Available online at www.sciencedirect.com

ScienceDirect

The Journal of Systems and Software 80 (2007) 1425–1437

The Journal of Systems and Software

www.elsevier.com/locate/jss

Experiences using systematic review guidelines

Mark Staples [a,*], Mahmood Niazi [a,b,1]

[a] National ICT Australia, Australian Technology Park, Eveleigh NSW 1430, Australia
[b] School of Computing and Mathematics, University of Keele, Staffordshire ST5 5BG, United Kingdom

Received 9 June 2006; received in revised form 15 September 2006; accepted 20 September 2006
Available online 29 November 2006

Abstract

Systematic review is a method to identify, assess and analyse published primary studies to investigate research questions. We critique recently published guidelines for performing systematic reviews on software engineering, and comment on systematic review generally with respect to our experience conducting one. Overall we recommend the guidelines. We recommend researchers clearly and narrowly define research questions to reduce overall effort, and to improve selection and data extraction. We suggest that "complementary" research questions can help clarify the main questions and define selection criteria. We show our project timeline, and discuss possibilities for automating and increasing the acceptance of 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 (Kitchenham, 2004). 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 a different type of scientific value than an ordinary literature survey. In finding, evaluating, and summarising all available evidence on a specific research question, a systematic review may provide a greater level of validity in its findings than might be possible in any one of the studies surveyed in the systematic review. However, systematic reviews require much more effort than ordinary literature surveys.

The following features differentiate a systematic review from a conventional literature review (Kitchenham, 2004):

- 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 of 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.

Procedures and systems for systematic reviews are well established in other disciplines, particularly in medicine

[*] Corresponding author. Address: National ICT Australia, Locked Bag 9013, Alexandria NSW 1435, Australia. Tel.: +61 2 8374 5549; fax: +61 2 8374 5520.

E-mail addresses: Mark.Staples@nicta.com.au (M. Staples), mkniazi@cs.keele.ac.uk (M. Niazi).

[1] Tel.: +44 1782 583081.

0164-1212/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2006.09.046

(Sackett et al., 2000). However, software engineering researchers have yet to come to a well understood consensus about the conduct and value of systematic reviews. To work towards this goal, Kitchenham (2004) has recently published guidelines for software engineering researchers performing 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 (Niazi and Staples, 2006) informed by Kitchenham's guidelines. This paper is an expanded version of a previously published paper (Staples and Niazi, 2006). Our systematic review was the first that we had 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.

Brereton et al. (2005) have also previously reported experiences about using the systematic review methodology for software engineering research. Throughout this paper we will compare and contrast our experiences with those of their reported lessons that are relevant to our study. (i.e. we do not address their lesson on L14 concerning meta-analysis.)

This paper is organised as follows. Section 2 sketches an overview of the systematic review process as described by Kitchenham. 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 (2004) 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. 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 (2004).

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 the nature of 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 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 that may be 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 describes selection as 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 combing 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 the guidelines (Kitchenham, 2004).

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 believed 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 organizations embark on CMM or CMM-based SPI initiatives?
2. Why do organizations not embark on CMM or CMM-based SPI initiatives?

Our motivation for examining these research questions was to provide evidence about the initial needs and barriers of organizations concerning SPI. There has been a call to understand business drivers for SPI "…to make SPI methods and technologies more … widely used" (Conradi and Fuggetta, 2002), and as a first step we should understand why existing SPI approaches are either chosen or not chosen by software-developing organizations. A better understanding of the needs of organizations and their constraints in implementing SPI can lead SPI researchers to improve SPI approaches and/or the transition methods for SPI approaches.

