically valuable to also extract data about geography and year of adoption. So we also added these attributes to our data model. In order to properly summarise data about individual reasons, we realised we also needed to distinguish a quote for a reason as an entity in our data model, so that we could categorise the quote and record the number of organizations associated with the reason. Our final data model is shown in Figure 3.

FIGURE 3: Final data model for our systematic review

During data extraction it also became clear that the individual motivators for SPI were different in kind from the organisational motivators for SPI. Although we had thought to include the studies of individual motivators for SPI only in a sensitivity analysis, we finally decided to exclude them from our study altogether. We also found some papers that listed problems faced by an organization, and then listed the fact of the organization's adopting of CMM-based SPI, but did not claim that the organization's adoption of CMM-based SPI was intended to address those problems. This was a subtle but important distinction – we decided that such papers should not be included in our systematic review, as we wanted to investigate (explicitly listed) reasons for adoption, as opposed to investigating the problems that are faced by organizations that adopt CMM-based SPI.

We also discovered after data extraction that we did not have a significant number of publications that provided reasons why organisations did not choose to adopt CMM or CMM-based SPI. We dropped this research question from our systematic review. Kitchenham notes that one of the possible uses for systematic review is to identify gaps in existing research, but we had not initially intended to use systematic review for this purpose!

3.4. Data synthesis

Kitchenham's guidelines are not entirely clear about the nature of the data extraction process – how much categorisation is done “on the fly” during data extraction, and how much is done in the data synthesis phase? We opted for trivial data extraction resulting in a list of quotes which were only minimally paraphrased (e.g. to separate reasons expressed in conjunctive phrases). We categorised these in the early parts of the data synthesis stage.

As planned in the protocol, two researchers independently classified and grouped these motivations. Each researcher invented a different list of categories, but a common list of categories was agreed upon. The independent classification of quotes into categories was repeated and checked with an inter-rater reliability check. This did not indicate good agreement, and differences in opinion were discussed and agreed largely between the two researchers, but in some cases with a third researcher arbitrating.

3.5. Reporting the Review

We used a report structure similar to that suggested by Kitchenham [6]. We agree with Kitchenham's suggestion that a conference or journal paper is unlikely to want to publish the full details of a systematic review. As

suggested, we have provided the full report as an institutional technical report [8], and are preparing a separate journal paper that will include an overview of the methodology and results of the systematic review.

4. DISCUSSION

This section contains a critique of Kitchenham's guidelines in light of our experiences as described in the previous section and discussed a few of the lessons we have learned.

4.1. Overall

In general we feel that the systematic review is an effective methodology capable of revealing new information about a research area. Overall, we can say that Kitchenham's guidelines provide a good framework for a process to identify, assess and analyse all available research relevant to a specific research question. The systematic review process is general enough to be applied to many research areas within software engineering research. Case studies are a very common methodology within software engineering research, and systematic review can discover and synthesise new results that are not readily apparent in any single case study. The two researchers who used the systematic review process were satisfied with the results and overall performance of the process, and would be willing to use the systematic review process again in the future. Our overall support for the methodology and guidelines agrees with others' experiences [2].

Kitchenham identifies formal planning as critical for systematic review in order to mitigate risks of researcher bias. We agree, but also found planning to be critical in supporting the practical conduct of a systematic review.

4.2. High Effort and Duration

Kitchenham acknowledges that systematic reviews take considerably more effort than ordinary literature surveys, and we fully agree that the effort required should not be underestimated. However, we have also found that systematic review takes considerably more calendar time too. The duration of a systematic review is long because of the large effort, but is exacerbated by the large number of review points: search term pilot reviews, protocol reviews, initial selection reviews, final selection reviews, data extraction reviews, and data analysis reviews. These joint reviews are all important to improve the quality of the systematic review and to reduce researcher bias, but they are often difficult to schedule among multiple independent researchers each with busy timetables. The systematic review protocol structure suggested by Kitchenham contains a project timetable, which could in principle help to address this issue, but which in practice is defined too far in advance of good knowledge about the actual effort required to conduct each phase of the systematic review.

4.3. Importance of Research Questions

The entire systematic review is driven by its research questions, and we agree with other researchers who say that their specification "...is the most critical element of a systematic review." [2] The clear definition of narrow research questions is critical to control the effort and duration of the systematic review. The research questions define the scope of the systematic review and significantly influence the ease of selecting publications, and extracting and analysing data. We found that it was very helpful to define complementary research questions that were not being investigated, so as to clarify the boundaries of our research question of interest. This directly improved and clarified our selection and data extraction process. As noted above, our research question was:

Why do organizations embark on CMM-based SPI initiatives?