These questions seemed prima facie to be amenable to investigation by systematic review, by aggregating information from surveys, case studies, and experience reports. For example, in the description of a case study about the outcomes of adopting CMM-based SPI, it is not uncommon for researchers to describe the initial motivation behind the choice to adopt the approach. For the second question, we similarly thought that it was possible that researchers could have discussed the reasons why CMM-based SPI was not chosen, perhaps leading the organization to adopt some other SPI approach. (We later discovered that reporting reasons for not choosing alternative SPI approaches is very uncommon in the research literature. So as we discuss below in Section 3.3, we were forced to discard the second research question.) In either case, the main data element that we hoped to extract from such case studies, experience reports, and surveys, was the reason or motivation given by organizations related to their CMM-based SPI adoption decision. As auxiliary variables, we would also attempt to extract data about organization size and industry, to see if these variables were related to the sort of reasons given by organizations.

As suggested in Kitchenham's guidelines, we tried to identify previous systematic reviews that either addressed our research questions or other questions in the same general area. We did identify a previous systematic review related to CMM (Goldenson and Herbsleb, 1995). However, that study did not address our research questions, 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 questions did not fit entirely comfortably with all of these viewpoints as suggested by Kitchenham. If we were to force our questions into this structure, the "populations" are organizations who have made a decision about adopting CMM or CMM-based SPI, and the "intervention" is best seen as the organization's motivation or key driver for our "outcome": adopting CMM or CMM-based SPI.

We decided to admit surveys, case studies, and experience reports, but decided to exclude expert opinions. Experience reports are similar to case studies, but are reported by members of the organization described in the publication. Both case studies and experience reports describe the situation and course of events within a specific organization. Expert opinions may be based on an expert's internal aggregation of prior practical experiences, but are expressed in general terms without reference to the explicitly described course of events within a specific organization. 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 actually 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. That is, we used iterative drafts of the protocol as the channel for communicating our search specifications. 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 organizational 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 very significant numbers of papers in other fields using the acronym "CMM" (e.g. "cutaneous malignant melanoma"). (The large number of these spurious results in some cases caused the maximum number of search results for the resource to be exceeded, thus "randomly" truncating the list of publications returned!) So for those resources we used a more specific search term:

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

Brereton et al. (2005) report two relevant lessons: "Current software engineering search engines are not designed to support systematic literature reviews. Unlike medical researchers, software engineering researchers need to perform resource-dependent searches."; and "There are alternative search strategies that enable you to achieve different sort of search completion criteria. You must select and justify a search strategy that is appropriate for your research question". We fully agree with both of these lessons – we found we were required to perform resource-dependent searches that not only varied in the concrete syntax for the search, but also varied in the search terms required.

Kitchenham's guidelines recommend searching "grey literature". In an attempt to discover some publications we broadened our list of resources from journal and conference databases to also include the SEI website.

We recorded the specific search string that we used for each resource. In principle, it might be considered a potential source of bias that we used different search terms for different resources. However, we do not believe that in our study, the use of our different search terms introduced any significant bias. Although we did not quantitatively analyze the issue, we believe that it very rare for publications about SEI's CMM or CMMI that use the acronyms

"CMM" or "CMMI" to not also provide their expansion, which includes the phrase "capability maturity".

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 organizations' explicitly stated reasons for adopting or not adopting CMM-based or CMMI-based SPI initiatives. This evidence could include publications that described why organizations had chosen one SPI approach over another.

The selection process was initially planned to be performed by one researcher as opposed to two, 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. Brereton et al. (2005) note that "All the medical standards emphasize that it is necessary to assess the quality of primary studies. However, it depends on the type of systematic literature review you are undertaking." We agree. For our systematic review, the data we were extracting for our research question was normally presented in the studies we found as part of the peripheral context for different research questions (typically concerning the benefits of successfully implementing CMM-based SPI). So for most of the studies we found, any specific measures used by the studies' authors to mitigate any threats to validity would not necessarily directly apply to the data we were extracting. Our general and weak notion of study quality was satisfactory for our study to make descriptive observations about study quality.

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

Brereton et al. (2005) reported that "Data extraction is assisted by having data definitions and data extraction guidelines form the protocol recorded in a separate short document". However, we did not encounter any problems with embedding data definition and data extraction guidelines within our protocol document.

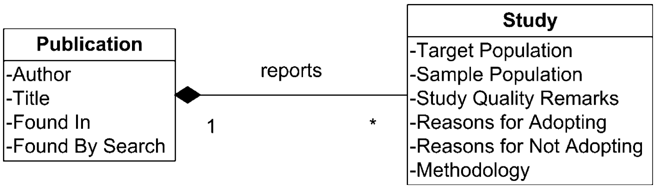

Fig. 1. Initial data model for our systematic review.

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 (Yin, 2002). 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.

As with Brereton et al. (2005), we would agree that "All systematic review team members need to take an active part in developing the review protocol". 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. Brereton et al. (2005) say that "There needs to be an agreed validation process separate from the protocol piloting activity". This agrees with Kitchenham's guidelines, and we would support it in principle, but in practice there is currently still a "chicken and egg" problem in software engineering concerning systematic reviews, with there being sparse expertise available to critically review systematic review protocols.

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

As discussed above in Section 3.1.2, we could not use precise search terms to create a small upper bound on the number of publications that might bear on our research questions. Our broad search terms identified 591 publications. The results of our 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.

3.2.2. Selection of primary studies

In the initial selection of publications from among our search results, we read and considered the titles and abstracts of the publications. In the final selection of publications from among our initial selection results, we also read and considered the bodies of the papers. 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.

Initially, 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 mutual agreement about its selection. This resulted in a list of 46 publications. An illustration of the selection process is shown in Fig. 2. Kitchenham's guidelines note that disagreements in selection should be discussed and resolved, which is what we have done. However, given the weak agreement of our independent selection results, a case could be made that we have essentially abandoned independent assessment as an approach for making a decision on final selection, and instead relied on a joint meeting to find mutual agreement. This is a weaker form of decision-making than independent assessment, as it is open to a greater risk of bias. However, it is nonetheless superior to decision-making by one individual researcher, and on balance we found it satisfactory.

The cause of some instances of disagreement in our independent selection process was due to possible problems of interpretation with our selection criteria. We found several studies discussing individual practioners motivations for SPI. Should these studies be included to investigate organizational motivations for SPI? (After all, these practioners were part of organizations involved with SPI.) We decided initially to select these studies only to include in a sensitivity analysis. These criteria interpretation issues were not discovered during the piloting process for our initial systematic review protocol. Later in the execution of our study, these issues of interpretation were recast as additional "complementary" research questions (discussed further in Section 4.3), to help clarify the scope of our main research question.

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, and the results were checked by a second researcher, largely as described in the protocol.

Table 1
Example of tabular format for search archive

ScienceDirect Search Term = (CMM OR CMMI) AND 'capability maturity'			
ID	Publication	Initial selection decision	Final selection decision
1	Title, Journal Title, Volume X, Issue Y1, Date, Pages N–M, Authors		