The complementary research questions that we defined were:

What motivates individuals to support the adoption of CMM-based SPI in an organization?

Why should organizations embark on CMM-based SPI initiatives?

What reasons for embarking on CMM-based SPI are the most important to organizations?

What benefits have organizations received from CMM-based SPI initiatives?

How do organizations decide to embark on CMM-based SPI initiatives?

What problems do organizations have at the time that they decide to adopt CMM-based SPI?

As seen above, an instance of our recommendation to define complementary research questions is to clearly identify the unit of analysis for the research question, i.e. to be explicitly clear whether you are studying organisations, teams, or individuals. We restricted our attention to organisational motivations for SPI rather than also considering individual practitioners' motivations for SPI. Kitchenham's systematic review guidelines do not explicitly mention the importance of defining the unit of analysis for the research question. Kitchenham et al. [7] mention that in empirical research in software engineering more generally it is important to define the experimental

unit, but justify this as mitigation against incorrectly inflating the sample size by multiple-counting organisations containing multiple individuals participating in the study. This for us was a second-order concern, as we found the main problem was that organisational and individual motivations for adopting CMM-based SPI were entirely different in character, and could not be easily translated or compared. Yin [10] discusses this reason for clearly defining the unit of analysis for case study research, and we find that it is true also for systematic review research.

4.4. Piloting and Modifying Protocols

We agree with others [2] who have found piloting to be very useful. However, we didn't know when we should stop the piloting process, and Kitchenham's guidelines did not give us specific guidance about the issue. The problems we encountered with inter-rater reliability may indicate that we stopped too soon! However later changes to the protocol are inevitable. Kitchenham notes that the stages in a systematic review are not strictly sequential, and that outputs from "earlier" stages may be refined or adapted in later stages. Our experiences support this: we refined selection criteria as late as the data extraction stage (to exclude studies of individual motivation for SPI), adapted our data extraction plans during the data extraction stage (to drop industry-type data, and add geography and year of adoption data), adapted our data analysis plans during the data analysis stage (to include analyses related to geography and year of adoption), and even dropped one of our research questions after the data extraction stage (as we did not find enough publications addressing why organisations chose not to adopt CMM-based SPI). Others have also reported [2] experiences with the revision of research questions during systematic review.

In light of all these changes, one might ask if there was any point to initially creating a protocol. We are reminded of Parnas and Clements' thoughts on why and how to fake a rational design process [9]. In particular, our original protocol provided useful guidance for us during the execution of the systematic review, our activity is closer to the original protocol than would have been if we did not have it, and the "standard procedure" of the protocol template improved our ability to perform our first systematic review and increased the quality of our work.

However, this poses a question for reporting – should the final report show the final ("fake") protocol design, or the full gory details of the initial protocol and the story and nature of all of subsequent changes? The nature and reason of the changes may reveal significant researcher bias, and so it is important to understand them. Kitchenham's guidelines partly address this question by suggesting that the course of the study selection be reported as a flow diagram to reveal how selection criteria changed throughout the course of the systematic review. We support that idea, but also suggest that in the full reporting of a systematic review only the final ("fake") protocol design be shown together with footnotes or other supplementary commentary that discuss the nature and reason of any changes made to the initial protocol during the course of the systematic review.

4.5. Automation Support

We used and benefited from only a basic level of automation to support our systematic review. In particular, we used simple tabular word processing documents to record lists of publication search results, file directories to store electronic copies of these publications, simple spreadsheet documents to record data extracted from selected publications and calculate inter-rater checking scores, and simple relational database tables and statistical analysis packages to analyse the extracted data. This seemed adequate for the performance of our systematic review.

What is the potential for advanced automated assistance for systematic review? The prospects seem very dim for the development of an "all singing, all dancing" repository or index to support all phases of a systematic review. The fundamental problem is that the research questions that could be addressed by systematic reviews in software engineering are conceptually complex, and are expressed in terms of ever-evolving theoretical models. It would be a "hard AI problem" to create a system to support automated selection or data extraction. Nonetheless, a generalised scientific ontology such as suggested by Hars [5] may be a step towards addressing this hard problem, and may provide a basis for improved interactive support tools for search and data extraction.