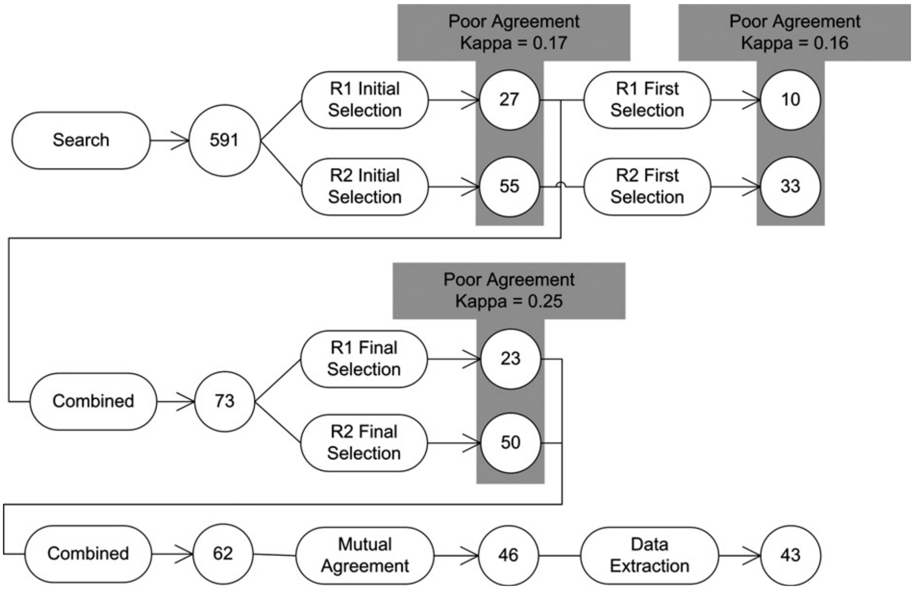

Fig. 2. Flow of selection and agreement in our systematic review.

We agree with Brereton et al.'s (2005) lesson that ''Having one reader act as a data extractor and one act as data checker may be helpful when there are a large number of papers to review''.

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 organization 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 Fig. 3.

During data extraction it also became clear that the individual motivators for SPI were different in kind from the organizational 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 being faced by organizations that adopt CMM-based SPI. At this point we decided to document these distinctions by creating a list of ''complementary'' research questions, i.e. research questions that were not within the scope of the systematic review. We discuss this further in Section 4.3.

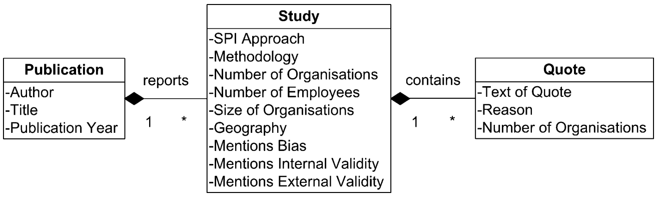

Fig. 3. Final data model for our systematic review.

We discovered after data extraction that we did not have a significant number of publications that provided reasons why organizations 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!

During data extraction, we identified two papers as being substantially equivalent – they reported the same results about the same study. The papers did not reference each other, their titles and abstracts were different, and their structures were slightly rearranged. We excluded one of the papers as a duplicate. Most of the papers we had found reported case studies, and so it was straightforward to determine that these were distinct cases. For two papers reporting interviews and surveys with multiple organizations, we were able to use geographical characteristics to determine that these also reported distinct cases.

We discovered that two papers reporting on many organizations did not give details about the number of organizations associated with each specific reason. Following Kitchenham's guidelines, we sent email inquiries to the authors of these two papers to ask for further details about their study. However, despite repeated queries, we did not receive enough information to enable us to use data from these two papers in our synthesis. We believe that a principle of scientific inquiry is that data supporting claims made in a report or paper should be made available for independent review and validation, if requested. Should the academic community also support the principle of making detailed data available for synthesis in a systematic review? We believe that it should, at the very least insofar as quality assessment is part of the systematic review process. Despite the lack of detailed data, it was possible to make qualitative comparisons between our summary findings and the high-level findings of the two studies.

Of the 46 publications we selected, only four scored "YES" for all three quality attribute attributes, i.e. explicitly mentioned the possibility of threats related to publication bias, internal validity, and external validity. Only three other publications studies scored "YES" for at least one quality attribute. So, only 15% of the publications we selected even partly addressed our very weak measures of publication quality. Such is the state of software engineering research. The four publications scoring "YES" for all three quality attributes included the two survey/interview studies that reported the bulk of the organizations in our systematic review. Brereton et al. (2005) note that "It is important to be sure how the quality assessment will be used in the subsequent data aggregation and analysis." We had loosely specified such a plan in our protocol – to determine if the findings of studies was associated with the quality scores of studies. However, given the generally low quality of the bulk of the studies we found, we decided not to compare the reasons given by studies on the basis of their quality scores. Instead, we compared the reasons given by studies on the basis of the methodologies used by the studies. We compared single-organization studies (case studies and experience reports) as a group against multi-organization studies (the survey and multi-organization interviews). We found (Niazi and Staples, 2006) a significant difference in the type of reasons given by organizations in these two groups – single-organization studies tended to report process-related reasons more frequently, whereas the multi-organization studies tended to report reasons related to product quality and project performance more frequently. For this and other reasons, we formed a belief that represented a bias in the SPI literature whereby process-related reasons tend to be reported more often than would be true for the total population of organizations that adopt CMM-based SPI. However, we believe that many of the broad conclusions of our systematic review hold in spite of this partial bias, as discussed in (Niazi and Staples, 2006).

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? Brereton et al. (2005) note that "IT and software engineering systematic reviews are likely to be qualitative in nature". We agree with this, and it may be beneficial for the guidelines to provide more guidance on the classification and summarization of qualitative information, and the positioning of these activities within systematic review protocols.

During the data extraction phase, we had 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). Fig. 3 shows the "Quote" class, with a "Text of Quote" attribute, which was the minimally-paraphrased quote from the paper. The number of organizations associated with each individual quote was also extracted as the "Number of organisations" attribute of the "Quote" class.

We had extracted 198 "Quote" objects. We categorised these quotes in the early parts of the data synthesis stage, as follows. 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 22 categories was agreed upon. An independent classification of quotes into categories was conducted and checked with an inter-rater reliability check. This check 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. The final classification for each "Quote" was stored in the "Reason" attribute shown in Fig. 3.

The definition of our data model (shown in Fig. 3) and the storage of our data in a relational database whose schema was based on this model, provided us with a structured mechanism to aggregate and query our extracted data. We were readily able to assemble summary lists of the frequency of reasons based on either number of publications or number of organizations, and were readily able to define queries whose results compared the numbers of organizations (and/or studies) giving each reasons by various characteristics of the studies and the organizations reported in the studies. These comparisons were then able to be analysed statistically by common tools, and depicted as charts. Brereton et al. (2005) found that "Tabulating the data is a useful means of aggregation but it is necessary to explain how the aggregated data actually answers the research questions". We would not disagree with this general principle, but do not think that it is a lesson particular to systematic review!

3.5. Reporting the review

We used a report structure similar to that suggested by Kitchenham (2004). 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. Brereton et al. (2005) found that "The software engineering community needs to establish mechanisms for publishing systematic literature reviews which may result in papers that are longer than those traditionally accepted by many software engineering outlets". However, we are not yet convinced that new mechanisms are required. As suggested by Kitchenham's guidelines, we have provided the full report as an institutional technical report (Niazi and Staples, 2006), and are preparing a separate journal paper that will include an overview of the methodology and results of the systematic review. We believe this a satisfactory approach for reporting systematic reviews.

3.6. Timeline

We did not keep detailed records on the *effort* spent on various stages of the systematic review. However, using email records, file creation dates, and project meeting notes, we have been able to reconstruct a timeline showing the *duration* in calendar days (excluding weekends) of major steps in the systematic review. This puts an upper bound on the effort required for each stage. This timeline is presented in Table 2.

Although we did not have comprehensive effort data for our study, we have been able to reconstruct coarse effort data for some of the sustained intensive tasks. Of the 23 days of duration shown in the table for the search and archiving task, one researcher devoted around 8 days of sustained effort to the task. One researcher undertook the 4 day duration initial selection task, and the other researcher undertook the 5 day duration independent selection task, but the effort for these tasks corresponded closely to full days of those durations.

As can be seen in the table, in terms of overall calendar time, after the decision to undertake the study, it took almost four months to derive a database of extracted and categorised data. After having derived a database of extracted and categorised data, it took around three months to finish the data synthesis and write the report. However, the effort for these activities was distributed sporadically and at lower levels across this duration, and mostly involved alternating serial effort from the researchers. The completion of the data synthesis task was delayed somewhat by waiting for responses from authors who were asked for further information about their primary studies. Publishing the technical report took a long duration (67 days) because of organizational delays, partly due to summer holidays.