Targeted automation might bring many more immediate benefits to specific stages in the systematic review process. We experienced many problems consistently and reliably searching for publications, and could significantly benefit from tools that unified disparate resources and provided a uniform search syntax interface. Our problems with search have also been experienced by other researchers [2]. Although our basic file and data management was adequate for our individual study, we believe that file and data management mechanisms targeted to support systematic review might provide large benefits for replicating or analysing others' systematic review studies. Targeted collaborative tools may allow problems with inter-rater reliability checks for selection or data extraction to be detected early during the execution of these stages, and so allow any systematic errors to be resolved earlier. Such tools may also reduce the duration required for systematic review by allowing joint reviews to proceed asynchronously.

5. CONCLUSIONS

In light of our experiences we would join with others [2] to commend Kitchenham's guidelines [6] to other researchers considering conducting a systematic review. The main lessons that we have learned are: to limit the scope (and hence effort) of the systematic review by choosing clear and narrow research questions; to define complementary research questions that are not being investigated by the systematic review; to clearly define the unit of analysis for the systematic review; and when writing the full report for the systematic review, to show the final protocol but also include full notes about changes that have been made since the initial protocol. We would appreciate more guidance about piloting protocols during their development. We do not understand how valid and reliable assessments of the quality of others' studies can be made in the context of a systematic review as suggested by the guidelines, and have used a much weaker form of quality assessment. Finally, we would also appreciate more (and more accessible) guidance on inter-rater reliability checks for systematic reviews – in particular on how to sample a significant number of items for partial checks, and on how to avoid repeatedly failing checks.

We support the publication of replications of existing systematic reviews. A published successful replication will serve two purposes: it will strengthen the claims made by the original review, and it will build confidence within the software engineering research community about the validity of systematic review methodology. Any published unsuccessful replications of systematic reviews will highlight fundamental misunderstandings within the software engineering community or contribute to methodological improvements for systematic reviews.

Finally, we support the creation and maintenance of a central index of systematic reviews for software engineering. This would (like the Cochrane Collaboration for systematic reviews in medicine [3]) enable researchers to more easily discover systematic reviews related to their research questions, and also to provide a nexus for the improvement of systematic review methodology and automation. Although we became aware of the recent work by Biolchini et al. [1] too late for it to impact our study, we believe that their approach could contribute to the operation of success of a central index of systematic reviews for software engineering.

6. ACKNOWLEDGEMENTS

Mark Staples is employed by National ICT Australia, and Mahmood Niazi was employed by National ICT Australia while conducting the work reported in this paper. National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

REFERENCES

- [1] Biolchini, J., Mian, P. G., Natali, A. C. C. and Travassos, G. H. (2005) *Systematic Review in Software Engineering*. Technical Report RT-ES679/05. Universidade Federal do Rio de Janeiro Program de Engenharia de Sistemas e Computação.
- [2] Brereton, P., Kitchenham, B., Budgen, D., Turner, D. and Khalil, M. (2005) *Employing Systematic Literature Review: An Experience Report*. Unpublished draft.
- [3] Cochrane-Collaboration. (2003) *Cochrane reviews' handbook*. Version 4.2.1.
- [4] Goldenson, D. R. and Herbsleb, J. D. (1995) *After the appraisal: A systematic survey of Process Improvement, Its benefits, And Factors That Influence Success*. Technical Report CMU/SEI-95-TR-009. Carnegie Mellon University Software Engineering Institute.
- [5] Hars, A. (2001) Designing Scientific Knowledge Infrastructures: The Contribution of Epistemology. *Information Systems Frontiers*, 3(1), 63-73.
- [6] Kitchenham, B. (2004) *Procedures for Performing Systematic Reviews*. Technical Report TR/SE0401, Keele University, and Technical Report 0400011T.1, National ICT Australia.
- [7] Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W. Hoaglin, D. C., El Emam, K. and Rosenberg, J. (2002) Preliminary Guidelines for Empirical Research in Software Engineering. *IEEE Transactions on Software Engineering* 28(8), 721-734.
- [8] Niazi, M. and Staples, M. (2006) *Systematic Review of Organizational Motivations for Adopting CMM-based SPI*. National ICT Australia Technical Report PA005957, February 2006.
- [9] Parnas, D. L. and Clements, P. C. (1986) A Rational Design Process: How and Why to Fake It. *IEEE Transactions on Software Engineering*, SE-12(2), 251-257.
- [10] Yin, R. K. (2002) *Case Study Research: Design and Methods*, Sage Publications.