Table 2
Timeline of systematic review project

Activity	Start Date	Duration (days, not weekends)
Decision to undertake systematic review	2 Feb 2005	–
Writing protocol	9 Feb 2005	37
Searching and archiving	10 Mar 2005	23
Initial selection	12 Apr 2005	4
Independent selection	11 May 2005	5
Joint selection	16 May 2005	3
Reason Extraction and classification	23 May 2005	3
Data synthesis	26 May 2005	40
Email inquiries to authors	1 Jun 2005	22
Writing report	29 Jul 2005	20
Publishing report	1 Dec 2005	67

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 (Brereton et al., 2005).

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.

Systematic review is yet to be widely known, understood, and accepted as a valid empirical methodology in software engineering. Our systematic review and this commentary go a small way towards establishing systematic review in these ways. Systematic review derives much of its force as a methodology by seeking to be a foundation for replicable studies. We believe that systematic review in software engineering could be significantly strengthened as a methodology by testing this claim – we recommend that researchers attempt replications of others' systematic reviews. Any subsequent differences in results between an original and replicated should be traced back to differences in the conduct of the studies, to determine if there are root causes within systematic review methodology or relevant guidelines that could be rectified.

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. However, we found in practice that it was hard for us to estimate the effort that would be 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" (Brereton et al., 2005). 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 publication, and extracting and analysing data. As noted above, our research question was:

- Why do organizations embark on CMM-based SPI initiatives?

We also 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. 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 organizations, teams, or individuals. We restricted our attention to organizational motivations for SPI rather than also considering individual practioners 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. (2002) 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 organizations containing multiple individuals participating in the study. This for us was a second-order concern, as we found the main problem was that organizational and individual motivations for adopting CMM-based SPI were entirely different in character, and could not be easily translated or compared. Yin (2002) 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

One of the lessons reported by Brereton et al. (2005) was "Piloting the research protocol is essential. It will find mistakes in your data collection and aggregation procedures. It may also indicate that you need to change the methodology you intend to use to address the research questions." We fully agree that piloting is very useful. We believe that the main benefit may be in gaining an increased level of common understanding of the protocol in the review team. Brereton et al. (2005) found that "Review team members must make sure they understand the protocol and the data extraction process". We agree, and piloting the protocol is extremely helpful in working towards common understanding of these issues.

However, we did not 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 organizations chose not to adopt CMM-based SPI). Another of the lessons from Brereton et al. (2005) was "Expect to revise your questions during protocol development, as your understanding of the problem increases experiences with the revision of research questions during systematic review". Our experiences fully agree with this.

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 (1986) thoughts on why and how to fake a rational design process. 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. We agree with Brereton et al. (2005) lesson that "Review teams need to keep a detailed record of decisions made throughout the review process", but we would amplify this and claim that these decisions must also be reported. 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 and 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 arbitrary systematic reviews. 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 (2001) may be a step towards addressing this hard problem, and may provide a basis for improved interactive support tools for search and data extraction.

Although a generic end-to-end tool may be hard to develop for the reasons given above, other kinds of general support can be realistically achieved. In medicine, the Cochrane-Collaboration (2003) for systematic reviews enables researchers to more easily discover earlier systematic reviews related to their research questions, and also provides a nexus for the improvement of systematic review methodology. Although we became aware of the recent work by Biolchini et al. (2005) 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.

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 (Brereton et al., 2005).

The initial selection task relies on a reading of the title and abstract of all found papers. We believe that many of the abstracts that we reviewed during our initial selection were low quality, and that this hampered our efforts. Brereton et al. (2005) similarly found that "The standard of IT and software engineering abstracts is too poor to rely on when selecting primary studies. You should also review the conclusions". We did not review the conclusions in our initial selection phase, but did review the conclusions (and the whole paper) in our final selection phase. It could be that our problems reaching agreement in the initial selection phase were partly due to the poor overall quality of abstracts. In our study, reviewing conclusions in addition to abstracts and titles in the initial selection phase would have very significantly increased the effort required. It is a distinct possibility that automatically generated summaries of the bodies of papers from the software engineering field may be higher quality (on average) than the abstracts hand-crafted by authors, and this idea may be worth exploring in future research on systematic review methodology.

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 (Brereton et al., 2005) to commend Kitchenham's (2004) guidelines 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.

In our systematic review, we were not able to obtain additional data about two previously published studies sufficient for us to integrate details of their results into a common synthesis. We would support a call for a principle of academic research that data supporting claims made in a report or paper should be made available by researchers upon request, for synthesis in systematic reviews. We believe such an obligation would not be a great additional burden over the existing moral obligation of researchers to archive data for the purpose of independent review in the event of inquiries about scientific propriety.

Finally, we support the creation and maintenance of a central index of systematic reviews for software engineering, in a manner similar to the Cochrane-Collaboration (2003) in medicine. A central and well-known site would provide not only a central index of existing systematic reviews, but also resources to help researchers in the practice of systematic reviews. These resources could include methodological guidelines, commentary on relevant tools, practical suggestions for searching resources, and information such as actual effort required for existing systematic reviews.

Acknowledgements

Thanks very much to anonymous referees whose comments on previous drafts of this paper have helped to improve it. 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

Biolchini, J., Mian, P.G., Natali, A.C.C., 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 Computaçào.

Brereton, P., Kitchenham, B., Budgen, D., Turner, D., Khalil, M. 2005. Employing Systematic Literature Review: An Experience Report. Unpublished draft.

Cochrane-Collaboration. 2003. Cochrane reviews' handbook. Version 4.2.1.

Conradi, R., Fuggetta, A., 2002. Improving Software Process Improvement. IEEE Software, July/August, 92–99.

Goldenson, D.R., 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.

Hars, A., 2001. Designing scientific knowledge infrastructures: the contribution of epistemology. Information Systems Frontiers 3 (1), 63–73.

Kitchenham, B., 2004. Procedures for Performing Systematic Reviews. Technical Report TR/SE0401, Keele University, and Technical Report 0400011T.1, National ICT Australia.

Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K., Rosenberg, J., 2002. Preliminary guidelines for empirical research in software engineering. IEEE Transactions on Software Engineering 28 (8), 721–734.

Niazi, M., Staples, M., 2006. Systematic Review of Organizational Motivations for Adopting CMM-based SPI. National ICT Australia Technical Report PA005957, February 2006.

Parnas, D.L., Clements, P.C., 1986. A rational design process: how and why to fake it. IEEE Transactions on Software Engineering SE-12 (2), 251–257.

Sackett, D.L., Straus, S.E., Richardson, W.S., Rosenberg, W., Haynes, R.B., 2000. Evidence-Based Medicine: How to Practice and Teach EBM, second ed. Churchill Livingstone, Edinburgh.

Staples, M., Niazi, M., 2006. Experiences Using Systematic Review Guidelines. In: Proceedings of 10th International Conference on Evaluation and Assessment in Software Engineering (EASE), Keele University, UK, 10–11 April 2006, BCS Electronic Workshops in Computing.

Yin, R.K., 2002. Case Study Research: Design and Methods. Sage Publications.

Mark Staples is a Senior Researcher in the Empirical Software Engineering program at National ICT Australia. Before joining NICTA, he worked for medium-sized software developing companies in roles covering verification, configuration management, product line development, and process improvement. His current research interests cover these areas of

experience. He holds a Ph.D. from the Computer Laboratory, University of Cambridge.

Mahmood Niazi is a Lecturer at School of Computing and Mathematics at Keele University. He has spent more than a decade with leading technology firms and universities as a process analyst, senior systems analyst, project manager, and lecturer. He has participated in and managed several software development projects. These have included the development of management information systems, software process improvement initiatives design and implementation, and several business application projects. He holds a Ph.D. from the Faculty of IT, University of Technology Sydney.